# SENSORS TO SIGN LANGUAGE: A NATURAL APPROACH TO EQUITABLE COMMUNICATION

*Thomas Fouts*⋆ *Ali Hindy*† *Chris Tanner*††

⋆ University of Michigan
† Stanford University
†† Harvard University

## ABSTRACT

Sign Language Recognition (SLR) aims to improve the equity of communication with the hearing impaired. However, SLR typically relies on having recorded videos of the signer. We develop a more natural solution by fitting a signer with arm sensors and classifying the sensor signals directly into language. We refer to this task as Sensors-to-Sign-Language (STSL). While existing STSL systems demonstrate effectiveness with small vocabularies of fewer than 100 words, we aim to determine if STSL can scale to larger, more realistic lexicons. For this purpose, we introduce a new dataset, *SignBank*, which consists of exactly 6,000 signs, spans 558 distinct words from 15 different novice signers, and constitutes the largest such dataset. By using a simple but effective model for STSL, we demonstrate a strong baseline performance on *SignBank*. Notably, despite our model having trained on only four signings of each word, it is able to correctly classify new signings with 95.1% accuracy (out of 558 candidate words). This work enables and motivates further development of lightweight, wearable hardware and real-time modelling for SLR.

***Index Terms***— Sign language, sign language recognition, EMG, service robotics, language modeling

## 1. INTRODUCTION

Although sign language is the primary mode of communication for hearing-impaired communities, only roughly one million people are fluent in American Sign Language, creating an inequality in communication. Most research in Sign Language Recognition (SLR) has been in gesture-recognition or Neural Sign Language Translation (NSLT), using Human Keypoint Estimation and video-to-text translation. Existing work on SLR or sign translation typically requires having a video feed that captures the speaker, which is appropriate for large events or remote telecommunication. However, for natural, in-person communication, having a video recording can be unrealistic and feel intrusive, whereas a slim, sleek armband would be more comfortable and feel more natural.

In this work, we deviate from previous wearable device approaches and introduce a novel armband design to map EMG muscle signals to signs. Ignoring facial gestures, our sensors measure the remaining four fundamental components of sign language: handshape, movement, location, and palm orientation. The goal is to directly classify the signer's hand movements into natural, textual language. Besides the practical reasons given above, working directly with motor features as input has several advantages over video recordings for any machine learning model. First, we do not require complex video processing to infer depth information and resolve visually ambiguous recordings. Second, there is less chance for noisy signals, as videos may suffer from poor quality and background artifacts. Sensor signals are grounded in motion and are directly recorded from hand movements, which allows for less variability across speakers, given that sensors are placed consistently (e.g., with a glove or armband). SLR with a glove is significantly limited by the intrusive nature of wearing such a visible translation aid. To minimize discomfort, alienation, and the stigma against ASL, a sleek armband that can be hidden by a sleeve is optimal for such a translation task.

Our contributions can be summarized as follows:

- We present an efficient and comfortable approach to SLR with a wearable device that maps EMG muscle sensors to text (STSL). (Section 3.2).
- We collect and publicly release one of the largest ASL datasets, which contains muscle sensor data and corresponding video footage of 6,000 words signed by 14 individual novice signers and 1 native signer. (Section 3).
- Our results demonstrate the feasibility of SLR using EMG muscle sensors (Section 5), and our included video dataset allows for future research to compare the different mediums.

We hope to inspire further multimodal SLR research that uses sensors, images, and textual language data. Novice sign-

Thomas and Ali conducted this research during high school.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

ers and native signers were used to compare the accuracy of sensor-to-text translation of different ASL proficiency levels.

## 2. RELATED WORK

ASL, despite being a complete, grammatical language, has received relatively little attention within the Service Robotics community. Hardware for ASL generally falls within two categories, based on their approaches and mediums. That is, some systems aim to learn a mapping either from (a) videos to text; or (b) glove-based hardware to text.

