Mohammed Nihal Paper f

by Adaickalam Dr. Vijayakumar

Submission date: 07-Jan-2025 02:47PM (UTC+0530)
Submission ID: 2560585884
File name: Mohammed Nihal Paper for Coference_.pdf (347.12K)
Word count: 2980
Character count: 18268

Voice translation using Whisper AI and Google Translate

Mohammed Nihal A S
School of Computer Science and Engineering
Presidency University
Bangalore, India
nihalm617@gmail.com

Mahesh G
School of Computer Science and Engineering
Presidency University
Bangalore, India
@gmail.com

Darshan A R
School of Computer Science and Engineering
Presidency University
Bangalore, India
darshan25ar@gmail.com

Sandesh W D
School of Computer Science and Engineering
Presidency University
Bangalore, India
wdsandesh@gmail.com

Rohan S Handral
School of Computer Science and Engineering
Presidency University
Bangalore, India
rohansafari@gmail.com

Vijayakumar Adaickalam
School of Computer Science and Engineering
Presidency University
Bangalore, India
vijaykumar.adaickalam@presidencyuniversity.in

***Abstract*— In this research, we examine the combination of Whisper API and Google Translate API within a web application created to transcribe audio in real-time and convert it into multiple Indian languages. The main goal is to assess how effective Whisper API is for transcription and Google Translate API is for translation in delivering precise and efficient translation services. Whisper API, created by OpenAI, is an advanced automatic speech recognition (ASR) system recognized for its reliability and exceptional accuracy. It can manage different accents, ambient sounds, and specialized terminology, making it a perfect option for live transcription. Conversely, the Google Translate API employs sophisticated neural machine translation (NMT) models, which have greatly enhanced translation precision compared to conventional statistical models. Google Translate offers support for a variety of languages, including several Indian languages, which makes it suitable for our application's need to deliver translations in the local languages of various Indian states. This web application, developed with the React framework, combines these two robust APIs to provide a smooth user experience. The app converts live audio to text via the Whisper API and subsequently translates the written content into the language chosen by the user with the Google Translate API. Furthermore, the app monitors the user's location and automatically identifies the main translation language based on the local language of the state they are in, thereby improving the user experience by making the service more pertinent and accessible.**

This document offers a comprehensive summary of the technical execution, encompassing the structure of the web application, the integration of the APIs, and the distinctive features that set our solution apart. Our results indicate that the integration of Whisper API with Google Translate API provides an enhanced option for real-time audio transcription and translation services, rendering it an essential resource for users in multilingual settings.

Keywords*— *Whisper AI, Google Translate, Automated Speech Recognition(ASR), Neural Machine Translation(NMT).

I. Introduction

Effective multilingual communication is essential in today's globalized society. This is particularly important in multilingual nations like India, which has hundreds of dialects spoken throughout its enormous territory in addition to 22 officially recognized languages. Real-time transcription and translation of spoken language can greatly improve communication, removing linguistic barriers and promoting mutual comprehension. Conventional transcription and translation techniques, such manual transcription followed by human translation, are frequently laborious and prone to mistakes. These approaches could produce errors and misunderstandings due to their inability to handle background noise, a variety of dialects, and the subtleties of other languages. Furthermore, traditional approaches are unable to deliver the instantaneous processing needed for real-time applications.

To overcome these difficulties, we designed a web application that leverages state-of-the-art technologies: Whisper AI for transcription and Google Translate for translation. OpenAI's Whisper AI is a state-of-the-art automated speech recognition (ASR) system renowned for its excellent accuracy and resilience. It is the perfect option for real-time transcribing because it can handle a variety of accents, background noise, and technical jargon.

Conversely, the Google Translate makes use of sophisticated neural machine translation (NMT) models, which outperform conventional statistical models in terms of translation accuracy. Google Translate is appropriate for the application's need to deliver translations in the regional languages of many Indian states because it supports a large number of languages, including numerous Indian languages.

