ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14921228

Real-Time Language Translator with Sign Language Recognition: A Multi-Modal Approach

Vasu Kapil[1]; Dheeraj[2]; Ritik Chauhan[3]; Amardeep Singh[4]

[1]**(23SCSE2160012)** School of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh.
[2]**(23SCSE2160031)** School of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh.
[3]**(23SCSE2160030)** School of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh.
Submitted To
[4]School of Computing Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh.

Publication Date: 2025/02/25

Abstract: This research explores the development of a real-time language translation system integrating speech recognition, text-based translation, and sign language recognition. The system employs Google Translate API for multilingual translation, MediaPipe Hands for sign language recognition, and SpeechRecognition for real-time voice input. The study aims to bridge communication gaps between spoken, written, and signed languages. The paper presents implementation details, experimental results, and future scope for improvement. Findings indicate promising accuracy in sign recognition and speech translation, highlighting the potential for real-world application in accessibility and communication enhancement.

How to Cite: Vasu Kapil; Dheeraj; Ritik Chauhan; Amardeep Singh (2025). Real-Time Language Translator with Sign Language Recognition: A Multi-Modal Approach. *International Journal of Innovative Science and Research Technology*, 10(2), 567-570. https://doi.org/10.5281/zenodo.14921228

I. INTRODUCTION

With advancements in artificial intelligence and real-time communication technologies, speech, text, and sign language can now be seamlessly transmitted and translated over the internet. The growing availability of cost-effective speech recognition and translation tools presents a significant opportunity for developing interactive, real-time multilingual communication systems. These systems can accept various input modes, including spoken language, typed text, and sign language, enabling broader accessibility.

This study emphasizes the importance of interactive disambiguation in real-time translation systems. Given the inherent complexities of speech and sign language recognition, ambiguities often arise during the translation process. Addressing these ambiguities early—particularly at the speech recognition stage—is crucial to ensuring accurate and meaningful translations. By integrating user feedback mechanisms and artificial intelligence-driven disambiguation techniques, the proposed system minimizes errors and enhances communication between individuals who rely on different language modalities.

➢ *The Need for Interaction in Translation*

The Language barriers pose significant challenges in global communication, particularly for individuals with hearing impairments. While numerous speech-to-text and text-based translation systems exist, sign language recognition remains an underdeveloped area. This study introduces a multi-modal translation framework that integrates speech, text, and sign language inputs to facilitate seamless communication. By leveraging artificial intelligence and deep learning models, this system provides an inclusive translation approach.

➢ *A Guiding Principle for Interactive Translation is:*

- *"Resolve Ambiguity Early to Prevent Errors from Spreading."*

The need for early intervention is especially critical in speech and sign language translation. While ambiguity is a challenge in text-based translation, it becomes even more pronounced in speech recognition, where factors such as pronunciation variations, background noise, and linguistic nuances introduce significant uncertainty. If these ambiguities are not addressed at the initial recognition stage, they can propagate through the entire translation process,

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14921228

leading to inaccurate or misleading outputs. Therefore, implementing interactive disambiguation strategies early on ensures a more reliable and accurate translation system.

II. METHODOLOGY

The system comprises three core modules: speech-to-text, text translation, and sign language recognition. The speech-to-text module utilizes Google's SpeechRecognition library to convert spoken words into text. The translation module employs the Google Translate API for real-time language conversion. The sign language recognition module leverages MediaPipe Hands to analyze hand gestures and map them to American Sign Language (ASL) letters. The integration of these modules into a graphical user interface using Tkinter enables user-friendly interaction.

- *Disambiguation for Language Translation*

We have emphasized the importance of interactive disambiguation in machine translation, particularly in the context of speech and sign language translation. A clearer understanding of this interaction can be gained by examining its role in both machine translation and speech recognition.

In interactive machine translation (MT) systems, the technology initially attempts to resolve ambiguities autonomously. However, when uncertainties persist, the system prioritizes the most critical issues and seeks user input for clarification. This is typically done through intuitive prompts, such as multiple-choice questions, allowing users to specify the intended meaning. By incorporating this interactive approach, translation accuracy is significantly improved, reducing errors that could otherwise propagate through the system.

