Using a “Computer” as an ASL translator to facilitate the “Interaction” between “Humans”

1. Abstract

There are over 500 million people who suffer from hearing impairment and cannot speak. We can build a translator that helps deaf people communicate, just like normal people do with the help of cutting-edge Deep Learning technologies. In this project, I have developed an automated English audio to ASL video translator using Neural Machine Translation with Attention Mechanism with fairly low latency. The translator takes input in form of Audio and generates a video containing American Sign Language gestures. The scope of the project is now limited to one language, however, future scopes are limitless and will be discussed.

2. Introduction

The number of people having a hearing impairment is expected to be doubled by 2050. According to the USA Bureau of Labor Statistics, the requirement of American Sign Language Translators is expected to have an increase of 24% in coming years. Hence, to tackle this sudden increase in the requirement, we can build a Translator for Sign Language using Deep Learning Techniques that can help deaf people interact with us easily and effectively. The project also performs some basic Qualitative Analysis by taking reviews from peers.

3. Related Work

Multiple attempts have been made to facilitate the communication between us and people having a hearing impairment. A company called Acesso para Todos has created an app “Hand Talk” [1] that helps deaf people understand the English sentence by generating a Sign Language Gesture using their avatar called “Hugo”. The implementation details of their ASL language conversion are unknown. However, as reported by a few of the users of the app on the Google Play store, this app sometimes fails to generate an ASL sentence that is grammatically correct according to some users of the app.

Text2Sign [2] makes an impressive attempt to generate American Sign Language from the input text with the help of Deep Learning Techniques such as Generative Adversarial Network(GAN). Dynamic GAN also makes an effort to improve existing results by creating Sign Language Video using Dynamic GAN. These solutions provide an interesting way to create Sign Language Video however, fail to

incorporate the complexity of ASL grammar. A few of the improvement approaches seem unrealistic and hard to implement. Also, as these approaches use GAN, the resulting video/image produced by the GAN may be blurry and the viewer may find it difficult to understand the generated output.

A few of the Universities in the USA such as Purdue University [4] and Boston University [5] are also working in this field and have also open-sourced their ASL images and videos databases.