The web application combines these two robust APIs to provide a smooth user experience. The application converts live audio into text through the Whisper AI API and subsequently translates the converted text into the user's preferred language via the Google Translate API. Moreover, the app monitors the user's location and automatically adjusts the main translation language to the local language of their current state, improving the user experience by making the service more pertinent and accessible.

This study seeks to highlight the benefits of utilizing the Whisper AI for transcription and the Google Translate for translation. We will present a comprehensive summary of the technical execution, covering the web application's architecture, the API integrations, and the cutting-edge

features that distinguish our solution. Additionally, we will examine the assessment outcomes of our method regarding transcription and translation precision, as well as user contentment.

The results indicate that the integration of Whisper AI and Google Translate provides an enhanced solution for live audio transcription and translation services. This method guarantees both great precision and a user-focused experience by addressing the linguistic variety in India. With this research, we seek to enhance the creation of more efficient and accessible translation services, ultimately promoting improved communication and understanding in a multilingual environment.

II. Related works

V. R. and I. A. Funcke (2023) introduce aiLangu, a system that allows speech transcription and translation in real time, thereby lowering language barriers. The writers stress the value of accessibility as well as the difficulties in creating reliable, latency-effective solutions that work in a variety of linguistic and auditory contexts. This study emphasizes how transcription and translation can be combined to facilitate smooth communication in multilingual environments. For real-time applications, the suggested framework acts as a fundamental manual for combining Whisper AI with translation programs like Google Translate. [1]

Y. Peng et al. (2023) investigate Whisper-style model training techniques with open-source toolkits and publicly accessible data. They utilized transformer-based architectures to replicate Whisper's functionality and leveraged frameworks such as PyTorch and Hugging Face's Transformers library for model development and experimentation. Their research demonstrates how Whisper AI's end-to-end architecture and strong multilingual capabilities make it an excellent choice for transcribing tasks. The authors address the difficulties in reproducing private models and suggest methods for using public datasets to duplicate Whisper's training paradigm. Whisper AI is a good option for real-time voice translation projects because of the insightful information this study offers about its architecture [2].

D. Wang et al. (2019) provided a thorough description of automatic speech recognition (ASR) systems from start to finish. They highlight the use of transformer-based models, such as the Transformer and Conformer architectures, which enhance sequence-to-sequence learning. The paper also discusses frameworks like TensorFlow and Kaldi that have been widely adopted in developing ASR systems. They talk about the switch from conventional pipeline-based ASR systems to end-to-end strategies that increase efficiency and simplify the design. The study explores a number of issues, including as scalability across languages, robustness in noisy situations, and the significance of extensive pretraining. Their research is highly relevant to the use of Whisper AI, an end-to-end ASR model, in real-time applications where high accuracy and resilience are required [3].

D. Macháček et al. (2024) look into what needs to be changed to make Whisper AI a real-time transcribing system. The study highlights the use of transformer-based architectures to optimize the model's latency and performance. Frameworks like PyTorch were employed to implement and evaluate the proposed enhancements, ensuring the system's scalability and real-time efficiency. In order to modify Whisper's architecture to satisfy the requirements of real-time applications, they address latency issues and provide optimization techniques. The study's emphasis on speeding up response times without sacrificing transcription quality offers a model for incorporating Whisper AI into speech translation systems, where a low latency is essential [4].

A popular global translation service, Google Translate continues to set the standard for accessibility and translation accuracy. It employs neural machine translation (NMT) techniques, leveraging transformer-based architectures such as the Transformer model to process and translate text efficiently. The platform utilizes TensorFlow as a primary framework for training and deploying its models The platform is an essential part of voice translation systems because of its frequent updates and support for more than 100 languages. Real-time speech-to-speech translation systems can be developed thanks to its API capabilities, which provide smooth integration into transcription pipelines [5].

