International Journal of Computer Sciences and Engineering
Vol.**11**, Issue.**12**, pp.**21-25**, December **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

VAARTALAP: Embedding Whisper-AI-like Model into a Video-Conferencing System to Aid Real-Time Translation and Transcription

Kunal Kashyap[1], Prashant Singh[2], Anjali Verma[3], Satya Mishra[4], Prachi Goel[5*]

[1,2,3,4]Undergraduate student, CSE Department, ADGIPS, New Delhi, India
[5]CSE Department, ADGIPS, New Delhi, India

**Corresponding Author: goyalprachi54@gmail.com*

Received: 04/Nov/2023; **Accepted**: 07/Dec/2023; **Published**: 31/Dec/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i12.2125

Abstract: In this world of digitalized communication, effective communication crosses regional boundaries and linguistic obstacles making the world more connected. The demand for seamless multilingual communication has never been more important as corporations, institutions, and individuals engage on a worldwide scale. This article explores a trailblazing initiative that uses real-time translation and transcription services offered by Whisper-AI to transform the world of video conferences. The goal of the research is to create an AI model that easily interfaces a translation and transcription-based model to work in a real-time video conferencing system. Participants may converse in real-time without any language barriers by utilizing cutting-edge voice recognition and translation technologies.

Keywords: Sound transcription, Sound translation, AI, Deep learning, Real-time, Language barrier, NLP.

1. Introduction

The widespread use of video conferencing equipment has completely changed how people and organizations communicate remotely. But even if these platforms bring individuals together from different linguistic origins, language barriers still stand in the way of inclusive and productive communication. Conventional methods of providing multilingual assistance frequently entail transcription and manual translation, which adds time to the process and reduces the flow of communication. Taking inspiration from the state-of-the-art open-source ASR system Whisper AI, created by OpenAI, this study aims to solve this problem. With its exceptional ability to properly translate spoken language across a wide range of dialects and languages, Whisper AI is a great choice for real-time translation and transcription services when used in video conferences.

By providing customers with an automatic and rapid solution, the proposed integration seeks to go beyond the constraints of the current language support capabilities.

Our goal is to enable video conferencing services with real-time speech transcription capabilities by utilizing a translation and transcription-based model, so that a written representation of the discussion may be produced. Concurrently, the connection facilitates dynamic translation services, enabling users to have smooth interactions in several languages.

2. Related Work

With the development of video conferencing technology, distant communication has seen a dramatic change that has made it possible for people and companies to communicate across geographic borders. But in an increasingly globalized world, the enduring problem of language variety has highlighted the need for creative solutions to promote efficient communication. In order to better understand the current state of research and technological developments in real-time translation and transcription, this literature review focuses on a translation and transcription-based AI model's integration with video conferencing systems.

2.1 Current Status of Video Conferencing Systems: In both personal and professional contexts, video conferencing has become indispensable. Even though these platforms make remote collaboration convenient, language barriers continue to be a significant obstacle that restricts their ability to effectively facilitate smooth interactions between users with different linguistic backgrounds.

2.2 Systems for Automatic Speech Recognition (ASR): In recent years, ASR systems have advanced significantly, with Whisper AI emerging as a well-known open-source option. ASR systems' accuracy, versatility, and multilingual capabilities have all been studied in this field of study, with a focus on how well they can convert spoken words into written text.

2.3 Real-Time Translation Technologies: A wealth of literature has been written about the difficulties and developments in this field. Systems that dynamically convert spoken language into the user's preferred language are the main emphasis in order to improve the inclusiveness and accessibility of communication platforms.

2.4 Integration of ASR in Video Conferencing: Research has looked into how to improve accessibility and user experience by integrating ASR systems into video conferencing frameworks. These studies frequently demonstrate the usefulness of turning spoken words into text during live discussions and the possibilities for real-time transcription services.

2.5 Whisper AI: The literature sheds light on Whisper AI's architecture, training techniques, and language support features. Its viability for inclusion into applications needing precise and flexible ASR capabilities has been established by evaluations of its effectiveness in identifying a variety of accents and languages.

2.6 Human-Computer Interaction (HCI) and User Experience: A lot of study has been done to understand how users interact with video conference interfaces and what their preferences are. The body of research highlights how crucial it is to reduce latency, design user interfaces for straightforward interaction, and provide a flawless user experience for users of transcription and translation functions.

