JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

REAL TIME SPEECH TRANSCRIPTION AND TRANSLATION

Kavitha B
Assistant Professor *Department of CSE, RNSIT Bengaluru, India*
kavitha.b@rnsit.ac.in

Shilpa
Department of CSE, RNSIT Bengaluru, India
1rn21cs144.shilpa@rnsit.ac.in

Ancy Thomas
Assistant Professor *Department of CSE, RNSIT Bengaluru, India*
ancythomas@rnsit.ac.in

Sampritha S
Department of CSE, RNSIT Bengaluru, India
1rn21cs134.samprithas@rnsit.ac.in

Shreya Gautam
Department of CSE, RNSIT Bengaluru, India
1rn21cs147.shreyagautam@rnsit.ac.in

Riya
Department of CSE, RNSIT Bengaluru, India
1rn21cs124.riya@rnsit.ac.in

***Abstract*—This paper introduces an advanced web-based platform designed to revolutionize speech translation and transcription. By leveraging the capabilities of OpenAI Whisper, the platform achieves high-accuracy transcription of spoken content, even in acoustically challenging environments. Whisper's robust multilingual support ensures reliable speech-to-text conversion across diverse languages and accents. The transcribed text is further processed using the Google Translate API, enabling seamless translation into multiple target languages, which is invaluable for fostering cross-linguistic communication in educational and professional contexts.To enhance accessibility, the Google Text-to-Speech (gTTS) module converts the translated text into natural-sounding audio, providing users with both visual and auditory outputs. The platform incorporates secure user authentication to safeguard user data and interactions, allowing users to input speech, view transcriptions and translations, and download processed audio files with confidence.The backend infrastructure is powered by MySQL, which efficiently manages the storage of audio files and metadata, ensuring streamlined data retrieval and scalability. This integration of cutting-edge AI technologies, intuitive design, and robust data management establishes a comprehensive solution for speech translation and transcription, making it an essential tool for global communication.**

***Index Terms*—Speech-to-text, Whisper AI, multilingual translation, gTTS, real-time transcription, NLP, user authentication, MySQL.**

I. INTRODUCTION

Speech translation and transcription are pivotal technologies that bridge communication gaps in our increasingly interconnected world. Speech transcription involves converting spoken language into written text, enabling accurate documentation and analysis of verbal communication. On the other hand, speech translation extends this capability by transforming the transcribed text from one language into another, facilitating seamless cross-linguistic communication. Advancements in Artificial Intelligence (AI) and Machine Learning have significantly enhanced the precision and efficiency of these processes. Tools like OpenAI Whisper excel in capturing nuanced speech patterns for transcription, while APIs like Google Translate offer dynamic multilingual text translations. When combined with text-to-speech technology, such as Google Text-to-Speech (gTTS), the output can be converted into audible formats in the desired language. This integration of technologies empowers various applications, including education, accessibility for differently-abled individuals, business communications, and cultural exchange, revolutionizing how people interact across linguistic barriers.

The project introduces a web-based platform for speech translation and transcription, aiming to bridge language gaps and enhance communication across linguistic boundaries. The system utilizes OpenAI Whisper for accurate speech-to-text conversion, capable of transcribing spoken words into text in realtime, even in noisy environments. Once transcribed, the text is translated into the desired language using the Google Translate API, offering robust translation capabilities for numerous languages. The Google Text-to-Speech (gTTS) module then converts the translated text back into natural-sounding audio, allowing users to listen to the translated content. A key feature of the system is its secure user authentication, ensuring privacy and restricted access to personal data. It also integrates a MySQL database to store metadata such as the original and translated audio files, transcriptions, and user-related information. This enables seamless management of user data and content, while allowing easy retrieval of stored translations and transcriptions. The platform's application spans various sectors, including education, where it enables real-time multilingual learning content, accessibility, assisting individuals with language barriers or hearing impairments, and global content localization, making it easier for businesses and institutions to reach international audiences. Its comprehensive functionality, combining speech-to-text, translation, and speech synthesis, makes it a versatile tool for enhancing cross-linguistic communication.

