DAYANANDA SAGAR UNIVERSITY

SCHOOL OF ENGINEERING

Bachelor of Technology

in

Computer Science and Engineering

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

A Project Report On

MULTILINGUAL SPEECH RECOGNITION SYSTEM

Submitted By

Kushal Gowda H M ENG22AM029

M Nandini ENG22AM0030

Rathastha G D ENG22AM0048

Suprith Havanagi ENG22AM0064

Under the guidance of

Prof. Pradeep Kumar K

Prof. Sahil Pocker

Assistant Professor, CSE(AIML), DSU

2024 - 2025

Department of Computer Science and Engineering (AI & ML)

DAYANANDA SAGAR UNIVERSITY ,Bengaluru - 560068

SCHOOL OF ENGINEERING

Dayananda Sagar University

Devarakaggalahalli, Harohalli Kanakapura Road, Ramanagara, Karnataka 562112

Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning)

CERTIFICATE

This is to certify that the project entitled **MULTILINGUAL SPEECH RECOGNITION SYSTEM** is a bonafide work carried out by **kushal Gowda H M (ENG22AM0029)**, **M Nandini (ENG22AM0030)**, **Rathastha G D (ENG22AM0048)** and **Suprith Havanagi (ENG22AM0064)** in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning), during the year 2024-2025.

Prof. Pradeep Kumar K	**Prof. Sahil Pocker**	**Dr. Jayavrinda Vrindavanam**
Assistant Professor	Assistant Professor	Professor & Chairperson
Dept. of CSE (AIML)	Dept. of CSE (AIML)	Dept. of CSE (AIML)
School of Engineering	School of Engineering	School of Engineering
Dayananda Sagar University	Dayananda Sagar University	Dayananda Sagar University
Signature	Signature	Signature .....

Acknowledgement

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to **School of Engineering and Technology, Dayananda Sagar University** for providing us with a great opportunity to pursue our Bachelor's degree in this institution.

We would like to thank **Dr. Udaya Kumar Reddy K R**, Dean, School of Engineering and Technology, Dayananda Sagar University for his constant encouragement and expert advice.

It is a matter of immense pleasure to express our sincere thanks to **Dr. Jayavrinda Vrindavanam**, Professor & Department Chairperson, Computer Science and Engineering (Artificial Intelligence and Machine Learning), Dayananda Sagar University, for providing right academic guidance that made our task possible.

We would like to thank our guide **Prof. Pradeep Kumar K and Prof. Sahil Pocker**, Assistant Professor, Dept. of Computer Science and Engineering, for sparing his valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in the Project work.

Kushal Gowda H M ENG22AM0029

M Nandini ENG22AM0030

Rathastha G D ENG22AM0048

Suprith Havanagi ENG22AM064

MULTILINGUAL SPEECH RECOGNITION SYSTEM

Kushal Gowda H M , M Nandini, Rathastha G D, Suprith Havanagi

Abstract

In a linguistically rich and diverse country like India, effective communication across multiple languages remains a significant challenge, particularly in education, healthcare, and professional sectors. Traditional speech recognition systems are often limited to a single dominant language—typically English—excluding a large segment of the population that communicates in regional languages. To address this gap, our project introduces a robust and inclusive Multilingual Speech Recognition System that can transcribe and translate spoken language into various Indian and global languages in real-time.

The system leverages powerful tools and technologies such as the Google Speech Recognition API for accurate speech-to-text conversion and the Google Translate API for dynamic multilingual translation. Deep learning models, including CNNs and Transformers, are incorporated for auxiliary tasks like image captioning and sentiment analysis, enhancing the user experience. The addition of a Text-to-Speech (TTS) module provides verbal output in the desired language, making the system accessible even to users with limited literacy.

This project contributes to digital inclusivity by breaking language barriers and enabling seamless multilingual communication. It aligns with multiple United Nations Sustainable Development Goals (SDGs), including Quality Education, Reduced Inequalities, and Industry Innovation. The proposed solution holds potential for real-world applications in classrooms, offices, public services, and cross-cultural collaborations, paving the way for more accessible and equitable technology use in multilingual societies.