3. Theory

A variety of programming languages, frameworks, and technologies would be needed to integrate an AI model for real-time translation and transcription into a video conference system. The following list of technologies is typical for these kinds of projects:

3.1 Languages Used in Programming: Python: Appropriate for managing backend functionality, Python is frequently used for AI and machine learning applications. JavaScript/TypeScript can be used for front-end programming, as web-based interfaces are used by the majority of video conferencing solutions. Along with these, HTML and CSS have been used.

3.2 Frameworks for Web Development: To create a dynamic and responsive user interface for the video conferencing system, Angular or React.js may be utilized. Django is a lightweight Python framework that is being used to construct the backend logic needed to manage the interactions with the translation and transcription-based model.

3.3 Web Real-Time Communication, or WebRTC: For real-time communication apps to be built directly within web browsers, WebRTC is an essential technology. It is necessary for the basic operation of the video conferencing system as it allows users to transmit voice and video to each other.

3.4 Translation Providers: To allow for the dynamic translation of recorded text into the user's selected language, utilize the Google Cloud Translation API or the Microsoft Translator API.

3.5 WebSocket: WebSocket may be used to provide transcriptions and translations as soon as they are completed for real-time communication between the server and clients.

3.6 Database: Use PostgreSQL, MySQL, or a NoSQL solution like MongoDB to store user preferences, settings, and maybe transcriptions.

3.7 Verification and Permission: To manage user identification and authorization securely, utilize OAuth or JWT (JSON Web Tokens).

3.8 Continuous Deployment/Continuous Integration (CI/CD): To automate the testing and deployment procedures for effective development workflows, use GitHub Actions, GitLab Continuous Integration, or Jenkins.

3.9 Optimizing Latency: The purpose of a content delivery network, or CDN, is to minimize latency and maximize the distribution of static assets to users in various geographic regions.

3.10 Frameworks for Testing: Jest for JavaScript and PyTest for Python: to carry out unit tests and guarantee the codebase's dependability.

4. Methodology

A methodical and planned strategy is used in the process of integrating an AI-model for real-time translation and transcription into a video conferencing system. The methodology is based on a phased development process that includes technical exploration, project planning, system architecture design, module design, development, testing, optimization, documentation, deployment, user training, monitoring, maintenance, and ongoing improvement based on user feedback.

4.1 Project planning:

Stage 1: Choosing a Technology stack- The integration framework is implemented using a technology stack that combines the flexibility of Python, for backend development and integration with the translation and transcription-based model. The frontend is built with React.js, ensuring a dynamic and responsive user interface. WebRTC facilitates real- time communication, while WebSocket ensures instant data transmission. The system also utilizes Google Cloud Translation API for dynamic language translation. This comprehensive technology stack forms the backbone of a robust and scalable solution for real-time translation and transcription within videoconferencing systems.

Stage 2: Selecting the approach- Python and TypeScript are used in a staged manner for backend and frontend development, respectively. Real-time transcription and translation are made possible using the AI based model that works in a similar manner as the OpenAI's Whisper-AI, while smooth communication is enabled via WebRTC and WebSocket. By helping with language translation, Google Cloud Translation API creates a productive foundation for multilingual video conferences.

4.2 Technical Exploration:

Analyzed the Whisper API documentation in detail to learn about the requirements, possibilities, and constraints for integration. Also, assessed Whisper AI's interoperability with the selected video conferencing system's current technological stack.

4.3 Development:

The development process will be iterative and staged, starting with requirements collecting and a thorough planning stage. The integration of AI model for real- time transcription will be part of the backend development process, which will use Django's flexibility. TypeScript will be used for frontend development at the same time, guaranteeing a responsive and dynamic user experience. WebRTCand WebSocket technologies will help the system communicate. Continuous testing will be carried out during development to guarantee system performance and module integrity. The implementation will take place gradually, with user input and iterations guiding improvements to create the most effective and inclusive multilingual video conferencing system possible.

5. Results and Discussion

The Vaartalap system has yielded notable outcomes, showcasing both advancements and considerations. The system's real-time translation and transcription functionalities demonstrated commendable performance in enhancing multilingual communication, fostering inclusivity, and improving overall user experience. Users reported increased efficiency in global collaborations and positive impacts on educational initiatives.

