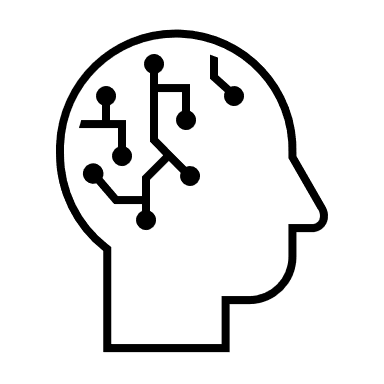
**TEAM PROPOSAL**

UTILISING DEEP LEARNING IN SIGN LANGUAGE RECOGNITION & INTERPRETATION

**AUTHORS:**

CHAN JOSHUA JUAN YIN

BRANDON LIN ZHAN HONG

DUONG NGOC YEN

PHAM VAN LONG PHUOC

YU XINHUI

WANG YUJING

**DATE OF SUBMISSION:** 13th JANUARY 2021

**Table of Contents**

[1. Background & Motivation 3](#_Toc61857267)

[2. Objective & Specific Aims 4](#_Toc61857268)

[2.1. Objective 4](#_Toc61857269)

[2.2. Specific Aims 4](#_Toc61857270)

[3. Methodology 5](#_Toc61857271)

[3.1. American Sign Language 5](#_Toc61857272)

[3.2. Building Models 5](#_Toc61857273)

[3.3. Phase 1: Interpretation of ASL Hand Gestures into English Words 6](#_Toc61857274)

[3.3.1. Data Preparation 6](#_Toc61857275)

[3.3.2. Selection of Video Frames 6](#_Toc61857276)

[3.3.3. Data Categorisation 7](#_Toc61857277)

[3.3.4. Hand Landmark Feature Tracking 7](#_Toc61857278)

[3.3.5. Spatial Scaling & Normalization 8](#_Toc61857279)

[3.3.6. Training the LSTM Neural Network 8](#_Toc61857280)

[3.3.7. Testing the Model 8](#_Toc61857281)

[3.4. Phase 2: Creation of Coherent & Meaningful Sentences through NLP 9](#_Toc61857282)

[3.4.1. Process Methodology 9](#_Toc61857283)

[3.5. Phase 3: Presentation of Synthesised Output on a GUI 10](#_Toc61857284)

[3.5.1. PC App Development 10](#_Toc61857285)

[3.5.2. Further Development 10](#_Toc61857286)

[4. Projected Timeline 11](#_Toc61857287)

[5. Proposed Budget Plan 12](#_Toc61857288)

[6. Risk Assessment 13](#_Toc61857289)

[7. References 15](#_Toc61857290)

# **1. Background & Motivation**

The ability to hear is often taken for granted. However, auditory perception is not universal trait. According to the World Health Organisation (WHO), more than 5% of the global population – approximately 466 million people – have some form of disabling hearing loss, and this number is estimated to grow to in excess of 900 million people by 2050 [1].

In lieu of verbal communication, members of Deaf cultures use sign languages as their primary means of communication within their communities. These languages are built upon the idea that the most useful tool a deaf person has to interact with his peers is their sense of sight [2]. Meaning is conveyed through the combined use of hand shapes, the orientation and movement of one’s hands, arms and/or body, and one’s facial expressions.

Naturally, as a result of the distinct difference in communication mediums – hearing people often utilise verbal means, whereas members of Deaf cultures use visual-manual means, enabling effective interaction between the two groups presents a unique challenge. As society as a whole consists primarily of hearing people, it is generally difficult for a hearing person and a deaf person to converse without the aid of a sign-language interpreter or technological accessibility, such as through texting. Some in the Deaf community may thus feel misunderstood or be discriminated against by those who do not know sign language [3].

