TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MACHINE LEARNING

- **TITLE:** Multilingual image and speech translator: A unified approach for language conversion from images and speech

By

Paramkusham hasini	-	20K91A05D5
Middela Naresh	-	21K95A0523
Muthyam saikiran	-	20K91A05C2

UNDER THE GUIDANCE OF -
MRS.TULASI RATNA MANI

INDEX

- Introduction
- Existing system
- Proposed System
- Functional Requirements
- Non-Functional Requirements
- Literature survey
- System Architecture
- Data Flow Diagrams
- UML Diagrams
- Modules
- Algorithm
- Output Screenshots
- Test Cases
- Conclusion
- Future Work

Introduction

- In an increasingly interconnected world, the ability to bridge linguistic barriers is paramount for effective communication and collaboration. The proliferation of digital media has introduced new challenges and opportunities in the realm of language translation, particularly with the integration of visual and audio information. The project addresses this pressing need by presenting a comprehensive solution that seamlessly translates both images and speech across multiple languages.
- The advent of deep learning and artificial intelligence has revolutionized the field of language processing, enabling the development of sophisticated translation systems. However, existing approaches often focus on either textual or audio data, overlooking the valuable contextual information conveyed through images. Conversely, standalone image recognition systems may struggle to interpret the nuanced meanings embedded in multilingual content.
- This project proposes a unified approach that synergistically combines image processing techniques with state-of-the-art speech recognition and natural language processing algorithms. By harnessing the complementary strengths of these modalities, our system aims to provide accurate and contextually relevant translations for a wide range of visual, audio and text inputs.

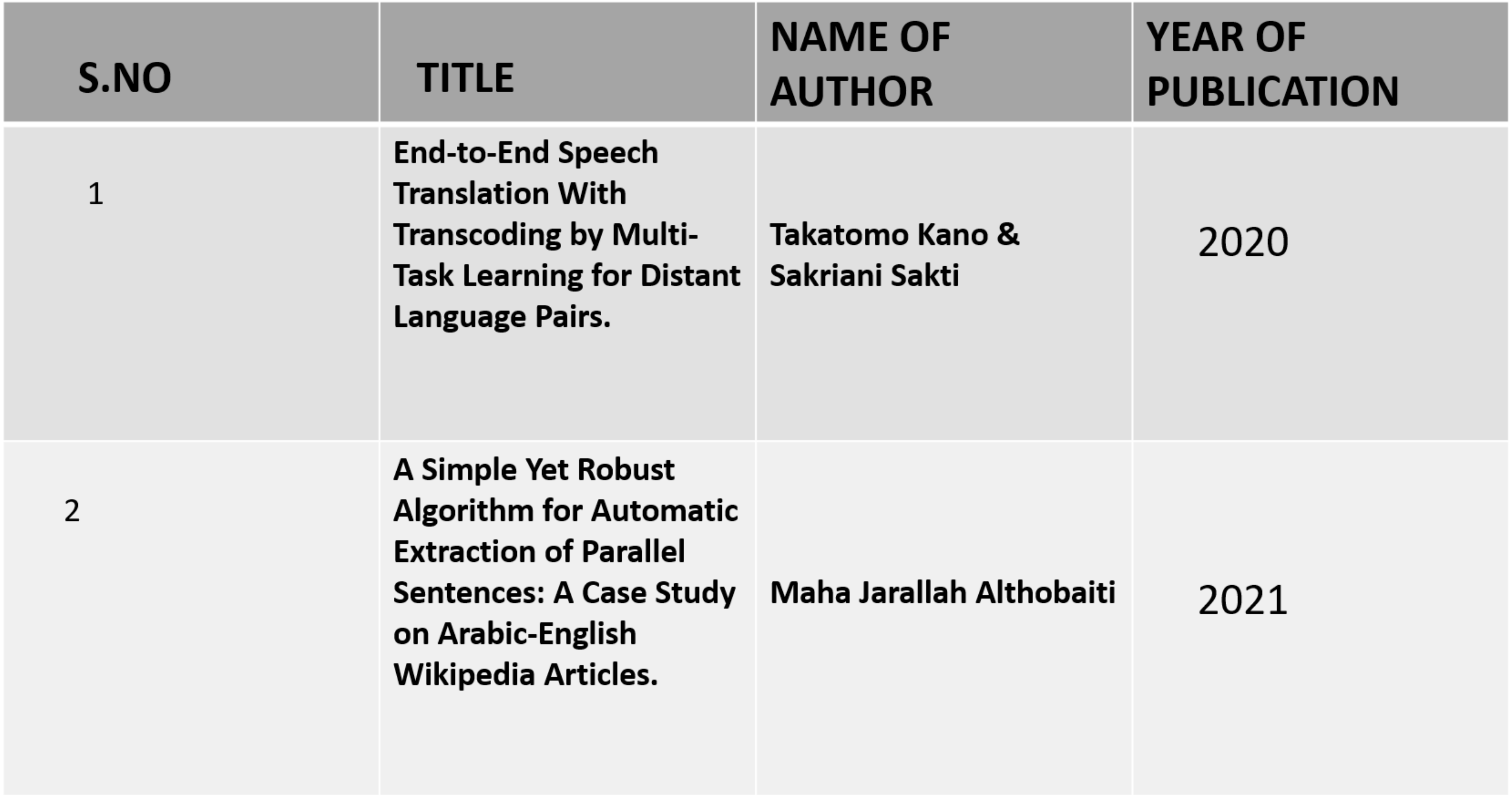
Existing System

- **Limited Accuracy in Image Recognition:** Despite advancements in image processing techniques, the system may still struggle with accurately recognizing text in complex or low-quality images. This limitation can result in errors or incomplete translations, particularly when dealing with handwritten text, distorted images, or non-standard fonts.
- **Speech Recognition Challenges:** While speech recognition technology has made significant strides, it may encounter difficulties in accurately transcribing speech in noisy environments or with speakers who have accents or speech impediments. In such cases, the accuracy of speech-to-text conversion may be compromised, leading to inaccuracies in the translation output.
- **Lack of Support for Multi-Languages:** The system's language support may be limited to single language, overlooking less widely spoken or minority languages. Users who speak or require translation for languages not supported by the system may experience difficulties in accessing accurate translations, thus limiting the system's inclusivity and accessibility.

Proposed System

The proposed system aims to revolutionize multilingual communication by offering a unified approach for translating content from both images and speech into multiple languages. Leveraging advanced image processing, speech recognition, and natural language processing techniques, this system seamlessly integrates diverse modalities to deliver accurate and contextually relevant translations. By breaking down linguistic barriers and embracing a holistic approach to language conversion, the proposed system empowers users to communicate effectively across linguistic boundaries.

LITERATURE SURVEY



S.NO	TITLE	NAME OF AUTHOR	YEAR OF PUBLICATION
1	End-to-End Speech Translation With Transcoding by Multi-Task Learning for Distant Language Pairs.	Takatomo Kano & Sakriani Sakti	2020
2	A Simple Yet Robust Algorithm for Automatic Extraction of Parallel Sentences: A Case Study on Arabic-English Wikipedia Articles.	Maha Jarallah Althobaiti	2021

S.NO	TITLE	NAME OF THE AUTHOR	YEAR OF PUBLICATION
3	**Entity Highlight Generation as Statistical and Neural Machine Translation.**	**Jizhou Huang**	2018
4	**RESHAPE: Reverse-Edited Synthetic Hypotheses for Automatic Post-Editing.**	**Wonkee Lee**	2022
5	**IDF-Sign: Addressing Inconsistent Depth Features for Dynamic Sign Word Recognition**	**Sunusi bala abdullahi**	2023

1. Title : End-to-End Speech Translation With Transcoding by Multi-Task Learning for Distant Language Pairs

- **Author :** Takatomo Kano & Sakriani Sakti

- **Abstract :** Directly translating spoken utterances from a source language to a target language is challenging because it requires a fundamental transformation in both linguistic and para/non-linguistic features.