However, challenges such as occasional inaccuracies in transcription, latency concerns, and the need for stable internet connectivity were identified. Privacy and security considerations were addressed through robust measures, but ongoing vigilance is crucial. Users appreciated the system's adaptability but faced a learning curve in fully leveraging its features.

The integration showcased promising applications across diverse industries, with educational and professional development opportunities being evident.

Feedback from users played a pivotal role in identifying areas for improvement, emphasizing the importance of a continuous feedback loop. These insights will inform future updates and iterations to address accuracy, latency, and language support concerns, ensuring a dynamic and responsive system that aligns with user expectations.

In conclusion, Vaartalap presents a positive outcome, marked by advancements in multilingual communication and global collaboration. The identified challenges underscore the need for ongoing refinement, emphasizing the project's commitment to evolving and addressing user needs in the ever-changing landscape of digital communication technologies.

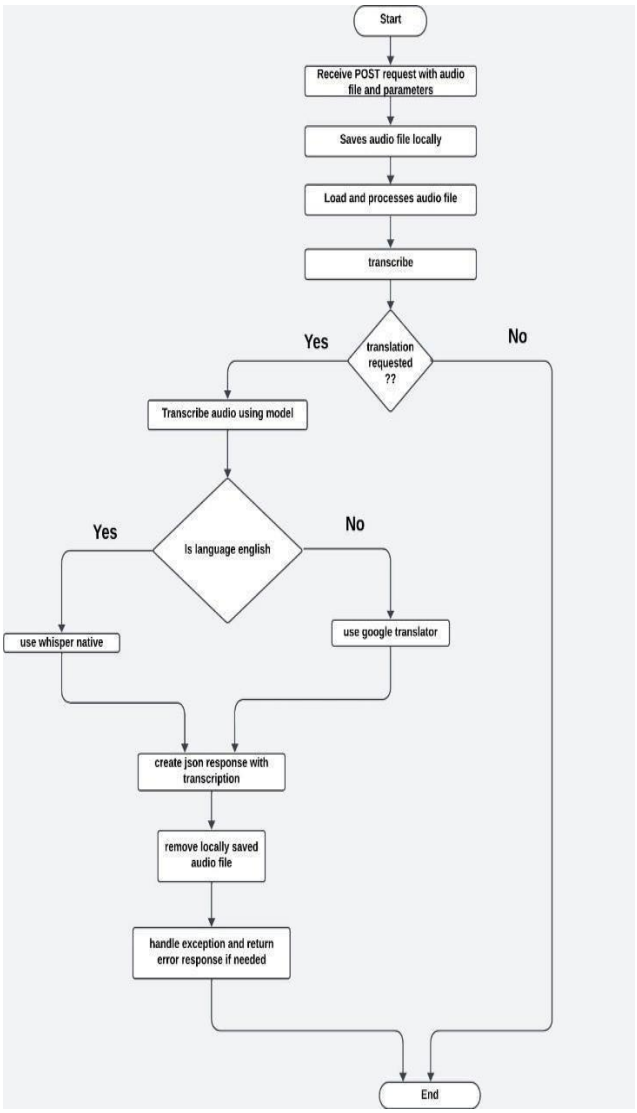

Figure 1. Flow Chart of AI Model

Fig.1: Displays the working flow of the AI model for translation and transcription, highlighting how the incoming audio is saved, transcribed and then checked if translation is asked for. If so, then it is translated based on the language.

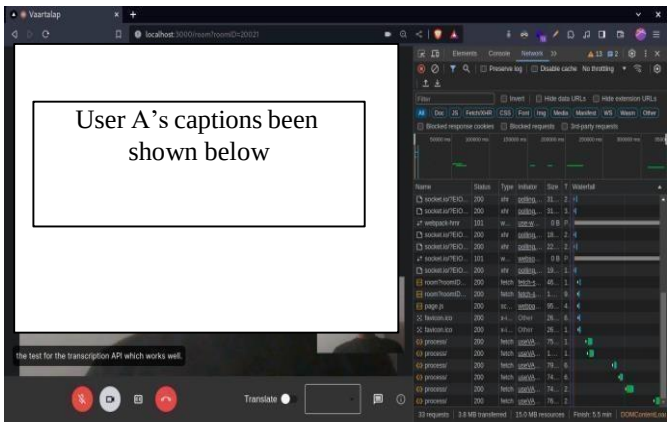

Figure 2. Captions on Vaartalap application

Fig. 2 display the captions in English as User A speaks in English.