II. LITERATURE SURVERY

Recent advancements in speech-to-speech (S2ST) and text-to-speech (TTS) technologies have marked a shift from traditional modular approaches to more efficient end-to-end systems. Innovations such as JANUS-III, ATR, Parrotron, and FastPitch have significantly enhanced multilingual capabilities, semantic translation, and expressive synthesis. Despite these strides, challenges such as achieving emotional depth, handling variability in spontaneous speech, and optimizing resource efficiency remain. Addressing these issues calls for improved multilingual datasets, advanced algorithms, and lightweight solutions to ensure broader accessibility. These technologies hold transformative potential, with societal benefits ranging from aiding the hearing-impaired to fostering global communication.

The paper "A Comprehensive Review of Text-to-Speech (TTS) Technologies" reviews the evolution and key types of TTS systems, including concatenative, formant synthesis, statistical parametric, neural, and hybrid approaches, highlighting their strengths, limitations, and applications. It details TTS mechanisms such as preprocessing, encoding, decoding, and vocoding, along with notable advancements in Bangla TTS, including syllabic unit selection and diphone concatenation. Recent developments like FastPitch and LightSpeech demonstrate how deep learning enhances expressiveness and efficiency, though challenges remain in achieving emotional naturalness. The review emphasizes the need for further research to address these challenges and expand TTS applications in education and accessibility, showcasing its transformative potential in human-computer interaction [1].

The paper "Speech To Speech Translation: Challenges and Future" examines the significance and potential of Speech-to-Speech Translation (S2ST) in overcoming language barriers. It highlights the three key components: Automatic Speech Recognition (ASR), Machine Translation (MT), and Speech Synthesis, explaining their sequential roles in converting spoken input from one language to another. The paper discusses advancements, challenges such as multilingual data limitations, and the complexity of natural language processing. It also explores applications in travel, education, and crisis response, emphasizing the need for further research to improve accuracy and accessibility [2].

The paper "The ATR Multilingual Speech-to-Speech Translation System" discusses a system enabling communication across languages, focusing on English and Asian languages like Japanese and Chinese. It integrates speech recognition, machine translation, and text-to-speech synthesis using a corpus-based statistical machine learning framework. A key feature is its multilingual database of over 600,000 sentences for travel-related conversations. The system employs advanced techniques like Hidden Markov Models (HMMs) and N-grams for speech recognition, combines example-based and statistical methods for machine translation, and uses the XIMERA framework for natural-sounding speech synthesis. It achieves quality comparable to a TOEIC score of 750, highlighting its practical potential [3].

The paper "Speech to Speech Translation Using Machine Learning" explores a system enabling real-time language translation for seamless communication. It integrates speech recognition, machine translation, and text-to-speech synthesis using tools like the Google Translate API and gTTS. A web application prototype built with Flask captures speech input, translates it into Marathi, and synthesizes the translated text into speech. The study addresses challenges such as translation accuracy, speaker variability, and data requirements, proposing custom models and expanded language support for improvements. It highlights the technology's potential in commerce, healthcare, and education while emphasizing the importance of preserving linguistic and cultural nuances [4].

The paper "Real-Time Speech-to-Speech Translation for PDAs" presents a handheld bilingual translation system enabling real-time communication between English and Iraqi Arabic. Developed under DARPA programs, it integrates automatic speech recognition, concept-based translation, and text-to-speech synthesis. Using optimized methods like compact acoustic models and entropy-pruned language models, the system achieves efficient performance on resource-limited devices. Testing on handhelds demonstrated near-real-time processing with competitive accuracy. Future enhancements, such as statistical machine translation and speaker adaptation, are suggested to expand its capabilities, highlighting its potential for military, humanitarian, and commercial applications [5].

The paper "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation" introduces Parrotron, an end-to-end model that converts spectrograms directly, improving speech intelligibility for hearing-impaired individuals and separating speech in noisy environments. It uses an encoder-decoder architecture with attention mechanisms and a vocoder, showing significant improvements in word error rates and intelligibility. The paper highlights Parrotron's potential in accessibility applications for speech impairments and suggests further expansions for other disorders [6].

