Instant Transcription and Translation Tool using OpenAI's Whisper ASR Model

Akarsh Ghale[1], **Janaki .K**[2], **Devaraj Verma .C**[3]

Department Computer Science and Engineering - Artificial Intelligence, Faculty of Engineering and Technology,
Jain Deemed-To-Be University, Karnataka, India

19btrci053[at]jainuniversity.ac.in
k.janaki[at]jainuniversity.ac.in

Abstract: ***In the fast-paced world of journalism, translation is becoming increasingly important in order to communicate news & events across the internet. However, traditional translation methods can often be slow and inaccurate, leading to potential miscommunication. With the help of new technology, such as machine translation, translations can be made more quickly and accurately. This is important in order to ensure that the information being communicated is correct and can be understood by the intended audience and also to avoid any possibilities of spreading misinformation that is becoming increasingly rampant during the digital age. As a result, we propose architecture to support the instant & accessible translation solution using OpenAI's state-of-the-art Automatic Speech Recognition and Translation AI model Whisper.***

Keywords: Translation, Transcription, Whisper, OpenAI, Automatic Speech Recognition

1. Introduction

In this era where we have the internet and social media, people are able to communicate with others across the globe without any barrier of time and space. With this, there is an increase in the use of video sharing platforms such as YouTube, Facebook and other social media websites which are creating a new trend in communication and making people more aware about global issues and critical events. As a result, translation has become integral due to the exponential rise of videos being uploaded on the internet in different languages that has resulted in creation of language barriers which leads to a widening "information gap".

Traditional translation methods relied either on manual effort that entailed hiring a competent translator or usage of expensive translation systems both of which are expensive, inaccessible and can often be slow and inaccurate, leading to potential miscommunication that further leads to increase in spread of misinformation. These methods were also specific to certain use-cases that limited their scope of applications. In order to reap the benefits of translation for smooth transferring of information and to maintain the integrity of the information, translation solutions shouldn't just be fast but also need to be accessible.

The earliest speech transcription and translation systems were rule-based and relied on text processing pipelines to translate text transcripts of speech. The pipeline approach was later replaced by a single, end-to-end trainable model, which achieved promising results despite the lack of training data. More recently, the trend has been to combine end-to-end models with pre-trained ASR (Automatic Speech Recognition) and MT(Machine Translation) models. This has led to a significant increase in translation quality.

2. Related Works

Many researchers have explored different architectures and models with the objective of overcoming language barriers in different settings.

1) Kristin N Dew, Anne M Turner and others have reviewed 27 papers and explored how machine translation is being developed to overcome language barriers in health settings. They have analyzed that for a variety of systems performance was best for translations of simple and less technical sentences and limited to translation from English to western European languages. They have also discovered translation accuracy concerns have limited the use of MT in clinical settings.
2) Dario Franceschini, Chiara Canton and others have presented a communication platform to tackle highly multilingual live speech translation for conferences and remote meetings live subtitling. Their tests showed that their system reached practical usability using PerVoice Architecture and proves its applicability in challenging settings. They have also shown that the current and future main challenges to address are related to improving speech recognition particularly for non native dialects and out-of-vocabulary words.
3) Bojar and Andrej have presented an automatic speech translation system with the objective of live subtitling of conference presentations. Their system was able to recognize English, Czech, and German Speech and present it translated simultaneously into 42 target languages.

3. Proposed Methodology

The main objectives of our proposed system are to address the following issues commonly encountered during a translation process:

- **Biased Translations -** Sometimes in a political crisis, having a local translator can lead to biased translations. It's essential to have trustworthy translations to maintain transparency & accuracy of information.
- **Cultural nuances-** It's hard to include the "cultural nuances" and overcome language barriers while translating news. Each language has its idioms, figures of speech, wordplay, etc. Hence, it is imperative to ensure that the translated message bears the same meaning as the original so that all who come across it can be on the same page.
- **Time pressure-** Timing is everything. Every online media outlet is unique, but for some media the publishing time is essential. It's important to be able to have fast and accurate translations on time to ensure accurate reporting & prevent spread of misinformation.
- **Inaccessible Resources-** As translation is difficult, the access to resources is expensive and limited thus making it inaccessible to the wider masses.