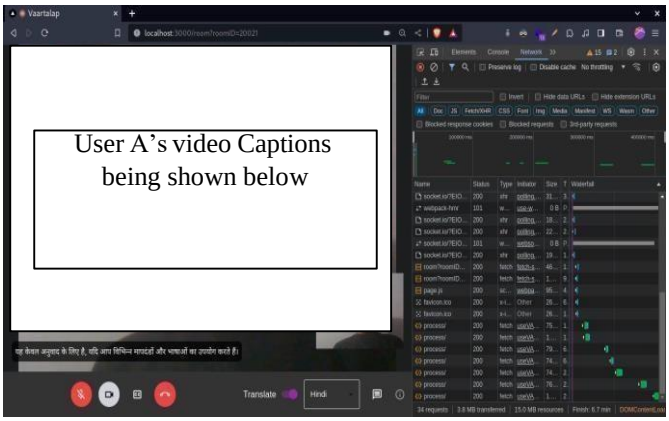

Figure 3. Displays translation to Hindi on Vaartalap application

Fig. 3 displays user A's audio being translated to Hindi asUser B has enabled translated captions in Hindi.

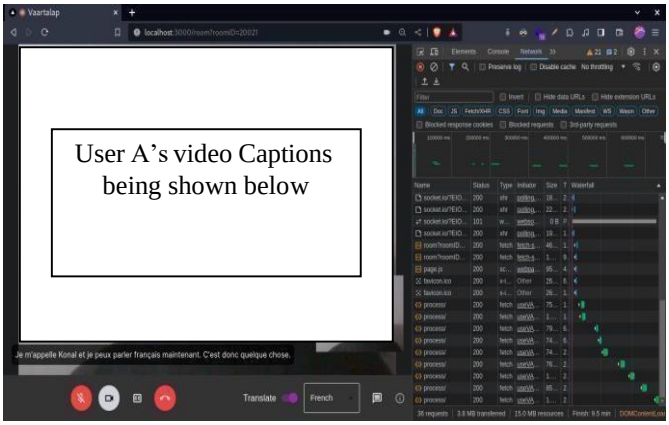

Figure 4. Displays translation to French on Vaartalap application

Fig. 4 displays user A's audio being translated to French as User B has enabled translated captions in French.

6. Conclusion and Future Scope

6.1 Conclusion

In summary, the incorporation of Vaartalap model for real-time translation and transcription into video conferencing systems is a novel approach with significant ramifications. This research has shed light on a variety of applications, including the promotion of inclusive communication platforms, user experiences, cooperation in education, and international commercial connections. The developments showcased here represent a fundamental change in the way we see and use multilingual communication tools.

The favorable results on inclusiveness, efficiency, and flexibility highlight Vaartalap's disruptive potential. Even though the system is now performing well, issues like accuracy and latency must be identified and fixed. These obstacles present chances for more innovation and improvement, attracting continued R&D projects to raise the effectiveness ofthe integration.

This research emphasizes the significance of technology in overcoming linguistic gaps and promoting a more connected world as we traverse the shifting environment of global communication. This integration is a leap forward in terms of technology as well as a good development, providing a window into the future of inclusive and fluid multilingual interactions online. As we direct the course of this integration, we see a day when language serves as a bridge to enhanced communication and mutual understanding rather than as a hindrance.

6.2 Future Scope

Future research and expansion of 'Vaartalap' for real-time translation and transcription into video conferencing systems appears promise. Important facets of the project's future scope are represented by the following areas:

6.2.1 Support for Advanced Languages:

Increase the scope of language support to include more regional and dialectal differences. In order to guaranteeprecise and nuanced translations across a variety of linguistic contexts, future development can concentrate on honing Vaartalap's capabilities.

6.2.2 Improved Precision and Decreased Latency: Invest in projects aimed at ongoing development to raise the precision of automated speech recognitionand lower latency even more. This entails making use of developments in real-time processing, neural networks, and machine learning.