The paper "Voice Recognition System: Speech-to-Text" discusses the design and implementation of a speech-to-text system that converts acoustic signals into text. It uses Mel Frequency Cepstral Coefficients (MFCC) for feature extraction, Hidden Markov Models (HMM) for acoustic modeling, and Vector Quantization (VQ) for feature matching. The system categorizes speech recognition based on speech utterance types, speaker models, and vocabulary size. Its innovative approach focuses on computational efficiency and accuracy, making it ideal for applications like home automation. The paper also highlights recent advancements and the potential of these systems in future automation solutions [7].

The paper "Speech to Text Using Deep Learning" presents a speech recognition model with multilingual capabilities for converting spoken language into text. It highlights the use of advanced natural language processing (NLP) and neural machine translation to improve accuracy and facilitate communication across language barriers. Key features include noise

reduction, enhanced accuracy through feature extraction, and integration with pre-trained datasets. The system enables voice command interactions for file operations, with applications in telecommunication, multimedia, and assistive technologies. Challenges such as varying accents and dialects are addressed, and the system is designed to improve accessibility and usability in human-machine interaction [8].

The paper "JANUS-III: Speech-to-Speech Translation in Multiple Languages" presents a speech-to-speech translation system designed for spontaneous dialogues, particularly for scheduling tasks. The system uses context-dependent acoustic models for speech recognition, dual parsing strategies to enhance accuracy, and an interlingua-based approach for semantic translation. Despite its high accuracy and adaptability to new languages and domains, the system faces challenges such as handling variability in spontaneous speech and requiring significant computational power and large datasets [9].

The paper "Speech-to-Speech Translation Using Deep Learning" presents a system that utilizes deep learning models and cloud services for speech-to-speech translation. It features a three-phase architecture: speech-to-text conversion, multilingual translation with neural machine translation, and high-quality audio synthesis using models like Tacotron2 and HifiGAN. The system performs well across multiple languages but faces challenges due to its reliance on cloud services and its computationally intensive nature [10].

The paper "End-to-End Speech-to-Text Translation Models" discusses the shift from traditional cascaded models to end-to-end integrated systems for speech-to-text translation. It highlights the advantages of reducing error propagation and resource costs, using models that combine Automatic Speech Recognition (ASR) and Machine Translation (MT). However, these systems require large datasets for training and involve complex design due to the integration of multiple tasks [11].

In conclusion, advancements in speech-to-speech (S2ST) and text-to-speech (TTS) technologies are shifting towards more efficient, end-to-end systems, enhancing multilingual capabilities and expressive synthesis. Innovations like JANUS-III, ATR, Parrotron, and FastPitch show great progress, but challenges such as emotional depth, speech variability, and resource optimization remain. Overcoming these hurdles requires improved datasets, refined algorithms, and lightweight solutions for broader accessibility. These technologies have transformative societal impacts, especially in aiding the hearing-impaired and enhancing global communication, with ongoing innovation needed to address current limitations and unlock new possibilities.

III. System Architecture

The system architecture integrates several components to provide seamless speech transcription and translation services. The user interacts with the frontend, developed using React, to either record or upload audio. The audio is sent to the backend server, powered by Flask, which processes the file. The audio is transcribed into text using Whisper AI, stored in a MySQL database, and then translated using the Google Translate API. The translated text is returned to the frontend and displayed to the user. Additionally, the system offers a text-to-speech feature using the gTTS library, converting the translated text back into audio, which is stored in the database for easy access. This architecture ensures scalability, flexibility, and a smooth user experience, handling speech-to-text, translation, and text-to-speech functionalities efficiently.

