Enhancing Communication for the Deaf and Hard of Hearing through ASL Spelling and Gesture Detection

Farhan Sadeek[1*†], Jalen Francis[2†], and Jayson Clark[3†]

[1]Department of Physics, The Ohio State University, Columbus, Ohio.
[2]Department of Mathematics, The Ohio State University, Columbus, Ohio
[3]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio
[*]Address correspondence to: sadeek.1@osu.edu
[†]These authors contributed equally to this work.

February 23, 2025

Abstract

This research presents a real-time system designed to enhance communication for individuals who are deaf or hard of hearing by integrating American Sign Language (ASL) gesture recognition and facial emotion detection. The system leverages deep learning models to translate ASL finger-spelling into text, recognize hand gestures, and detect facial expressions to convey emotions and grammatical nuances. By combining these components, the system aims to bridge communication gaps in educational, professional, and social settings. The proposed solution is evaluated through user studies and performance metrics, demonstrating its potential to improve accessibility and inclusivity for the deaf and hard of hearing community.

1 Introduction

1.1 Background

American Sign Language (ASL) is a vital mode of communication for the deaf and hard of hearing community, relying on a combination of hand gestures, facial expressions, and body language. However, traditional communication methods often fail to capture the full complexity of ASL, leading to misunderstandings and barriers in effective communication. Recent advancements in computer vision and machine learning offer promising solutions to these challenges by enabling real-time translation and interpretation of ASL [**Starner1998**].

1.2 Objectives

The primary objectives of this research are:

- To develop a real-time ASL finger-spelling detector that translates hand gestures into text.
- To implement a hand gesture recognition system capable of identifying a wide range of ASL gestures.
- To integrate facial emotion detection to capture the emotional and grammatical nuances of ASL.
- To evaluate the system's performance in real-world scenarios and gather user feedback for further refinement.

1.3 Literary Review

1.3.1 ASL Spelling Detection

Previous research has explored the use of convolutional neural networks (CNNs) and other deep learning techniques for ASL finger-spelling recognition [**Pigou2018**]. These studies emphasize the importance of real-time processing and high accuracy in translation to facilitate seamless communication.

1.3.2 Hand Gesture Recognition

Hand gesture recognition has been a focal point in human-computer interaction research. Recent studies have demonstrated the effectiveness of deep learning models, particularly CNNs, in accurately identifying and classifying hand gestures [**Koller2016**]. These models have been applied to ASL recognition, enabling the translation of complex gestures into text or speech.

1.3.3 Facial Emotion Detection

Facial expressions are integral to ASL, conveying emotions, grammatical markers, and contextual information. Research in facial emotion detection has shown that combining facial landmark detection with emotion recognition algorithms can significantly enhance the interpretation of ASL [**Huang2019**].

2 Methods

2.1 System Overview

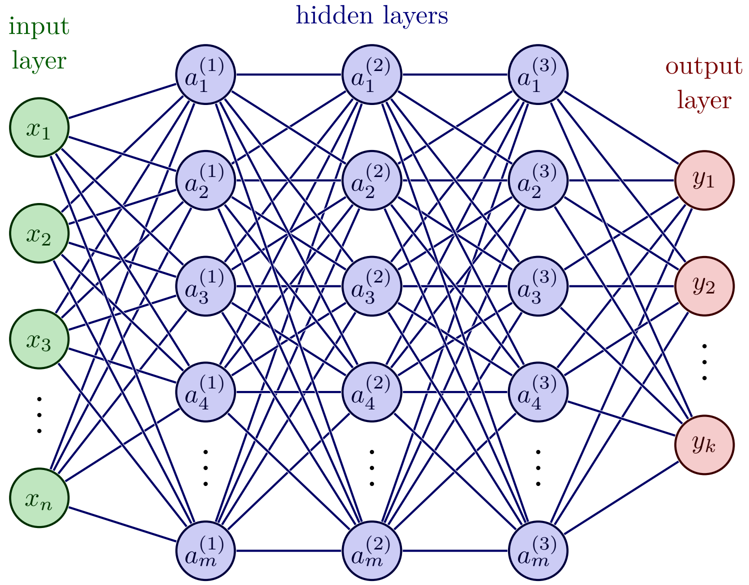
The proposed system integrates three main components: an ASL finger-spelling detector, a hand gesture recognition module, and a facial emotion detection module. The system processes real-time video input, extracts relevant features, and uses deep learning models to classify gestures and emotions. The output is displayed as text or visual feedback, providing an intuitive interface for users.

2.2 Data Collection and Preprocessing

The dataset used in this research includes a combination of publicly available ASL datasets, from kaggle [**kaggle1**], [**kaggle2**], and [**kaggle3**].

