Modeling Global Body Configurations in American Sign Language

Nicholas Wilkins[1], *Beck Cordes Galbraith*[2], *Ifeoma Nwogu*[1]

[1]Rochester Institute of Technology, Rochester, NY, USA [2]Sign-Speak, USA

npw3202@rit.edu, beck.cordes.galbraith@gmail.com, ion@cs.rit.edu

Abstract

In this paper we consider the problem of computationally representing American Sign Language (ASL) phonetics. We specifically present a computational model inspired by the sequential phonological ASL representation, known as the Movement-Hold (MH) Model. Our computational model is capable of not only capturing ASL phonetics, but also has generative abilities. We present a Probabilistic Graphical Model (PGM) which explicitly models holds and implicitly models movement in the MH model. For evaluation, we introduce a novel data corpus, ASLing, and compare our PGM to other models (GMM, LDA, and VAE) and show its superior performance. Finally, we demonstrate our model's interpretability by computing various phonetic properties of ASL through the inspection of our learned model.

Index Terms: Sign Language, ASL, Phonetics, Probabilistic Graphical Model, Language Model

1. Introduction

American Sign Language (ASL) is the fourth most commonly used language in the United States and is the language most commonly used by Deaf people in the United States and the English-speaking regions of Canada [1]. Unfortunately, until recently, ASL received little research. This is due, in part, to its delayed recognition as a language until William C. Stokoe's publication in 1960 [2]. Limited data has been a long-standing obstacle to ASL research and computational modeling. The lack of large-scale datasets has prohibited many modern machine-learning techniques, such as Neural Machine Translation, from being applied to ASL. In addition, the modality required to capture sign language (i.e. video) is complex in natural settings (as one must deal with background noise, motion blur, and the curse of dimensionality). Finally, when compared with spoken languages, such as English, there has been limited research conducted into the linguistics of ASL.

We realize a simplified version of Liddell and Johnson's Movement-Hold (MH) Model [3] using a Probabilistic Graphical Model (PGM). We trained our model on ASLing, a dataset collected from three fluent ASL signers. We evaluate our PGM against other models to determine its ability to model ASL. Finally, we interpret various aspects of the PGM and draw conclusions about ASL phonetics. The main contributions of this paper are

1. A PGM which models the the body configuration in ASL
2. A video dataset containing over 800 phrases comprising of over 4,000 signs.

1.1. The Movement-Hold Model

The Stokoe Model was the first attempt to analyze the internal structure of signs and challenge the notion that signs are holistic iconic gestures. Stokoe analyzed signs using three parameters: location, handshape, and movement. He assumed each parameter was produced simultaneously [4]. The foundational Stokoe's model is unable to recognize sequential contrasts occurring between signs because it does not require the parameters of a sign to be produced in a prescribed order [5].

In response to the limitations of the Stokoe Model, Liddell and Johnson published their Movement-Hold (MH) Model. The MH Model partitions each sign into sequential hold and movement segments. A bundle of articulatory features is associated with each segment and contains information regarding the segment's handshape, location, palm orientation, and non-manual markers (e.g. facial expressions). A segment with no changes to the articulation bundle is a hold segment and can be represented with an H for a full-length hold or an X for a shorter hold. A segment in which some of the articulatory features are transitioning is a movement segment and can be represented with an M. The MH Model recognizes at least nine sign structures, including Hold (H), X M H, and H M X M H, but not H M. The MH Model was novel in its approach to modeling the sequential aspect of signs [6].

The MH Model is a significant departure from previous models because its basic units of movements and holds are produced sequentially, allowing for precise representation of sequences within signs.