4. Technical Part (3-4 pg)

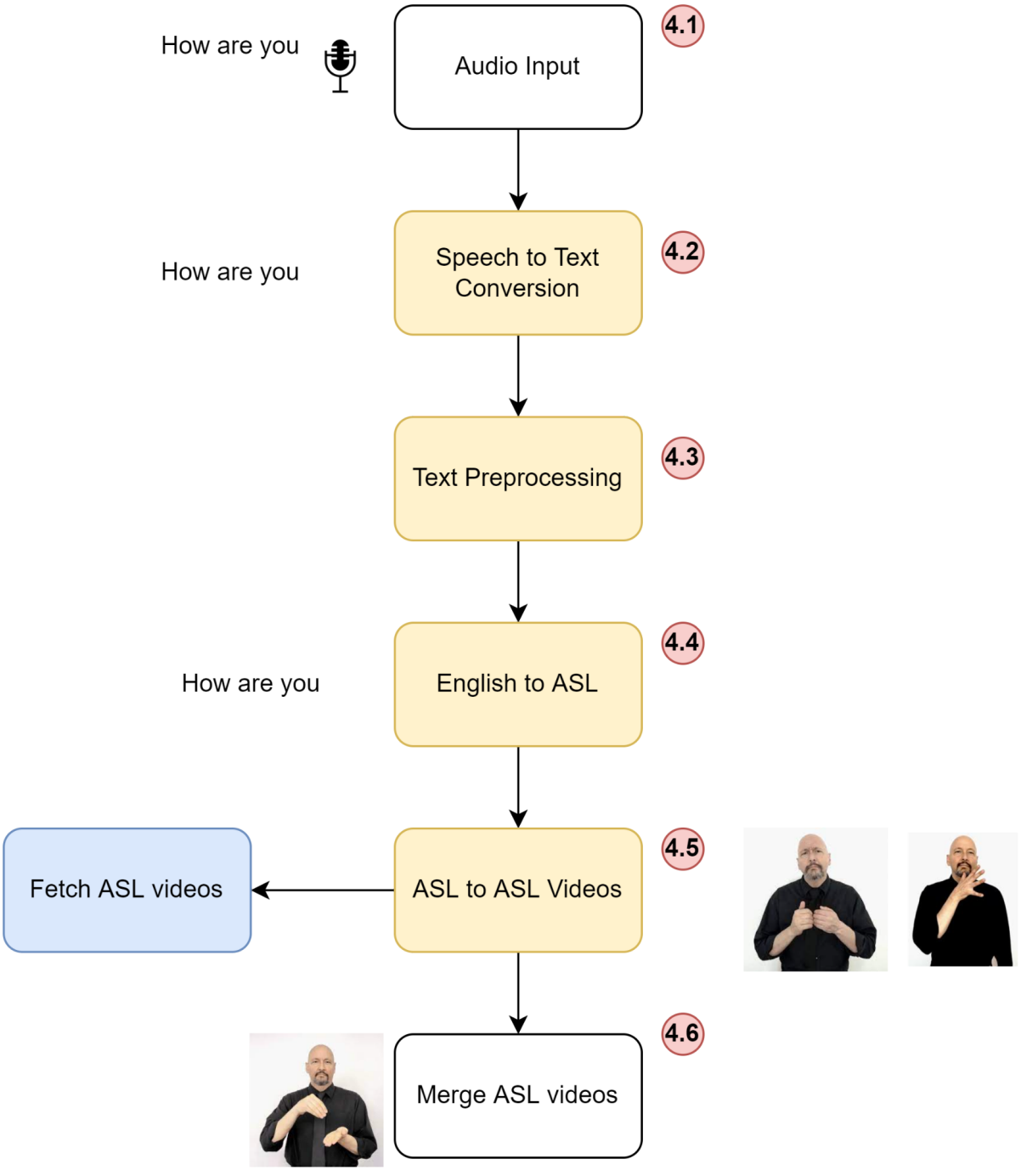

Figure 1

Figure 1 displays the flowchart of the project. All the steps are described below in detail.

4.1 Audio Input

For taking input, the microphone of the computer is used. speech_recognition module turns on the microphone and listens for 7 seconds. (More details can be found in the Demo file attached with the project)

4.2 Speech to Text

Recorded Audio file is converted into Text using speech_recognition's recognize_google API.

4.3 Preprocessing

Pre-processing of the English sentence is performed both while recording the audio and converting it to text to avoid noise and redundant details.

4.4 English to ASL

English to ASL conversion is done using Attention Mechanism[8]. Attention Mechanism is built on Neural Machine Translation which is used for Language Translation. It assigns waits to previous words while translating the current word to the expected language. Basically, it maintains the context of previous words while performing Machine Translation which improves the overall accuracy of machine translation.

4.5 ASL to ASL video

ASL gesture for each of the words generated is scraped from Signing Savvy | ASL Sign Language Video Dictionary website. The sentence is divided into words and each word is searched on the Signing Savy website for finding corresponding American Sign Language gesture. There are chances that a word might have multiple possible meanings and multiple gestures on the website. Hence, to handle this scenario efficiently, the scraping logic is coded in such a way that when there are multiple similar words found, the logic selects the first matching word and considers that word for generating ASL gesture video. The logic also stores the collected video data under a folder ASL_Gestures. When the same word occurs in the next sentences, it directly gets the stored gesture for that particular word. In

this way, we can optimize the execution time. Each gesture videos is stored at 24 fps which reduces the disk utilization of the logic because of smaller video sizes.

4.6 Merge ASL video

Generated videos of various ASL words are merged according to their places in the grammar and the final video is written to the disk. The final video is also displayed to the user on the rendered HTML page as shown in the demo video.

