

Accepted at the Gesture Workshop'99, March 17–19, 1999,  
Gif-sur-Yvette, France.

See Appendix B for some comments that did not make it into the paper.

# Toward Scalability in ASL Recognition: Breaking Down Signs into Phonemes

Christian Vogler and Dimitris Metaxas

Department of Computer and Information Science,

University of Pennsylvania, Philadelphia, PA 19104-6389

cvogler@gradient.cis.upenn.edu, dnm@central.cis.upenn.edu

## Abstract

In this paper we present a novel approach to continuous, whole-sentence ASL recognition that uses phonemes instead of whole signs as the basic units. Our approach is based on a sequential phonological model of ASL. According to this model the ASL signs can be broken into movements and holds, which are both considered phonemes.

This model does away with the distinction between whole signs and epenthesis movements that we made in previous work [14]. Instead, epenthesis movements are just like the other movements that constitute the signs.

We subsequently train Hidden Markov Models (HMMs) to recognize the phonemes, instead of whole signs and epenthesis movements that we recognized previously [14]. Because the number of phonemes is limited, HMM-based training and recognition of the ASL signal becomes computationally more tractable and has the potential to lead to the recognition of large-scale vocabularies.

We experimented with a 22 word vocabulary, and we achieved similar recognition rates with phoneme- and word-based approaches. This result is very promising for scaling the task in the future. We plan to conduct more experiments that will demonstrate that using phonemes can improve both recognition rates and computational complexity.

## 1 Introduction

Gestures are destined to play an increasingly important role in human-computer interaction in the future. Humans use gestures in their everyday communication with other humans, not only to reinforce the meanings that they convey through speech, but also to convey meaning that would be difficult or impossible to convey through speech alone. Surely, to make human-computer interaction truly natural, computers must be able to recognize gestures in addition to speech. Furthermore, gesture recognition is an important part of virtual reality environments, where the user must be able to manipulate the environment with his hands.

Closely related to the field of gesture recognition is the field of sign language recognition. Because sign languages are the primary mode of communication for many deaf people, and because they are full-fledged languages in their own rights, they offer a much more structured and constrained research environment than general gestures. Thanks to linguistic research since the early 1960s, the properties of sign languages, especially of American Sign Language (ASL), have become well-understood. For these reasons, sign language recognition offers an appealing test bed for researching

the more general problems of gesture recognition. Last but not least, working sign language recognition systems would also make the interaction of deaf people with their surroundings easier.

Possibly the most significant property of sign languages is that signs do not consist of unanalyzable wholes. They can be broken down into parts in a systematic manner, much like words in spoken languages can be broken down. Such a breakdown is an essential prerequisite for building truly scalable systems with large vocabularies (or gesture sets).

Yet, to date, research on systematically breaking down signs into their constituent parts for recognition purposes has been sketchy. If such research addressed the problem at all, it followed the early transcription system of ASL by Stokoe [12]. This system has several shortcomings, the most serious of them being that it treats all aspects of signs as occurring in parallel. More recent research in the late 1980s and early 1990s has shown that sequentiality is a very important feature of sign languages, and that it should in fact be the base for a good phonological model of ASL [7, 2].

In this paper we explore the possibilities of basing continuous, whole-sentence ASL recognition on a sequential phonological model. Our focus is strictly on phonology. We do away with the distinction between whole signs and epenthesis movements that we made in previous work [14], and unify them in a single phonological framework. Epenthesis movements are just like the movements that constitute signs. Although morphology, syntax, and semantics are important aspects of sign language recognition, they are beyond the scope of this paper. For simplicity, we do not address handshapes and nonmanual features, such as facial expressions, in this paper either. However, this is not a limitation, because they can be expressed in terms of phonemes as well.

We begin with an overview of related work, then proceed to a discussion of ASL phonology, and show how Hidden Markov Models can be used to capture statistical variations in sign movements. We then provide preliminary experiments with a 22 sign vocabulary to validate our assumptions about phonological modeling of ASL. Finally, we provide a discussion of open research questions.