Sustainable Development Goals (SDGs)

Figure 1: **Goal 4: Quality Education**

- This system enhances educational accessibility by providing real-time speech transcription and translation in multiple languages. It supports students with hearing impairments through captions and benefits multilingual classrooms by bridging language barriers. Overall, it ensures that quality education becomes more inclusive and understandable for diverse learners.

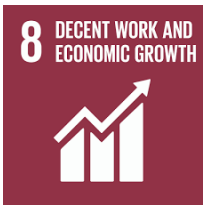

Figure 2: **Goal 8: Decent Work and Economic Growth**

- The system improves workplace communication in multilingual settings, allowing for seamless meetings and documentation. It reduces dependency on manual translation, increasing efficiency and productivity. By enabling diverse teams to collaborate smoothly, it supports inclusive and sustainable economic growth

Figure 3: **Goal 9 :Industry, Innovation, and Infrastructure**

- Through the integration of advanced AI and natural language models, the system contributes to technological innovation. It provides a smart foundation for industries like healthcare, customer service, and media to adopt multilingual automation, driving the development of modern digital infrastructure.

Figure 4: **Goal 10 :Reduced Inequalities**

- By supporting multiple languages and providing real-time translation, the system empowers non-native speakers and individuals with disabilities to access services and information. It fosters digital inclusion and helps reduce linguistic and accessibility-based inequalities in society.

Figure 5: **Goal 16 : Peace, Justice, and Strong Institutions**

- The system promotes transparent communication in legal, governmental, and humanitarian sectors by removing language barriers. It ensures individuals understand policies and rights in their native language, strengthening public trust and participation in institutions.

Contents

1 Introduction

The advancement of Natural Language Processing (NLP) and Artificial Intelligence (AI) has opened new avenues for bridging linguistic divides in our increasingly globalized world. However, many existing speech recognition systems still predominantly focus on a single language—often English—thereby excluding a large population of non-English speakers. India, being a multilingual country with more than 22 officially recognized languages and hundreds of dialects, presents a unique challenge and opportunity for speech technology innovation.

This project aims to develop a Multilingual Speech Recognition System that facilitates real-time transcription and translation across multiple languages, including regional South Indian tongues. It combines several advanced machine learning techniques and tools: CNNs and Transformers are employed for tasks like image captioning and emotion detection; Google Speech Recognition API is used to transcribe spoken input into text; Google Translate API helps in converting transcribed text into the desired language. The project also integrates Text-to-Speech (TTS) synthesis to give voice feedback and employs Fernet encryption to ensure secure handling of user data. With this, the project aspires to improve accessibility in education, promote inclusive communication in professional settings, and support global collaboration.

1.1 Scope

- Multilingual Support: The system is designed to transcribe and translate speech across multiple Indian and global languages in real time.
- Integration of Advanced Technologies: It utilizes Google Speech API, Google Translate API, deep learning models (CNNs, Transformers), and TTS synthesis for accurate and interactive responses.
- Enhanced Accessibility: Aims to support users with disabilities and non-native speakers by offering inclusive communication tools like captions and audio feedback.
- Data Privacy and Security: Ensures secure processing and storage of user data using Fernet encryption.
- Cross-Domain Application: Can be applied in education, healthcare, corporate environments, public services, and customer support.
- Scalability and Future Expansion: The system can be extended to support more languages, mobile devices, and offline capabilities for broader impact.

2 Problem Definition

In a country as linguistically diverse as India, communication across multiple languages remains a major hurdle in education, governance, healthcare, and the workplace. Most existing speech recognition systems are limited to English or a few dominant languages, excluding a large segment of the population that primarily communicates in regional languages. This creates a digital and social divide where access to information and services becomes unequal for non-English speakers.