H_1	H_1	H_1	H_2	H_2	E	E	E
H_1	H_1	H_1	E	E	E	E	E
H_1	H_2	H_3	H_4	H_4	H_4	E	E

Figure 1: *Some example sign structures realized by our model. Each sign is represented in eight frames instead of twenty-five for easier viewing. Each H_i is a distinct body configuration with E being an ending token that is used for padding. The first example illustrates an HMH. The second demonstrates an H. Finally, the third illustrates an $XMXMXMH$*

1.2. Prior Works

While many computational approaches to modeling sign language, including the previously published Interspeech sign language recognition paper [7] focus on representing signs holistically. A prior approach from speech recognition that gained popularity within sign language modeling is the use of the Hidden Markov Models (HMM) and its variants to model subunits of signs. Influenced by the MH Model, Theodorakis et al. [8] used HMMs to represents signs which they segmented into dynamic and static portions. Fang et al. [9] modeled Chinese Sign Language using HMMs and temporal clustering. Vogler et al. [10] trained HMMs to recognize movement and hold phonemes in an attempt to increase the scalability of ASL recognition systems. Bauer et al. [11] take a different approach in their attempt to create an automatic German Sign Language recognition system, using a K-means algorithm to define subunits. With all four of these papers focusing on creating a sign language recog-

nition system, there is more focus on recognition success over the phonological model. However, those algorithms do not always identify recognizable phonemes. For this reason, building from the ground up, we focus specifically on creating a phonological model and interpreting the model. Additionally, our model is entirely unsupervised, probabilistic, and capable of being trained end-to-end.

2. Methodology

We aim to computationally model the phonetic structure within ASL. We restrict ourselves to a simplified form of the MH Model. Our primary deviations from the MH Model are

1. we are modelling the global positions (specifically the wrist, elbow, shoulder, and head positions), and
2. we are explicitly modeling holds and implicitly modelling movement between the holds and compressing it into a single frame.

2.1. Dataset

ASLing is an American Sign Language corpus produced by three fluent signers of American Sign Language. Two signers identify as Deaf, and one signer identifies as Hard of Hearing. At the time of writing, ASLing contained over 800 phrases. The dataset had over 1,200 signs, with each sign produced an average of about 3 times. The phrases were chosen because they each contained at least one word with a high frequency in English. Each sign was glossed and each phrase was translated into English. Each phrase was tagged with a level of noise (none, low, medium, high, or broken) that described the performance of our pose predictor. The signers were instructed to record with as much of their upper body in frame as possible, in a location with good lighting, and while wearing contrastive clothing. The data was collected in phrases instead of individual signs to allow for more natural signing.

2.2. Feature Extraction and Data Preparation

The data collected was stored in a video format. Each sign was extracted from the videos (resulting in $M = 1515$ signing portions). From each frame of the signing portions, 2D estimates of key points were extracted using a Convolutional Pose Machine [12] (specifically the head, shoulders, elbows, and wrist locations). These poses were then normalized (setting the head to the origin and setting the unit to the average distance between the head and shoulders) resulting in $D = 14$ features. Finally, the sign were padded to a length of $P = 25$ frames with a vector of zeros (as this was the length of the longest sign). Note that the vector of zeros was used as the end token.

2.3. Phonetic Model

Through the MH Model, the production of signs can be treated as a sequential process (see Figure 1 for an example). The entire process can be posed as a Probabilistic Graphical Model (PGM). PGMs have several benefits, including their ease of training and interpretability. Additionally, PGMs can be used to generate synthetic data by sampling from the model [13]. We chose specifically to use a Dynamic Bayesian Network to model the sequential aspect of signing. Specifically, let

N =The number of body configurations prototypes
M =The number of signs
P =The number of frames in each sign
D =The dimensionality of frame features
μ_i =The i^{th} idealized body configuration prototypes
σ =The amount of dispersion around each dimension from the idealized body configuration prototypes during sign production
π =The probability of each body configuration prototype being at the start of a sign
$T_{i,j}$ =The transition probability of moving from body configuration prototype i to body configuration prototype j
c_i^w =The body configuration prototype chosen for frame i in sign w
x_i^w =The observed production of the chosen body configuration prototype for frame i in sign w