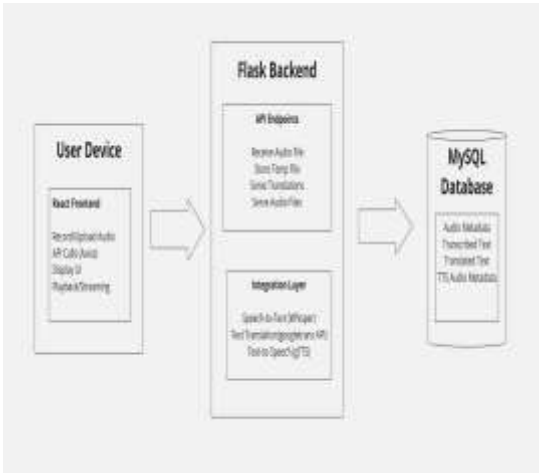

Fig. 1. System Architecture

IV. Methodology

The methodology of the project involves the integration of several advanced technologies to create a seamless platform for speech transcription, translation, and synthesis. The process begins with speech input, where the user's spoken content is captured via a microphone or other audio input devices. This audio is then processed using OpenAI Whisper, a robust AI-based speech-to-text model, which transcribes the speech into accurate text, even in noisy environments. Whisper's multilingual support ensures that the transcription is reliable across various languages and accents.

Once the speech is transcribed into text, the platform utilizes the Google Translate API to translate the text into the desired language. The Google Translate API is integrated to support multiple languages, ensuring that users can access translated content in real time. After translation, the platform uses the Google Text-to-Speech (gTTS) module to convert the translated text back into natural-sounding audio, enabling users to both read and listen to the content in their preferred language.

The system is built with a secure user authentication module, ensuring that only authorized users can access and interact with the platform. This is important for maintaining privacy and securing user data, especially when dealing with sensitive information. All user data, including speech files, transcriptions, translations, and metadata, is stored in a MySQL database, which efficiently manages and organizes the data for quick retrieval and further processing.

Throughout the development, emphasis was placed on ensuring the scalability and efficiency of the system, allowing

it to handle multiple users and requests simultaneously. The platform is tested under various real-world conditions to ensure high performance in terms of transcription accuracy, translation speed, and speech synthesis quality. This methodology combines cutting-edge technologies with a robust backend system to create an intuitive and highly functional platform for real-time speech translation and transcription.

A. Whishper AI

Whisper AI, developed by OpenAI, plays a pivotal role in the success of the Speech Transcription and Translation System. This advanced speech-to-text model is trained on a large dataset of diverse audio sources, enabling it to handle various accents, dialects, and noisy environments with high accuracy. Its robustness ensures precise transcription of spoken words into text, forming the foundation for subsequent translation and synthesis processes.

Whisper AI demonstrates its efficiency by seamlessly transcribing uploaded or recorded audio into text, regardless of background noise or speech variations. Its ability to maintain high transcription quality makes it suitable for real-world applications, such as multilingual communication, education, and cross-cultural exchanges. Whisper AI's adaptability and performance underline its importance in creating reliable, user-friendly speech processing systems, making it a cornerstone technology in this project.

B. Flow Diagram

The Speech Translation Application follows a streamlined process to convert speech into text, translate it, and generate audio. The process begins when a user uploads or records an audio file (in .mp3 or .wav format), which is then verified and saved. The audio is passed to the OpenAI Whisper AI model for transcription, converting speech into text. The text is then translated using the Google Translate API into the specified target language. Following translation, the system employs Google Text-to-Speech (gTTS) to convert the translated text into speech, which is saved as an .mp3 file. The application stores metadata in a MySQL database and provides a downloadable link for the translated audio. Finally, the process completion is logged, offering transparency and error handling throughout the workflow.

This structure ensures that the prediction model is trained on diverse algorithms and tested rigorously to produce reliable and accurate predictions for unseen data.

C. Algorithm

The listed algorithms are applied to train and test the dataset.

The algorithm for the speech translation application is designed to provide a seamless experience for users requiring real-time speech translation. The process begins on the frontend, where users interact with a React-based interface. Users can either upload an audio file in WAV or MP3 format or record speech directly through their browser using the device's microphone. React efficiently manages the application state using hooks like useState and useRef, ensuring smooth functionality for audio uploads, recording, and language selection.