- Traditional speech-to-speech translation approaches concatenate automatic speech recognition (ASR), text-to-text machine translation (MT), and text-to-speech synthesizer (TTS) by text information. The current state-of-the-art models for ASR, MT, and TTS have mainly been built using deep neural networks, in particular, an attention-based encoder-decoder neural network with an attention mechanism. Recently, several works have constructed end-to-end direct speech-to-text translation by combining ASR and MT into a single model.

2. Title : A Simple Yet Robust Algorithm for Automatic Extraction of Parallel Sentences: A Case Study on Arabic-English Wikipedia Articles

- **Name of the author :** Maha Jarallah Althobaiti

- **Year of publication :** 2021

- **Abstract :** Parallel corpora are vital components in several applications of Natural Language Processing (NLP), particularly in machine translation. In this , we present a novel method for automatically creating parallel sentences from comparable corpora.
- The method requires a bilingual dictionary as well as an enough word-vectorization method. We use Arabic and English Wikipedia as a comparable collection to apply our proposed method and construct a parallel collection between Arabic and English. During our study, we compared two methods of word vectorization, word embedding and term frequency-inverse document frequency, in terms of their usefulness in computing similarities between well-formed and syntactically ill-formed sentences.

3. Title:-Entity Highlight Generation as Statistical and Neural Machine Translation

- **Author:-** Jizhou zhang
- **Introduction:-**Entity highlights are essential for applications such as web search results enrichment, entity recommendation in web search engines, and named entity disambiguation. The article focuses on the task of entity highlight generation, aiming to create a short and characteristic description from a descriptive sentence about an entity.

4. Title:-Reverse-Edited Synthetic Hypotheses for Automatic Post-Editing

- **Author:-** Wonkee lee

- **Introduction:-**Synthetic training data has been widely employed in training Automatic Post-Editing (APE) models due to perceived insufficiency in human-created data. This paper addresses a limitation in the widely-used synthetic APE dataset, eSCAPE, which overlooks the minimal editing property of genuine data, potentially affecting APE model performance. We propose RESHAPE, a new synthetic APE dataset generated through back-translation, integrating stochastic sampling during decoding to maintain output diversity. Experimental results demonstrate that RESHAPE outperforms eSCAPE, showing a higher resemblance to genuine APE data and significantly improving APE model performance.

5. Title:- Hiformer Sequence Modeling Networks With Hierarchical Attention Mechanisms.

- **Author:-** sunusi bala abdullahi
- **Introduction:-**The attention-based encoder-decoder structure, exemplified by the Transformer, has achieved remarkable success in sequence modeling tasks such as machine translation (MT) and automatic speech recognition (ASR). However, the conventional Transformer's self-attention mechanism primarily focuses on intra-layer integration and lacks explicit modeling of inter-layer information relationships. This paper introduces Hiformer, a sequence modeling structure equipped with a hierarchical attention mechanism, addressing this limitation. Hiformer considers both inter-layer and cross-coder hierarchical information, enhancing structured prediction performance. Experimental results on MT and ASR tasks demonstrate the effectiveness of Hiformer over the Transformer.

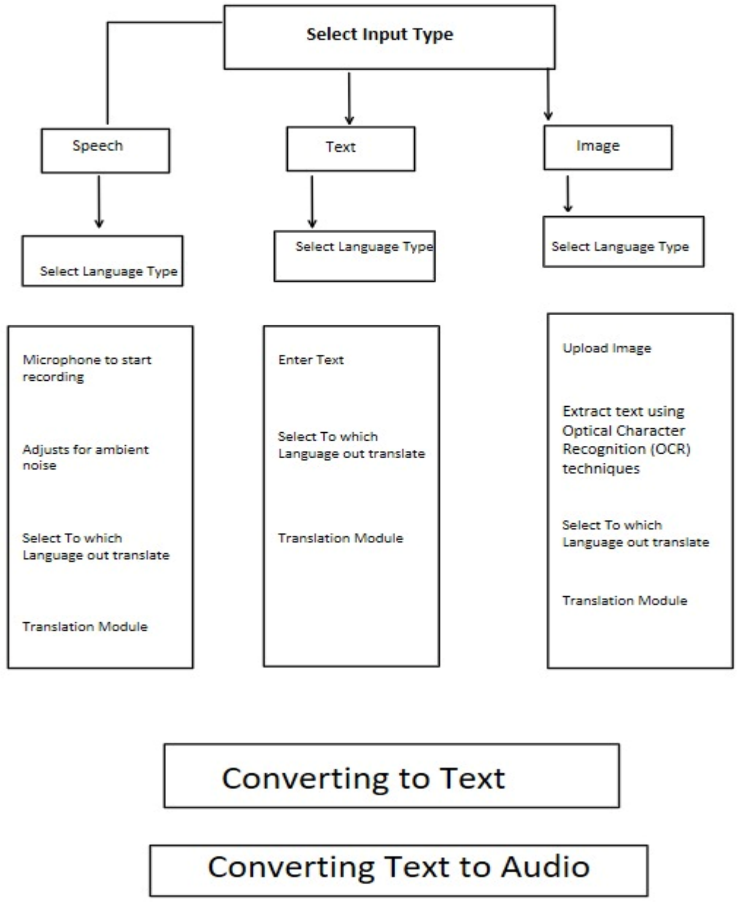
Functional Requirements

- **Image-to-Text Conversion**: The system should accurately convert text contained within images into machine-readable text for translation.
- **Speech Recognition**: The system must be capable of accurately transcribing spoken language into text for translation.
- **Language Translation**: The system should translate text from both images and speech into the desired target language(s) with high accuracy and contextual relevance.
- **Cross-Modal Fusion**: The system should integrate information from images and speech to enhance translation accuracy and capture contextual nuances.
- **Multilingual Support**: The system must support translation between multiple languages to cater to diverse linguistic needs.
- **User Interface**: The system should provide an intuitive and user-friendly interface for users to input images, speech, or text, select languages, and view translated output.

Non-Functional Requirements

- **Performance**: The system should provide fast and responsive translation services, with minimal latency, to enhance user experience.
- **Accuracy**: The system must achieve high accuracy in both image-to-text conversion and speech recognition to ensure reliable translation output.
- **Scalability**: The system should be scalable to accommodate a growing user base and increasing translation demands without compromising performance or reliability.
- **Reliability**: The system should operate reliably under various conditions, including network disruptions, to ensure continuous availability of translation services.
- **Robustness**: The system should be resilient to noise, distortions, and variations in image and speech inputs to maintain translation accuracy across diverse scenarios.
- **Compatibility**: The system should be compatible with a wide range of devices and platforms, including desktop computers, mobile devices, and web browsers, to maximize accessibility for users.
- **Adaptability**: The system should be adaptable to different domains, user preferences, and linguistic contexts, allowing for personalized translation experiences.
- **Documentation and Support**: The system should provide comprehensive documentation and user support resources to assist users in effectively utilizing its features and functionalities.