There have been attempts in the past to overcome the communication barrier. The introduction of cochlear implants since the 1970s has helped many users gain auditory perception to a certain degree, but their use is not widespread and has faced resistance from members of Deaf communities; cochlear implants are seen as a threat to their culture, and discrimination by a hearing majority. It will, as such, take time for these devices to become more widely accepted and utilised. Additionally, while the use of text messages as an alternative means of communication has grown with the increasing prevalence of mobile devices, this ‘solution’ merely forces communication through a secondary medium, and does not truly address the root of the problem – a lack of comprehension of the other party’s primary language. The problem thus persists; there exists a need for improved methods of interaction.

# **2. Objective & Specific Aims**

In order to overcome the communication barrier, this Team believes a better approach to ameliorating this problem is to enable hearing people to better comprehend sign language. We seek to do so via a computer application which, via the use of deep learning models capable of interpreting American Sign Language (henceforth referred to as ASL) in near real-time and forming coherent sentences through Natural Language Processing, will be able to interpret a video input of communication in ASL into American English.

## **2.1. Objective**

To improve ease of communication between ASL users and non-ASL users through the use of deep learning.

## 

## **2.2. Specific Aims**

* To develop an interpreter, utilising deep learning models, capable of interpreting ASL in video format and output interpretations in American English in near real-time
* To provide an avenue through which future parties can build upon our work to greater benefit the Deaf community

# **3. Methodology**

## **3.1. American Sign Language**

There exist over 200 distinct sign languages, with markedly different grammar and lexicons [2]. Among them, American Sign Language, henceforth referred to as ‘ASL’, ranks as one of the most commonly used, especially in North America, West Africa, and Southeast Asia. Due to readily-available resources, we decided on using ASL for our deep learning model.

## **3.2. Building Models**

In ASL, each hand gesture or each consecutive motion of hand gestures correlates to one single word in spoken and written American English (henceforth referred to as ‘English’). A series of hand gestures form a sequence of words (or a ‘vector’ of words in the context of Natural Language Processing, henceforth referred to as ‘NLP’). Such word sequences, although expressed in a different grammatical and syntax structure from, can be further interpreted to form coherent and meaningful sentences that can be understood as English.

Our interpreter will convert a 30-second to 1-minute video file into coherent and meaningful English sentences. To do so, our interpreter will first utilise deep learning models to interpret ASL hand gestures into a sequence of English words; next, it will construct English sentences from individual ASL words through NLP techniques; finally, Graphical User Interface (henceforth referred to as ‘GUI’), will then display the synthesised result for ease of user viewing.

We will build our product in 3 phases:

* **Phase 1: Interpretation of ASL Hand Gestures into English Words**
* **Phase 2: Creation of Coherent & Meaningful Sentences through NLP**
* **Phase 3: Presentation of synthesised output on a GUI**

## **3.3. Phase 1: Interpretation of ASL Hand Gestures into English Words**

In Phase 1, we will focus on building a deep learning model which will interpret ASL hand gestures into separate English words. To do so, hand landmarks will first be extracted as numerical values. These values will subsequently be fed into the classification model. Below are the steps to implement that model:

### **3.3.1. Data Preparation**

In order to train our deep learning model for video classification, datasets of ASL video files will need to be imported. Ideally, we would want short video clips of approximately 1-2 seconds, each performing an ASL hand gesture correlating to a particular English word.

There are various sources of ASL gesture video datasets freely available on the Internet. To improve the accuracy of our model, we would want our model to be trained on as large a combined dataset as possible. Our desired size of combined dataset is 20,000 samples. In due course of this project, we will continue searching and adding relevant datasets and apply data augmentation methods to increase the size of our datasets.

At the time of this writing, we will be using the following datasets:

* **Word-Level American Sign Language** (‘WLASL’) features 2,000 common different words in ASL [4].
* **The American Sign Language Lexicon Video Dataset** (‘Boston ASLLVD’ or ‘ASLLVD’), owned by Boston University, has a labelled 9,800-video dataset of more than 3,300 English words [5].
* **MSR-Action3D** comprises 336 video files corresponding to 12 English words [6].