To address translation ambiguities, interactive disambiguation techniques are employed. The system first attempts to resolve uncertainties automatically. If ambiguities persist, users are prompted with multiple-choice options to clarify meaning. This approach minimizes errors that could propagate through the translation pipeline.

For speech recognition, a hybrid method is used to improve accuracy. Instead of relying solely on continuous speech recognition, the system incorporates an option for word-by-word dictation. This method allows users to confirm or correct individual words before proceeding, reducing overall recognition errors and ensuring a more accurate translation outcome.

To address structural ambiguities, the system provides users with a set of possible interpretations in the source language. These interpretations follow predefined linguistic patterns tailored to the specific type of ambiguity detected. For instance, if a sentence has multiple meanings depending on how words are grouped, the system presents the user with clear choices to select the intended meaning.

For example, in a scenario where a sentence could imply two different actions, such as whether a worker is digging both the road and the asphalt or digging the road and then paving it, the system prompts the user to select the correct interpretation. Similarly, for lexical ambiguities, the system offers multiple definitions of a word in the source language, allowing the user to clarify the intended meaning before proceeding with the translation.

This interactive approach ensures that translations are both accurate and contextually appropriate, reducing potential misinterpretations.

- *Disambiguation for Speech and Sign Recognition*

The benefits of interactive disambiguation in speech translation have been emphasized, yet several factors still require consideration, such as the style of user interaction and the timing of intervention.

For instance, the conventional method of selecting from an n-best list of possible transcriptions for an entire spoken sentence, as seen in some speech recognition systems, may not be ideal in practical use. While such an approach may work well in controlled demonstrations, it could become cumbersome in spontaneous conversations, where users must repeatedly choose the correct transcription from numerous similar alternatives.

A more user-friendly alternative could involve collapsing redundant elements across multiple candidates and displaying a structured selection menu, where users can confirm their intended phrase using a pointing device or another intuitive selection method. Another possible improvement is to present only the system's best-guess transcription and provide a quick correction mechanism for any misrecognized words.

An even more interactive method would be to intervene frequently and at an early stage, ensuring errors do not accumulate and impact subsequent steps. A prime example is **isolated-word speech recognition**, as seen in various modern dictation tools. In this approach, users dictate one word at a time, confirming or correcting each entry before proceeding to the next. This method offers significant benefits, as it guarantees that the preceding portion of the text is accurate, allowing the language model to generate more reliable predictions for upcoming words. Additionally, knowing the precise boundaries of each word considerably reduces recognition errors. When ambiguity arises, the system presents multiple word choices, and users can select the correct one via a touch interface or voice commands such as "Option one" or "Option two."

Although isolated-word input may feel unnatural and slower compared to fluid speech, it has distinct advantages, especially in ensuring robustness and reliability. Modern dictation systems can accurately recognize words spoken by new users without requiring extensive voice training. Furthermore, this method supports a broad vocabulary range, allowing recognition of uncommon words across multiple domains. For example, some advanced dictation

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14921228

programs can correctly transcribe complex historical texts, such as Lincoln's Gettysburg Address, without prior exposure to the speaker's voice. Unlike traditional speech recognition models that struggle with domain-specific limitations, isolated-word recognition reduces dependency on specialized training and lexicon constraints.

For speech translation, a system based on isolated-word input offers additional advantages beyond accuracy. Since users must speak more slowly and clearly, spontaneous speech disfluencies (such as fillers and hesitations) are minimized. This results in syntactically cleaner sentences that are easier to process and translate. Additionally, because dictated speech is typically more concise than conversational speech, translation models require less interaction to generate an accurate output.

While such a system may not entirely eliminate the need for translation model customization, its ability to handle a wider range of inputs with minimal domain-specific adaptation makes it a compelling option for real-time speech and sign language translation applications.