System Architecture



Select Input Type

Speech

Text

Image

Select Language Type

Select Language Type

Select Language Type

Microphone to start recording

Adjusts for ambient noise

Select To which Language out translate

Translation Module

Enter Text

Select To which Language out translate

Translation Module

Upload Image

Extract text using Optical Character Recognition (OCR) techniques

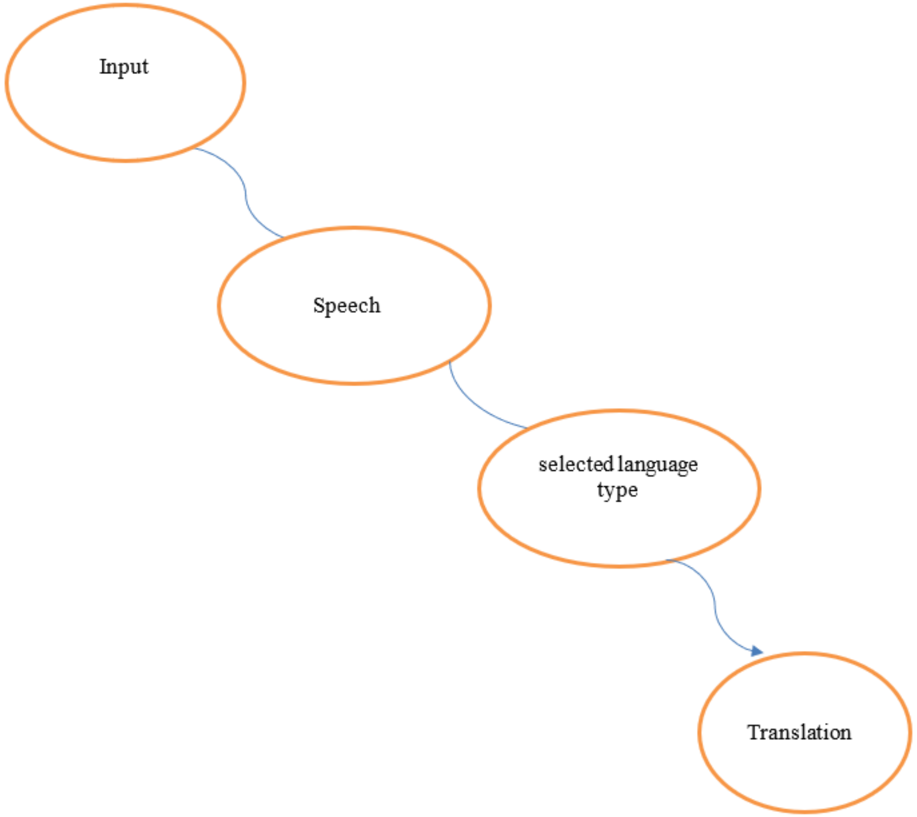
Select To which Language out translate

Translation Module

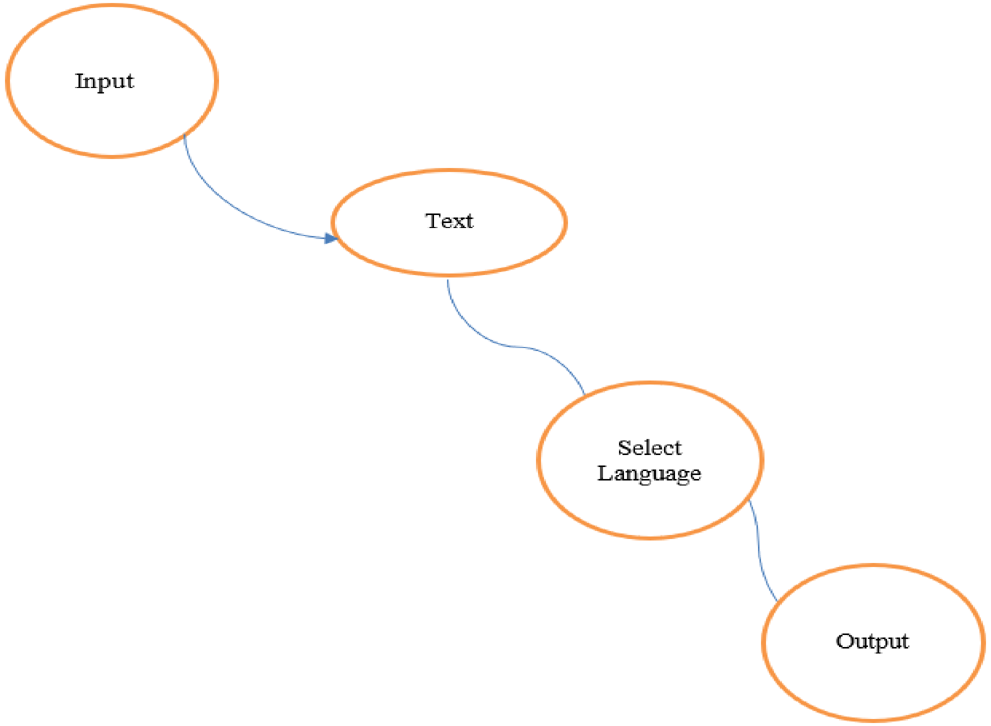
Converting to Text

Converting Text to Audio

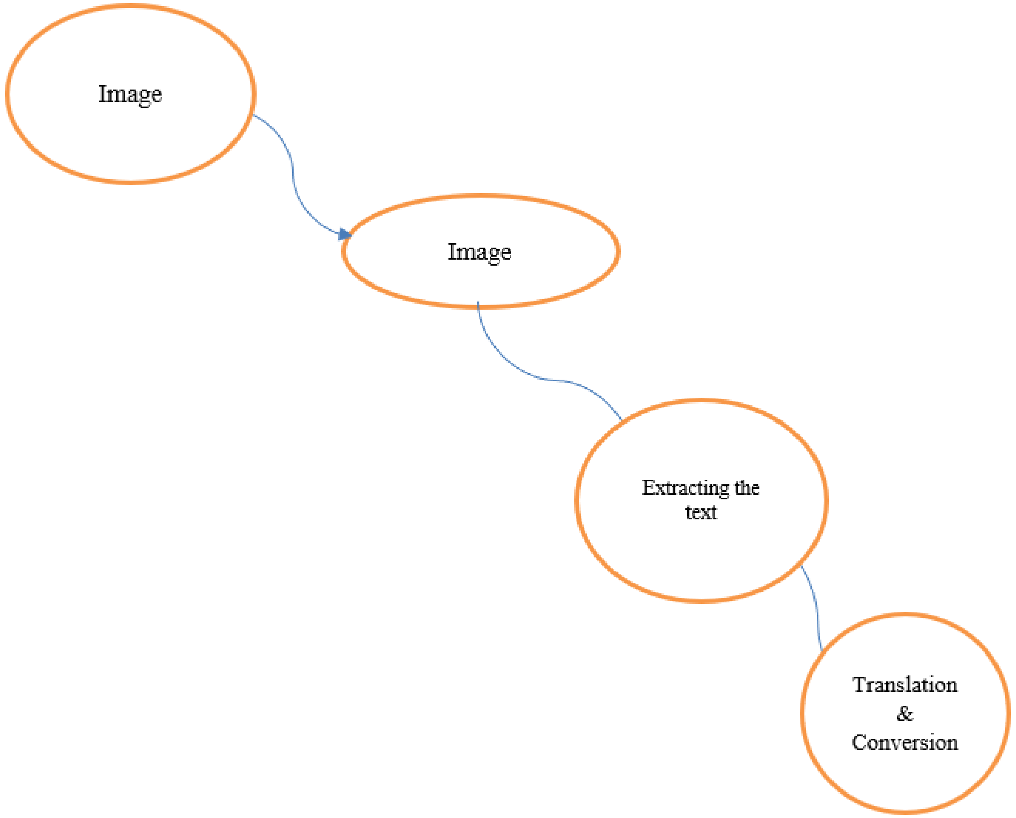
Data Flow Diagram-0



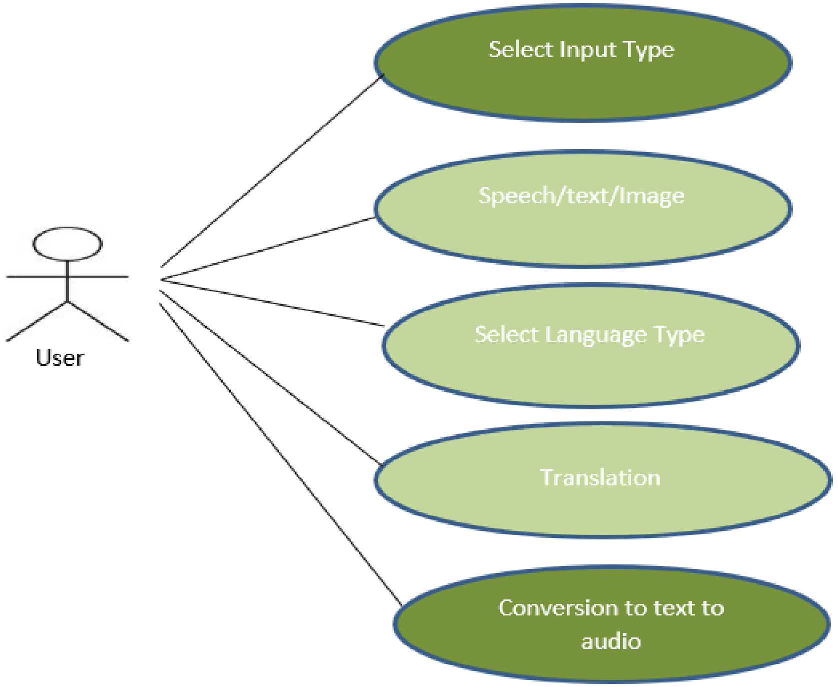
Data Flow Diagram-1



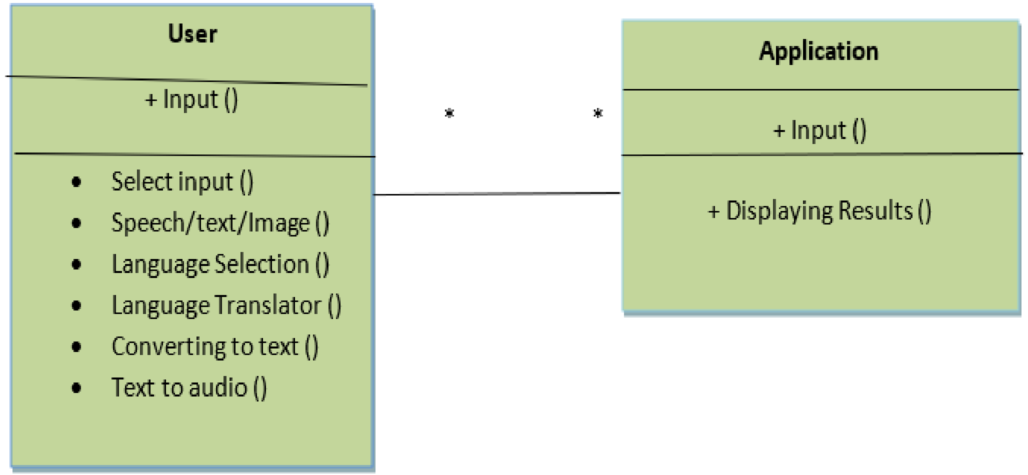
Data Flow Diagram-2



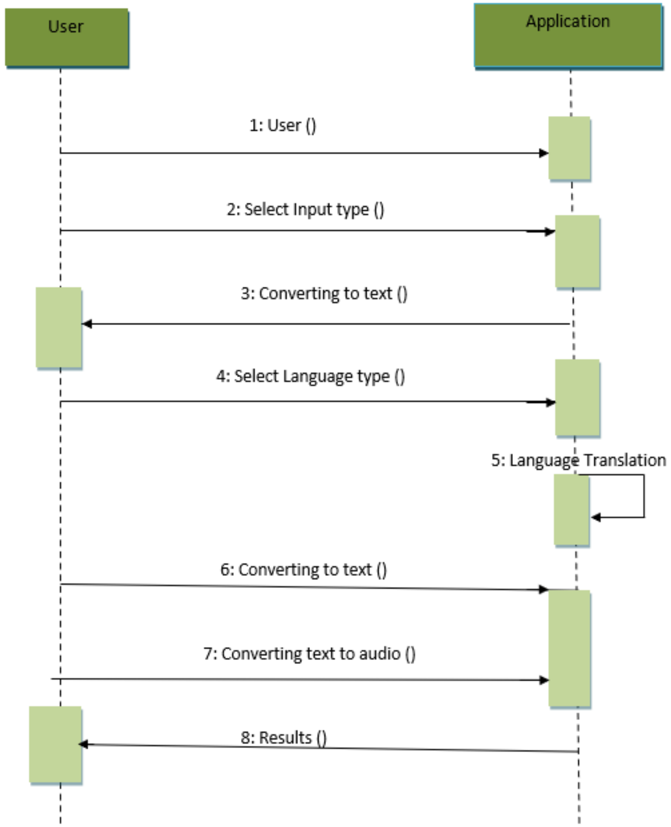
Use Case Diagram