## 2 Related Work

Most previous work has focused on isolated sign language recognition with clear pauses after each sign. These pauses make it a much easier problem than continuous recognition without pauses between the individual signs, because explicit segmentation of a continuous input stream into the individual signs is very difficult. For this reason, work on isolated recognition often does not generalize easily to continuous recognition.

M. B. Waldron and S. Kim use neural networks to recognize a small set of isolated signs [16]. They use Stokoe's transcription system [12] to separate the handshape, orientation, and movement aspects of the signs. M. W. Kadous uses Power Gloves to recognize a set of 95 isolated Auslan signs with 80% accuracy, with an emphasis on computationally inexpensive methods [5]. R. Erensthteyn and colleagues use neural networks to recognize fingerspelling [3].

Kirsti Grobel and Marcell Assam use HMMs to recognize isolated signs with 91.3% accuracy out of a 262-sign vocabulary. They extract the features from video recordings of signers wearing colored gloves. [4]

There is still relatively little work on continuous sign language recognition. Most of it is based on Hidden Markov Models (HMMs). HMMs offer the advantage of being able to segment a data stream into its constituent signs implicitly, thus bypassing the difficult problem of segmentation.

T. Starner and A. Pentland use a view-based approach with a single camera to extract two-dimensional features as input to HMMs with a 40-word vocabulary and a strongly constrained sentence structure consisting of a pronoun, verb, noun, adjective, and pronoun in sequence [11]. They assume that the smallest unit in sign language is the whole sign and make no further effort to break the signs down into their constituent parts.

Y. Nam and K. Y. Wahn [8] use three-dimensional data as input to HMMs for continuous recognition of a very small set of gestures. They introduce the concept of movement primes, which make up sequences of more complex movements.

R. H. Liang and M. Ouhyoung use HMMs for continuous recognition of Taiwanese Sign Language with a vocabulary between 71 and 250 signs. [6] They work with Stokoe's model [12] to detect the handshape, position, orientation, and movement aspects of the running signs. Unlike other work in this area, they do not use the HMMs to segment the input stream implicitly. Instead, they perform explicit segmentation based on discontinuities in the movements. They perform the integration of the handshape, position, orientation, and movement aspects at a higher level than the HMMs. The sequential aspects of sign language also manifest themselves only at that higher level.

C. Vogler and D. Metaxas use HMMs for continuous ASL recognition with a vocabulary of 53 signs and a completely unconstrained sentence structure [14, 15]. In [15] they use whole-word context-dependent modeling for the HMMs, which segment the input stream implicitly. They couple this approach with a purely computer-vision based analysis that segments the input stream explicitly and extracts its geometric properties to back up the HMM modeling. In [14] they drop whole-word context-dependent modeling in favor of modeling transitions between signs explicitly. These transitions are known as **movement epenthesis** and are an integral part of ASL phonology. However, they still use whole signs as the smallest units of ASL.

This paper is an extension of the work done in [14]. Our goal is to abandon the notion of whole signs as the smallest units of ASL and replace them with phonemes. We strive to treat the aspects of ASL phonology at the HMM level as comprehensively as possible, including the sequential aspects. We now summarize the relevant linguistic research in ASL.

### 3 American Sign Language Phonology

Before we review what is known about ASL phonology, a quick note about terminology. Although sign languages appear to be radically different from spoken languages, the differences are largely in appearance, rather than in the underlying concepts. Most concepts from spoken language linguistics readily carry over to sign language linguistics. For this reason we follow the established terminology of spoken language linguistics.

A **phoneme** is defined to be the smallest contrastive unit in a language; that is, a unit that distinguishes one word from another. In ASL, an example of such a phoneme would be the downward movement in the sign for "good." Phonemes are especially interesting for recognition purposes, because their number is limited in any language,

as opposed to an unlimited number of words that can be built from the phonemes. This limited set of phonemes helps keeping speech recognition tractable. We attempt to show that they can also help keep ASL recognition tractable.

