Indic Subtitler: A customizatable open-source platform for generating subtitles and language learning

Kurian Benoy[1], Aldrin Jenson[2], Nabeel Boda[3], and Manu Madhavan[1]

[1] Indian Institute of Information Technology, Kottayam
kurian1998.emoa2114@iiitkottayam.ac.in
[2] Thoughtworks, Bengaluru
[3] Amity University, Noida

Abstract. This project introduces a new method for creating subtitles using advanced ASR models and Large Language Models (LLMs), such as WhisperX, Faster-Whisper, Vegam-Whisper, and SeamlessM4T. It aims to generate precise and accurate subtitles, complete with timestamps, in various languages. The project leverages the capabilities of LLMs to ensure high-quality translations and precise timing, thereby improving the accessibility and usability of multimedia content across linguistic barriers. The project is publicly available, encouraging collaboration and innovation in the research community. Moreover, a user-friendly web application has been developed and deployed on top of this framework, offering a seamless user interface (UI) and user experience (UX). This application makes it easy to access and use the subtitle generation tool and has already attracted over 500+ visitors to date. By combining LLMs with intuitive web technology, this project not only advances machine translation and accessibility in multimedia communication but also encourages widespread adoption and usage among diverse user groups.

Keywords: Subtitle generation · Language Models (LLMs) · Multilingual subtitles · Accessibility

1 Introduction

India is blessed with a vast array of languages throughout the country, boasting 22 scheduled languages as outlined in the 8th Schedule of the Indian Constitution [7]. Each state in India typically has its own official language, although some languages are shared across states. Hindi serves as the official language of India, yet it is quite common for individuals in the southern regions to not comprehend it. The majority of Indians prefer communicating and engaging with content in their native languages. This preference has led to the rise of numerous popular YouTubers in India, each with over 10 million subscribers, who produce content exclusively in their native languages. Such trends highlight the remarkable diversity within the country. Moreover, there is a widespread interest in watching

popular films from various linguistic backgrounds, exemplified by movies like Drishyam from the Mollywood industry (Kerala), which attract audiences from Hindi, Tamil, and Telugu-speaking regions, among others. In today's digital age, the accessibility and usability of multimedia content are crucial. The proliferation of fast and affordable internet has enabled individuals to explore diverse cultures, access educational resources, and engage with an extensive range of content.

Nevertheless, seamless access to digital content is continuously challenged by language barriers and the needs of individuals who are deaf. A significant portion of video content remains inaccessible, either because it is presented in unfamiliar languages or not in the viewer's native tongue. Additionally, individuals with hearing loss face further challenges in accessing audio content.

Subtitles have been identified as an effective strategy to overcome these challenges. They aid in understanding content in multiple languages and serve as an essential accessibility tool for individuals with auditory disabilities. However, creating accurate subtitles in a vast array of languages remains a significant challenge. Currently, subtitles are manually created, often only when the content garners significant interest. This leads to a stark imbalance between the amount of content produced and the subtitles generated for respective languages.

To address this discrepancy, we propose a solution-based web application that generates subtitles in 12 Indian languages and 8 European languages. The application transcribes video content from the source language to produce subtitles in that language and translates the source video's subtitles into 19 additional languages. This approach supports approximately 400 different language combinations.

Our methodology involves the use of models like WhisperX[6], Faster-Whisper[9], Vegam-Whisper[8], and SeamlessM4T[10]. The web application allows users to generate video subtitles in the desired target language via uploading video/audio in five formats, including wav, mp3, mp4, webm, and M4A, or by directly pasting YouTube video link.

The main contributions of this work are threefold. First, we provide the capability to support audio of any length, overcoming the limitations of models like SeamlessM4T[10], which generally only support up to 30 minutes. Second, we have developed a generative user interface similar to ChatGPT, as opposed to the traditional request/response process. This allows subtitles for even hour-long videos to be displayed progressively. The generation starts from language identification of source video/audio and starts showing subtitles from the first 30 seconds onward, offering users real-time accuracy reviews. Lastly, the web application is open-source and easily deployable, with a plug-and-play architecture that facilitates easy integration of new models.

2 Related Work

Whisper paper [5] was a revolutionary paper by OpenAI team. The model trained was an encoder-decoder network with almost 680,000 hours of audio data in 99

languages. The model was trained in transformers architecture and the paper [5] support tasks like English Speech Recognition, Multilingual Speech Recognition, Speech translation (X-¿ En) and (En-¿ X), language identification.