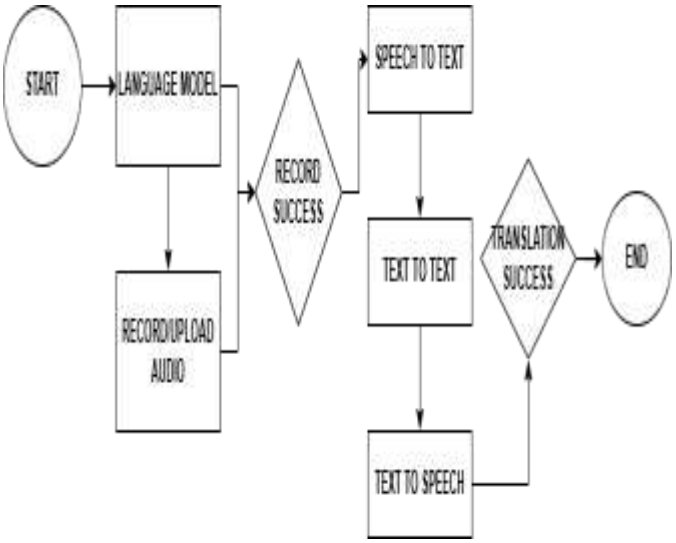

Fig. 2. System Workflow

When the user submits the audio for translation, a POST request is sent to the Flask backend. The Flask server is equipped with CORS support to handle cross-origin requests and leverages advanced AI and cloud tools to process the audio. The audio file is first saved in a predefined directory. It is then passed to the Whisper AI model, which performs speech-to-text transcription with high accuracy. Whisper outputs the transcribed text, which is analyzed by the Google Translate API. This API detects the original language and translates the text into the user's selected target language.

The translated text undergoes further processing through the gTTS library to generate a natural-sounding audio file in the target language. The generated speech is saved as an MP3 file, and metadata, including the file name, original speech text, transcribed text, detected language, and translated text, is securely stored in a MySQL database.

Once the translation process is complete, the Flask server constructs a JSON response that includes the recognized speech, translated text, and a downloadable URL for the translated audio file. This response is sent back to the React frontend, where the user can view the transcribed and translated text alongside an audio player for the synthesized speech.

The backend also includes additional endpoints for user management and data retrieval. Users can sign up or log in, with credentials securely stored in a MySQL database using hashed passwords for security. A dedicated endpoint retrieves previously processed audio files and their associated metadata, enabling users to access past translations.

V. RESULTS

The Speech Transcription and Translation Application is an innovative platform designed to facilitate seamless communication across languages through its intuitive user interface and powerful backend functionalities. The application begins with a Sign-Up Page, which offers a visually appealing and user-friendly interface for new users. This page includes input

fields for name, email, and password, enhanced by real-time validation for user input errors, such as invalid email formats or weak passwords. The form's semi-transparent overlay on a full-screen background image creates a clean and professional aesthetic, ensuring ease of navigation and accessibility.

The Login Page provides a secure entry point for returning users. It incorporates client-side validation using regular expressions to ensure the accuracy and security of the provided credentials. Upon successful login, users are directed to the main interface; otherwise, appropriate error messages are displayed. The page maintains a modern, responsive design that ensures usability across various devices.

The File Upload Page serves as the hub for processing audio inputs. Users can either upload pre-recorded audio files or record audio directly within the application using their device's microphone. A dropdown menu allows users to select their desired target language for translation, accommodating a wide range of linguistic preferences. The interface dynamically adapts based on user actions, ensuring a smooth workflow.

Once an audio file is uploaded or recorded, the File Upload Result Page displays the outcomes of the backend processing. This includes the recognized speech from the uploaded audio, its translated text, and an audio player for listening to the translated speech. The backend processes involve transcription using Whisper AI, translation using Google Translate, and text-to-speech conversion through gTTS, ensuring high accuracy and efficiency.

The Audio Stored Page organizes and displays previously processed audio files and their metadata in a user-friendly table. Each entry includes details such as the filename, detected language, original speech text, translated text, and a playback option for the translated audio. The page also includes error handling mechanisms, such as displaying messages when no files are found or when data fetching encounters issues.