### 3.1 Stokoe's system

W. Stokoe realized that signs can indeed be broken down into smaller parts [12]. He used this observation for devising a transcription system. This transcription system assumes that signs can be broken down into three parameters (phonemes), which consist of the location of the sign (**tabula** or **tab**), the handshape (**designator** or **dez**), and the movement (**signation** or **sig**).

A fundamental assumption of this system is that the tab, dez, and sig contrast only simultaneously. That is, variations in the sequence of these parameters within a sign are considered not to be significant. Many other transcription systems are based on the Stokoe system, such as [9].

### 3.2 Segmental Models

S. Liddell and R. Johnson argued convincingly against Stokoe's assumption that there was no sequential contrast in ASL. They went even further and made sequential contrast the basis of ASL phonology [7]; that is, instead of emphasizing the simultaneous occurrence of phonemes in ASL, they emphasized sequences of phonemes. Such models are called **segmental models**.

S. Liddell and R. Johnson describe two major classes of segments in their Movement-Hold model in [7], which they call **movements** and **holds**. Movements are defined as those segments during which some aspect of the signer's configuration changes, such as a change in handshape, a hand movement, or a change in hand orientation. Holds are defined as those segments during which all aspects of the signer's configuration remain stationary; that is, the hands remain stationary for a brief period of time.

Signs are made up of sequences of movements and holds. Some common sequences are *HMH* (a hold followed by a movement followed by another hold, such as "good"), *MH* (a movement followed by a hold, such as "sit"), and *MMM* (three movements followed by a hold, such as "chair"). Attached to each segment is a **bundle of articulatory features** that describe the hand configuration, orientation, and location. In addition, movement segments have features that describe the type of movement (straight, round, sharply angled), as well as the plane and intensity of movement.

Although the Movement-Hold model has some shortcomings, such as the absence of nonmanual features and the presence of redundancy, its basic sequential structure has been accepted [2]. In addition, a sequential phonological model is ideally suited for a Hidden Markov Model recognition framework, see Section 4.

In this paper we follow the ideas of the Movement-Hold model, but focus only on the movement types and the locational features. We also add wrist rotation movements and the directions in which the movements take place to the description of the movement types. In the Movement-Hold model wrist rotations and directions of movement are implicit in the articulatory bundles; however we found it impractical to model our recognition framework in this way.

In Table 1 and Figure 1 we give a partial overview of the different descriptions of movements and locations that we used. In addition, the locations can be modified with the distance from the body, and with the vertical and horizontal distance from the basic location.

If a location does not touch the body, it can be prefixed with one of these distance markers: *p* (proximal), *m* (medial), *d* (distal), or *e* (extended), in order of distance to the body. If a location is centered in front of the body, the distance marker is suffixed with a 0. If the location is at the side of the chest, the distance marker is suffixed with a 1, and if the location is to the right (or left) of the shoulder, the distance marker is suffixed with a 2. For example, *d-1-TR* means a location a comfortable arm’s length away from the right side of the trunk (torso). Further markers describe the vertical offset to the basic location and whether the location is on the same side or opposite side of the body as the hand. These are described in detail in [7].

Movement	Transcriptions used
straight	<i>strAway</i> , <i>strToward</i> , <i>strDown</i> , <i>strUp</i> , <i>strLeft</i> , <i>strRight</i> , <i>strDownAway</i> , <i>strDownRightAway</i>
short straight	<i>strShortUp</i> , <i>strShortDown</i>
circle in vertical plane	<i>rndVP</i>
wrist rotation	<i>rotAway</i> , <i>rotToward</i> , <i>rotUp</i> , <i>rotDown</i>

Table 1: Partial list of movements. Note that the description of the movements deviates from the approach used by the Movement-Hold model.

