

# Adapting Hidden Markov Models for ASL Recognition by Using Three-dimensional Computer Vision Methods

Christian Vogler and Dimitris Metaxas  
Department of Computer and Information Science  
University of Pennsylvania  
Philadelphia, PA 19104  
cvogler@gradient.cis.upenn.edu, dnm@central.cis.upenn.edu

## ABSTRACT

*We present an approach to continuous American Sign Language (ASL) recognition, which uses as input three-dimensional data of arm motions. We use computer vision methods for three-dimensional object shape and motion parameter extraction and an Ascension Technologies Flock of Birds interchangeably to obtain accurate three-dimensional movement parameters of ASL sentences, selected from a 53-sign vocabulary and a widely varied sentence structure. These parameters are used as features for Hidden Markov Models (HMMs). To address coarticulation effects and improve our recognition results, we experimented with two different approaches. The first consists of training context-dependent HMMs and is inspired by speech recognition systems. The second consists of modeling transient movements between signs and is inspired by the characteristics of ASL phonology. Our experiments verified that the second approach yields better recognition results.*

## 1. INTRODUCTION

Sign language and gesture recognition have important applications in virtual reality, where gesturing would certainly provide a natural and efficient way for human-computer interaction. Unlike general gestures, sign language is highly structured, which makes the recognition problem easier, since its structure can be used to form constraints and exploit context. Thus, sign language recognition provides a good starting point for researching more general gesture recognition problems. It is even conceivable that the user could learn a subset of sign language to interact with the computer. In addition, a functional sign language recognition system could facilitate the tedious process of transcribing conversations for sign language research tremendously, as well as facilitating interaction between deaf and hearing people.

This paper presents a new approach to continuous American Sign Language (ASL) recognition. We use computer vision methods for three-dimensional object shape and motion parameter extraction and an Ascension Technologies Flock of Birds interchangeably to obtain accurate three-dimensional movement parameters of ASL sentences from a 53-sign vocabulary

and a widely varied sentence structure. The parameters are used as features for Hidden Markov Models (HMMs), which have successfully been used in speech recognition [6].

Obtaining accurate three-dimensional data, however, forms only the first step toward good recognition performance. For continuous recognition, a major problem are coarticulation effects, that is, the pronunciation<sup>1</sup> of a sign is influenced by the preceding and following signs. We discuss two approaches to solving this problem and provide experimental results. The first one consists of training context-dependent HMMs and is inspired by speech recognition systems. The second one consists of modeling transient movements between signs and is inspired by the characteristics of ASL phonology [7]. The second approach is new and appears to perform better than the first one.

ASL is the primary mode of communication for many deaf people in the United States. Its linguistic properties are the best-understood of all sign languages in the world, thanks to a large user base and early pioneering research, which makes ASL an appealing choice for research in automatic sign language recognition. The arm and hand movements form the most versatile and at the same time easiest to capture aspect of ASL, so our approach concentrates on these. Like all sign languages, ASL makes extensive use of space, which makes three-dimensional methods essential for accurate parameter estimation. The use of space also has important grammatical functions, most notably establishing subjects and objects [12]. However, we choose to ignore them at present to keep the problem domain manageable. For the same reason, we do not address independence across many different signers.

The ability to recognize continuous sentences, without the introduction of artificial pauses, has a profound influence on the naturalness of the human-computer interface. Although recognition of isolated signs is of interest itself, a gesture or sign language interface stands a higher chance of becoming accepted if it does not require the user to introduce unnatural pauses

---

<sup>1</sup>We choose to follow the established terminology of spoken language linguistics where applicable.

during interaction. Besides, only a continuous recognition system will ever stand a chance of aiding in the transcription of sign language conversations.

In the following sections, this paper discusses related work, then briefly describes the data collection process with the Flock of Birds and computer vision methods. After giving an overview on HMM theory, it addresses the coarticulation problem and two alternatives toward a solution, and gives experimental results to back up the discussion.

## 2. RELATED WORK

Previous work on sign language recognition focuses primarily on fingerspelling recognition and isolated sign recognition. Some work uses neural networks [2], [13]. For this work to apply to continuous ASL recognition, the problem of explicit temporal segmentation must be solved. HMM-based approaches take care of this problem implicitly. Mohammed Waleed Kadous uses Power Gloves to recognize a set of 95 isolated Auslan signs with 80% accuracy, with an emphasis on computationally inexpensive methods [14].

There is very little previous work on continuous ASL recognition. Thad Starner and Alex Pentland use a view-based approach 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]. Annelies Braffort describes ARGo, an architecture for LSF recognition based on linguistic principles and HMMs [1], but provides limited experimentation results. Yanghee Nam and Kwang Yoen Wahn [9] use three-dimensional data as input to HMMs for continuous recognition of a very small set of gestures.