In addition to these, we have the hyperparameters $\alpha, \mu_\mu, \mu_\sigma, \sigma_\mu, \sigma_\sigma$. We specifically chose $\alpha = 1, \mu_\sigma = 1, \sigma_\sigma = 10, \mu_\mu = 0, \sigma_\mu = 10$. Our full model can then be stated as

$$\sigma|\mu_\sigma, \sigma_\sigma \sim LogNormal(\mu_\sigma, \sigma_\sigma) \tag{1}$$
$$T_n|\alpha \sim Dir(\alpha \mathbf{1}_N) \tag{2}$$
$$\mu_0 = \mathbf{0}_D \tag{3}$$
$$\mu_{n>0}|\mu_\mu, \sigma_\mu \sim \mathcal{N}(\mu_\mu, Diag(\sigma_\mu)) \tag{4}$$
$$\pi|\alpha \sim Dir(\alpha \mathbf{1}_N) \tag{5}$$
$$c^w_{f=0}|\pi \sim Categorical(\pi) \tag{6}$$
$$c^w_{f>0}|T, c^w_{f-1} \sim Categorical(T_{c^w_{f-1}}) \tag{7}$$
$$x^w_f|\mu, c^w_f, \sigma \sim \mathcal{N}(\mu_{c^w_f}, Diag(\sigma)) \tag{8}$$

This process is shown in the plate diagram in Figure 3. Note that the 0th body configuration is defined as a special end phoneme (which is indicated by the 0 feature vector).

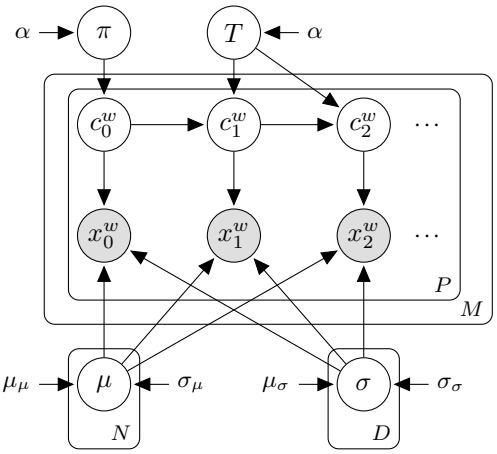

Figure 3: *The plate diagram of our PGM. The non-shaded nodes are latent variables. Shaded nodes are observed variables.* $\alpha, \mu_\mu, \sigma_\mu, \mu_\sigma, \sigma_\sigma$ *are hyperparameters.*

2.3.1. Training

We wish to obtain estimates of $\theta = \pi, T, \mu, \sigma$. With these parameters, one can run the generative process in full and derive

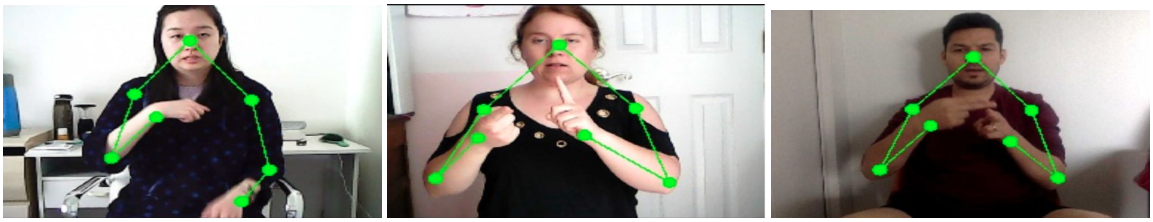

Figure 2: *Example frames from our database, ASLing. The green points indicate the features extracted by our Convolutional Pose Machine: the algorithm used to extract body keypoint locations.*

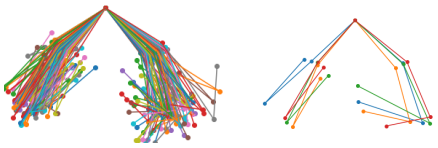

Figure 4: *A sampling of body poses exhibited in the data (left) and the learned body configuration prototypes from the 5 body configuration prototype model (right).*

maximum a posteriori (MAP) estimates of the model conditioned on data. As we wish only to obtain MAP estimates of these, we will use Expectation Maximization (EM).

To do this, we follow the standard EM procedure:

1. Calculate the optimal cluster assignments given the data and our current estimation of θ.
2. Calculate an updated (MAP) estimate of θ using our cluster assignments and prior on θ.