An open-source translation program called LibreTranslate has more customization options than Google Translate but still provides comparable functionality. It uses neural machine translation (NMT) techniques and leverages frameworks such as OpenNMT and PyTorch for its implementation, ensuring flexibility and efficiency in translation tasks. Without depending on proprietary APIs, it gives developers the freedom to create translation systems. By ensuring that the system can operate independently of commercial platforms, LibreTranslate is incorporated into voice translation initiatives, offering a financially viable alternative for research and development [6].

B. Raj and R. Olivier (2022) talk about how resilient Whisper AI is to suspicious cases. Their study reveals weaknesses in Whisper's transcriptional capacities under precisely calibrated disturbances. This study emphasizes how crucial it is to strengthen transcription models' resilience, especially in real-time speech translation systems where adversarial noise may affect efficiency. In order to guarantee the dependability and security of systems that use Whisper AI, these issues must be resolved [7].

M. Aiken (2021) offers a comprehensive assessment of Google Translate's accuracy in a variety of languages and fields. The report identifies enduring difficulties in less popular or domain-specific translations while highlighting advancements in translation quality, especially in widely used languages. The importance of extensive training data in enhancing translation accuracy and dependability is shown by Aiken's work. This work highlights the significance of integrating transcription accuracy with strong translation skills and provides insightful information about Google Translate's possible integration with Whisper AI for multilingual voice translation systems [8].

III. PROPOSED APPROACH

A. Initialization and Regional Language Detection:

On startup, the application request location permission from the user, once the permission is granted, the state that the user is currently in, is automatically detected and the regional language mapped to that state will be set as the

primary language for the user. The location is encoded employing the reverse geocoding services provided by Nominatim API, it generates an address from a coordinate given as latitude and longitude. For example, if the user is in Karnataka, the state is detected as Karnataka and since the regional language is Kannada the primary language in the app is set to Kannada. With this feature, the human input can be minimized and the user experience can be improved.

B. Audio Recording and permission handling:

When the user wants to start translating, the app prompts the user to provide microphone access, upon granting permission, the application starts recording the audio input from the user and transmitted to the server for processing, which is later sent to Whisper AI via an API call for transcription.

C. Transcription using Whisper AI:

Whisper AI is a general-purpose speech recognition model developed by OpenAI. It is trained on a large dataset of diverse audio of over 680,000 hours, is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification. An encoder-decoder transformer forms the core architecture of Whisper. The audio input that is given is resampled to 16,000 Hz and converted to an 80-channel log-magnitude Mel spectrogram using 25ms window period with a 10ms stride. This Mel spectrogram in given as an input to the encoder, which processes it. Sinusoidal positional embeddings are added and further processed by a series of Transformer encoder blocks. The output is layer normalized. A standard decoder transformer block which has the same width as that of the transformer blocks in the encoders is used. It uses learned positional-embeddings and tied input-output token representations.

Several speech processing tasks, such as multilingual speech recognition, speech translation, spoken language identification, and voice activity detection, are used to train a transformer sequence-to-sequence model. Many steps of a conventional speech-processing pipeline can be replaced by a single model since these tasks are simultaneously represented as a series of tokens that the decoder must anticipate. A collection of unique tokens is used in the multitask training style as task specifiers or categorisation goals.

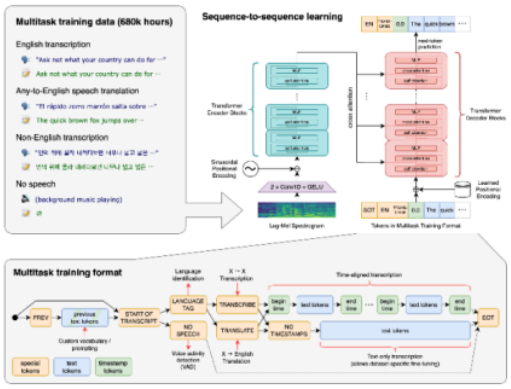


Fig 1. Architecture of Whisper AI by OpenAI

D. Translation using Google Translate:

The transcribed text, now in the original language of the speaker, is sent to the Google Translate API for translation. The Google Translate API processes this transcription and converts it into the desired regional language, as determined by the initial location detection.