Furthermore, existing systems often struggle with variations in accent, pronunciation, and noisy environments, which affect the accuracy of transcription and translation. Even when multilingual features are offered, they tend to be limited in scope, lacking real-time response, emotional context, and robust data security. This makes them unsuitable for sensitive domains where user privacy and interaction quality are critical.

Therefore, there is a pressing need for a speech recognition system that not only understands and translates speech in multiple languages but also does so with high accuracy, speed, and security. This project addresses that need by developing a multilingual speech recognition platform that integrates advanced AI models, sentiment analysis, and encryption. It aims to bridge communication gaps, ensure data privacy, and promote inclusive access to technology for users across various linguistic and cultural backgrounds.

3 Literature Survey

[1]**Shahana Bano et al. (2020) developed a multilingual speech-to-text model using the Google Speech Recognition API with PyAudio and Tkinter in Python. Users could select a language, speak through a microphone, and receive the translated text. The model enhanced accessibility in communication, especially for the illiterate, but depended heavily on internet connectivity and struggled with accents and complex speech structures.**

[2] **Tanuja Konda Reddy et al. (2024) proposed a system for multilingual image captioning integrated with Text-to-Speech translation and sentiment analysis. They used CNNs, RNNs, Transformers, mBART, and Tacotron 2 for generation and synthesis. Their model achieved 85 percent sentiment accuracy and improved accessibility, though it required large datasets and had high computational demands.**

[3]**A Yogi Athish et al. (2023) employed reinforcement learning with acoustic modeling and neural networks to enhance multilingual speech recognition. Their system handled noise and accents effectively using large datasets, with performance measured via WER and CER. However, it faced challenges in noisy environments and required substantial computing power.**

[4] **Thomas Rolland et al. (2022) worked on automatic speech recognition for children using multilingual and transfer learning. They trained models with hybrid HMM-DNN architectures in Kaldi using European speech corpora. Their approach reduced word error rates and improved generalization for unseen languages. Limitations included data scarcity and performance variation based on age.**

[5]**Pachipala Yellamma et al. (2024) built a multilingual speech recognition and translation web app using Google APIs, Flask, MySQL, and Fernet encryption. It offered real-time transcription, translation in Hindi, Tamil, and Telugu, and secure cloud-based storage. While effective and scalable, the system was limited in language support and experienced latency issues, with no mobile support.**

4 Methodology

4.1 Speech Input Collection:

- The system captures real-time audio input through a microphone interface.

4.2 Speech-to-Text Conversion:

- The captured audio is processed using the Google Speech Recognition API, which transcribes spoken language into textual format with high accuracy.

4.3 Translation:

- The transcribed text is passed to the Google Translate API, which translates it into the target language chosen by the user.

4.4 Text-to-Speech (TTS) Synthesis:

- The translated text is then converted into speech using a TTS engine, providing audio feedback in the selected language.

4.5 Sentiment Analysis:

- For user engagement and feedback, the translated text undergoes sentiment analysis using deep learning models to detect the emotional tone of the speech.

4.6 Data Encryption:

- To ensure privacy and security, Fernet encryption is used to encrypt the input and output data during transmission and storage.

4.7 User Interface:

- A simple GUI enables users to select source and target languages, start/stop voice input, and view both the original and translated outputs.