To compute the optimal cluster assignments, we calculate $P(Z|X)$:

$$P(c_0^w = i|x_0^w) \propto e^{\frac{-1}{2}(x_0^w-\mu_i)^T diag(\sigma)^{-1}(x_0^w-\mu_i)}\pi_i \tag{9}$$

$$P(c_f^w = i|c_{f-1}^w = j, x_f^w) \propto e^{\frac{-1}{2}(x_f^w-\mu_i)^T diag(\sigma)^{-1}(x_f^w-\mu_i)}T_{j,i} \tag{10}$$

From this, one can trivially compute the optimal cluster assignments ($Z = c$) given θ in $\mathcal{O}(MPN)$.

To compute the optimal estimate of θ given our cluster assignments and priors we use calculate $P(\theta|X, Z)$

$$P(\pi|c_0^w = i) \propto \frac{\pi_i}{B(\alpha)}\prod_j \pi_j^{\alpha_j - 1} \tag{11}$$

$$P(T_j|c_f^w = i, c_{f-1}^w = j) \propto \frac{T_{j,i}}{B(\alpha)}\prod_k (T_{j,k})^{\alpha_k - 1} \tag{12}$$

$$P(\mu_i|x_f^w, c_f^w = i) \propto e^{\frac{-1}{2}(x_f^w-\mu_i)^T diag(\sigma)^{-1}(x_f^w-\mu_i)} * e^{\frac{-1}{2}(\mu_i-\mu_\mu)^T diag(\sigma_\sigma)^{-1}(\mu_i^w-\mu_\mu)} \tag{13}$$

$$P(\sigma|\mu_i, x, c) \propto P(x, c, \mu|\sigma) * \prod_i \frac{1}{\sigma_i \sigma_{\sigma_i}} exp\left(-\frac{\ln(\sigma_i) - \mu_{\sigma_i}}{2\sigma_{\sigma_i}}\right) \tag{14}$$

where $P(x, c, \mu|\sigma)$ was left unexpanded in the interest of space. These can each be optimized numerically (we chose to optimize these in the order of π, T, μ, σ). These two steps were repeated until convergence.

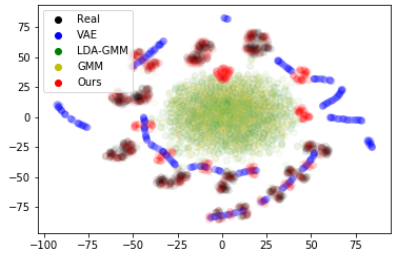

Figure 5: *A T-SNE visualization of the various models presented in this paper*

3. Results and Evaluation

After this model was trained to convergence, we performed the following analysis

3.1. Evaluation

We chose to evaluate our model by determining how capable it was of producing signs. We compared our model against several other models:

1. A GMM-LDA model of sign language
2. A GMM model of sign language
3. A (Seq2Seq) Variational Auto Encoder (VAE) [14]

It should be noted that there was not enough data to successfully train a generative adversarial network[15] (GAN). The former two models can be seen as ablations of our model. The VAE was chosen to determine how our model compared with similar deep models. The probabilistic models are described in further detail at the end of this section.

Each of these algorithms were evaluated on its ability to fool a descriminator. We chose to use a simple neural network as a descriminator (specifically a GRU network). This network was trained on the binary problem of distinguishing between generated data (from the generative processes listed above) and real data. The results are shown in Table 1. In this evaluation, our model outperformed all other models by a wide margin. This was likely because the ablated versions were incapable of capturing the dynamic nature of ASL.