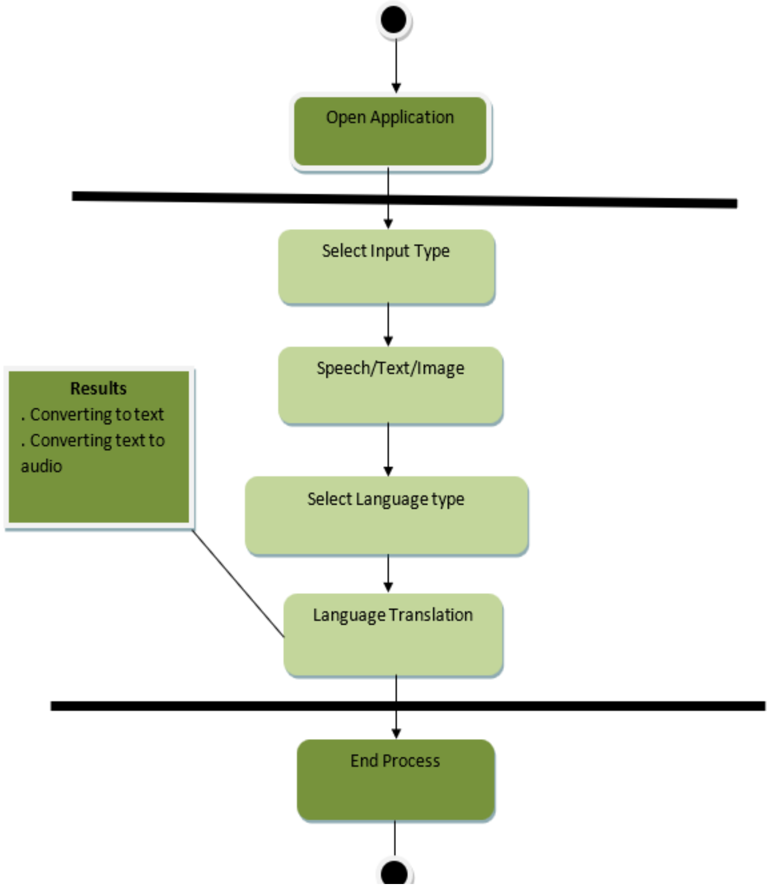
Class Diagram



Sequence Diagram



Activity Diagram



Modules

- **Image Processing Module**: Processes uploaded images to extract text using optical character recognition (OCR) techniques.
- **Speech Recognition Module:** Transcribes spoken language input into machine-readable text using speech recognition algorithms.
- **Translation Engine Module:** Translates text obtained from images and speech inputs into target languages, incorporating contextual understanding and supporting multiple languages.
- **Cross-Modal Fusion Module:** Integrates information from images and speech inputs to enhance translation accuracy and capture contextual nuances.
- **User Interface Module:** Provides an intuitive interface for users to interact with the system, select languages, and view translated output.