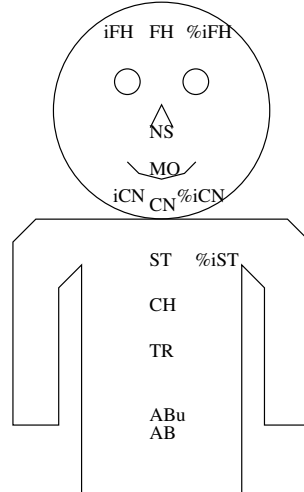


Figure 1: Partial list of body locations used in the Movement-Hold Model

### 3.3 Phonological Processes

S. Liddell and R. Johnson also describe several phonological processes in ASL [7]. A phonological process changes the appearance of an utterance through well-defined rules in phonology, but does not change the meaning of the utterance. In order to achieve robustness, a recognition system must be able to cope with such processes.

The most basic, and at the same time also most important phonological process is called **movement epenthesis**. It consists of the insertion of extra movements between two adjacent signs, and it is caused by the physical characteristics of sign languages. For example, in the sequence “father read,” the sign for “father” is performed at the forehead, and the sign for “read” is performed in front of the trunk. Thus, an extra movement from the forehead to the trunk is inserted that does not exist in either of the two signs’ lexical forms (Figure 2).



Figure 2: Movement epenthesis. The arrow in the middle picture indicates an extra movement between the signs for “FATHER” and “READ” that is not present in their lexical forms.

Movement epenthesis poses a problem for ASL recognizers, because the appearance of the movement depends on which two signs appear in sequence. We handle this problem by modeling such movements explicitly. Ideally, these movements should be captured by the same phonemes as we use for the movements within signs. Unfortunately, epenthesis movements are not as well-defined and researched as the movements that constitute the actual signs. Therefore, we choose to model each epenthesis movement as a separate phoneme for the time being. We do not yet model any other phonological processes in ASL, such as hold deletion and metathesis (which allows for swapping of the order of segments under certain circumstances).

We now cover briefly how to model ASL phonemes with Hidden Markov Models.

## 4 Hidden Markov Models

There are always statistical variations in the way that humans perform movements, even if they perform two identical signs successively. A recognition system must be able to handle these variations. Hidden Markov Models (HMMs) are a state-based statistical model especially suitable for modeling a signal over time. They have been used successfully in speech recognition, and more recently in gesture and sign language recognition.

The underlying idea is to have a distinct HMM for each phoneme. These HMMs are trained to yield the maximum probability for the signal representing their respective phoneme. For the recognition task, compute the probabilities that parts of the input

signal could have been generated by the HMMs and pick the most probable HMM as the recognized phoneme. For a thorough discussion of HMM theory see [10], and for a thorough discussion of the details of using HMMs for sign language recognition, see [13, 14].

#### 4.1 Phoneme Modeling with HMMs

Training HMMs that represent movement and hold phonemes is a straightforward process. However, from looking at the phonetic transcriptions of ASL signs, it becomes clear that many signs start with a movement phoneme (that is, they follow the *MH*, *MMMH*, or *MHMH* pattern). Since we classify movement phonemes only by their type and direction of movement, which can take place anywhere in the signing space, we do not get a good estimate of a sign's location until we encounter the first hold segment. Particularly for the *MMMH* pattern, this can lead to unnecessary classification errors for the sign's location.

This problem can be alleviated by adding HMMs that do not have a phonetic equivalent in the Movement-Hold model. Their sole purpose is to obtain an estimate of the location at the beginning of signs that begin with a movement segment. They are different from hold models in that they do not require the hand to remain stationary for any length of time.

Training the HMMs representing the epenthesis phonemes is more complicated than training the movement and hold HMMs. The reason is that there are many more epenthesis models than models of any other kind of phonemes. In the worst case, there must be an epenthesis phoneme from every location in the signing space to every other location in the signing space. Just for the 20 major body locations defined by the Movement-Hold model, this would yield  $20^2 = 400$  phonemes.

Fortunately, we can reduce the number of epenthesis models by taking advantage of the similarities between many of the epenthesis phonemes. For example, for practical purposes, there is no difference between a movement from the side of the forehead to the chest, and the center of the forehead to the chest (*iFH* to *CH*, and *FH* to *CH*, respectively). Thus, these two phonemes can be covered by a single model. Applying such optimizations allowed us to cut the number of epenthesis models into half. Future work should express epenthesis models completely in terms of the movements that already exist in ASL, so as to ameliorate this problem even more.