With the advent of Whisper [5], previous works like ESPNET [4] , Kaldi [3] etc became obsolete became usage started moving to newer advances in Speech Recongition and Speech Translation domain came in place.

3 Methodology

This section outlines the architecture and components of our web application designed for subtitling, which supports 20 languages directly and facilitates speech transcription and translation across 400 language combinations.

3.1 Machine Learning Models

Seamless M4T Our approach includes leveraging Meta's recently released Seamless Communication technology[10], specifically the SeamlessM4Tv2large model. This model inherently supports transcription in nearly 12 Indian languages and in total supports 101 languages, enabling us to transcribe audio in these languages and translate subtitles into various other languages. Further insights into SeamlessM4T are available in the referenced paper[10].

Faster-Whisper To enhance efficiency, we utilize a re-implementation of OpenAI Whisper model[5], known as faster-whisper[9], which employs CTranslate2 for faster inference. This version achieves up to four times the speed of the original Whisper model with equivalent accuracy and reduced memory usage. It also supports 8-bit quantization for both CPU and GPU, maintaining compatibility with all 99 languages recognized by Whisper.

WhisperX WhisperX[6] encompasses a dynamic speech recognition model engineered for generating Time-Accurate Speech Transcriptions of extended audio. This model attains unparalleled timing precision and word-level timestamps due to the innovative features within its architecture.

According to the WhisperX paper[6], its architecture comprises key elements such as:

1. Voice Activity Detection
2. A Cut and Merge mechanism that segments speech into 30-second intervals
3. Parallel processing of speech chunks through Whisper and Phoneme models
4. CTC (Connectionist Temporal Classification) forced alignment to produce sentence-level and word-level transcripts, offering notably precise timestamps in comparison to the Whisper model[5].

Vegam Whisper To enhance performance across various Indian languages, fine-tuning have been applied to the Whisper Architecture[5], enabling the creation of specialized Automatic Speech Recognition (ASR) models for those languages.

Specifically, in the case of Malayalam, the model 'thennal/whisper-medium-ml' stood out upon benchmarking. It underwent optimization with faster-whisper[9] leading to the inception of the Vegam-whisper[8] model family, tailored for improved processing speeds and accuracy.

3.2 User Interface Design

For the development of our web application's user interface, which is presented to end-users via a publicly available website, we selected a contemporary technology stack that includes the following components:

1. Next.js: Chosen for its server-side rendering capabilities, which enhance the performance and SEO of our application. Next.js also offers a simplified page routing system that streamlines the development process.
2. Tailwind CSS: Utilized for its utility-first approach to styling, allowing for rapid UI development with a focus on responsive design. This framework significantly reduces the time required to implement custom designs.
3. DaisyUI: Integrated as a plugin for Tailwind CSS, DaisyUI provides a comprehensive library of pre-designed components. This accelerates the development of visually appealing interfaces while maintaining consistency across the application.

3.3 Back-end Infrastructure

Our back-end infrastructure is designed to underpin our machine learning models with robust processing power. For the framework, we have chosen FastAPI due to its high performance and ease of use for building APIs. For hosting services, we employ modal.com, which has been selected for its capability to host machine learning models in a serverless environment while also offering access to GPUs. This choice is further justified by modal.com's attractive offer of up to $30 in free monthly credits.

1. FastAPI: Chosen for its efficiency and simplicity in API development, facilitating rapid and reliable back-end services.
2. modal.com: Selected as our hosting platform due to its support for serverless deployment and GPU access, essential for machine learning tasks, along with its cost-effective pricing model. They provide GPUs starting from the range of T4 to the latest H100 GPU.

3.4 Database Management

For the storage of subtitles within our application, we initially utilize local storage, aligning with our project's open-source nature. This approach facilitates straightforward subtitle display and management in the current phase. Looking ahead, as we contemplate the development of an AI-based Software as a Service (SaaS) platform, our strategy involves adopting Supabase. Supabase represents a compelling open-source alternative to Firebase, offering a scalable solution for our future database management needs.

4 Results and Discussion

In this section, we discuss the results of our research and the implications of our main contributions. Our work has led to significant advancements in the field of audio processing and subtitling, particularly in terms of handling audio of any length, developing a generative user interface, and creating an open-source web application for subtitling in Indian languages.