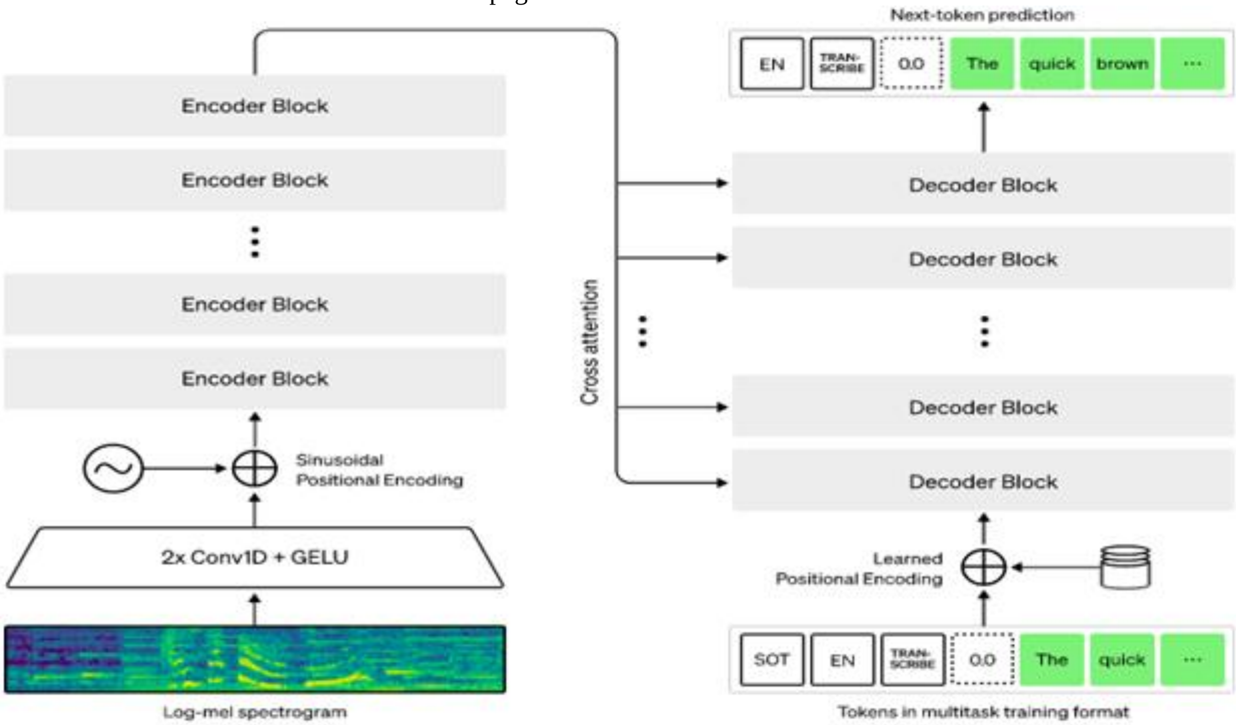

Figure 1: Proposed architecture of Instant Transcription and Translation Tool using OpenAI's Whisper ASR Model

The proposed solution includes adding subtitles in required language which are transcripted in an instant. When there is an audio track present on the video, it is automatically translated into required language through Open AI's whisper model that can then translate it into text form. This enables the viewers to read what is being said on screen without having to download any files or apps for that matter just like Google translate does now.

3.1 Model Selection

This paper entails the use of OpenAI's Pre-trained Whisper model. Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. This model has shown great robustness and accuracy in understanding diverse accents and technical language even in environments that contained background noises. It can transcript multiple languages and can then translate those into English thereby making it a perfect selection for our use-case where diversity is paramount.