6.2.3 Combining Emerging Technologies with Integration:

Investigate opportunities for collaboration using cutting edge technologies like augmented reality (AR), natural language processing (NLP), and artificial intelligence (AI) to build more intelligent and immersive multilingual communication experiences invideo conference settings.

6.2.4 Personalization and User Choices:

Provide tools that let users alter settings and language options to create a more unique experience. Putting machine learning algorithms into practice to adjust to unique speech patterns and preferences might be one way to do this.

6.2.5 Combining Augmented and Virtual Reality: For an even more dynamic and immersive multilingual communication experience, look at integrating Vaartalap into virtual and augmented reality settings. This may find use in online conferences, language instruction, and cooperative virtual environments.

Conflict of Interest

The authors have no financial or personal conflicts of interest to disclose.

Authors' Contributions

Author-1:

Played a key role in protocol development.
Mainly contributed in application development.

Author-2:

Contributed to the study design.
Contributed to manuscript preparation

Author-3:
Drafted the initial version of the manuscript.
Conducted a thorough literature review.
Conceived and developed the study.
Author-4:
Contributed to data interpretation
Provided valuable intellectual input.
Author-5:
Supervised the entire research process.
Revised the manuscript critically for important intellectual content.
All Authors:
Collaboratively reviewed and edited the manuscript.
Approved the final version for submission.

References

[1]. A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large- Scale Weak Supervision", *arXivpreprint*, arXiv:2212.04356, **2022.** doi10.48550/ arXiv.2212.04356

[2]. E. Cho, C. Fügen, T. Herrmann, K. Kilgour, M. Mediani, C. Mohr, J. Niehues, K. Rottmann, C. Saam, S. Stüker, and A. Waibel. 2013. "A real- world system for contemporaneous restatement of German lectures", In the Proceedings of the 2013 *INTERSPEECH*, Lyon, France, pp.**3473-3477, 2013.**

[3]. N. Arivazhagan, C. Cherry, I. Te, W. Macherey, P. Baljekar and G. Foster, "Re-Translation Strategies for Long Form, Simultaneous, Spoken Language Translation," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp.**7919-7923, 2020.** doi: 10.1109/ICASSP40776.2020.9054585.

[4]. Rothman and A. Gully, "Mills for Natural Language Processing" *Second Edition*, Packt Publishing, UK, ch. 2, 2022, ISBN 9781803247335

[5]. T. Chen, W. Wang, W. Wei, X. Shi, X. Li, J. Ye and K. Knight, "DiDi's Machine Restatement System for WMT 2020", In the Proceedings of the 2020 *Workshop on Statistical Machine Translation (WMT)*, pp.**105-112, 2020.**

AUTHORS PROFILE

Kunal Kashyap earned his diploma in Digital Electronics from Ambedkar Institute of Technology. He is currently an undergraduate student of B Tech in Computer Science and Engineering from Dr. Akhilesh Das Gupta Institute of professional studies, GGSIPU. He is a full-stack developer with proficiency in MERN stack.

Prashant Singh earned his diploma in Computer Engineering from Guru Nanak Dev Institute of Technology. He is currently an undergraduate student of B Tech in Computer Science and Engineering from Dr. Akhilesh Das Gupta Institute of Professional Studies, GGSIPU. He is a Quality Assurance Tester having tested on projects like Yatra.com using selenium.

Anjali Verma earned her diploma in Computer Engineering from Ambedkar Institute of Technology. She is currently an undergraduate student of B Tech in Computer Science and Engineering from Dr. Akhilesh Das Gupta Institute of Professional Studies, GGSIPU. She is a data analyst, working with Python.

Satya Mishra earned her diploma in Computer Engineering from Guru Nanak Dev Institute of Technology. She is currently an undergraduate student of B Tech in Computer Science and Engineeringfrom Dr. Akhilesh Das Gupta Institute of Professional Studies, GGSIPU. She is an ML Engineer, working with Python in technologies like, Neural Network, Deep learning etc.

Prachi Goel earned her B Tech and M Tech from MDU, Rohtak in 2018 and 2021 respectively. She is currently working as an Assistant Professor in department of Computer Science and Engineering at Dr. Akhilesh Das Gupta Institute of Professional Studies, GGSIPU. She has published two research papers in international journals. Her main research work focuses on Cryptography Algorithms, Network Security, Big Data Analysis, Data Mining and Computational Intelligence based education.