Pseudo Code

```
import streamlit as st
import pytesseract
import speech_recognition as spr
from googletrans import Translator
from gtts import gTTS
import os
import time
from PIL import Image

# Function to convert image to text
def image_to_text(image):
    text = pytesseract.image_to_string(image)
    return text

# Function to recognize speech input
def recognize_speech(recognizer, microphone, timeout=3):
    with microphone as source:
        st.write("Get ready to speak! Recording will start in:")
        for i in range(3, 0, -1):
            st.write(i)
            time.sleep(1)

        st.write("Speak now to initiate the Translation!")
        recognizer.adjust_for_ambient_noise(source, duration=0.2)
        audio = recognizer.listen(source, timeout=timeout)

        try:
            text = recognizer.recognize_google(audio).lower()
            return text
        except spr.WaitTimeoutError:
            return ""
```

Pseudo Code

```
# Function to take input text from the user
def input_text():
    return st.text_input("Enter the text to be translated: ")

# Function to translate text to the selected language
def translate_text(text, from_lang, to_lang):
    translator = Translator()
    translation = translator.translate(text, src=from_lang, dest=to_lang)
    translated_text = translation.text
    return translated_text

# Function to convert translated text to speech
def text_to_speech(text, lang):
    tts = gTTS(text=text, lang=lang, slow=False)
    tts.save("captured_voice.mp3")
    st.audio("captured_voice.mp3", format="audio/mp3")

# Main function to run the Streamlit app
def main():
    recognizer = spr.Recognizer()
    microphone = spr.Microphone()

    # Set title and page layout
    st.title("Language Translator")
    st.image("your.png", use_column_width=True)  # Add your image URL here

    # Select boxes for language selection and input method
    from_lang_option = st.selectbox("Select the language you are speaking:", ["English", "Hindi", "Tamil", "Telugu", "Kannada", "Malayalam", "Urdu"])
    to_lang_option = st.selectbox("Select the language you want to translate to:", ["English", "Hindi", "Tamil", "Telugu", "Kannada", "Malayalam", "Urdu"])
    input_method = st.radio("Select the input method:", ["Speech", "Text", "Image"])  # Added "Image" option

    # Default text value
    text = ""
```

Pseudo Code

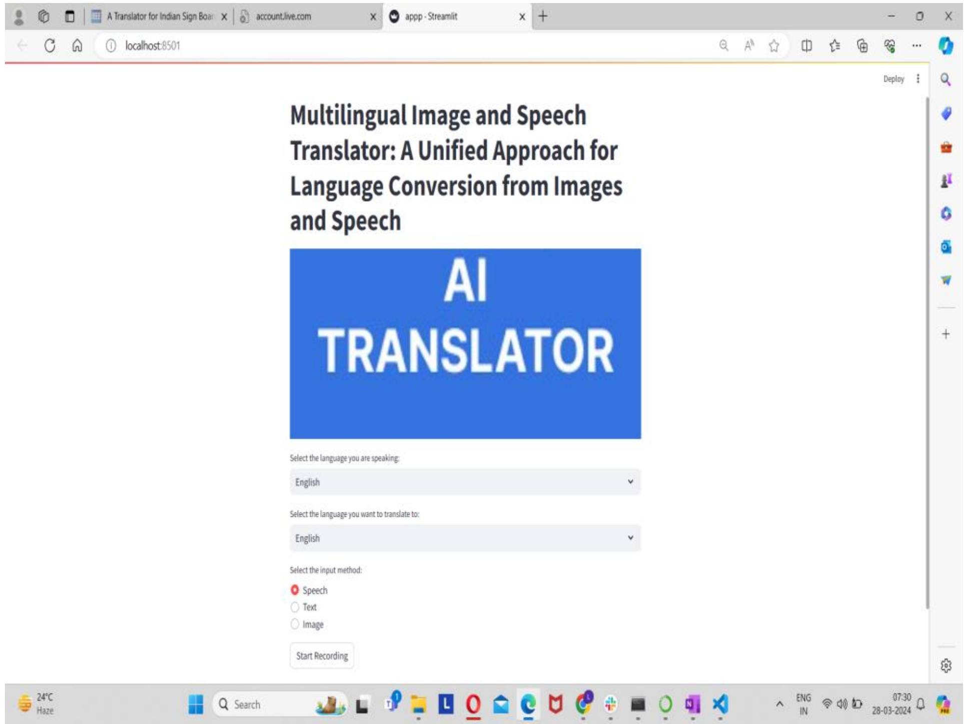
```
# Mapping language options to language codes
    language_codes = {"English": "en", "Hindi": "hi", "Tamil": "ta", "Telugu": "te", "Kannada": "kn", "Malayalam": "ml", "Urdu": "ur"}
    from_lang = language_codes.get(from_lang_option)
    to_lang = language_codes.get(to_lang_option)

    # Based on input method, either recognize speech, take input text, or process image
    if input_method == "Speech":
        if st.button("Start Recording"):
            text = recognize_speech(recognizer, microphone)
            st.write("Recording Stopped!")
    elif input_method == "Text":
        text = input_text()
    elif input_method == "Image":
        uploaded_file = st.file_uploader("Upload image", type=["jpg", "jpeg", "png"])
        if uploaded_file is not None:
            image = Image.open(uploaded_file)
            st.image(image, caption="Uploaded Image", use_column_width=True)
            text = image_to_text(image)

    # Translate the input text and display the translated text
    if text:
        translated_text = translate_text(text, from_lang, to_lang)
        st.write("Translated Text:", translated_text)
        text_to_speech(translated_text, to_lang)

if __name__ == "__main__":
    main()
```

Output Screenshots



Multilingual Image and Speech Translator: A Unified Approach for Language Conversion from Images and Speech

Select the language you are speaking:

English

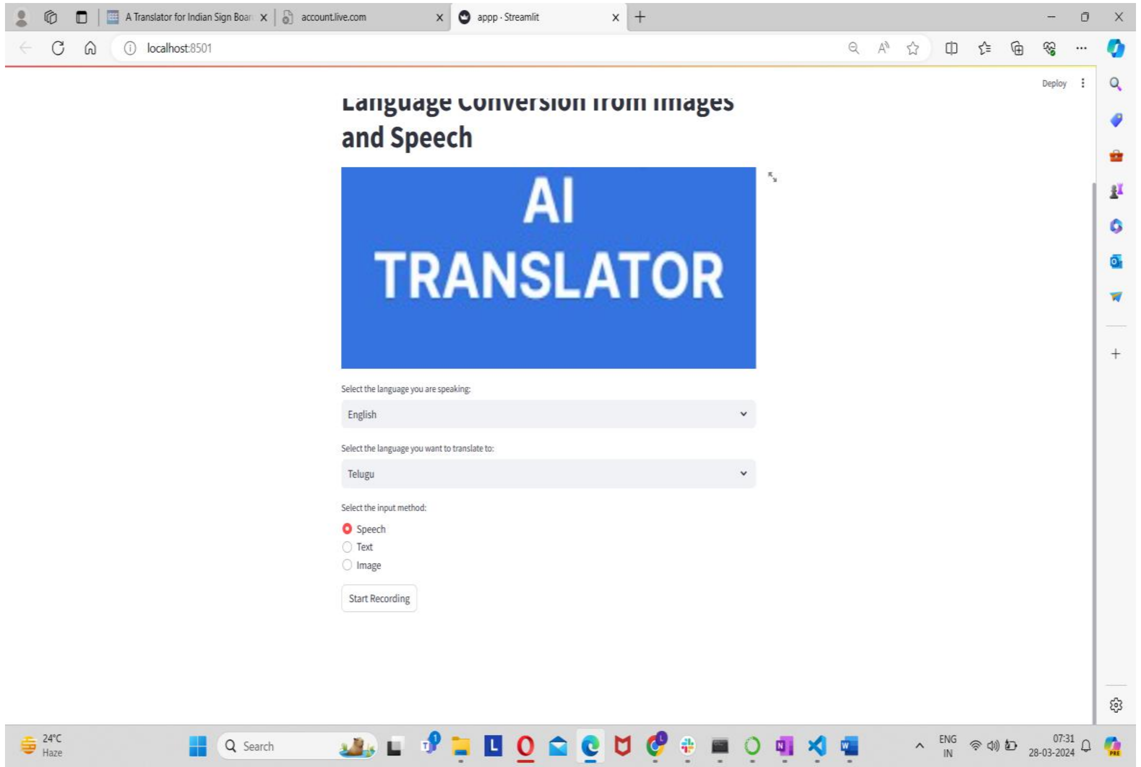
Select the language you want to translate to:

English

Select the input method:

● Speech
○ Text
○ Image

Start Recording



Language Conversion from Images and Speech

Select the language you are speaking:

English

Select the language you want to translate to:

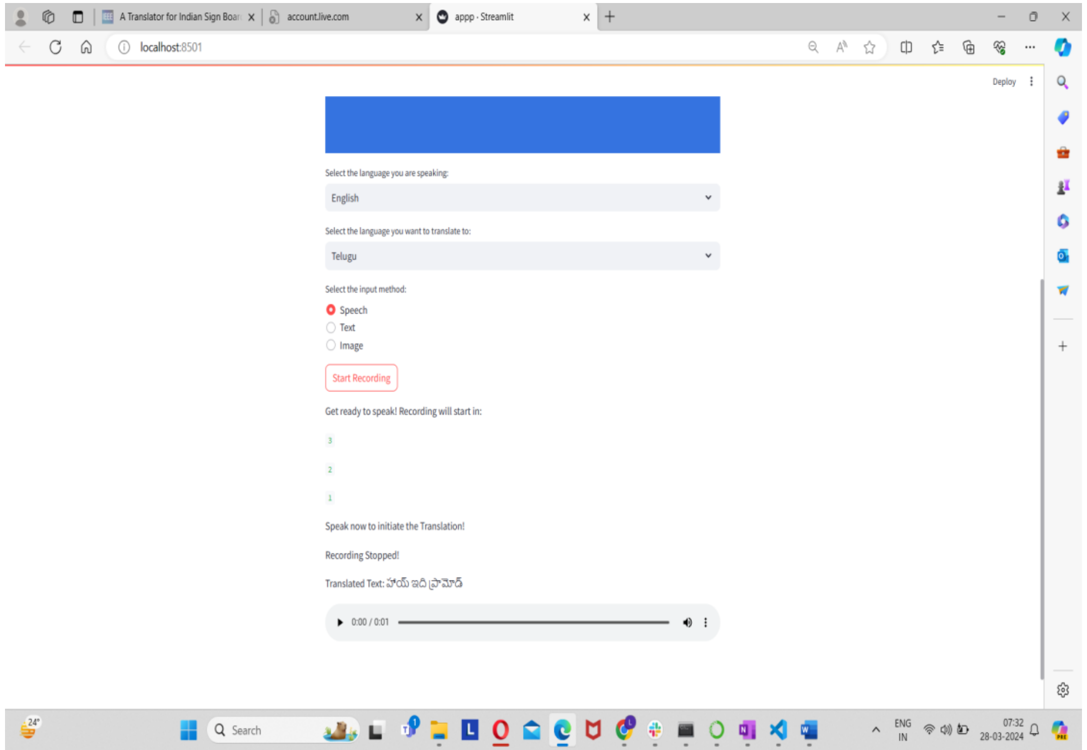
Telugu

Select the input method:

● Speech
○ Text
○ Image

Start Recording

Output Screenshots



Select the language you are speaking:

English

Select the language you want to translate to:

Telugu

Select the input method:

● Speech
○ Text
○ Image

Start Recording

Get ready to speak! Recording will start in:

3

2

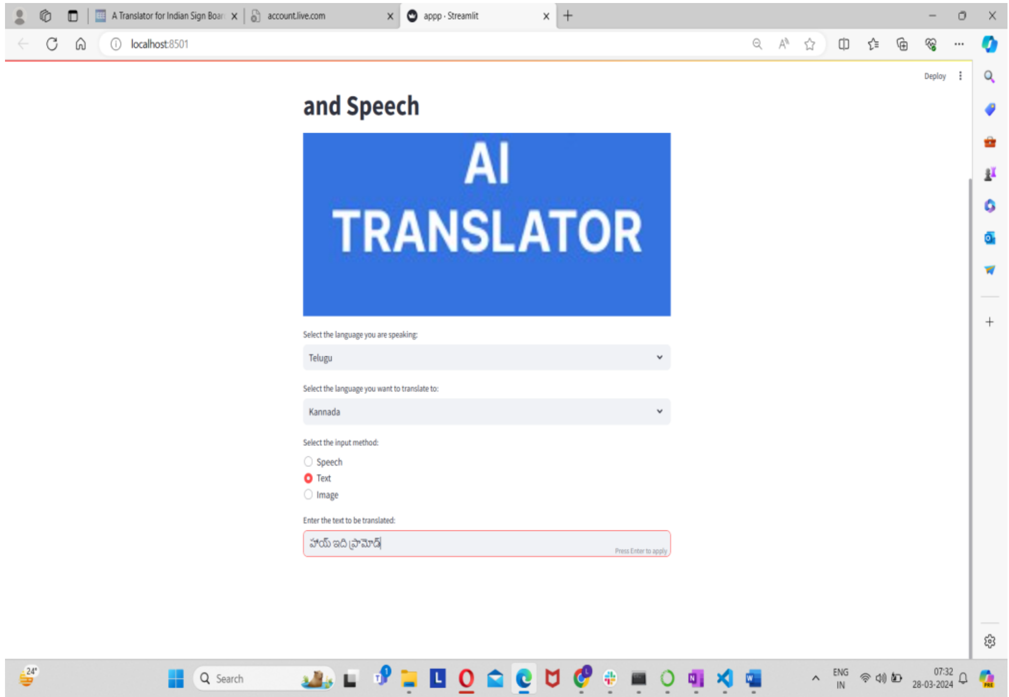
1

Speak now to initiate the Translation!