Google translate uses a transformer model to translate the transcript. A transformer model consists of an encoder and a decoder. Encoders in Transformers are neural network layers that process the input sequence and produce a continuous representation, or embedding, of the input. The decoder then uses these embeddings to generate the output sequence. The encoder typically consists of multiple self-attention and feed-forward layers, allowing the model to process and understand the input sequence effectively.

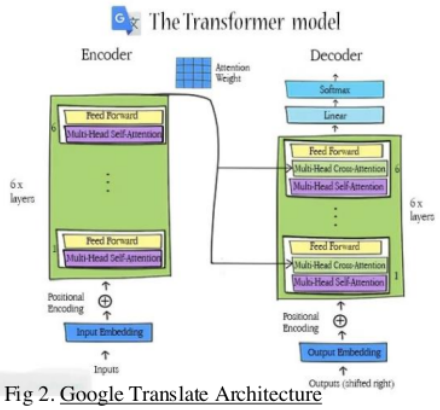


Fig 2. Google Translate Architecture

In this model, both encoder and decoder have 6 stacked layers. Each layer has sub-layers of multi-head self-attention and feed-forward. In the decoder, there is an additional sub-layer of multi-head cross-attention. The cross-attention performs an attention function over the output of the encoder layers. This attentional mechanism draws global dependencies between input and output.

This two-step process ensures a high degree of accuracy and contextual relevance in the final product by combining the benefits of Whisper AI for transcription and Google Translate for translation. Once the translated text has been formatted and returned to the frontend for display, the user can view both the original transcription and its translated version.

E. Database Integration

With the transcript in original language and the translation now available, a record in created in the database. MongoDB a scalable NoSQL database is used due to its flexibility. This enables the user to access their

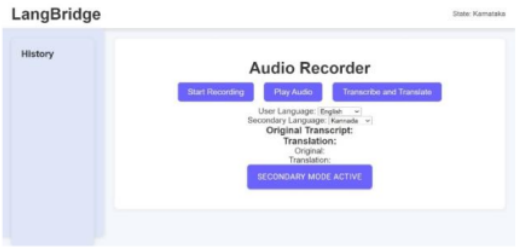
respective translation history by preserving a persistent storage layer, making it easier to reuse and refer to as needed.

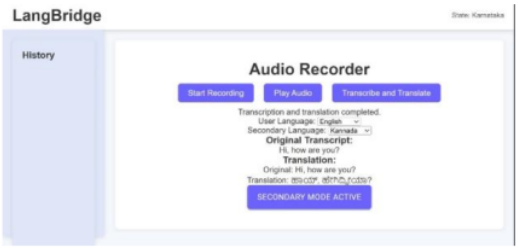
F. User Interface and Output

The frontend, which was created with React, is intended to offer a smooth and user-friendly interface. Managing user preferences, letting users interact with stored data and displaying transcriptions and translations in real-time are some of the key features. The state management features of React guarantee seamless data transfer between components, and API calls to the backend are optimized for dependability and speed. The user interface is designed to be straight forward but efficient, accommodating a wide range of users with different degrees of technological expertise.

The proposed methodology demonstrates a user-friendly approach to solve translation challenges, in a multilingual environment. Integrating Whisper AI and Google Translate APIs into a seamless workflow, the application addresses the needs of the users in a linguistically diverse region such as India.

IV. RESULT





An audio input is taken in selected language, the audio then is transcribed by whisper ai for accurate and fast transcription. The transcribed text is then translated google translate API, as google translate's linguistic features offers an extremely effective and adaptable answer for instantaneous language translation and translated text is displayed below translation.

V. CONCLUSION

In this study, we investigated the combination of the Whisper API for live audio transcription and the Google Translate API for translation to create a precise and effective language translation service. Our system proficiently integrates these two robust technologies to transcribe spoken material and convert it into a chosen Indian language, incorporating a location-based language detection feature to improve user experience.