### 2.1. Video-to-Text

The biggest challenge in SLR is the lack of publicly available labeled data. Labeling large datasets is laborious, so most datasets are either (a) large, but lacking labels [14, 8], or (b) sufficiently labeled but too small [5, 13]. For example, TV broadcasts include large amounts of data for SLR, but they are often not annotated. Recent papers have tried addressing this issue. For example, Forster et al. [4] released RWTH-PHOENIX-Weather 2012 and its extended 2014 version, providing gloss annotations for videos. However, these datasets include 7 different signers who are not equally represented throughout the data, which introduces variability between the signs and makes translation more difficult.

Deep Learning, particularly seq2seq models, has demonstrated strong performances in generating high-quality neural sign language translations (NSLT) [3, 14] We additionally experimented with combining seq2seq models with convolutional neural networks (CNNs), which are traditionally used for image processing due to their ability to robustly capture meaningful representations. The latest NSLT research focuses on directly learning video-to-text translations, by learning tokenization and using glosses in order to overcome the need for fine-grained annotations [14].

Tokenization is the pre-processing step that aims to capture and create meaningful units of input data. This important step often affects the success of translations. Orbay et al. [14] state that tokens can only be learned from videos with semi-supervised data, and annotation at the gloss level is costly and scarce. Similarly, glosses (brief notation for certain gestures) also require supervised data, and although gloss-to-text translations are more effective than sign-to-text translation, annotated gloss data is scarce since annotation is a time-consuming process and limited video data exists [11, 10].

### 2.2. Sensor-to-text Translation

Other SLT tasks focus on using 3-D modeling or gloves; however, these works [13, 17] often deal with only 50-60 words from 1-5 signers, whereas our dataset consists of over 6,000 words from 15 signers (14 novice, 1 native). These sensor-based recognition tasks also require either skeletal modeling [16] or a Kinect V2 device [5], whereas our method would ideally only require a thin sleeve to be worn. A glove is visible when communicating and may isolate the ASL speaker, whereas an arm sleeve can be easily concealed. Further, current SLR research mainly focuses on finger-spelling, which is a basic form of sign language that is rarely used in practice and is unrealistic for natural, in-person communication.

Gloves can be used to predict finger orientation using flex sensors, and the hand's position is calculated using an accelerometer and a gyroscope. While flex sensors can accurately measure finger displacement, the practicality of these sensors is significantly limited. Gloves are intrusive, not only for their visible nature, but because the flex sensors create resistance as the signer moves their fingers – making the overall experience feel unnatural. Our custom built sensors, which could be concealed within an arm sleeve, avoid this problem by directly measuring the EMG data from the muscle groups in the forearm that control each finger's movement.

The EMG sensors on the forearm can accurately predict finger movements, eliminating the need for a glove. Further, an accelerometer and gyroscope can predict the hand's position and motion in 3D-space, all while being concealed under a sleeve. Myoware sensors were used in the past to create an armband that translates ASL, yet this technology has been discontinued and is no longer in the current canon of research.

## 3. SENSOR DATA COLLECTION

We introduce a large, fully annotated dataset with 15 unique signers. Of these signers, 14 are equally represented throughout the dataset, as they all signed the same 22 distinct words, 7 times each. In addition, one volunteer signed over 558 unique words 7 times each. Specifically, the Handspeak Public Dictionary (a dictionary of over 8,000 videos from expert signers) served as the basis for our novice signers to learn sign language, as they aimed to perfectly mimic the expert signers.

### 3.1. Sensor Features

For recording the hand movements during signing, we use custom-built sensors modeled after the Muscle SpikerShield from Backyard Brains, and the schematics are included in the Appendix. These sensors are affordable, easy to use, and are adjustable for the user's sensitivity and settings. To minimize variability across signers, we used wet electrodes that could be secured in the correct position for each signer. Each sensor generates a value by calculating the potential difference between the two electrodes A and B on the muscle:

$$\Delta V = V_B - V_A$$

Each arm was equipped with 5 sensors and (2) 6-axis gyroscope accelerometer mpu6050s, which communicate via $I^2C$ with an Arduino Mega. 15 floating point values are generated per arm: 5 from each of the sensors, and 10 from the mpu6050. Additionally, 5 values are generated by the mpu6050's accelerometer: the acceleration in the x-direction ($A_x$), y-direction ($A_y$), and the z-direction ($A_z$), as well as the angle in 3D-space at which the sensor is accelerating in the x-direction ($\theta_x$) and y-direction ($\theta_y$). These 15 values from each arm are generated every $\frac{1}{15^{th}}$ of a second in order to match the frame rate of the video dataset.