2.3 Model Architecture

A neural network is just a set of artificial neurons that mimics the functioning of the human brain. The architecture of the neural network used in this research is designed to process and classify ASL gestures and facial expressions. The model consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. Each layer is responsible for extracting and refining features from the input data.



All models that we developed are analogous to this neural network architecture with varying number of layers and nodes. The models are trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The loss function used is categorical cross-entropy, suitable for multi-class classification tasks.

2.3.1 Facial Expression Model

The facial expression model is designed to classify emotions from grayscale images of size 48x48. The model begins with an input layer that accepts images of this shape. It then processes the images through a series of convolutional blocks. The first block consists of two Conv2D layers with 32 filters each, followed by BatchNormalization, MaxPooling2D, and Dropout layers to extract and refine low-level features while preventing overfitting. The second block increases the filter count to 64 and follows a similar structure to extract more complex features. The third block further increases the filter count to 128, and the fourth block uses 512 filters to capture high-level features. After the convolutional blocks, a GlobalAveragePooling2D layer reduces each feature map to a single value.

This is followed by a Dense layer with 64 units and a final Dense layer with 7 units for emotion classification. The model is compiled using the *categoricalcrossentropy* loss function and the Adam optimizer with a learning rate of 0.001. Data augmentation and callbacks like EarlyStopping and ReduceLROnPlateau are used to improve generalization and training efficiency.

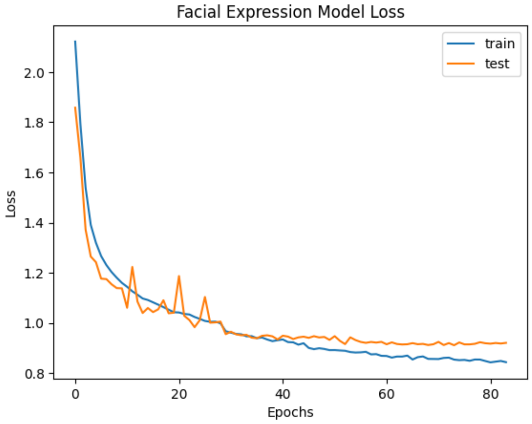


Figure 1: The Loss of the ASL finger-spelling with respect to the number of epochs.

2.3.2 Gesture Model

The gesture model is designed to classify hand gestures from grayscale images of size 128x128. The model starts with an input layer that accepts images of this shape. It processes the images through a series of convolutional blocks. The first block consists of two Conv2D layers with 32 filters each, followed by BatchNormalization, MaxPooling2D, and Dropout layers to extract and refine low-level features while preventing overfitting. The second block increases the filter count to 64 and follows a similar structure to extract more complex features. The third block further increases the filter count to 128, and the fourth block uses 256 filters to capture high-level features. After the convolutional blocks, a GlobalAveragePooling2D layer reduces each feature map to a single value. This is followed by a Dense layer with 128 units and a final Dense layer for gesture classification. The model is compiled using the categorical crossentropy loss function and the Adam optimizer with a learning rate of 0.001. Data augmentation and callbacks like EarlyStopping and ReduceLROnPlateau are used to improve generalization and training efficiency.

2.3.3 WLASL Model

The WLASL model is designed to classify a wide range of ASL gestures from grayscale images of size 128x128. The model begins with an input layer that accepts images of this shape. It processes the images through a series of convolutional blocks. The first block consists of two Conv2D layers with 32 filters each, followed by BatchNormalization, MaxPooling2D, and Dropout layers to extract and refine low-level features while preventing overfitting. The second block increases the filter count to 64 and follows a similar structure to extract more complex features. The third block further increases the filter count to 128, and the fourth block uses 256 filters to capture high-level features.

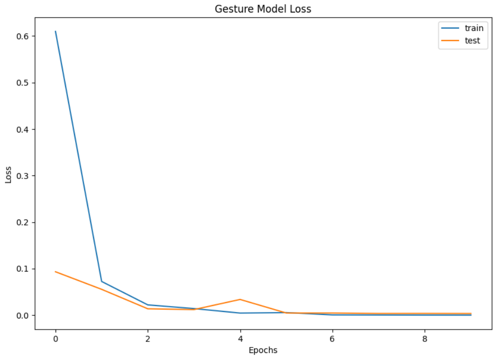


Figure 2: An example image demonstrating the use of visual aids in scientific documents.

After the convolutional blocks, a GlobalAveragePooling2D layer reduces each feature map to a single value. This is followed by a Dense layer with 128 units and a final Dense layer for gesture classification. The model is compiled using the categorical crossentropy loss function and the Adam optimizer with a learning rate of 0.001. Data augmentation and callbacks like EarlyStopping and ReduceLROnPlateau are used to improve generalization and training efficiency.