In addition to these quantitative experiments, we also used T-SNE[16] to project simulated data from each model to two

Model	Ours	GMM	GMM-LDA	VAE
Binary Cross Entropy of Descriminator	$\mathbf{0.65 \pm 0.01}$	0.05 ± 0.01	0.03 ± 0.01	0.04 ± 0.02

Table 1: *Binary Cross Entropy (BCE) of a Discriminator trained on distinguishing Synthetic data from Real data. A low BCE means the synthetic generator was unable to adequately fool a discriminator. Note the ability to fool a discriminator corresponds with the ability to produce life-like signs.*

$$
\begin{aligned}
\sigma|\mu_\sigma, \sigma_\sigma &\sim LogNormal(\mu_\sigma, \sigma_\sigma) \\
\mu_0 &= \mathbf{0}_D \\
\mu_{n>0}|\mu_\mu, \sigma_\mu &\sim \mathcal{N}(\mu_\mu, Diag(\sigma_\mu)) \\
\psi_t|\beta &\sim Dir(\beta \mathbf{1}_T) \\
\theta_w|\alpha &\sim Dir(\alpha \mathbf{1}_M) \\
z_w^f|\theta &\sim Categorical(\theta_w) \\
c_f^w|\psi, z_w^f &\sim Categorical(\psi_{z_w^f}) \\
x_f^w|\mu, c_f^w, \sigma &\sim \mathcal{N}(\mu_{c_f^w}, Diag(\sigma))
\end{aligned}
$$

$$
\begin{aligned}
\sigma|\mu_\sigma, \sigma_\sigma &\sim LogNormal(\mu_\sigma, \sigma_\sigma) \\
\mu_0 &= \mathbf{0}_D \\
\mu_{n>0}|\mu_\mu, \sigma_\mu &\sim \mathcal{N}(\mu_\mu, Diag(\sigma_\mu)) \\
\pi|\alpha &\sim Dir(\alpha \mathbf{1}_N) \\
c_f^w|\psi, z_w^f &\sim Categorical(\psi_{z_w^f}) \\
x_f^w|\mu, c_f^w, \sigma &\sim \mathcal{N}(\mu_{c_f^w}, Diag(\sigma))
\end{aligned}
$$

Figure 6: *The model specifications for the LDA-GMM (left) and GMM (right). f is the frame index, t is the topic index, and T is the number of topics. α, β are concentration hyperparameters, π is the relative frequency of the different body configuration prototypes, and ψ, θ are the word distributions of the topics and the topic choice for the signs respectively*

dimensions (shown in Figure 5). From this, we can observe that the LDA-GMM and GMM models underfit (likely due to the high bias of the model). It is also evident that the VAE did not fit the data well. Finally, one can observe that our model produced several signs which did not coincide with real data. This was likely due to our model implicitly modeling movement within the MH Model. Even so, all signs present in the corpus had similar signs generated by our model. This implies that our model was capable of modeling all body configurations present in the corpus.

3.1.1. GMM-LDA and GMM

The GMM-LDA model we used is a topic model[17] over the body configurations used within a sign. In this model, the topics are distributions of body configuration prototypes and the words are the body configuration prototypes themselves. The GMM model is similar to the GMM-LDA. Their model specifications are given in Figure 6.

3.2. Interpretation

To interpret the model, we chose to fit a simplified model containing only five body configuration prototypes (including the end state). Note that without loss of generality, this section assumes the signer is right-hand dominant. The body configuration prototypes discovered appear to be the following body configurations (where the listed location is the location of articulation for the right hand): end state, head, right chest, left hand, and neutral signing space (i.e. the location used for fingerspelling). We examined π to determine that the most common starting position was the right chest area. The least common starting position was the head (we are not counting the end state in the least common starting positions). In addition, one can determine the average length of each hold. This can be done by examining the self transition probabilities $T[j, j]$ and running Geometric Trials (where the first transition away from the given state is a success):

$$\mathbb{E}(\text{length of hold } i) = \frac{1}{1 - T[i, i]} \tag{15}$$

From this, one can determine most holds last for approximately 8.16 frames (with the standard deviation between the