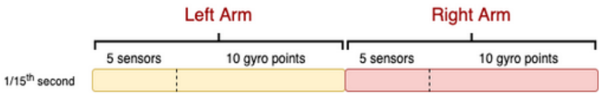

**Fig. 1**: One vector of the muscle sensor input

Note that in these figures, there are 10 gyro points for each arm; however, since two of these values are the sum of two other values multiplied by the elapsed time (which is consistent across all of our data), we remove them from our data analysis. Hence, the initial 30 values are reduced to a vector of 26. For each sign, a sequence of variable length (time) is generated. Our longest sign took 8.9 seconds, so all inputs to our model had a shape of (133,26) – shorter signs were padded to include 133 vectors.

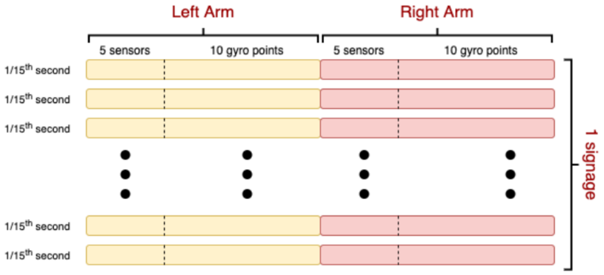

**Fig. 2**: One full sequence of the muscle sensor input

### 3.2. Recording Signs

Our dataset, *SignBank*, consists of two parts:

- Part 1: 558 unique words signed 7 times by one novice signer desiring to learn ASL by precisely mimicking an expert signer
- Part 2: 22 unique words signed 7 times by 14 novice signers (desiring to learn ASL) and 1 native signer.

This yields a grand total of exactly 6,000 signed words, which are amongst the most popular ASL words, and our complete vocabulary is listed in the Appendix.

In order to test the precision of our recordings, we inspect the recorded features visually. To discern the distinctness of each signed word, we projected each 26-length vector (corresponding to $\frac{1}{15^{th}}$ of a second) to a 2D representation, via PCA, while using the z-axis to represent time. In Figure 3, we show the full time-series signing of four randomly selected words. We see that they have reasonably distinct representations, even in 3D. For all of our models and experiments, we used the full 26-length vectors.

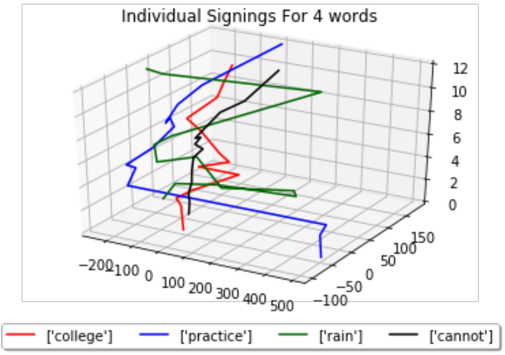

**Fig. 3**: Four distinct words (PCA-reduced to 3D)

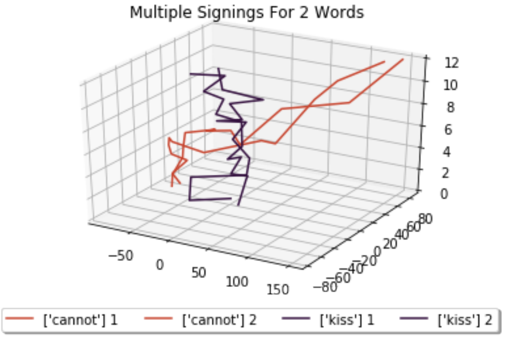

**Fig. 4**: Multiple signing for two words

As seen in Figure 4, signings of the same word (e.g., "cannot") have relatively consistent paths when represented in 3D-space. However, there is enough intra-word variation (e.g., "kiss") that the problem is not trivially easy.