### **3.3.2. Selection of Video Frames**

Videos of ASL hand gestures of approximately 1 to 2 seconds will be inputted, each corresponding to a particularly English word. Each video file is comprised of a number of video frames; we will impose standardization across the number of frames used in modelling. To do so, we will use random temporal cropping to select a certain number of frames as a form of temporal data augmentation for our model. The specific number of frames used will be determined by experimentation.

### **3.3.3. Data Categorisation**

After being pre-processed, the combined video dataset will be split in 3 folders:

* Training Set (80%): Used for training the model.
* Development Set (10%): Used for hyperparameter tuning.
* Test Set (10%): Used for testing the model.

Data will be stored on a shared Google Drive and on an external hard drive, both of which combined will provide adequate file storage space.

### **3.3.4. Hand Landmark Feature Tracking**

In order to analyse hand gestures, we will need to track hand landmarks. To do so, we will be using MediaPipe, a framework developed by Google Research in 2019 [7]. This solution can provide real-time body, facial and hand tracking and extract important landmark features. In the context of hand tracking, the output value of every hand image is a list of 21 landmark features, each in a form of (x, y, z), where the ‘x’ and ‘y’ components are positional coordinates in the normalized form from 0.0 to 1.0, and the ‘z’ component represents the landmark depth. Landmark depth represents the degree of confidence in determining the specific position of each landmark.

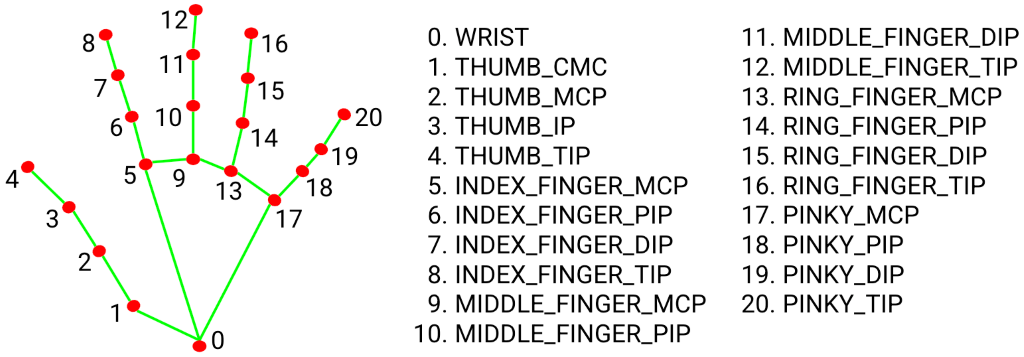


Fig. 1: Hand Landmarks with associated labels [7]

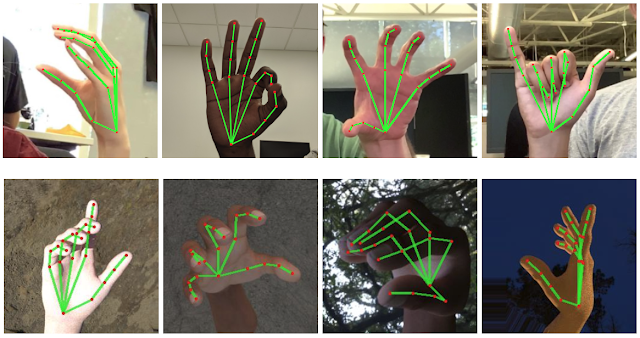


Fig. 2: Tracked 3D Hand Landmarks are represented by nodes and lines [7]

The tracked landmarks of each video frame are extracted as dimensional tensors and stored for use as inputs for the subsequent classification step.

### **3.3.5. Spatial Scaling & Normalization**

There are several methods of special scaling for landmark keypoint data. These include scaling by shoulder angle and neck distance. In terms of data augmentation, we can introduce random translation and random rotation. We will adopt appropriate methods through experimentation.