4.1 Accuracy of models

In terms of results, we haven't made any new novelties to improve WER or CER of ML models which we have used like SeamlessM4T[10], faster-whisper[9], WhisperX[6] and vegam-whisper[8]. We have the same accuracy as reported by these models. Some ideas are explored at the moment which in pre-processing and post-processing stage can improve the accuracy of the base models. This will be discussed in the conclusion section.

4.2 Our Contribution

Audio of Any Length Our first major contribution is the ability to support audio of any length. This is a significant improvement over existing models like SeamlessM4T[10], which are typically limited to handling audio up to 30 seconds long. Our model's ability to process longer audio files opens up new possibilities for applications in various fields, such as transcription services, language translation, and media entertainment. The ability to handle longer audio files also means that our model can be used in real-world scenarios where audio files often exceed the 30-second mark.

Generative User Interface The second contribution of our work is the development of a generative user interface, similar to ChatGPT. Unlike the traditional request/response process, our interface allows subtitles for even hour-long videos to be displayed progressively. The generation starts from language identification of the source video/audio and starts showing subtitles from the first 30 seconds onward. This feature offers users real-time accuracy reviews, enhancing user experience and improving the efficiency of subtitle generation. The generative user interface also allows for more interactive and dynamic interactions, which can be particularly useful in live streaming or real-time video conferencing scenarios.

Open-Source Web Application for Indian languages Lastly, we have developed an open-source web application that is easily deployable. The plug-and-play architecture of our application facilitates easy integration of new models, making it a versatile tool for developers and researchers. The open-source nature of our application also promotes transparency and reproducibility, two key aspects of scientific research. By making our application open-source, we hope to foster a collaborative environment where researchers and developers can contribute to the continuous improvement of our model.

In conclusion, our work has made significant strides in overcoming the limitations of existing models, developing a more interactive user interface, and promoting open-source research. We believe that our contributions will pave the way for more advanced and user-friendly audio processing tools in the future.

5 Use Cases

5.1 Facilitating Language Acquisition through Comprehension of Unfamiliar Languages

5.2 Enhancing Accessibility for Individuals with Disabilities, Including the Deaf and Blind Communities

6 Conclusion and Future plans

The exploration of automatic speech recognition and translation for Indic languages, as embodied by the "Indic Subtitler" project, is a testament to the ongoing innovations in machine learning and natural language processing. By drawing on the capabilities of models like SeamlessM4T and fine-tuned Whisper, alongside developments aimed at enhancing processing speed and efficiency, this project contributes valuable insights to the discourse on linguistic inclusivity and accessibility in the digital age. It aligns with a burgeoning body of work dedicated to ensuring that technology serves as a bridge, rather than a barrier, to global communication and understanding.

References

1. S. Alharbi et al., "Automatic Speech Recognition: Systematic Literature Review," *IEEE Access*, vol. 9, pp. 131858–131876, 2021, doi: 10.1109/ACCESS.2021.3112535. Keywords: Systematics; Software; Automatic speech recognition; Quality assessment; Nails; Databases; Speech recognition; automatic speech recognition; ASR systematic review; ASR challenges.
2. Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case, and Paulius Micikevicius. "OpenSeq2Seq: extensible toolkit for distributed and mixed precision training of sequence-to-sequence models," May 2018.
3. Julian Linke, Saskia Wepner, Gernot Kubin, and Barbara Schuppler. "Using Kaldi for Automatic Speech Recognition of Conversational Austrian German," arXiv:2301.06475 [cs.CL], January 2023.

4. Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. "ESPnet: End-to-End Speech Processing Toolkit," arXiv:1804.00015 [cs.CL], April 2018.
5. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey and Ilya Sutskever. "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv:2212.04356 [cs.CL], October 2022.
6. Max Bain, Jaesung Huh, Tengda Han and Andrew Zisserman. "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," Proc. Interspeech 2023, March 2023.
7. 8th Schedule of India Constitution. https://www.mha.gov.in/sites/default/files/EighthSchedule$_1$9052017.*pdf*
8. Benoy Kurian, Swathra Malayalam Computing Project https://huggingface.co/smcproject/vegam-whisper-medium-ml
9. Guillaume Klein, SYSTRAN https://github.com/SYSTRAN/faster-whisper
10. Barrault, L., Chung, Y. A., Meglioli, M. C., et al "SeamlessM4T-Massively Multilingual and Multimodal Machine Translation," AI Meta Publications, August, 2023 https://ai.meta.com/research/publications/seamlessm4t-massively-multilingual-multimodal-machine-translation/