## 3. DATA COLLECTION

We collected data for a 53-sign vocabulary, which follows the dialect used in the Philadelphia, PA, area, given in Table 1. The total number of sentences collected was 486, with a total of 2274 signs. Each sentence was between 2 and 12 signs in length. We collected the three-dimensional parameters of the sentences in two different ways: with computer vision methods, and with an Ascension Technologies Flock of Birds, a magnetic sensor system. We outline the methods briefly in the next two subsections.

### Data Collection with Computer Vision

The computer vision approach consists of two parts. The first part consists of an automated, active, integrated approach that identifies the parts of a moving articulated object (such as a signer) and estimates their shape and motion from a set of controlled experiments.<sup>2</sup> This approach estimates the shapes

<sup>2</sup>The set of experiments is the same across humans with different anthropometric dimensions

| Category   | Signs used   |
|------------|--|
| Nouns      | America, Christian, Christmas, book, brother, chair, college, family, father, friend, interpreter, language, mail, mother, name, paper, president, school, sign, sister, teacher |
| Pronouns   | I, my, you, your, how, what, where, why  |
| Verbs      | act, can, give, have, interpret, like, make, read, sit, teach, try, visit, want, will, win   |
| Adjectives | deaf, good, happy, relieved, sad   |
| Other      | if, from, for, hi  |

Table 1: The complete 53 sign vocabulary

of the three-dimensional parts by using information from three orthogonally positioned cameras [3], [4]. The algorithm uses deformable models that are fitted to the image data from each of the three active views.

The second part consists of tracking the identified parts and estimating reliable three-dimensional position and orientation parameters [5], [8]. The use of the three orthogonally positioned cameras alleviates problems with occlusion and degenerate views that occur during unconstrained movements. In addition, the algorithm uses theorems from projective geometry to establish a correspondence between the occluding contours and the points on the three-dimensional deformable models. A physics-based modeling approach transforms the correspondences into generalized forces that are applied to the model, so as to estimate its rotational and translational parameters. Figure 1 demonstrates an example result from the application of this computer vision method.

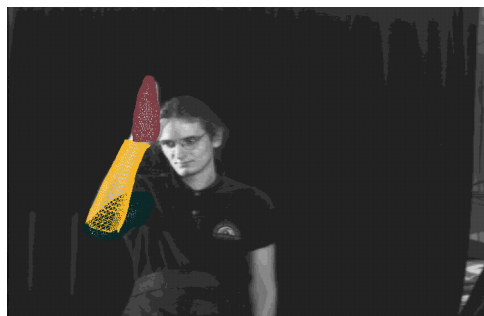


Figure 1: Fitting the models to the sign for “FATHER.”

Our computer vision method gives us three-dimensional wrist position coordinates and orientation parameters, which we use as features (and parameters derived from position and orientation).

### Data Collection with the Flock of Birds

Although the computer vision approach provides accurate estimates and does not require attaching ob-

trusive equipment to the signer’s body, it puts a burden on computational resources. Therefore, for faster prototyping and quick changes to the system, we use an Ascension Flock of Birds to collect most of the training and the test data. The Flock of Birds consists of a magnet and sensors that are capable of detecting their positions and orientations relative to the magnetic field, yielding accurate three-dimensional wrist positions and orientations, among other things.

There is a correspondence between the coordinates established by the computer vision approach and the Flock of Birds, which can be made explicit by a simple alignment of coordinate systems. We verified the interchangeability of the two methods by running a few test examples obtained with vision methods on HMMs trained with Flock of Birds data.

## 4. HMM OVERVIEW

Hidden Markov Models are a type of statistical model. Formally, an HMM  $\lambda$  consists of  $N$  states and a transition matrix, which gives the probabilities  $a_{ij}$  that a transition from state  $S_i$  to state  $S_j$  is taken at discrete time intervals. A vector of probabilities  $\pi_i$  denotes the probability of the HMM  $\lambda$  initially being in state  $S_i$ . Each state has assigned an output probability distribution function  $b_i(\mathbf{O})$ , which gives the probability of  $S_i$  generating observation  $\mathbf{O}$  under the condition that the system is in  $S_i$ .  $b_i(\mathbf{O})$  can be either a discrete or a continuous density.

There are three basic problems associated with HMMs. The first problem consists of estimating  $P(O|\lambda)$ , that is, the probability that HMM  $\lambda$  has generated the observation sequence  $O = \mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_T$ . The forward-backward algorithm solves this problem in  $O(N^2T)$  time [10]. This problem is equivalent to finding the most likely word for an unknown data sequence, under the assumption that each word has assigned an individual HMM.

The second problem consists of finding the most likely state sequence  $S = S_1, S_2, \dots, S_T$  through a given HMM, given an observation sequence  $O$ , that is,  $\max\{P(S|O, \lambda)\}$ . The Viterbi algorithm is a dynamic programming algorithm that finds this probability in  $O(N^2T)$  time and recovers the state sequence  $S$  at the same time [10]. This makes the algorithm useful for continuous word recognition, as the recovery of the state sequence implies the automatic recovery of word boundaries in a recognition network consisting of concatenated HMMs, one for each word. Thus, no explicit segmentation of the data into individual words is necessary.

