# Connect – A Sign Language Translation Tool

Jinal Pawar
*Department of Electronics Engineering*
Vidyalankar Institute of Technology
*Mumbai, India*
jinal.pawar@vit.edu.in

Rashi Raut
*Department of Electronics Engineering*
Vidyalankar Institute of Technology
*Mumbai, India*
rashi.raut@vit.edu.in

Sumeeth Moolya
*Department of Electronics Engineering*
Vidyalankar Institute of Technology
*Mumbai, India*
sumeeth.moolya@vit.edu.in

Yash Jadhav
*Department of Electronics Engineering*
Vidyalankar Institute of Technology
*Mumbai, India*
yash.jadhav@vit.edu.in

Akhil Masurkar
*Department of Electronics Engineering*
Vidyalankar Institute of Technology
*Mumbai, India*
akhil.masurkar@vit.edu.in

***Abstract*—Connect is a real-time sign language translation tool. It uses Mediapipe's Hand tracking solution to obtain 21 3D hand landmarks from an image. The 3D parameters are squashed into a single parameter by finding their euclidean distance. A new 'angle' parameter is added. A dataset of ASL alphabets is put through this system to create a 42-D database (21 distance & 21 angle parameters) to train ML models. Out of all the models tried, Random Forest provides the best accuracy of 89%. From the result, it can be observed that it is possible to produce a fully generalizable translator for more than just letters of the ASL.***

***Keywords*—Sign Language Translation.**

## I. INTRODUCTION (*HEADING 1*)

Though sign language speakers number in millions, most of the society remains inaccessible to them, making getting through day-to-day life unnecessarily difficult. Hard of hearing (HoH) and non-speaking people are often treated as a liability by a society which believes that they cannot work or perform tasks as efficiently as the hearing and speaking public. Lack of proper communication methods as well as previously deep-rooted prejudices against disabilities impact various aspects of their life. A study in USA suggests that in 2017 while employment rates for hearing people was 75.8%, it lied at a low 53.3% for deaf people [1]; whereas 89.4% of hearing adults completed their high school education while the number lied at 83.7% for their deaf peers [2]. Most places do not teach sign language in schools or other institutions, which alienates sign language speakers from the rest of the world. Interactions with non-sign language speakers require methods like lip-reading or a human translator to make communication possible. However effective these methods are, they are rather difficult to dispense in times of an emergency or even otherwise. A system that can recognize and translate Sign Language can solve this problem. Our aim is to make communication between two groups of sign language speakers and non-sign language speakers easier. A vision-based approach to our solution attempts to reduce the requirement of human translators and increase dependency on a readily available device for translation.

## II. LITERATURE

A real time sign-language translation tool is an important aid in facilitating communication between HoH, non-speaking & hearing, speaking public. A few technological advances made in this field of research are stated below.

### A. *ASLAN*

'ASLAN' [3] is a 3D printed robotic arm that can translate words into sign language for HoH people, created by a Belgian team of scientists at the University of Antwerp. It'd take written and spoken words as input and sign the words using 'Fingerspelling' - a subset of sign language where words are signed letter-by-letter. The prototype needs a computer for processing. However, they intend to create the final product compact enough to be carried in a rucksack.

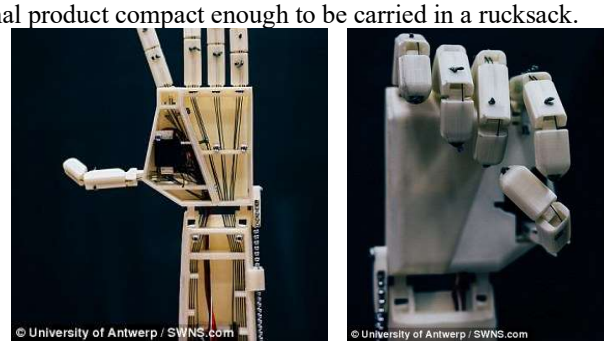


Fig. 1. ASLAN – A robotic arm for sign language translation

This surely acts as a step towards an inclusive society, however its accessibility remains questionable as the device can cost about 400 EUR ($560), which though cheap when compared to usual 3D-printed items of this scale, it could still be expensive for the common person. Even If the financial and portability concerns were taken care of, and the device's function expands from simply fingerspelling to other signs as well, it'd still only facilitate one-way communication as it does not appear to translate signs back to spoken or written words.