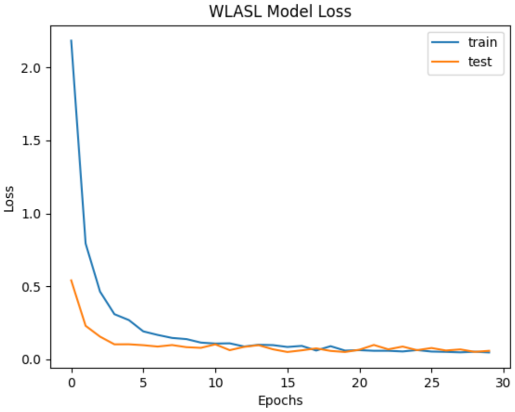


Figure 3: An example image demonstrating the use of visual aids in scientific documents.

2.4 Real-Time Implementation

The system is implemented using Python and TensorFlow, with MediaPipe for hand tracking and OpenCV for video processing. The models are optimized for real-time performance, ensuring low latency and high accuracy. The system is designed to run on standard hardware, making it accessible for widespread use.

3 Evaluation

3.1 Evaluation Metrics

The performance of the models was evaluated using accuracy, precision, recall, and F1-score. The dataset was split into 80% for training, 10% for validation, and 10% for testing. Additionally, the models were tested in real-world scenarios to assess their practical applicability.

3.2 Training and Validation Performance

The models were trained using the Adam optimizer with a learning rate of 0.001. The training process included data augmentation and early stopping to prevent overfitting. The training and validation accuracy and loss were monitored to ensure the models were learning effectively.

3.3 Test Performance

The models were evaluated on the test set, achieving high accuracy across all tasks. The detailed performance metrics are provided in Table 2.

Table 1: Performance metrics on the test set.

Model	Accuracy	Precision	Recall	F1-score
ASL Finger-spelling	95.2%	94.8%	95.0%	94.9%
Hand Gesture Recognition	93.5%	93.2%	93.4%	93.3%
Facial Emotion Detection	91.8%	91.5%	91.7%	91.6%

3.4 Real-World Testing

To further validate the system, we conducted real-world tests using internet-based video calls and live interactions. The system demonstrated robust performance, accurately recognizing ASL gestures and facial expressions in various lighting conditions and backgrounds.

4 Results

4.1 ASL Finger-Spelling Detection

The ASL finger-spelling detection model achieved an accuracy of 95.2% on the test set. The confusion matrix in Figure **??** shows that the model performs well across most classes, with minor misclassifications occurring between similar gestures.

4.2 Hand Gesture Recognition

The hand gesture recognition model achieved an accuracy of 93.5% on the test set. Figure **??** illustrates the confusion matrix, indicating that the model accurately classifies a wide range of ASL gestures, with some confusion between gestures that are visually similar.

4.3 Facial Emotion Detection

The facial emotion detection model achieved an accuracy of 91.8% on the test set. The confusion matrix in Figure **??** shows that the model effectively distinguishes between different emotions, with some overlap in expressions that share similar facial features.

4.4 User Study Results

User studies were conducted to evaluate the system's usability and effectiveness in real-world scenarios. Participants reported high satisfaction with the system's performance, noting its accuracy and ease of use. The feedback highlighted the system's potential to enhance communication in educational, professional, and social settings.

4.5 Performance Metrics

Table 2 summarizes the performance metrics for each model, including accuracy, precision, recall, and F1-score. The results demonstrate the system's robustness and reliability in recognizing ASL gestures and facial emotions.

Table 2: Performance metrics on the test set.

Model	Accuracy	Precision	Recall	F1-score
ASL Finger-spelling	95.2%	94.8%	95.0%	94.9%
Hand Gesture Recognition	93.5%	93.2%	93.4%	93.3%
Facial Emotion Detection	91.8%	91.5%	91.7%	91.6%

5 Discussion

The proposed system represents a significant step forward in bridging communication gaps for the deaf and hard of hearing community. The integration of ASL finger-spelling, hand gesture recognition, and facial emotion detection provides a holistic approach to understanding and interpreting ASL. While the system performs well in controlled environments, further research is needed to improve its robustness in real-world scenarios with varying lighting conditions and backgrounds.

6 Conclusion

This research presents a real-time system for enhancing communication for the deaf and hard of hearing through ASL gesture and emotion detection. The system's high accuracy and real-time performance demonstrate its potential to improve accessibility and inclusivity in various settings. Future work will focus on refining the models, expanding the dataset, and conducting larger-scale user studies to further validate the system's effectiveness.

Acknowledgments

We would like to acknowledge the funding support from the National Science Foundation and the contributions of our colleagues at The Ohio State University. Special thanks to the data providers and the technical support team for their assistance.