

Translating Sign Language to Speech

Group 1



November 3, 2024

David Plotkin

Mervin McDougall

Francisco del Castillo

# Objective

Thanks to the increased investment and development of Artificial Intelligence and Machine Learning, we are now able to interact with computers in a new way. We can enunciate commands to compose an email or respond to a text message via our digital assistants using verbal prompts. We can also initiate and hold verbal conversations with our smart devices like Siri, Alexa, and Google Assistant when we have pressing questions about the weather, traffic, or the latest headline news. Moreover, we can use these digital assistants to translate real-time communication with someone who speaks a foreign language. These features free a user from the use of keyboards and bottomless menus allowing him or her to interact with the technology naturally. However, there are some groups for which the benefits of these technological advances are currently unreachable. Our project aims to bridge this technological divide by allowing smart devices to translate sign language in real-time and give a voice to those who are unable to speak.

Currently, none of the major manufacturers of personal assistants provide a means of translating from sign language to speech. The most well-known personal assistant, Siri, requires a user who has speech disabilities to type their response using an on-screen keyboard and have Siri convert it to a voice (Apple). The other well-known assistant Alexa, from Amazon, requires a similar means of interaction, requiring a user to make use of an on-screen keyboard which can be operated remotely or by touchscreen to engage with Alexa. It should be noted that Amazon limits its text-to-speech facility to Fire Tablets only (Amazon). Lastly, Google’s Assistant, with Android-enabled devices, does not provide a means of directly translating from sign language to speech (Google). But it does provide a setup video for your Android device which is presented using ASL (American Sign Language) (Google, 2021).

It is our goal to provide those without speech the ability to communicate in a more organic and second-nature manner without being tethered to a keyboard allowing them to engage in coherent, fluid communication with other employees at work, a private doctor via a video chat, or even a treasured family member at a Thanksgiving table.

# DataSetDescription

Our dataset consists of the ASL Alphabet dataset from Kaggle (Nagaraj, 2018). The dataset is a collection of images, containing 87,000 images of American Sign Language gestures. The data set has 29 classes; letters A – Z (26 classes), as well as “space”, “delete”, and “nothing” gesture. These 3 classes are very helpful in real-time applications, and classification, and provide a means of controlling what is ultimately generated. Each of the alphabet classes contains 3,000 images (size 200x200 pixels RGB). The following image to the right is an example of the letter A found in the training set.

The image dataset will be preprocessed to create a second dataset of dynamic coordinates of gestures. This dataset will be generated using a third-party API called MediaPipe (Google, 2024). A call on the API will return coordinates of the hand and finger locations by overlaying a geometric “skeleton” on the hand as shown in the image below. These coordinates will then be used to train our model to recognize hand gestures. Each observation in this new dataset will have 21 points each containing (x, y, z) coordinate values. The final data set will contain 29 classes with each class having 3,000 image observations, for a total of 87000 observations. Each image will generate 21 points with (x, y, z) corresponding coordinates for a total of 1,827,000 data points.

A screenshot of a computer

Description automatically generated

# Methods

A screen shot of a hand

Description automatically generatedPreprocessing the image data is crucial to our approach. Therefore, we will be using Google’s MediaPipe API, more specifically, we will use the Hand Landmark Detection feature, which includes an object detection component that identifies hands and places landmarks at key points on the hand (Google, 2024). The output from this API provides the x, y, and z landmark coordinates of a hand gesture, which we will use as inputs to train a fully connected neural network for classifying the letter being signed. The image to the left is an example output of a processed MediaPipe image which illustrates the landmarks and the landmark coordinates of 1 point.

Our neural network model will incorporate dropout layers, batch normalization layers, as well as encoder and decoder components to optimize learning and prevent overfitting. To capture input images in real-time, we will use OpenCV (the Open Computer Vision library) to interface with a computer’s webcam and pass the images through the pipeline (OpenCV, 2024). Lastly, we will integrate Google’s text-to-speech API to articulate the sentence translated from the gestures (Google, 2024).