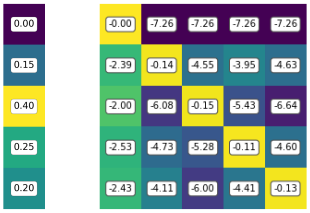

Figure 7: *The initial probabilities and (natural) log of transition probabilities learned by our simplified phonetic model. Note the high probability of self transitions. Self transitions mean that one stayed in the body configuration prototype and thus naturally maps to the concept of a Hold (see Figure 1)*

lengths of the different body configuration prototypes being only 0.98 frames). This means that each hold is approximately 800 milliseconds. Note that the holds derived by the non-simplified model are substantially shorter (in the order of 400 milliseconds). We can also determine the popularity of each body configuration prototype by evaluating the expected number of times each will arise within a sign:

$$\mathbb{E}(\text{count of } j) = \sum_{i=0}^{19} P(c_i = j) = \sum_{i=0}^{19} (\pi T^i)[j] \tag{16}$$

This leads to the conclusion that the "ending state" will be be in an average of 11 frames. Additionally, the most common pose is the right chest (showing up an average of 2.91 frames per sign) and the least common is the head (showing up an average of 1.86 times per sign). We can also examine the probability of any given body configuration prototype ending a sign by examining the probability of transitioning to the 0th prototype. Therefore, the right chest is the most common ending body configuration prototype and the left hand is the least common ending body configuration prototype. Finally, we can examine the amount of variation during production between each joint compared to the body configuration prototype. From examining this, one can observe that the head and shoulder exhibit the least variation. The elbows and left hand exhibit slightly more variation. The right hand exhibits the most variation. This is to be expected as the right hand performs the most motion when signing.

4. Conclusions

In this paper we successfully modeled the phonetic structure of ASL using a simplified Movement-Hold Model. We evaluated our model against multiple ablated versions and a deep counterpart. Finally, we interpreted various aspects of the model to demonstrate its utility and interpretability. Nevertheless, our model did have several limitations. First, we did not explicitly model movement. This likely caused the artifacts seen in Figure 5. Additionally, our model did not model the local movement

of the hand (specifically handshape and palm orientation). We plan on not only addressing these issues in future works, but also modeling non-manuals and incorporating 3D to allow for easier interpretation.

5. References

[1] S. Davies, S. O'Brien, and M. Reed, Mar 2001. [Online]. Available: https://www.uvm.edu/~vlrs/doc/sign_language.htm

[2] W. C. Stokoe, "Sign language structure: An outline of the visual communication systems of the american deaf," *The Journal of Deaf Studies and Deaf Education*, vol. 10, no. 1, pp. 3–37, 01 2005.

[3] S. Liddell and R. Johnson, "American sign language: The phonological base," *Sign Language Studies*, vol. 64, pp. 195–278, 01 1989.

[4] C. Valli, C. Lucus, and K. J. Mukrooney, *The Stokoe System*, 4th ed. Clerc Books, 2005, p. 23–26.

[5] ——, *The Concept of Sequentially in the Description of Signs*, 4th ed. Clerc Books, 2005.

[6] C. L. Clayton Valli, *The Liddell and Johnson Movement-Hold Model*, 4th ed. Clerc Books, 2005, p. 34–38.

[7] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," vol. 1, 01 2007, pp. 2513–2516.

[8] S. Theodorakis, V. Pitsikalis, and P. Maragos, "Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition," *Image and Vision Computing*, vol. 32, 08 2014.

[9] G. Fang, X. Gao, W. Gao, and Y. Chen, "A novel approach to automatically extracting basic units from chinese sign language," vol. 4, pp. 454–457, 2004.

[10] C. Vogler and D. Metaxas, "Toward scalability in asl recognition: Breaking down signs into phonemes," in *Gesture-Based Communication in Human-Computer Interaction*, A. Braffort, R. Gherbi, S. Gibet, D. Teil, and J. Richardson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 211–224.

[11] B. Bauer and K.-F. Kraiss, "Towards an automatic sign language recognition system using subunits," vol. 2298, pp. 64–75, 04 2001.

[12] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4724–4732.

[13] D. Koller and N. Friedman, *Probabilistic graphical models principles and techniques*. MIT Press, 2012.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes." in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://dblp.uni-trier.de/db/conf/iclr/iclr2014.html#KingmaW13

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," Jun. 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[16] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[17] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 2003, 2003.