The application also provides an error screen when something goes wrong. Possible reasons for an error are issues in microphone recording, internet connection, issues in Speech to Text conversion etc.

5. Validation (1 page)

Finding people who are actually suffering from hearing loss was a bit difficult task for me. Hence, I have used a Qualitative Evaluation Technique for analyzing ease of use and possible issues, and bottlenecks in the project. The developed project was given to 5 persons for collecting their reviews. 2 of them had a basic knowledge about common Sign Language words (I provided a pdf containing gestures for commonly used Sign Language words and their corresponding gestures) and the other 3 did not have any idea about the grammar as well as the gestures. The possible words and sentences supported by the project were provided to the users.
Out of 5 persons, 4 persons found the UI interesting and easy to use. 1 person faced an issue while converting Spoken English sentences to text for 1 sentence out of 5 testing sentences. (Issue in Speech2Text). For 2 persons, the system displayed the Sign Language video of the wrong English word (Issue in English to ASL-text conversion). For 1 person, the system could not find a Sign Language Video for a few words in the spoken sentence (Issue in Video Library used).

The following table represents the users' reviews in a few criteria decided.

User	Ease of Use	Execution Time	Interactive UI	Accuracy	Ease of Setup
1	4	3	4	3	3
2	4	3	4	3	4
3	5	5	3	3	4
4	4	5	4	4	3
5	4	4	5	3	4

As we can see, the system received a rating of 3 in Accuracy. This suggests that the Model used for converting English Sentences needs to be trained effectively. Also, some users reported that the words supported by the systems were limited and they found it a bit difficult to form sentences using those words. This limitation is because of the availability of the free Datasets of American Sign Language. This issue can be solved by collecting more sentences using some paid websites or other possible sources.

6. Future Work

This project explores a new way of communicating with people having hearing impairment or deafness with the help of Deep Learning and NLP techniques. The accuracy of the Language Model used in the project can be improved by collecting more ASL text data. Due to a lack of time and resources, the project fails to explore other possible ASL grammar datasets and ASL gesture video datasets. The accuracy of the project heavily depends on the availability of the Sign Language Video on the selected website. Instead of scraping the video from the website, a more intuitive approach such as generating a gesture video using GANs [2] can be used incorporated in order to reduce the dependency on the ASL video libraries and websites. An avatar-based solution can also be provided to show the results in a more intuitive and understandable way.

As of now, I only aim to translate from one audio language to one sign language. However, there are over 7000 spoken languages today and tens of other sign languages. Each language consisting hundreds of domains to work for; education, entertainment, communication, etc.

One more area that can be explored here is chatbots. Deaf people don't really have friends to talk to. This project can turn into a potential business model. With the right set of people and necessary resources, this can be a boon for the deaf community,

7. Conclusion

Hence, through this project, I got to learn a lot about machine translation. I tried to address an existing problem in the market and got an amazing opportunity to explore the domain in a deep and different manner. It is difficult but possible to model sign language and translate audio to sign language. The latency is a little higher than expected. Hence, the model can not be used in real-time. However, it can be improved by implementing an efficient video generation mechanism.

References

[1] Hand Talk App | Sign Language Translator in the palm of your hand

[2] Stoll, S., Camgoz, N.C., Hadfield, S. et al. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. Int J Comput Vis 128, 891–908 (2020). https://doi.org/10.1007/s11263-019-01281-2

[3] Bala, Natarajan & R., Elakkiya. (2021). Dynamic GAN for High-Quality Sign Language Video Generation from Skeletal poses using Generative Adversarial Networks. 10.21203/rs.3.rs-766083/v1.

[4] RVL-SLLL American Sign Language Database (purdue.edu)

[5] DAI - ASLLVD (bu.edu)

[6] https://pubmed.ncbi.nlm.nih.gov/16177267/

[7] Word Level English to Marathi Neural Machine Translation using Encoder-Decoder Model | by Harshall Lamba | Towards Data Science

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).