### B. *Microsoft's Live Captioning*

Live Captioning is a widely used method to convert spoken text to written to help HoH people understand the speech as well. Due to the pandemic, this technique has certainly proven a necessity as everything has been moved online and seminars, lectures etc might not have live sign language interpreters as before. Live Captioning removes the need for a human sign language interpreter. A blog by Microsoft [4] details the use of Microsoft Translator in a biology class in Rochester Institute of Technology. This translator, with the help of AI, creates captions for live speech by converting raw spoken language into fluent, punctuated text. It supports high quality translations into more than 60 languages. However, it isn't entirely accurate; for example, it can sometimes miss the difference between 'I' and 'Eye'. Though it doesn't address two-way communication either, Live Captioning surely is a practical approach to make

public spaces such as classrooms, seminars and even online spaces such as Youtube, etc more accessible for HoH people. The community of individuals who are deaf and hard of hearing recognized this cleaned-up and punctuated text as an excellent tool to access spoken language and not just ASL. This Microsoft Translator app looks promising since it receives the captions in real time within the language of their choice because there was no waiting time and he/she can get the information at the same time as their hearing peers, and will help them to stay up with the rest of their class.

### C. *Mediapipe's Hands: On-Device Real-Time Hand Tracking*

Mediapipe created a real-time on-device hand tracking solution [5] that predicts a hand skeleton of a human from a single RGB camera. It utilizes an ML pipeline that consists of two models working together:

- A palm detector that will be operating on a full input image and locating palms via an oriented hand bounding box.
- A hand landmark model that operates on the cropped hand bounding box provided by the palm detector and returning high-fidelity 2.5D landmarks.

After running palm detection over the whole image, the subsequent hand landmark model performs precise landmark localization of 21 2.5D coordinates inside the detected hand regions via regression. The model learns a consistent internal hand pose representation and is robust even to partially visible hands and self-occlusions. The model has three outputs:

1. 21 hand landmarks consisting of x, y, and relative depth.
2. A hand flag indicating the probability of hand presence in the input image.
3. A binary classification of handedness, for example, left or right hand.

### D. *Desired Approach*

From the above examples, it is clear that though the need to bridge this communication gap is being realized, we still do not have a solution that facilitates two-way communication as ultimately desired. Hence, we want to develop a solution that can address this problem. Our project - Connect - aims to translate sign language into audio and text as well as spoken words into text. To accomplish the former, we use Mediapipe's 'On-Device Real-Time Hand Tracking' solution.

## III. WORKING

Our proposal is based on creating an app capable of translating Sign Language to audio as well as text in real time. To facilitate communication the other way around, we also include capability to convert speech to text to be displayed on app screen. To do so, we have organized our app into two sections, or modes:

- MODE 1: Sign
- MODE 2: Speak

Here, Mode 1 translates signs of a person speaking in sign language into audio and text in real time whereas Mode 2 translates speech into text and displays it on screen.

To focus on creating proof of concept, we have only designed and created a Desktop GUI for this app. In future, it can be developed into an API for a mobile app as well.