### **3.3.6. Training the LSTM Neural Network**

The extracted values obtained from the previous step will be used as the inputs for the Long Short-Term Memory (‘LSTM’) Neural Networks to classify English words. The outputted words will be further saved as inputs for NLP.

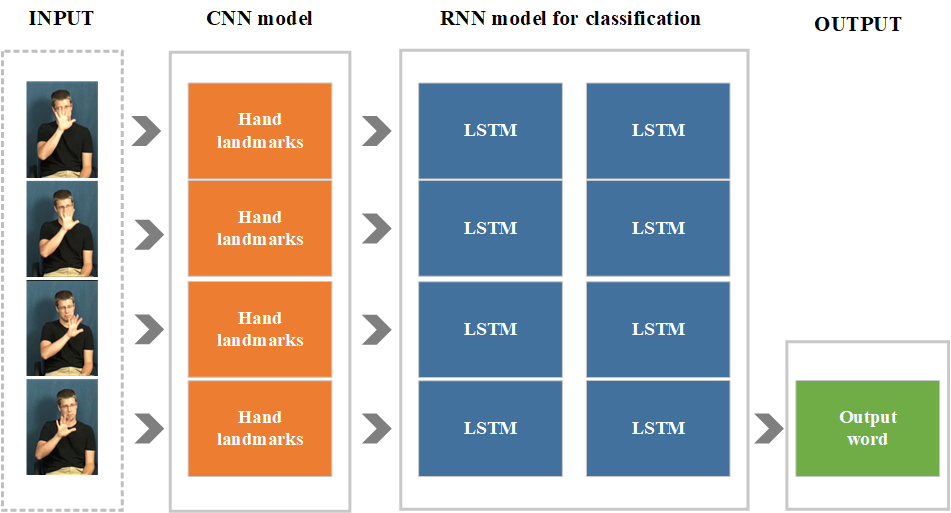


Fig. 3: Simple illustration of progression in Phase 1

### **3.3.7. Testing the Model**

The final step in this phase would be to test the model and determine its prediction accuracy. Alterations and adjustments will be made to improve our model as we progress in the project.

## **3.4. Phase 2: Creation of Coherent & Meaningful Sentences through NLP**

The speech habits in sign language differ from those in normal languages. As such, sentences transcribed directly from ASL gestures may not follow the same grammatical rules or sentence structures as seen in English. Hence, we will utilize NLP in order to form complete, cohesive sentences with appropriate grammar and tense for improved transcribing.

### **3.4.1. Process Methodology**

We must first determine the starting word of the sentence from a set of words transcribed from ASL. Our model will take into consideration parts of speech and morphology to evaluate the possibility of each word being the first word. The three words with the highest evaluation score will be passed into the sentence generation model as the beginning word for a specific sentence.

Given the first word, and a series of subsequent words potentially in random order, the sentence generation model will produce a possibility matrix which will display the probability of these words coming after the previous word. This process takes into consideration the meanings and sequence of previous words. The model then predicts the following words, one by one, using the probability matrix.

We may adapt an existing pre-trained NLP model, BERT [8], in our project.



Fig. 4: Next-word and previous-word attention patterns learned by BERT [9]

The generated sentences will be fed into a sentence evaluation model which will examine whether the sentences have appropriate structure and grammar as in English, and make corrections accordingly. The corrected sentences will be the final output.

## **3.5. Phase 3: Presentation of Synthesised Output on a GUI**

The transcribed result of our combined model will require a platform through which users may view our transcription. We decided on creating a computer application, or ‘PC App’, to serve as a GUI, through which users may input and receive data. The aim of the app is to provide a convenient transcription service to meet the real-life communication needs between the Deaf communicating in ASL and English users who do not understand ASL.

### **3.5.1. PC App Development**

The development of a PC App for MacOS and Windows 10 will facilitate both near real-time transcribing of ASL through deployment of our trained model. We will use Python’s Tkinter GUI library as the application’s foundation, and utilise OpenCV to access connected video capture devices. Users may record videos of ASL gestures through a webcam or similar video capture device connected to their PC, or import video files from external sources for transcription.