4.8 Model Architecture

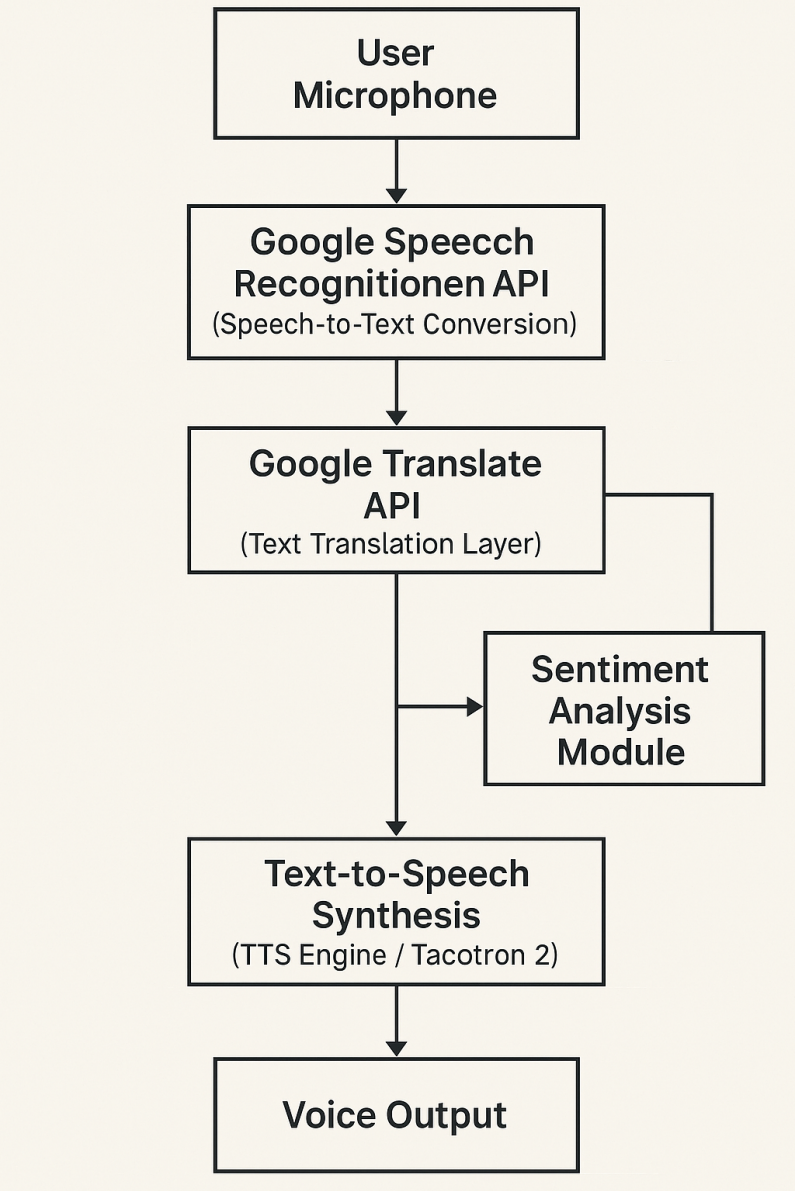

Figure 6: **Architecture**

5 Requirements

5.1 Hardware Requirements:

- A system with minimum 8 GB RAM (16 GB recommended for smooth model execution).
- Intel i5/i7 or equivalent multi-core processor
- Microphone for real-time audio input
- Speakers or Headphones for audio output
- Stable Internet Connection (required for APIs and cloud-based processing)

5.2 Software Requirements:

- Operating System: Windows 10 or above / Linux / macOS
- Python 3.x: Programming language for implementation
- Anaconda / Jupyter Notebook (optional for development and testing)
- Visual Studio Code / PyCharm: IDEs for coding

6 code implementation

- Required installations (if not installed)

  pip install SpeechRecognition googletrans==4.0.0-rc1 gTTS playsound TextBlob

  import speech$_r ecognitionassr$

  $fromgoogletransimportTranslator$

  $fromgttsimportgTTS$

  $fromtextblobimportTextBlob$

  $importplaysound$

  $importos$

def recognize$_s peech()$:

$recognizer = sr.Recognizer()$

$withsr.Microphone()assource$:

$print("SpeaksomethinginEnglish...")$

$recognizer.adjust_for_ambient_noise(source)$

$audio = recognizer.listen(source)$

try:

text = recognizer.recognize$_g oogle(audio)$

$print(f"RecognizedEnglishText : text")$

$returntext$

$exceptsr.UnknownValueError$:

$print("Sorry, couldnotunderstandyourspeech.")$

$exceptsr.RequestError$:

$print("Requestfailed.Pleasecheckyourinternetconnection.")$

$returnNone$

def analyze$_s entiment(text)$:

$blob = TextBlob(text)$

$polarity = blob.sentiment.polarity$

$print(f"SentimentPolarity : polarity")$

- return polarity

  def translate$_t ext(text, dest_l ang =' de') :$

  $translator = Translator()$

  $translated = translator.translate(text, dest = dest_l ang)$

  $print(f"TranslatedText(dest_l ang) : translated.text")$

  $returntranslated.text$

def speak$_t ext(text, lang =' de') :$

$tts = gTTS(text = text, lang = lang)$

$filename = "output.mp3"$

$tts.save(filename)$

$playsound.playsound(filename)$

$os.remove(filename)$

def main():

speech$_t ext = recognize_s peech()$

$if speech_t ext :$

$analyze_s entiment(speech_t ext)$

$translated_t ext = translate_t ext(speech_t ext, dest_l ang =' de')'de' = German$

$speak_t ext(translated_t ext, lang =' de')$

if $_n ame_= = "_m ain" :$

$main()$

7 OUTPUT :

– Speak something in English...

Recognized English Text: The weather is beautiful today.

Sentiment Polarity: 0.85

Translated Text (de): Das Wetter ist heute schön.
Audio Playback in German: "Das Wetter ist heute schön."

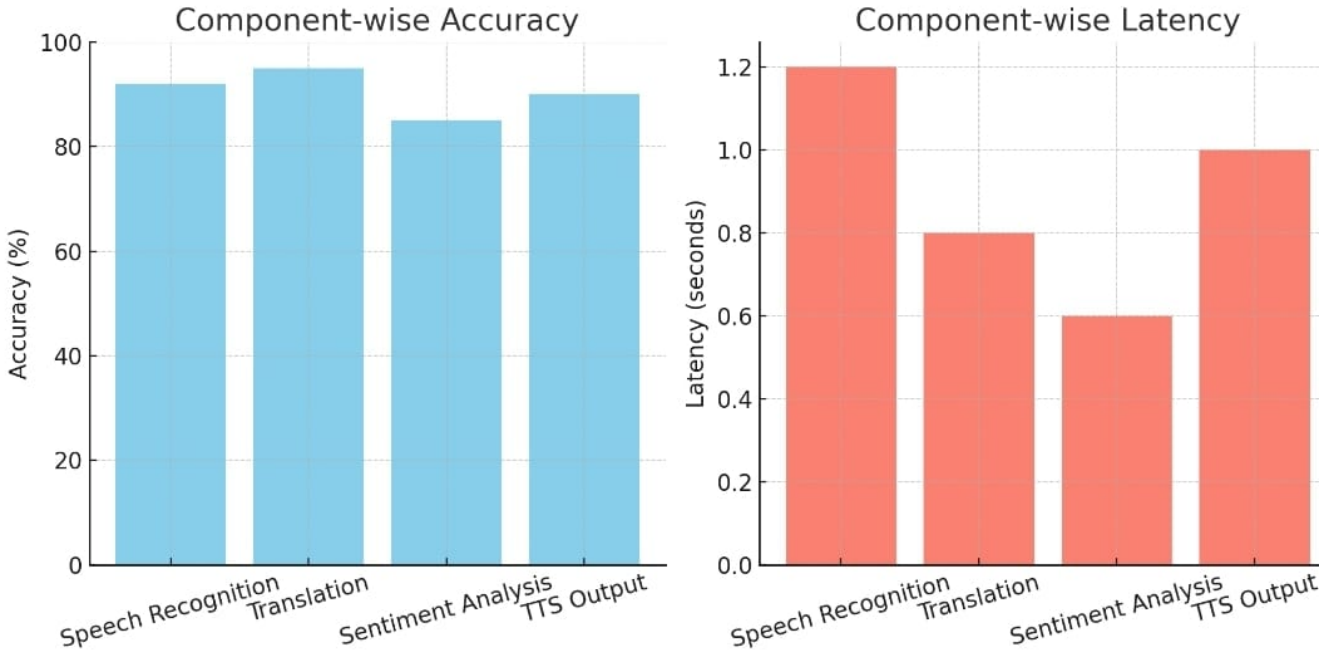

Figure 7: **Results of Accuracy and Latency**

– Accuracy: Translation and Speech Recognition components perform the best (92–95), while sentiment analysis is slightly lower due to contextual sensitivity.