[4]This model follows an end-to-end approach that utilizes Transformers architecture implementation that contains two components, an encoder and decoder Transformer functions. The input audio is split into chunks of 30 seconds and converted into a Log-mel spectrogram which is then passed into the encoder. The decoder function, on the other hand, is trained to predict the corresponding text caption. A special token is utilized as a way to instruct the model to perform a variety of tasks supported by the model such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation. Fig 2 displays the architecture of the model discussed above.

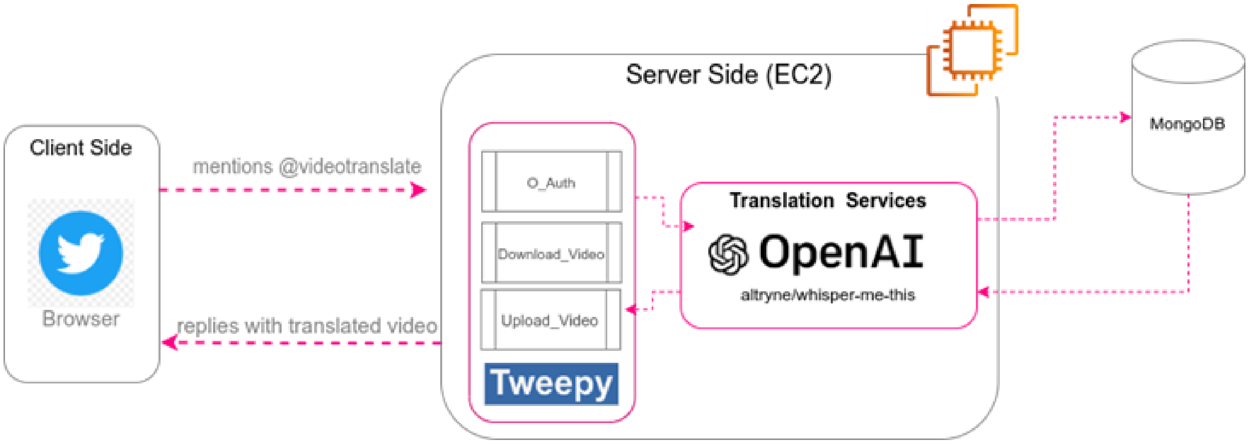

Figure 2: OpenAI Whisper model architecture by OpenAI [4]

3.2 Server Side Architecture

As shown in Fig 1, the model is hosted on Amazon AWS EC2 instance with specifications displayed in Table 1 and communicates with client side through the use of APIs supported by Tweepy [6] which is an open source library specifically built to access Twitter APIs using Python. The following are the system specifications for EC2 instance:

Table 1: Specifications of Amazon AWS EC2 instance for backend hosting.

EC2 specifications	
Instance Type	g5.xlarge
Operating System	Ubuntu 20.04 LTS
Memory	16GiB
vCPUs	4

Libraries Utilized for supporting the architecture:

Table 2: Softwares & Libraries utilized in the architecture

Software & Libraries used	
Python 3.7	Back-end Language.
Tweepy	Python Library for accessing the twitter API.
altryne/ Whisper-me-this	Open Source library for automating transcription, translation and subtitling a video using OpenAI's whisper.
MongoDB	NoSQL Database

This is a simple example of how a video can be translated instantly when a user comes across a video content in another language. It also demonstrates how multiple services can work together to provide a better user experience.

The architecture has three core functions built in:

1) Receives notification that a user has mentioned the bot using '@videotranslate translate' and begins the Download_Video event where it downloads the video of the post where it was mentioned and handles the translation service.
2) Translation Service is responsible for extracting the audio and is passed into the Whisper model which takes care of identifying the language, transcripting it into English and generating a subtitle. The Generated subtitle is burnt to the video. All of the following is supported by the open source library Whisper-me-this[5].
3) Begins the Upload_Video event where the bot posts uploads the video with translated subtitles in English with a reply on the user's comment.