We will assess the neural network model’s performance using metrics such as F1 score and AUC. We consider the project as successful if we can create a complete gesture-to-speech software solution that is practical for communication and can be easily downloaded onto any computer.

# Risks / Concerns

The risks and concerns of our project include the following:

* **MediaPipe Hand Landmark Detection:**

Google’s MediaPipe can sometimes misinterpret hand recognition in poor lighting conditions. During the creation of our coordinate dataset, there may be images where MediaPipe fails to detect hands due to lighting and could result in poor model performance in real-world applications. To mitigate this risk, we will closely monitor our model’s performance on the test set throughout training. Additionally, we will ensure consistent, well-lit conditions and a uniform background for the poster day presentation.

* **Overfitting Risk:**

Due to limited variations in the rotational angles of our images, the model may only correctly classify gestures if the hand is straight, with no wrist rotation, and oriented downward. To reduce the risk of overfitting, we will consider adding more diverse data if time permits.

* **Model Complexity and Resource Demands:**

Image recognition often requires substantial computational resources. With 87,000 images of 200x200 pixels being transformed into a coordinate dataset there is a possibility that requests made on the Google MediaPipe API could be throttled. Therefore, we will be monitoring and managing our usage limits when preprocessing the image dataset for the training phase.

* **Time Constraints for Project Data Flow**

Given the project’s complexity, managing the data flow between Google MediaPipe, our ML-based ASL recognition, and the Text-to-Speech output will require a significant time investment. We may face constraints as we approach the poster presentation day. To mitigate this, we will begin working immediately and strategically distribute tasks among team members to optimize our timeline.

* **Surveillance**

Our project includes different moving parts. The most notable is the use of Google’s MediaPipe which requires making an API call and sharing an image over the internet. During regular usage, an image frame of a user making a sign will be shared. It is always possible that the image can be intercepted over the internet exposing the user to surveillance and other privacy issues.

* **Age Restriction**

In many states in the U.S., there are age restrictions on photography or videography of minors. There are also restrictions on the venue. Use of this technology will need to comply with the prevailing laws of one's state and the onus will fall on the user to use the technology responsibly.

* **Geographic Restrictions**

Our tool is restricted to the use of ASL (American Sign Language). Other sign language systems exist around the world, often considering differences in the alphabet or characters. This tool can only be used for translation of American Sign Language.

# **References**

Amazon. (n.d.). *What is Tap to Alexa?* Retrieved from Amazon: https://www.amazon.com/gp/help/customer/display.html?nodeId=G69UDERKYBL55DW9

Apple. (n.d.). *Accessibility features for speech on iPhone.* Retrieved from Apple Support: https://support.apple.com/guide/iphone/overview-of-accessibility-features-for-speech-iph8b6c223ac/18.0/ios/18.0

Google. (2021, 12 23). *Accessibility in Our Products & Features.* Retrieved from Google: https://about.google/belonging/disability-inclusion/product-accessibility/?q=speech#module-modal-embed-accessibility-android-pixel\_\_link-anchor

Google. (2024, November 3). *Cloud Text-to-Speech*. Retrieved from Google Cloud: https://cloud.google.com/text-to-speech?hl=en

Google. (2024, May 21). *Hand landmarks detection guide*. Retrieved from Google AI for Developers: https://ai.google.dev/edge/mediapipe/solutions/vision/hand\_landmarker

Google. (n.d.). *Accessibility in Our Products & Features — Google.* Retrieved from Google: https://about.google/belonging/disability-inclusion/product-accessibility/

Nagaraj, A. (2018). *ASL Alphabet [Data set]*. Retrieved from Kaggle: Akash Nagaraj. (2018). ASL Alphabet [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/29550

OpenCV. (2024, November 3). *OpenCV*. Retrieved from OpenCV: https://opencv.org/