The single greatest advantage of breaking down the signs into the individual phonemes is that it limits the number of HMMs that need to be trained. There is only a finite number of distinct phonemes, whereas the number of possibilities to combine them into words is practically unlimited. Although there is no real benefit in modeling phonemes as opposed to whole signs for small-scale applications, it is the only way to make large-scale applications possible. The benefits become particularly obvious when context-dependent HMMs are used. Using a HMM for every possible sequence of two phonemes is tractable. Using a HMM for every possible sequence of two signs is not, even if the vocabulary is as small as 150 signs, because the number of required models is the square of the vocabulary size.

## 4.2 Local Features and Global Features

Recognition performance depends significantly on the features that are extracted from the input signal. Some features that we use are extremely localized; that is they characterize the signal only in the immediate vicinity of a specific point in time. Both the position of the hands in the signing space and the velocities of the hands are examples of local features. They do not reveal anything about the behavior of the signal just a hundred milliseconds from the time at which they are sampled.

But particularly ASL movement phonemes describe geometric properties of the signal on a more global level, such as movements along a straight line, or along an arc. Thus, it is desirable to have a quantitative measure of some of the signal’s global properties. An example of such a measure is how well the signal fits a line or a plane within a specific time interval.

This measure can be easily computed by estimating the covariance matrix over the points in the time interval and taking its eigenvalues. If the largest eigenvalue is significantly larger than the other two eigenvalues, the signal fits a line well. If the two largest eigenvalues are nearly equally large, and significantly larger than the smallest eigenvalue, the signal fits a plane well. These relationships can be quantified with two numbers by taking the square roots of the two largest eigenvalues, and normalizing them such that the sum of the square roots of all three eigenvalues is 1.

## 5 Experiments

We designed several experiments to verify that breaking down signs into phonemes is a viable approach in ASL recognition. Our vocabulary consisted of 22 signs with the phonetic transcriptions listed in Table 3 in Appendix A. Note that the phonemes beginning with an “M” are movement phonemes, phonemes beginning with an “H” are hold phonemes, and “phonemes” beginning with an “S” denote the additional HMMs mentioned in Section 4.1 along with the locations they are to estimate.

We collected 499 sentences of different length, with 1610 signs overall, with an Ascension Technologies MotionStar™ magnetic tracking system. This system gave us three-dimensional positions and orientations of the hands and other body parts at 60 frames per second.

We split the 499 sentences into 400 training examples with 1292 signs and 99 test examples with 318 signs. No part of the test examples was used for any part of the training of the HMMs. We conducted three different types of experiments, one of which was a control experiment that measured the performance of word-level HMMs along with movement epenthesis modeling. This control experiment was similar to the one conducted in [14]. The other two experiments tested the performance of the phoneme-level HMMs, one without global features, and one with global features.

To keep the experiments simple, we looked only at features extracted from the right hand. In all cases, the local features were the right hand’s position in space, relative to the signer’s base of the spine, and the right hand’s velocities. The global features consisted of the two largest normalized eigenvalues, as described in Section 4.2. The results are given in Table 2. We use word accuracy as our evaluation criterion. It is computed by subtracting the number of insertion errors from the number of correctly spotted signs.

The results indicate that the phoneme-level HMMs did not perform significantly worse than the word-level HMMs. They also indicate that global features are a valuable characterization of the signal. Both the breakdown of signs into movement and hold phonemes, and the research on global features look promising.

Type of experiment	Word acc.	Details
word-level	91.82%	H=296, D=10, S=12, I=4 N=318
phoneme-level, local features	88.36%	H=286, D=14, S=18, I=5, N=318
phoneme-level, global features	91.19%	H=294, D=8, S=16, I=4, N=318

Table 2: Results of recognition experiments. H denotes the number of correct signs, D the number of deletion errors, S the number of substitution errors, I the number of insertion errors, and N the total number of signs in the test set.