III. RESULTS AND DISCUSSION

Preliminary testing of the system demonstrates high accuracy in speech-to-text conversion and text translation. Sign language recognition, however, presents challenges due to gesture variations and environmental factors such as lighting and camera angles. Experimental results indicate an average recognition accuracy of 85% for static ASL alphabet gestures. Further enhancements, including machine learning-based gesture classification, can improve recognition performance. Additionally, user feedback suggests that the system effectively facilitates communication for non-verbal individuals and supports real-time multilingual conversations.

➢ *System Implementation the Implementation Consists of:*

- **Speech Recognition Module:** Uses Python's SpeechRecognition library to capture and convert spoken language into text.
- **Text Translation Module:** Implements the Google Translate API to translate input text into a selected target language.
- **Sign Language Recognition Module:** Employs MediaPipe Hands for gesture detection and classification into ASL letters.
- **Graphical User Interface (GUI):** Built using Tkinter, the interface allows users to input text, translate, and recognize sign language in real-time.

IV. CONCLUSIONS AND FUTURE WORK

The research highlights the feasibility of integrating speech, text, and sign language into a unified translation system. While current accuracy rates are promising, future work will focus on incorporating dynamic sign language recognition, expanding the training dataset, and enhancing real-time processing capabilities. The integration of deep learning-based models can further refine gesture recognition accuracy. This study contributes to the advancement of accessible communication technologies and encourages further exploration in multi-modal translation systems.

ACKNOWLEDGEMENTS

I sincerely appreciate the support and guidance of my colleagues and mentor, whose insights and discussions have been invaluable to the development of this real-time language translation and sign recognition project.

REFERENCES

[1]. "Direct Speech to Speech Translation Using Machine Learning", December 2020
[2]. S. Venkateswarlu, D. B. K. Kamesh , J. K. R. Sastry and Radhika Rani, " Text to Speech Conversion", 23 September 2020
[3]. Chris Piech, Sami Abu-El-Haija, "Auto-Translation for Localized Instruction", Sep 2019
[4]. Sagar Patil, Mayuri Phonde, Siddharth Prajapati , "Multilingual Speech and Text Recognition and Translation using Image", April-2020
[5]. Bapna, A. 2019. Googleblog. Accessed 06.02.2022 https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html
[6]. Belval 2022. Github. Accessed 05.02.2022 https://github.com/Belval/pdf2image
[7]. Bradski, G. & Kaehler, A. 2008. Learning OpenCV, Computer Vision with the OpenCV Library. Sebastopol. O'Reilly Media, Inc.
[8]. Riverbank Computing. Accessed 27.03.2022 https://riverbank computing.com/s oftware/pyqt/intro
[9]. Glyph & Cog LLC 2011. Mankier. Accessed 05.02.2022 https://www.mankier.com/1/pdftoppm
[10]. Google 2006. Announcing Tesseract OCR. Accessed 05.02.2022 https://web.archive.org/web/20061026075310/http://google-code-updates.blogspot.com/2006/08/announcing-tesseract-ocr.html
[11]. Han, S. 2020. Googletrans Documentation. Accessed 06.02.2022 https://py-googletrans.readthedocs.io/en/latest/
[12]. Harwani, B. 2018. Qt5 Python GUI Programming Cookbook: Building responsive and powerful cross-platform applications with PyQt. Birmingham, UK: Packt Publishing Ltd.
[13]. Konica Minolta 2018. Accessed 05.02.2022 https://www.konicaminolta.com.au/news-insight/blog/how-optical-character-recognition-works
[14]. Lee, M. 2022. Pypi. Accessed 05.02.2022 https://pypi.org/project/pytesseract/
[15]. Bowen, L. & Caswell, I. 2020. Googleblog. Accessed 06.02.2022 https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html
[16]. Lutz, M. 2001. Programming Python, 2nd edition. Sebastopol. O'Reilly media.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14921228

[17]. Och, F. 2006. Googleblog. Accessed 05.02.2022 https://ai.googleblog.com/2006/04/statistic al-machine-translation-live.html
[18]. OpenCV, 2022. Accessed 06.02.2022 https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.ht ml