A MongoDB [7] database is utilized in order to store processing and other back-end logs necessary for error-processing and to keep track of usage metrics.

All the softwares and libraries utilized above are outlined in Table 2.

4. Results and Discussion

We have found that this architecture supported by the OpenAI Whisper is practical and highly applicable. The model addresses the common issues discussed in the paper.

Model retained robust performance in speech recognition with consistency and accuracy when faced with environments where the speech quality was low and filled with audio disturbances.

Based on empirical observations, whisper has turned out to be really good at understanding cultural nuances where there is a lot of slang and specific terms with references that were not easy to translate. In these cases, the model performed exceptionally well in retaining those references and nuances in the translation process.

This architecture solves the time latency and accessibility issue compared to traditional processes as the user can just simply mention the Twitter bot and then get served with a translated subtitled video within a few minutes making it especially useful in environments that have time pressures.

As the model handles all the process of translation without little to none input, and drastically cheap ensures high accessibility that is imperative in order to support adoption.

5. Conclusion

In this study, we proposed an architecture for instant translation using Open AI's state-of-the-art Whisper Model. The proposed solution successfully tackles the common issue faced during the translation process as outlined in the paper and is a viable solution keeping accessibility, usability and instancy in mind. This architecture can be used to decrease the information gap due to language barriers on the internet.

As the model is trained on an unsupervised dataset, it picks up bias that is inherently built into the data. Although this bias could prove as an advantage in retaining the cultural nuances and references but could also incur potential issues especially in use during a high-stakes environment for instance, in international diplomacy.

In order to tackle this, a standardized reporting could be utilized to uncover all the possible biases and could prove to be an effective way to generate awareness about its limitations so prior steps could be undertaken to minimize its effects.

Although the translation accuracy is adequate, the model does face some difficulties in transcribing and translating low-resource languages. This can be addressed with newer versions of the model with increased capabilities with future releases by OpenAI.

This study focuses on Twitter as a central focus, however this architecture is flexible enough to be used on other social media platforms as long as the platforms support API services.

References

[1] Dew, Kristin N., and Anne M. Turner. Development of machine translation technology for assisting health communication: A systematic review. vol. 85, Journal of Biomedical Informatics, 2018. Development of machine translation technology for assisting health communication: A systematic review, https://www.sciencedirect.com/science/article/pii/S1532046418301448. Accessed 16 Nov 2022.

[2] Dario Franceschini, Chiara Canton, Ivan Simonini, Armin Schweinfurth, Adelheid Glott, Sebastian Stüker, Thai-Son Nguyen, Felix Schneider, Thanh-Le Ha, Alex Waibel, Barry Haddow, Philip Williams, Rico Sennrich, Ondřej Bojar, Sangeet Sagar, Dominik Macháček, and Otakar Smrž. 2020. Removing European Language Barriers with Innovative Machine Translation Technology. In Proceedings of the 1st International Workshop on Language Technology Platforms, pages 44–49, Marseille, France. European Language Resources Association.

[3] Bojar, Ondřej; Macháček, Dominik; Sagar, Sangeet; Smrž, Otakar; Kratochvíl, Jonáš; Polák, Peter; Ansari, Ebrahim; Mahmoudi, Mohammad; Kumar, Rishu; Franceschini, Dario; Canton, Chiara; Simonini, Ivan; Nguyen, Thai-Son; Schneider, Felix; Stüker, Sebastian; Waibel, Alex; Haddow, Barry; Sennrich, Rico; Williams, Philip (2021). ELITR Multilingual Live Subtitling: Demo and Strategy. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Online, 1 April 2021, ACL Anthology.

[4] OpenAI. "Introducing Whisper." OpenAI, 21 September 2022, https://openai.com/blog/whisper/. Accessed 16 November 2022.

[5] https://github.com/altryne/whisper-me-this

[6] https://www.tweepy.org/

[7] https://www.mongodb.com/