## 6 Discussion and Future Work

We showed that it is possible for phoneme-level HMMs to achieve ASL recognition performance comparable to word-level HMMs. However more work needs to be done to establish the validity of the results, they are already very important. The entire question of whether it is possible to scale ASL recognition to large vocabularies hinges on this result. We also showed that analyzing the input stream for global features has the potential to make a large impact on recognition performance.

There are, however, many questions that still need to be resolved. In the experiments described in this paper, we have looked only at the right hand. The left hand should remain as independent from the right hand as possible, both from a linguistic point of view and a technical point of view. Liddell and Johnson argue that the two hands are more or less independent from each other, as well as that the articulatory bundles are relatively independent from each other [7]. From a technical point of view, if the two hands were dependent on each other, it would cause a combinatorial explosion of different phonemes. It seems that the answer to these questions lies in using several HMMs in parallel, either independently, or as Coupled Hidden Markov Models [1]. The hand configuration and orientation features could be incorporated in a similar way.

Future research should also look at ways to express the epenthesis phonemes in terms of phonemes that occur during regular signs, so as to cut down on the number of distinct phonemes. Finally, training biphone or triphone context-dependent HMMs, analogous to speech recognition, might be a way to improve recognition performance even further.

### Acknowledgments

This work was supported in part by a NSF Career Award NSF-9624604, ONR-DURIP'97 N00014-97-1-0385 and N00014-97-1-0396, ONR Young Investigator Proposal, and NSF IRI-97-01803.

### References

- [1] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

- [2] Geoffrey R. Coulter, editor. *Current Issues in ASL Phonology*, volume 3 of *Phonetics and Phonology*. Academic Press, Inc., San Diego, CA, 1993.
- [3] R. Erenshsteyn and P. Laskov. A multi-stage approach to fingerspelling and gesture recognition. To appear in the proceedings of the Workshop on the Integration of Gesture in Language and Speech, Wilmington, DE, USA, 1996.
- [4] Kirsti Grobel and Marcell Assam. Isolated sign language recognition using hidden Markov models. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 162–167, Orlando, FL, 1997.
- [5] Mohammed Waleed Kadous. Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. In Lynn Messing, editor, *Proceedings of WIGLS. The Workshop on the Integration of Gesture in Language and Speech*, pages 165–174, Applied Science and Engineering Laboratories Newark, Delaware and Wilmington, Delaware, October 1996.
- [6] R.-H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, pages 558–565, Nara, Japan, 1998.
- [7] Scott K. Liddell and Robert E. Johnson. American Sign Language: The phonological base. *Sign Language Studies*, 64:195–277, 1989.
- [8] Yanghee Nam and Kwang Yoen Wohn. Recognition of space-time hand-gestures using hidden Markov model. To appear in ACM Symposium on Virtual Reality Software and Technology, 1996.
- [9] Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, and Jan Henning. *Ham-NoSys. Version 2.0; Hamburg Notation System for Sign Languages. An introductory guide*, volume 5 of *International Studies on Sign Language and Communication of the Deaf*. Signum Verlag, Hamburg, 1989.
- [10] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989.
- [11] Thad Starner and Alex Pentland. Real-time American Sign Language recognition from video using Hidden Markov Models. Technical Report 375, MIT Media Laboratory, 1996.
- [12] William C. Stokoe. *Sign Language Structure: An Outline of the Visual Communication System of the American Deaf*. Studies in Linguistics: Occasional Papers 8. Linstok Press, Silver Spring, MD, 1960. Revised 1978.
- [13] Christian Vogler and Dimitris Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. Technical report, MS-CIS-98-21, Department of Computer and Information Science, University of Pennsylvania.
- [14] Christian Vogler and Dimitris Metaxas. Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 156–161, Orlando, FL, 1997.
- [15] Christian Vogler and Dimitris Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 363–369, Mumbai, India, 1998.
- [16] M. B. Waldron and Soowon Kim. Isolated ASL sign recognition system for deaf persons. *IEEE Transactions on Rehabilitation Engineering*, 3(3):261–71, September 1995.