## 4. EXPERIMENTS

STSL is a classification task, whereby our number of classes is equal to the number of distinct words in our dataset. The goal is to learn a mapping from sensors to words such that when

| | **Part 1** | | **Part 2** | |
|---|---|---|---|---|
| **Model** | Dev | Test | Dev | Test |
| LSTM | 90.5 | 85.5 | 85.4 | 84.2 |
| CNN + Bi-LSTM | 97.1 | 93.2 | 87.6 | 86.6 |
| Transformer | 95.2 | 94.1 | 89.9 | 85.6 |
| **Bi-LSTM (Ensemble)** | **97.4** | **95.1** | **93.4** | **92.3** |

**Table 1**: Our Models' Performances (Accuracy Percentage). Note that Part 1 of the dataset contains 558 unique words and Part 2 of the dataset contains 22 unique words.

| Paper | # of Distinct Signs | Method | Accuracy (%) |
|---|---|---|---|
| Rizwan et al. [15] | 10 | Glove | 94.2 |
| Bajpai et al. [2] | 9 | Glove | 90.0 |
| Haidar et al. [6] | 33 | Glove | 91.25 |
| Amor et al. [1] | 7 | Myo Armband | 91.7 |
| Madushanka et al. [12] | 12 | Myo Armband | 94.4 |
| **Our Model (BI-LSTM, ens.)** | **22** | **EMG Armband** | **92.3** |

**Table 2**: SLST systems. Accuracies are not directly comparable across systems, as they all concern different datasets.

presented with sensor data from a new signing, our model can accurately predict the signed word.

For part 1 of the dataset: each of the 558 words has been signed exactly 7 times. We randomly divided our data into a training set (4 signings of each word), development set (1 signing of each word), and test set (2 signings of each word). For all experiments, we optimized our hyperparameters on the development set by uniformly sampling, varying the number of hidden layers from 1 to 3, the number of hidden units from 64 to 1024, the number of epochs from 10 to 300, and the dropout rate from 0 (no dropout) to 0.4. We used Adam [9] as our optimizer.

For part 2 of the dataset: signs were grouped together on a user-basis and randomly divided into a training set (8 users corresponding to 1,232 signs), development set (5 users corresponding to 770 signs), and a test test (2 users corresponding to 308 signs). To measure the consistency of our models, we separately repeated this randomization process 2 more times and ran our models on all three sets of users.

The hyperparameters of our best system are listed in the Appendix. All of our models listed in Table 1 (except the ensemble) trained on a 2020 MacBook Pro without a GPU in 10-30 minutes. Alternatively, they trained in 2-10 minutes when using a free Google Colab environment.

### 4.1. Models

Since we are working with time-series data, a natural choice for a baseline model is an LSTM [7], which is in suit with Wu et al. [17] RNN approach to this task. Since it is permissible to look at the full time-series of data for a particular signing, we also tried a Bidirectional LSTM. Next, to better capture patterns across recordings of a timing, and in hopes of being more robust to varying lengths for the same signed word, we used CNNs. We then combined a Bidirectional LSTM with a CNN, in hopes of reaping the benefits from both models. We also experimented with a Transformer [19], as they are known to outperform LSTMs. For our best performing model, we also conducted an ensemble approach ($k$-fold cross-validation across unique signs for Part 1 of the dataset, and random splits of non-testing users for Part 2 of the dataset).

## 5. RESULTS

We optimized each model's hyperparameters on the development set, then evaluated its results on the test set. In Table 1, we see that our ensembled approach yields the best results. Moreover, our accuracy is comparable to existing systems (Table 2, where we present our results on Part 2 of our dataset), despite our classification task being much harder. For example, the most recent related work [6] selects from 33 potential classes, whereas our model selects from 558 potential classes. Further, Amor et al. [1] had a small vocabulary with each word signed an average of 54 times, which makes the learning task much easier than our setup of having 7 signings per word. This suggests that our model provides the most compelling results to date.

Our most frequently misclassified words (e.g., chase, study, toilet, boy, delicious) are listed in the Appendix. Upon inspection, we observe that our model has difficulties when signed hands are mostly stationary. This was particularly a trend for the left hand. Conversely, when the speed of the sign significantly increases, the signs become increasingly more difficult to learn, especially when the total range of motion is small.