Recording Stopped!

Translated Text: హాయ్ ఇది ప్రామోద్

0:00 / 0:01



and Speech

Select the language you are speaking:

Telugu

Select the language you want to translate to:

Kannada

Select the input method:

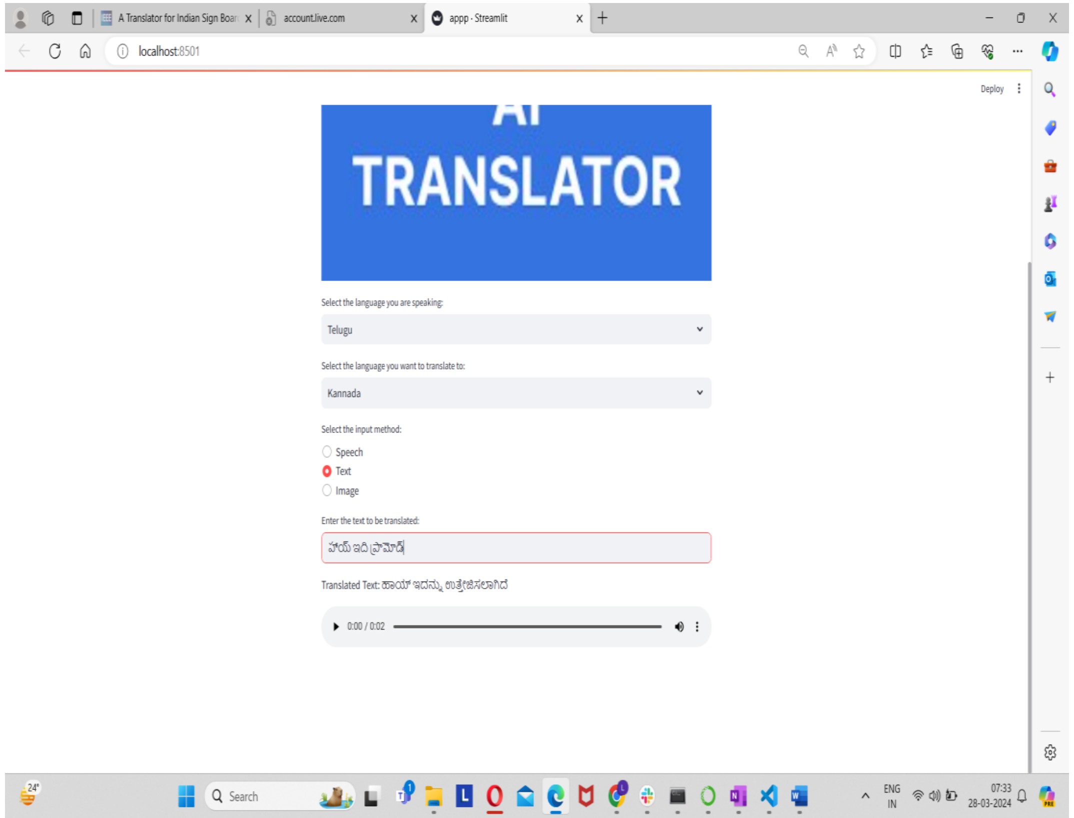
○ Speech
● Text
○ Image

Enter the text to be translated:

హాయ్ ఇది ప్రామోద్

Press Enter to apply

Output Screenshots



localhost:8501

Select the language you are speaking: Telugu

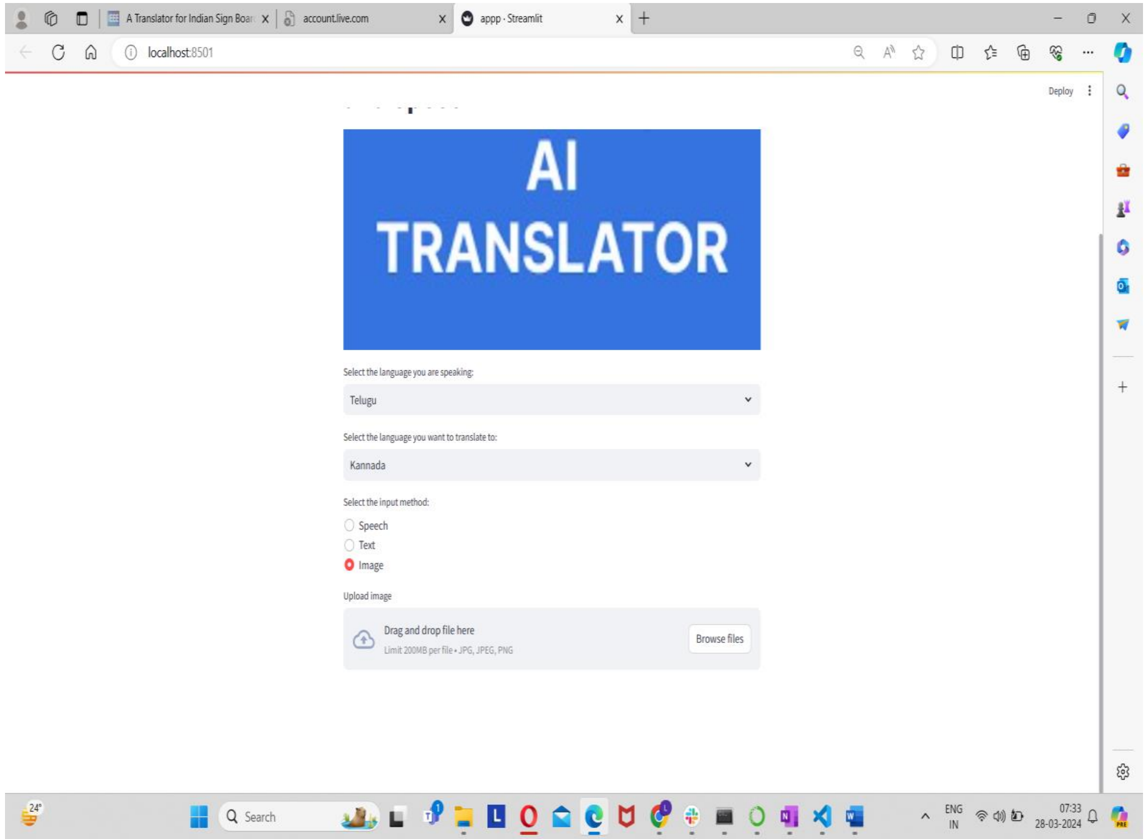
Select the language you want to translate to: Kannada

Select the input method:

○ Speech
● Text
○ Image

Enter the text to be translated: హాయ్ ఇది ప్రామోద్

Translated Text: ಹಾಯ್ ಇದನ್ನು ಉತ್ತೇಜಿಸಲಾಗಿದೆ



localhost:8501

Select the language you are speaking: Telugu

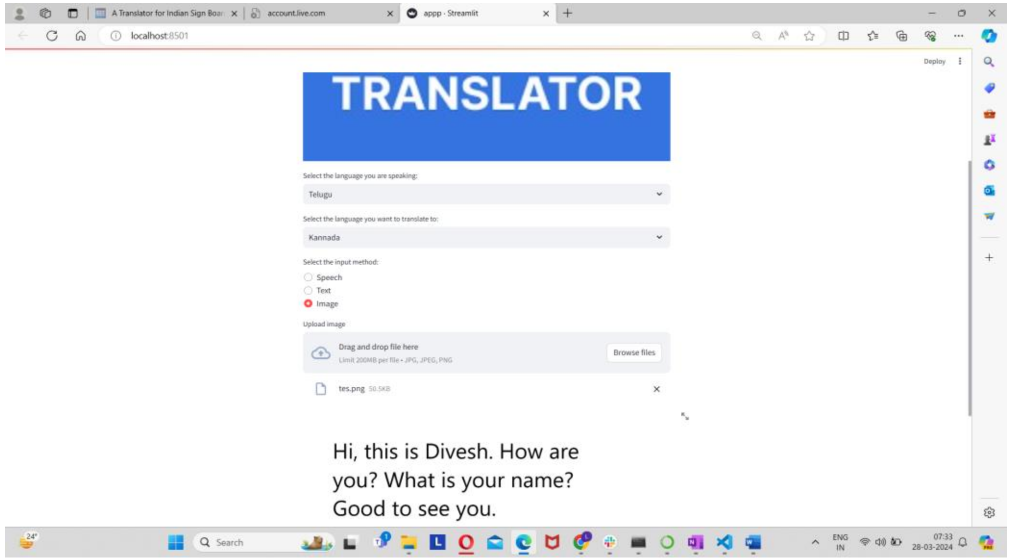
Select the language you want to translate to: Kannada

Select the input method:

○ Speech
○ Text
● Image

Upload image

Output Screenshots



Select the language you are speaking:

Telugu

Select the language you want to translate to:

Kannada

Select the input method:

○ Speech
○ Text
● Image

Upload image

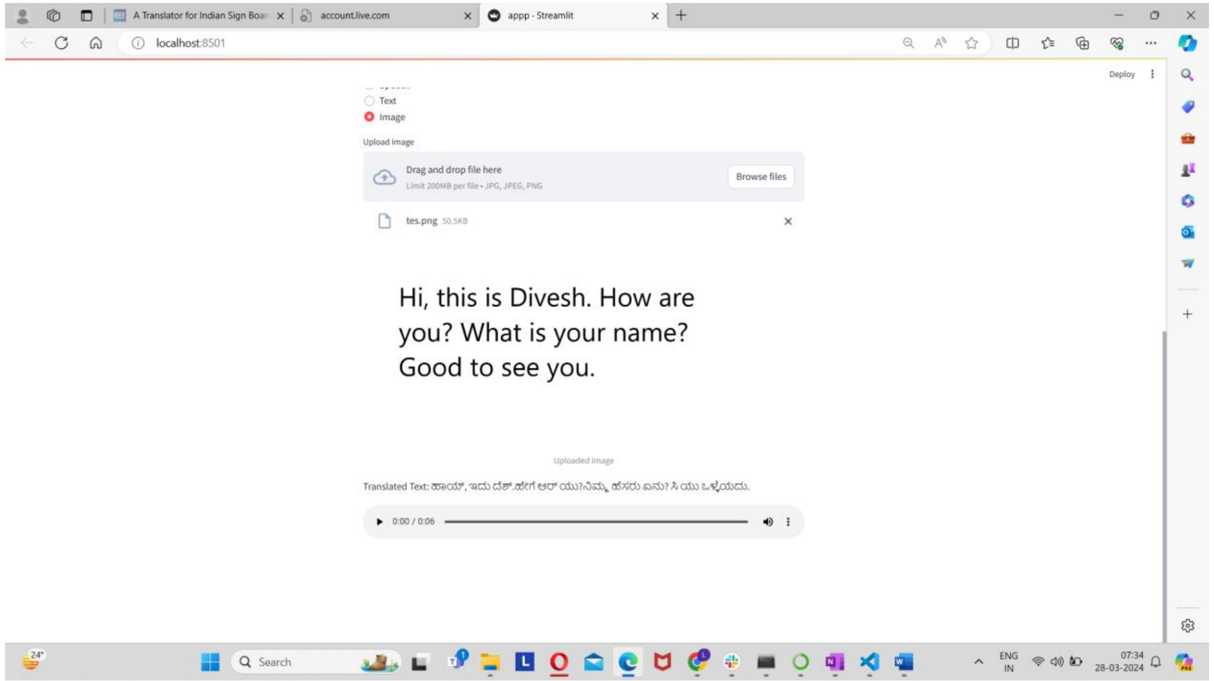
Drag and drop file here

Limit 200MB per file • JPG, JPEG, PNG

Browse files

tes.png 50.5KB

Hi, this is Divesh. How are you? What is your name? Good to see you.



○ Text
● Image

Upload image

Drag and drop file here

Limit 200MB per file • JPG, JPEG, PNG

Browse files

tes.png 50.5KB

Hi, this is Divesh. How are you? What is your name? Good to see you.

Uploaded Image

Translated Text: ಹಾಯ್, ಇದು ದೆಶ್.ಹೇಗೆ ಆರ್ ಯು?ನಿಮ್ಮ ಹೆಸರು ಏನು? ಸಿ ಯು ಒಳ್ಳೆಯದು.

0:00 / 0:06

Test Cases

- **Image Processing Test Cases**:
    - Upload an image with clear text in English.
    - Upload an image with handwritten text in a different language.
    - Upload a low-resolution image with text.
    - Upload an image with text overlaid on a complex background.
    - Verify that the extracted text matches the text in the input image accurately.
- **Speech Recognition Test Cases**:
    - Record clear spoken sentences in English.
    - Record spoken sentences with background noise.
    - Record spoken sentences with various accents.
    - Record spoken sentences in different languages.
    - Verify that the transcribed text matches the spoken input accurately.
- **Translation Test Cases**:
    - Translate extracted text from images to English.
    - Translate transcribed speech to English.
    - Translate extracted text from images to different target languages.
    - Translate transcribed speech to different target languages.
    - Verify that the translated text accurately conveys the meaning of the source text.

Test Cases

- **Cross-Modal Fusion Test Cases**:
  - Test the integration of information from images and speech inputs.
  - Verify that contextual understanding enhances translation accuracy.
  - Verify that cross-modal fusion captures nuances and cultural subtleties in the translation output.
- **User Interface Test Cases**:
  - Test the user interface for uploading images.
  - Test the user interface for recording and transcribing speech.
  - Test language selection functionality for source and target languages.
  - Verify that translated output is displayed clearly and accessible to users.
- **Performance Test Cases**:
  - Test system performance with a large number of concurrent users.
  - Test system response time for image processing, speech recognition, and translation tasks.
  - Verify that the system operates reliably under different network conditions and loads.

Conclusion

The project represents a significant advancement in the field of multilingual communication. By integrating image processing, speech recognition, and natural language processing techniques, this unified system breaks down linguistic barriers and facilitates effective language conversion from diverse modalities. Through cross-modal fusion and contextual understanding, the system ensures accurate and contextually relevant translations, empowering users to communicate seamlessly across linguistic boundaries. The user-friendly interface and adaptive learning mechanisms further enhance the system's usability and effectiveness, fostering global connectivity and understanding.

Future work

- **Domain-Specific Adaptation**: Investigate methods to adapt translation models to specific domains or industries, allowing for more accurate translations of specialized content.
- **Expansion of Language Support**: Expand language support to include more languages and dialects, ensuring inclusivity and accessibility for users worldwide.
- **Integration of Multimodal Inputs**: Explore the integration of additional modalities, such as gestures or facial expressions, to enhance translation accuracy and user experience.
- **Real-Time Translation**: Develop real-time translation capabilities to enable instantaneous language conversion during live conversations or video interactions.
- **User-Centric Design Improvements**: Continuously gather user feedback to refine the user interface and incorporate user preferences, improving overall usability and satisfaction.

Thank You