## A Phonetic Transcriptions

Sign	Transcription
I	$S-\{p-0-CH\} M-\{str_{Toward}\} H-\{CH\}$
man	$H-\{FH\} M-\{str_{Down}\} M-\{str_{Toward}\} H-\{CH\}$
woman	$H-\{CN\} M-\{str_{Down}\} M-\{str_{Toward}\} H-\{CH\}$
father	$S-\{p-0-FH\} M-\{str_{Toward}\} M-\{str_{Away}\} M-\{str_{Toward}\} H-\{FH\}$
mother	$S-\{p-0-CN\} M-\{str_{Toward}\} M-\{str_{Away}\} M-\{str_{Toward}\} H-\{CN\}$
interpreter	$S-\{m-1-CH\} M-\{rot_{Down}\} M-\{rot_{Up}\} M-\{rot_{Down}\} S-\{p-1-CH\} M-\{str_{Down}\} H-\{m-1-TR\}$
teacher	$S-\{m-1-CH\} M-\{rot_{Away}\} M-\{rot_{Toward}\} M-\{rot_{Away}\} S-\{p-1-CH\} M-\{str_{Down}\} H-\{m-1-TR\}$
chair	$S-\{m-1-TR\} M-\{str_{ShortDown}\} M-\{str_{ShortUp}\} M-\{str_{ShortDown}\} H-\{m-1-TR\}$
try	$S-\{p-1-TR\} M-\{str_{DownRightAway}\} H-\{d-2-AB\}$
inform	$H-\{iFH\} M-\{str_{DownRightAway}\} H-\{d-2-TR\}$
sit	$S-\{m-1-TR\} M-\{str_{ShortDown}\} H-\{m-1-TR\}$
teach	$S-\{m-1-CH\} M-\{rot_{Away}\} M-\{rot_{Toward}\} M-\{rot_{Away}\} H-\{m-1-CH\}$
interpret	$S-\{m-1-CH\} M-\{rot_{Down}\} M-\{rot_{Up}\} M-\{rot_{Down}\} H-\{m-1-CH\}$
get	$S-\{d-0-CH\} M-\{str_{Toward}\} H-\{p-0-CH\}$
lie	$S-\{iCN\} M-\{str_{Left}\} H-\{\%iCN\}$
relate	$S-\{m-1-TR\} M-\{str_{Left}\} H-\{m-0-TR\}$
dont-mind	$H-\{NS\} M-\{str_{DownRightAway}\} H-\{m-1-TR\}$
good	$H-\{MO\} M-\{str_{DownAway}\} H-\{m-0-CH\}$
gross	$S-\{ABu\} M-\{rnd_{VP}\} M-\{rnd_{VP}\} H-\{ABu\}$
sorry	$S-\{\%iSTu\} M-\{rnd_{VP}\} M-\{rnd_{VP}\} H-\{\%iSTu\}$
stupid	$S-\{p-0-FH\} M-\{str_{Toward}\} H-\{FH\}$
beautiful	$S-\{p-0-FH\} M-\{rnd_{VP}\} H-\{p-0-iFH\}$

Table 3: Phonetic transcriptions of the 22 sign vocabulary. The phonemes beginning with “M” denote movements, the phonemes beginning with “M” denote holds, and the phonemes beginning with “S” denote special HMMs designed to estimate locations at the beginning of a sign.

## B Additional Notes

The following points came up after the original paper deadline, and as a consequence did not make it into the paper itself:

- The transcriptions for “teacher” and “interpreter” in Appendix A are incorrect. The  $S\{-l-CH\}$  segments should be replaced with  $S\{m-l-CH\}$ . This change improved recognition accuracy from 91.19% to 91.82%, which is identical to the accuracy achieved by the word-level HMMs.
- As described in Section 4.1, we had to add segments that are not in the original description of the Movement-Hold model. In Appendix A these are denoted by the letter “S.” It seems that these segments are very similar in function to the “X” segments that appear in the latest, as of yet unpublished, description of the Movement-Hold model.