To save on Random Access Memory (RAM) usage and allow the interpreter to function more smoothly, the video files will be split into shorter segments, which will, in turn, be broken down into frames for gesture recognition.

To run our combined model, we intend to explore two options:

* Same-Device Transcription: Deploy our trained model within the PC App. This guarantees data security and swift access to data as the transcription service is performed directly on the same PC.
* Transcription via External Server: Video data is sent from and to the PC via the Internet. This has the advantage of access to the external server’s greater computational power, enabling faster and/or improved transcriptions as compared to same-device ones.

Upon transcription, the application will produce the transcription as an on-screen output for the user.

### **3.5.2. Further Development**

Given the successful deployment of the PC App, we intend to devote additional time and resources into two possible developments:

* **Mobile App Development** – Similar to the PC App, a mobile application, built on platforms such as TensorFlow Lite and Core ML (iOS), would enable us to bring near real-time transcription to phones, increasing the convenience of the transcription service.
* **Receiving User Data** – Our application may be upgraded to allow users to contribute their own ASL gesture video files for use as training data.

# **4. Projected Timeline**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task / Milestone** | **Duration [DD/MM/YY]** | | **Remarks** |
| Start Date | Completion Date |
| **Part 1: Research & Study** | | | |
| Preliminary Research & Studying of Subject Matters | 21/12/20 | 17/01/21 | - |
| Drafting of Team Proposal | 25/12/20 | 13/01/21 | - |
| Submission of Team Proposal | - | 13/01/21 | - |
| **Part 2: Interpretation of ASL Hand Gestures into English Words** | | | |
| Sourcing & Organisation of Dataset(s) | 04/01/21 | 24/01/21 | - |
| Preliminary Coding of Deep Learning Algorithm | 04/01/21 | 24/01/21 | - |
| Coding & Concurrent Testing of Hand Gesture Interpretation Model | 25/01/21 | 28/02/21 | - |
| Completion of Hand Gesture Interpretation Model | - | 28/02/21 | - |
| **Part 3: Natural Language Processing (NLP) Model & Final Testing** | | | |
| Coding & Concurrent Testing of NLP Model | 01/03/21 | 04/04/21 | - |
| Completion of NLP Model | - | 04/04/21 | - |
| Model Integration & Refinement | 05/04/21 | 18/04/21 | - |
| Final Testing | 11/04/21 | 25/04/21 | - |
| **Part 4: Final Report & Presentation** | | | |
| Drafting of Final Report | 05/04/21 | TBD | To be updated when details are made available |
| Preparation for Final Presentation | 12/04/21 | TBD |
| Submission of Final Report | - | TBD |
| Final Presentation | - | TBD |

Fig. 5: Projected Project Timeline

# **5. Proposed Budget Plan**

We will endeavour to, as much as possible, rely on freely available resources such as Google Drive, Google Colab Notebook, Google Cloud Services, Jupyter Notebook, etc.

Should there be a need for additional resources, funds provided for our project will be allocated as shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Item / Service** | **Pricing w/ Remarks** | **Allocated Funding [SGD]** |
| 1. | Access to relevant additional ASL Datasets | TBD | 100.00 |
| 2. | Google Cloud Compute Services [GPUs] | For NVIDIA® Tesla® T4 GPU:  USD 0.37 (~SGD 0.50) per GPU, per hour | 400.00 |

Fig. 6: Proposed Budget Plan Table

In terms of computational infrastructure, we will be utilising Google Cloud Compute [10], which will facilitate cloud computing for our data analytics and machine learning needs. Google Cloud provides USD 300 in free credit for newly-registered Google accounts for the first 90 days after initial registration, as well as access to a general purpose machine per month for free. Should the free credit be fully-utilised, there is no autocharge, thereby avoiding the risk of overspending. In the event that the free credit is used up before the end of our project, we will move our code and database to another newly-registered Google account eligible for free credit and machine provisioning by Google Cloud. This will allow us to work on our project with optimized Graphics Processing Units (GPUs) infrastructure without the need for additional funding.