The working of Connect can be understood through this flowchart:

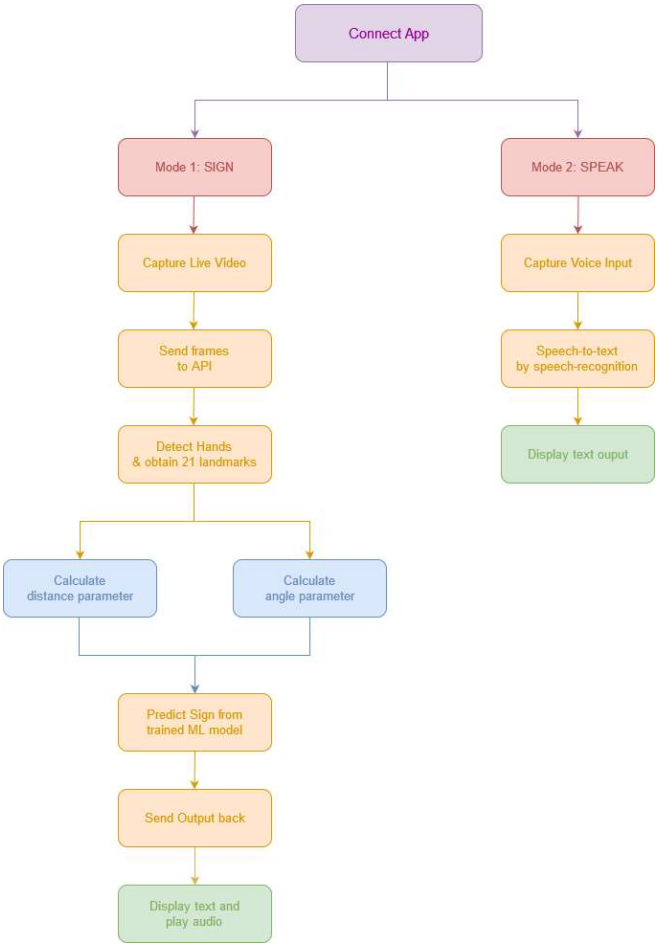


Fig. 2. Connect working overview

## IV. METHODOLOGY

Our project's methodology can again best be understood separately through the two modes.

### A. *Mode 1*

Mediapipe's "On-device Real-time Hand Tracking Solution" makes use of two ML models to achieve accurate hand tracking with landmarks. The number of hands to be detected can be varied by their "handedness" parameter. For our purposes, we limited the number of hands to two to achieve accurate desired results.

The second ML model of Mediapipe's solution provides us with 21 landmarks on each hand detected in the image. Each landmark is given an id going from 0 to 20 as shown in the figure. The ids remain the same throughout any hand movements. Each id carries three parameteres - 'X', 'Y', 'Z' - corresponding to usual location parameters where 'X' and 'Y' represent the location with respect to X and Y axes while 'Z' represents the depth.

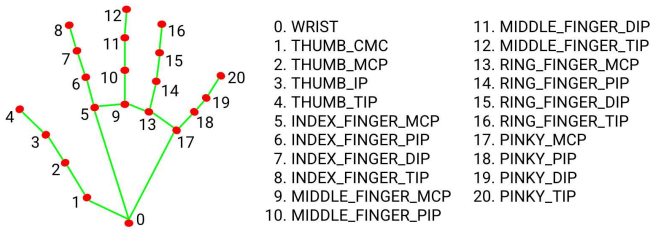


Fig. 3. Hand Landmarks ids and locations

To focus on creating a proof of concept, we will be creating a final product equipped to detect letter of ASL. Focusing on singular letters allows us to create a strong foundation to further build upon our project. Hence, we began by using an ASL alphabets dataset with about 30 images for each alphabet and digit. These datasets are then put through Mediapipe's solution where their ML models detect the palms and provide 21 landmarks for each hand in the image. Using this, we have created our own database where each alphabet contains 21 landmarks' information detected from at least 20 images from the original dataset.

*1) Preprocessing*

We determined that the most logical approach would be to try and squash our 3 parameter-ed landmark to a single metric using Euclidean distance formula. We did so because training our model directly on X, Y and Z parameters of landmarks could lead to inaccurate results as a change in position of the hand with respect to the dataset (for example, the hand is positioned in lower right corner of the image whereas all hands in dataset were centered) would classify our input as a different or incorrect sign. Hence, it makes more sense to convert the 3 parameters into a single metric which judges them in relation to each other, as opposed to individually. We then normalized the data by making the 0th landmark the origin.