Finally, the About Us Page provides a comprehensive overview of the application's purpose, features, and mission. It explains the core functionalities, from speech-to-text conversion to real-time translation and text-to-speech playback. The page emphasizes the use of advanced AI tools like Whisper AI, Google Translate, and gTTS, highlighting their integration into the app's workflow. The mission section underscores the app's goal to break down linguistic barriers, making it an essential tool for travelers, students, and professionals needing real-time language support.

This application combines robust technology with user-centric design, offering a seamless and engaging experience for individuals seeking to communicate effortlessly across languages.

VI. Conclusion

The integration of advanced AI and NLP technologies enables seamless real-time multilingual communication. By leveraging Whisper AI for accurate speech-to-text transcription, Google Translate for language detection and translation, and gTTS for text-to-speech synthesis, the system establishes a robust and efficient speech-to-speech translation workflow. It ensures high accuracy and usability, even in challenging environments, making it ideal for diverse applications in business, education, and cross-cultural communication.

The Flask-based web application provides a user-friendly interface, allowing users to upload audio, translate speech, and download the output while storing metadata in a MySQL database for easy retrieval. This feature-rich design enhances productivity and accessibility, catering to both individual and professional use cases.

Looking ahead, several enhancements can further improve the system's capabilities. Real-time translation during live conversations can be integrated, accommodating various accents and dialects for a more natural interaction. Improvements in speech synthesis can enhance the human-like quality of audio output, making interactions more engaging. Additionally, incorporating deep learning-based contextual understanding can refine translation accuracy by considering linguistic nuances. Expanding support for industry-specific terminologies will further enhance applicability in fields such as healthcare, tourism, education, and emergency response.

Overall, AI's potential to unify and connect diverse cultures by breaking language barriers, fostering inclusivity, and paving the way for a globally interconnected future is immense. As technology evolves, such systems will become even more indispensable, driving innovation and collaboration across regions and industries.

References

[1] Md. Jalal Uddin Chowdhury and Ashab Hussan, "A review-based study on different Text-to-Speech technologies," *arXiv preprint arXiv:2312.11563*, 2023.

[2] Sandeep Dhawan, "Speech To Speech Translation: Challenges and Future," *International Journal of Computer Applications Technology and Research*, vol. 11, no. 3, pp. 36–55, 2022.

[3] Satoshi Nakamura, Kiyohiro Shikano, Tetsunori Kobayashi, Masaki Kohda, and Ryuichi Matsumoto, "The ATR Multilingual Speech-to-Speech Translation System," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 5, no. 1, pp. 1–12, 1997.

[4] Mrs. R. Y. Totare, S. K. Gupta, and P. R. Deshmukh, "Speech to Speech Translation Using Machine Learning," *International Journal of Novel Research and Development*, vol. 9, no. 4, pp. 1–6, 2024.

[5] R. Prasad, U. Germann, D. Nahamoo, and J. F. Jarrett, "Real-Time Speech-to-Speech Translation for PDAs," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. IV–129–IV–132, 2007.

[6] Fadi Biadsy, Ron J. Weiss, Kevin Wilson, and Michael L. Seltzer, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 106–110, 2019.

[7] Prerana Das, A. K. Sahoo, and T. K. Pradhan, "Voice Recognition System: Speech-to-Text," *Journal of Applied and Fundamental Sciences*, vol. 1, no. 2, pp. 1–8, 2015.

[8] L. Nalini Joseph, R. Balasubramaniam, and R. Suganya, "Speech to Text Using Deep Learning," *International Journal of Novel Research and Development*, vol. 9, no. 4, pp. 1–6, 2024.

[9] Alon Lavie, Lori Levin, Alex Waibel, and Stephan Vogel, "JANUS-III: Speech-to-Speech Translation in Multiple Languages," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 261–264, 1997.

[10] R. K. Sharma, P. T. Mahapatra, and K. R. Mehta, "Speech-to-Speech Translation Using Deep Learning," *International Journal of Novel Research and Development*, vol. 9, no. 4, pp. 1–6, 2024.

[11] Nivedita Sethiya and Chandresh Kumar Maurya, "End-to-End Speech-to-Text Translation Models," *arXiv preprint arXiv:2401.13891*, 2024.