## 6. CONCLUSION

Most Sign Language Recognition (SLR) research focuses on video-to-text translation, which is unrealistic for natural in-person communication with an American Sign Language (ASL) speaker. Instead, we focused on measuring one's signing directly, via the Sensors-to-Sign-Language (STSL) classification task. Whereas past STSL systems concern small vocabularies of fewer than 100 words from few subjects, we demonstrated that STSL can scale to much larger lexicons: we created the largest dataset for this task, which includes 6,000 individual signings and spans 558 distinct signed words from 15 unique subjects. The accuracy of our models prove not only the promise of STSL, but also the potential of using EMG sensors to teach sign language. Given the challenging size of our classification task (558 classes), our strong models outperform all existing work, which will hopefully motivate future work within this important but generally overlooked language.

## 7. REFERENCES

[1] Amina Ben Haj Amor, Oussama El Ghoul, and Mohamed Jemni. 2019. Sign language hand shape recognition using myo armband. 2019 7th International conference on ICT Accessibility (ICTA).

[2] Dhananjai Bajpai and Vijay Mishra. 2016. Low costfull duplex wireless glove for static and trajectory based American Sign Language Translation to multi-media output. 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN).

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neu-ral sign language translation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, page 7784–7793. IEEE.

[4] Forster, Jens I& Schmidt, Christoph I& Hoyoux, Thomas I& Koller, Oscar I& Zelle, Uwe I& Piater, Justus I& Ney, Hermann. (2012). RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus.

[5] Ivan Gruber, Marek Ryumin, Dmitry, and Alexey Karpov. 2018. Sign language numeral gestures recognition using convolutional neural network. In Interactive Collaborative Robotics, Lecture Notes in Computer Science, page 70–77. Springer International Publishing.

[6] Galib Ibne Haidar and Hasin Ishraq Reefat. 2020.Glove based American Sign Language interpretation using Convolutional nNural Network and Data Glass. 2020 IEEE Region 10 Symposium (TENSYMP).

[7] Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. Neural Computation, 9 8):1735–1780.

[8] Hamid Reza Vaezi Joze and Oscar Koller. 2019. MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. arXiv:1812.01053[cs].

[9] Diederik P. Kingma and Jimmy Ba. 2014. Adam:A method for Stochastic Optimization. Cite arxiv:1412.698 0 Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[10] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on Human Keypoint Estimation. Applied Sciences, 9(13):2683.

[11] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Ben Swift, Hanna Suominen, and Hongdong Li. 2020. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. arXiv:2010.05468 [cs].

[12] A.l.p Madushanka, R.g.d.c Senevirathne, L.m.h Wijesekara, S.m.k.d Arunatilake, and K.d Sandaruwan. 2016. Framework for Sinhala Sign Language Recognition and translation using a wearable armband.2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer).

[13] Ben Mcinnes. 2014.South African Sign Language dataset development and translation: A Glove-Based Approach.

[14] Alptekin Orbay and Lale Akarun. 2020. Neural sign language translation by learning Tokenization. arXiv:2002.00479 [cs]. ArXiv: 2002.00479.

[15] Shaheer Bin Rizwan, Muhammad Saad Zahid Khan, and Muhammad Imran. 2019. American Sign Language translation via smart wearable glove technology. 2019 International Symposium on Recent Advances in Electrical Engineering (RAEE).

[16] Stephanie Stoll, Necati Cihan Camgoz, Simon Had-field, and Richard Bowden. 2020. Text2sign: Towards sign language production using neural ma-chine translation and generative adversarial net-works. International Journal of Computer Vision, 128(4):891–908.

[17] J. Wu, L. Sun, and R. Jafari. 2016. A wearable system for recognizing American Sign Language in real-time using IMU and surface EMG sensors. IEEE Journal of Biomedical and Health Informatics ,20(5):1281–1290.

[18] Jan Zelinka and Jakub Kanis. 2020. Neural sign language synthesis: Words are our glosses. page 3395–3403.

[19] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. ”Attention Is All You Need.” ArXiv.org. December 06, 2017.