In addition to that, we also created a new 'angle' parameter by choosing two arbitrary landmarks (0th and 7th in our case) to form the base line and calculate the angle of each landmark with respect to this point.

Thus, after the above two preprocessing steps, we end up with a 42-D database of landmark information.

*2) Training ML Models*

Once we had our desired database, we began the model selection and training process. We used the SKLearn library for this purpose. Since our data was imbalanced, we opted for Stratified K-fold cross validation to ensure a proportionate distribution of the data into test and training sections. We tried out SVM, KNN, Random Forest and XGBoost models and obtained the following accuracies.

| ML models | Kernel | Accuracy |
|---|---|---|
| SVM | RBF | 61.0389 |
| | Polynomial | 59.74 |
| | Linear | 79.22 |
| KNN | N = 1 | 77.92 |
| | N = 2 | 70.129 |
| | N = 3 | 71.428 |
| | N = 4 | 70.129 |
| | N = 5 | 71.42 |
| | N = 6 | 72.727 |
| | N = 7 | 68.83 |
| | N = 8 | 70.102 |
| | N = 9 | 66.23 |
| | N = 10 | 66.23 |
| Random Forest | - | 76.623 |
| XGBoost | | 72.72 |

Fig. 4. Accuracies

To improve our accuracies further, we used a larger dataset. Hence, for round 2 of training, we opted for a dataset with 3000 images per alphabet [6]. However, here we ran into a problem that seems unsolvable at our hands. Mediapipe's algorithm is unable to detect hands from all the images. This leads us to have an even more imbalanced dataset with minimal change in accuracies.

| Letter | Sample count |
|---|---|
| Z | 2554 |
| R | 328 |
| L | 237 |
| W | 216 |
| C | 205 |
| I | 187 |
| B | 186 |
| S | 182 |
| K | 178 |
| E | 177 |
| Y | 172 |
| D | 169 |
| X | 154 |
| V | 149 |
| A | 149 |
| O | 132 |
| U | 128 |
| F | 121 |
| H | 119 |
| G | 119 |
| P | 87 |
| M | 80 |
| N | 66 |
| T | 55 |
| Q | 53 |
| J | 17 |

Fig. 5. Imbalanced Dataset

| ML models | Kernel | Accuracy |
|---|---|---|
| SVM | RBF | 77.73 |
| | Polynomial | 80.064 |
| | Linear | 88.665 |
| KNN | N = 1 | 88.1028 |
| | N = 2 | 85.93 |
| | N = 3 | 86.093 |
| | N = 4 | 85.04 |
| | N = 5 | 84.807 |
| | N = 6 | 83.19 |
| | N = 7 | 81.913 |
| | N = 8 | 81.913 |
| | N = 9 | 81.270 |
| | N = 10 | 81.67 |
| Random Forest | - | 89.95 |
| XGBoost | - | 87.62 |

Fig. 6. Accuracies with new dataset

### B. *Mode 2*

This can be achieved through many ways using a number of renowned speech recognition libraries available publicly on the internet. We have used Google's speech recognition library to convert speech to text. In Future, we can also explore other options and settle on the most accurate option available to us.