The Whisper API, boasting strong features and a vast amount of training data, excels at transcribing various speech styles and accents, delivering impressive accuracy in real-time transcription. This is especially important for our system, which must manage diverse linguistic subtleties among users from various areas. Utilizing Whisper's models, we can guarantee that the transcription method is both fast and accurate, providing a dependable basis for subsequent translation.

The Google Translate API provides extensive language support and advanced machine learning models, enabling accurate translations during the translation process. Its broad range of Indian languages and live translation features render it an important resource for multilingual applications, especially in varied linguistic environments such as India. Furthermore, incorporating location-based language detection enhances the translation process by automatically choosing the most relevant language according to the user's geographical position, minimizing translation delays and boosting user satisfaction.

The integration of Whisper's transcription precision with Google Translate's linguistic features offers an extremely effective and adaptable answer for instantaneous language translation. The system illustrates how utilizing these cutting-edge AI technologies together can provide a more precise, responsive, and user-focused translation experience than conventional methods, positioning it as an optimal solution for various multilingual communication needs..

REFERENCES

[1] V. R. and I. A. Funcke, “aiLangu – Real time Transcription and Translation to Reduce Language Barriers,” KTH Royal Institute of Technology, 2023.

[2] Y. Peng et al., “Reproducing Whisper Style Training Using An Open-Source Toolkit And Publicly Available Data,” 2023, doi: 10.1109/ASRU57964.2023.10389676.

[3] D. Wang, X. Wang, and S. Lv, “An overview of end-to-end automatic speech recognition,” Symmetry. 2019, doi: 10.3390/sym11081018.

[4] D. Macháček, R. Dabre, and O. Bojar, “Turning Whisper into Real-Time Transcription System,” 2024, doi: 10.18653/v1/2023.ijcnlp-demo.3.

[5] [Online]. Available: https://play.google.com/store/apps/details?id=com.google.android.apps.translate&hl=en&gl=US.

[6] [Online]. Available: https://github.com/LibreTranslate/LibreTranslate.

[7] R. Olivier and B. Raj, “There is more than one kind of robustness: Fooling whisper with adversarial examples,” arXiv:2210.17316, 2022.

[8] M. Aiken, “An Updated Evaluation of Google Translate Accuracy,” School of Business Administration, University of Mississippi, 2021.

[9] Birkenbeuel J, Joyce H, Sahyouni R, et alGoogle translate in healthcare: preliminary evaluation of transcription, translation and speech synthesis accuracyBMJ Innovations 2021;7:422-429.

[10] B. Naidoo and F. Ghayoor, "Developing an Automatic Speech-to-Speech Translator Mobile Application using AWS," 2023 IEEE AFRICON, Nairobi, Kenya, 2023, pp. 1-5, doi: 10.1109/AFRICON55910.2023.10293612.

Mohammed Nihal Paper f

ORIGINALITY REPORT

12%	10%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.lavivienpost.com Internet Source	2%
2	www.scaler.com Internet Source	2%
3	Submitted to Christ University Student Paper	1%
4	Submitted to Hanoi University Student Paper	1%
5	Submitted to British University In Dubai Student Paper	1%
6	link.springer.com Internet Source	1%
7	github.com Internet Source	1%
8	Submitted to University of Sunderland Student Paper	<1%
9	Submitted to Queen Mary and Westfield College Student Paper	<1%

10	Submitted to School of Business and Management ITB Student Paper	<1%
11	arxiv.org Internet Source	<1%
12	Submitted to Middlesex University Student Paper	<1%
13	www.coursehero.com Internet Source	<1%
14	www.diva-portal.org Internet Source	<1%
15	"Chinese Computational Linguistics", Springer Science and Business Media LLC, 2019 Publication	<1%
16	H . Anandakumar, R. Arulmurugan, Chow Chee Onn. "Big Data Analytics for Sustainable Computing", Mobile Networks and Applications, 2019 Publication	<1%
17	ogxkdlk.farmaciasantostefano.it Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		