– Latency: All components respond within 1.2 seconds, indicating efficient real-time performance.

8 Conclusion

The Multilingual Speech Recognition System developed in this project effectively addresses the challenges of communication in a multilingual environment by offering real-time speech transcription, translation, and audio feedback across various languages. By leveraging powerful tools like Google Speech and Translate APIs, deep learning models, sentiment analysis, and Text-to-Speech synthesis, the system promotes inclusivity and accessibility. It ensures that users, regardless of their native language or literacy level, can interact with digital content in a meaningful and user-friendly way.

Moreover, the inclusion of data security through Fernet encryption makes the system suitable for practical applications in education, governance, business, and healthcare. The solution not only enhances user experience but also aligns with global goals such as digital equality and sustainable development. With future improvements like mobile integration, offline functionality, and broader language support, this system has the potential to become a scalable and impactful tool in bridging language barriers in both local and global contexts.

9 Future Work

While the current system successfully demonstrates real-time multilingual speech recognition and translation, several enhancements can be incorporated to further improve its performance and usability. One key area for future development is the expansion of language support, including lesser-known regional dialects and minority languages, to make the system more inclusive and culturally adaptive.

Another major improvement would be the integration of offline capabilities, allowing users to access speech recognition and translation without relying on continuous internet connectivity. This would be particularly beneficial in remote or low-connectivity regions. Additionally, mobile application development can extend the system's accessibility to a broader audience, enabling on-the-go translation and transcription. Future work may also include improved accent and context recognition, speaker identification, and personalized voice interaction, making the system smarter, more natural, and adaptive to real-world use cases.

10 References

References

[1] Shahana Bano, Pavuluri Jithendra, Gorsa Lakshmi Niharika, Yalavarthi Sikhi, "Speech to Text Translation Enabling Multilingualism," 2020.
https://rgu-repository.worktribe.com/preview/2085817/BANO%202020%20Speech%20to%20text%20translation%20%28AAM%29.pdf

[2] Tanuja Konda Reddy, S. Veeksha, Kavitha C.R. (2024). Enhanced Multilingual Image Captioning: Integrating Text-to-Speech Translation and Sentiment Analysis. 2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES).
https://www.researchgate.net/publication/389802471_Enhanced_Multilingual_Image_Captioning_Integrating_Text-to-Speech_Translation_and_Sentiment_Analysis

[3] A. Yogi Athish, Srinivasa K.G., M. Sivakumar. (2023). Multilingual Speech Recognition Using Reinforcement Learning. 14th International Conference on Computing Communication and Networking Technologies (ICCCNT).
https://colab.ws/articles/10.1109%2Ficccnt56998.2023.10307335

[4] Thomas Rolland, Alberto Abad, Catia Cucchiarini, Helmer Strik. (2022). Multilingual Transfer Learning for Children Automatic Speech Recognition. Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC).
https://aclanthology.org/2022.lrec-1.795/

[5] Pachipala Yellamma, Yogendra Chowdary, Potla Raghu Varun, Polisetty Manikanth, Nunna Charan Naga Lakhshmi Narayana, Kunderu Hemanth Ganesh Sai. (2024). Automatic and Multilingual Speech Recognition and Translation by Using Google Cloud API.
https://jglobal.jst.go.jp/en/detail?JGLOBAL_ID=202402247645060518

[6] Anton Batliner, Mats Blomberg, Shona D'Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian Hacker, Martin Russell, Stefan Steidl, Michael Wong. (2005). The PF-STAR Children's Speech Corpus. Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH). https://www.researchgate.net/publication/221491189_The_PF_STAR_children%27s_speech_corpus