## V. Results

Thus, as concluded in Methodology, with our current database, Random Forest appears to have the best accuracy. Hence, we have selected it as the final model to be integrated into our algorithm. With that, we then developed a Desktop GUI using Tkinter and here is the final product:

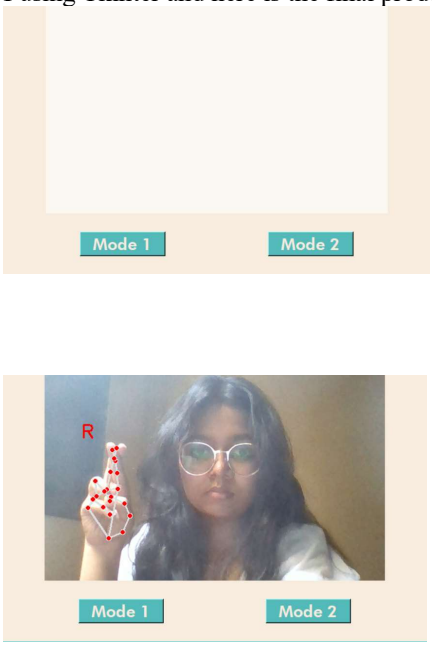


Fig. 7. Connect Desktop GUI

## VI. Limitations

An ML model's accuracy is highly dictated by the dataset used to train the model. Though we have tried out a variety of ASL and ISL datasets, a larger and more well-rounded dataset could improve the model's accuracy vastly. Along with a better dataset, Mediapipe's HANDS solution also raises an issue as it sometimes detects multiple hands in a single handed image and as discussed earlier, sometimes fails to recognize hands entirely. For example, out of 30 images of sign 'A', Mediapipe's solution could detect a hand and provide landmarks for 20. However, 2 out of those 20 images were detected to have two hands even when the image clearly contained only one hand.

Aside from dataset, processing frames in real time also brings up a few obstacles as video processing requires higher processing power.

Sign languages differ from region to region. ASL (American Sign Language) is much different than ISL (Indian Sign Language) or BSL (British Sign Language). Hence, the app would need to be trained in signs of those languages to identify correctly. Also, Many local sign languages do not use English as their base language. Here, more complications arise as it'd have to first translate to its base language before translating to English or other languages.

However, these limitations can slowly be worked upon and minimized or eliminated entirely with time in future.

## VII. Conclusion

Since our project can now recognize letters, it can be further developed to integrate recognition of words through fingerspelling – a subset of sign language where words are spelled out letter by letter. Through this horizon of possibilities for our project widens hugely. After accomplishing fingerspelling detection, the project can proceed to training on signs with movement. This would require video processing.

With good proven results, this project could potentially be used as a bridge between sign language and non-sign language speakers by adapting it to suit digital as well as physical environments. For example, it could be developed into plugins to provide captions for sign language based videos, or for participants speaking sign language in real time in digital classrooms, etc. It could also be used in real life as a translator for in-person seminars, etc.

Though the challenges faced by HoH/non-speaking people cannot be eradicated until disabilities are completely destigmatized and accessibility is made a priority in society, we hope Connect can offer a way to shorten this gap by making communication easier.

## VIII. References

[1] Garberoglio, C. lou, Palmer, J. L., Cawthon, S., & Sales, A. (n.d.). Deaf People and Employment in the United States: 2019.

[2] Garberoglio, C. lou, Palmer, J. L., Cawthon, S., & Sales, A. (n.d.). DEAF PEOPLE AND EDUCATIONAL ATTAINMENT IN THE UNITED STATES: 2019.

[3] ASLAN robot arm translates words into sign language for deaf people. (n.d.). Retrieved April 28, 2022, from https://www.dailymail.co.uk/sciencetech/article-5517971/amp/ASLAN-robot-arm-translates-words-sign-language-deaf-people.html

[4] *AI technology helps students who are deaf learn - The AI Blog*. (n.d.). Retrieved April 28, 2022, from https://blogs.microsoft.com/ai/ai-powered-captioning/

[5] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M. (2020). *MediaPipe Hands: On-device Real-time Hand Tracking*. http://arxiv.org/abs/2006.10214

[6] *Significant (ASL) Sign Language Alphabet Dataset | Kaggle*. (n.d.). Retrieved April 28, 2022, from https://www.kaggle.com/datasets/kuzivakwashe/significant-asl-sign-language-alphabet-dataset?resource=download