Over the air Voice Translation

Yash Bhandare , Winne Har Mark Trovinger and Atharva Atre
Department of Computer Science
Project Advisor and Sponsor: Dr. Zesheng Chen

Feel free to try out the application at:
https://10.13.11.172:5051
https://10.13.11.153:5051

Introduction

When it comes to unsolved problems in deep learning, real-time translation is at the top of the list. The ability to translate in real-time, with a high degree of accuracy, a person's speech, has long been a goal of researchers in natural language processing. With the release of OpenAI's Whisper model, the goal is closer than it ever has been. The goal of this project is to leverage Whisper, along with other open-source tools, to build an application for OTA translation.

Methodology

In this project, we built an application which has functionalities of over the air voice translation from English to 30+ different languages and vice-versa. We also have the functionality to auto-generate English subtitles for non-English videos. First we take the input of 'Language tag' along with the audio. Then for the duration of the transcribe timeout, the audio is divided into smaller segments and , text tokens are generated. These test tokens are then fed to out to our translate model which converts it into the desired language

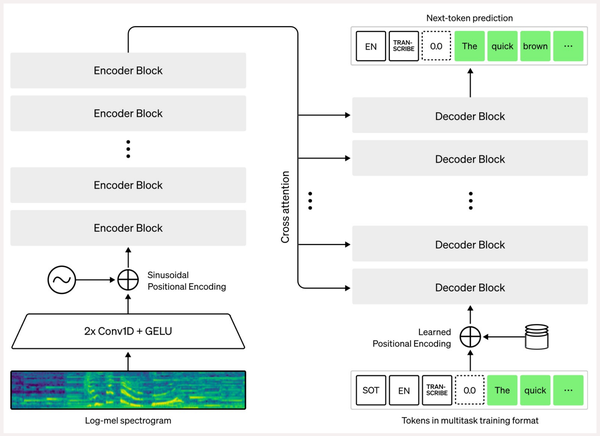

Figure 1. Overview of the Whisper AI mode

Dataset	wav2vec 2.0 Large 960h	Whisper Large	RER (%)
LibriSpeech test-clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.7**	72.7
Fleurs (English)	14.6	**4.6**	68.5
Common Voice	29.9	**9.5**	68.2
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.6**	61.1
WSJ	7.7	**3.1**	59.7
VoxPopuli (English)	17.9	**7.3**	59.2
AMI-IHM	37.0	**16.4**	55.7
CallHome	34.8	**15.8**	54.6
Switchboard	28.3	**13.1**	53.7
CORAAL	38.3	**19.4**	49.3
AMI-SDM1	67.6	**36.9**	45.4
LibriSpeech test-other	6.2	**5.6**	9.7
Average	29.5	**12.9**	55.4

Figure 2. Training Dataset

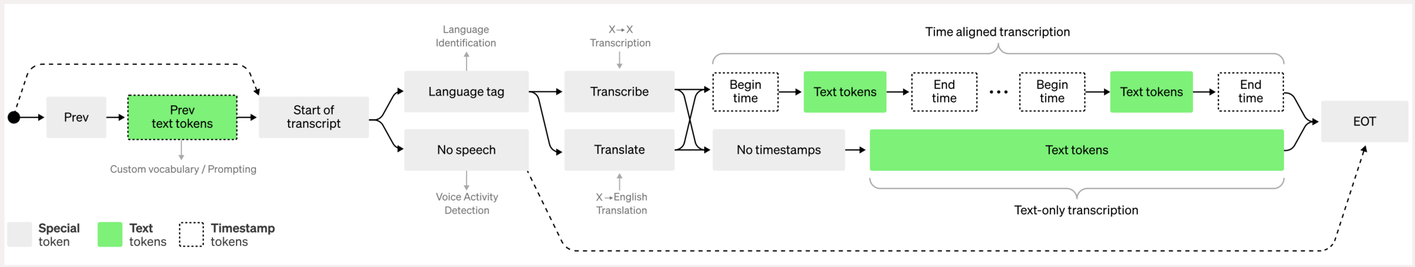

Figure 3. Overview of translate and transcribe system design

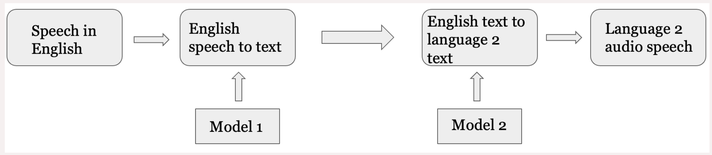

Figure 4. Block Diagram of System

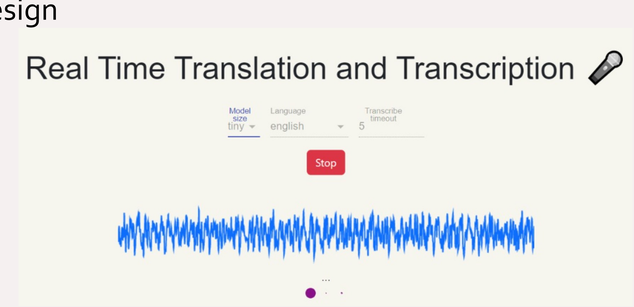

Figure 5. User Interface used to record and translate/transcribe audio

Technology and Features

To support several deep learning features and models, we used Flask micro-framework as our backend. For our frontend, we are using React , which has JavaScript and HTML/ CSS. We are using the open-source whisper AI model for translation and transcription.

Conclusion & Future

In conclusion, we created an application that hosts over the air voice translation as well as transcription. The webapp also has the functionality of generating subtitles of videos in different languages. This can can be used by Educational Institutions for understanding lectures / videos from a different language. It can also be used for over the air translations for people with no common language to communicate to each other in real time. There are several improvements which need to be made , ranging from translation time to accuracy across different languages. We hope to identify and try to improve on the model in the future.