# **6. Risk Assessment**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Risk Identification | | | | Risk Evaluation | | | | Risk Control | |
| 1a | 1b | 1c | 1d | 2a | 2b | 2c | 2d | 3a | 3b |
| S/N | Activity | Hazard | Possible  unwanted  consequences | Existing Control Measures,  if any | Likelihood | Severity | Risk | Additional  Control | Persons  & Dateline  to act |
| 1. | Physical Team Meetings | COVID-19 Infection | Becoming infected with COVID-19 | 1. Requirement for all to wear masks. 2. SafeEntry & NTU QR Code for check in/check out at all venues. 3. Mandatory safe distancing | Low | Moderate | Low | 1. Have online meetings as opposed to physical ones to minimise risk of exposure. 2. If unwell, excuse oneself from meetings and see a doctor. 3. Maintain safe distancing of at least 1 metre apart regardless of circumstance. | All Team Members |

Fig. 7: Risk Assessment Table

In light of the COVID-19 pandemic, the primary hazard this Team will face during the project is the possibility of coming into close proximity with an individual infected with COVID-19 and subsequently becoming infected. To mitigate the risks for transmission, this Team will endeavour to conduct online meetings in lieu of physical meetings to the fullest extent. Should there be a need for a physical meeting, this Team will adhere to safe distancing measures as required by Nanyang Technological University’s guidelines, wearing masks and maintaining at least 1 metre distance between individual members.

**7. References**

|  |  |
| --- | --- |
| [1] | World Health Organisation, “Deafness and hearing loss,” 1 March 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. [Accessed 6 January 2021]. |
| [2] | National Institute on Deafness and Other Communication Disorders, “What is American Sign Language (ASL) | NIDCD,” National Institute on Deafness and Other Communication Disorders, March 2019. [Online]. Available: https://www.nidcd.nih.gov/health/american-sign-language. [Accessed 9 January 2021]. |
| [3] | J. E. Nash and A. Nash, Deafness in society, Massachusetts: Lexington Books, 1981. |
| [4] | D. Li and H. Li, “Welcome to WLASL Homepage | WLASL,” GitHub, Inc., 20 Jan 2020. [Online]. Available: https://dxli94.github.io/WLASL/. [Accessed 9 January 2021]. |
| [5] | Boston Unversity, Rutgers University, “American Sign Language Linguistic Research Project (ASLLRP) Data Access Interface (DAI),” Boston Unversity, 2008. [Online]. Available: http://dai.cs.rutgers.edu/dai/s/signbank. [Accessed 9 January 2021]. |
| [6] | W. Li , “Dr Wanqing Li (UOW) | MSR Gesture3D,” Google, 7 November 2011. [Online]. Available: https://sites.google.com/view/wanqingli/data-sets/msr-gesture3d. [Accessed 9 January 2021]. |
| [7] | Google LLC, “Hands - mediapipe,” Google LLC, 2020. [Online]. Available: https://google.github.io/mediapipe/solutions/hands. [Accessed 9 January 2021]. |
| [8] | K. Toutanova, K. Lee, M.-W. Chang and J. Devlin, “[1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Cornell University, 24 May 2019. [Online]. Available: https://arxiv.org/abs/1810.04805. [Accessed 9 January 2021]. |
| [9] | J. Vig, “Deconstructing BERT, Part 2: Visualizing the Inner Workings of Attention,” KDnuggets, March 2019. [Online]. Available: https://www.kdnuggets.com/2019/03/deconstructing-bert-part-2-visualizing-inner-workings-attention.html. [Accessed 9 January 2021]. |
| [10] | Google LLC, “Compute Engine: Virtual Machines (VMs) | Google Cloud,” Google LLC, [Online]. Available: https://cloud.google.com/compute/. [Accessed 9 January 2021]. |