The third problem consists of adjusting the parameters of an HMM  $\lambda$ , such that, given an observation sequence  $O$ ,  $P(O|\lambda)$  is maximized. The Baum-Welch reestimation algorithm [10] estimates the parameters toward a local maximum in polynomial time. This problem corresponds to training HMMs with a set

of labeled data to enable them to recognize future unknown test data. Because the algorithm provides only local maxima, the performance of HMMs depends critically on their initial parameter estimates before the Baum-Welch algorithm is applied.

## 5. ADAPTING HMMs FOR ASL RECOGNITION

The previous sections described the data collection process and gave an overview on the theory of HMMs. This section describes how HMMs can be adapted for ASL recognition.

It is tempting to look into research in speech recognition and make use of these results for ASL recognition purposes. There are parallels: Both processes aim to recognize language conveyed through a medium — auditory in the case of speech and visual in the case of ASL —; both are time-varying processes, which show statistical variations, making HMMs a plausible choice for modeling the processes; and both must devise ways to cope with context and coarticulation effects.

However, there are also important differences. Speech signals are well-suited for analysis in the frequency domain, whereas ASL signals, due to their spatial nature, do not show such a suitability. Consequently, it is unclear whether results from research in speech recognition on preprocessing and feature vectors apply to ASL recognition at all.

Possibly even more important, the nature of cross-word coarticulation effects in ASL is very different from spoken languages. The sequence of the signs “FATHER READ” provides an example. In its lexical form, the former sign is performed with a 5-handshape by repeatedly tapping the forehead with the thumb, while the latter sign is performed in neutral space in front of the chest. The consequence is the insertion of an extra movement between the two signs, when they are performed in sequence, which is not indicated by their lexical entries. The Movement-Hold phonological model [7] refers to this phenomenon as *movement epenthesis*.

The presence of movement epenthesis greatly complicates the coarticulation problem, since it introduces a great variety of movements that are not present in the signs’ lexical forms. From a phonological point of view, it *inserts extra phonemes*, instead of merely affecting the pronunciation of adjacent phonemes. It can also implicitly cause *deletion of phonemes*, such as the deletion of the final hold in the sign for “GOOD” when it occurs in the sequence “GOOD IDEA.” We now describe two different attempts to overcome the coarticulation problem in ASL recognition.

### Modeling Context-Dependent HMMs

Our first attempt to overcome the coarticulation problem consists of training context-dependent

HMMs. It is inspired by speech recognition, which frequently uses triphone context-dependent HMMs. The analog in ASL is to train bi-sign or even tri-sign context dependent HMMs.<sup>3</sup> Training tri-sign context dependent models requires a prohibitive number of models, even with the small 53-sign vocabulary. Therefore, we limited training to bi-sign context dependent models, which account for the sign immediately preceding the one for which we trained each HMM. This kind of modeling still requires  $53^2 = 2809$  models to cover all possible contexts. After using various possibilities for tying the parameters, we were left with 337 unique HMMs. In particular, our training method ties the transition matrices and output states of all HMMs for a sign, where the preceding signs end in similar locations. We determined the similarities of the locations by hand from the lexical forms of the signs, and determined the effects of tying experimentally. The experiments showed that the tying of the transition matrices was particularly critical for recognition accuracy.

The entire training session followed this scheme. First of all, it assigned the global means and variances uniformly to all states of the HMMs, the reason being that it is notoriously difficult to label boundaries between signs by hand. Then it performed two rounds of context-independent embedded reestimation [15] to gain more reliable initial estimates before tying the parameters. The tying was followed by several more rounds of embedded reestimation until the HMM’s parameters converged.

Because movement epenthesis introduces a large variety of movements of different duration, the HMM topology must be flexible enough to accommodate some variance in the length of different signs. These considerations led to the general topology depicted in Figure 2, with Gaussian densities with diagonal covariance matrices as the output probabilities. We fine-tuned the topology through experimenting.

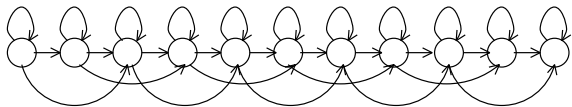


Figure 2: Topology of the context-dependent model. The arcs that skip states allow the modeling of variabilities in the duration of different signs.

This kind of context-dependent modeling for ASL recognition, however, has some inherent problems. First, it is linguistically implausible, because it fails to model movement epenthesis properly. Second, by using signs as the basic phonetic unit, the number of states used in the HMM recognition network grows roughly with order  $O(W^2)$ , where  $W$  is the number

<sup>3</sup>The analogy is not entirely correct; signs are not the smallest unit in ASL phonology, but this topic is beyond the scope of this paper.

of signs in the vocabulary, as the number of possible contexts itself grows with order  $O(W^2)$ . Considering that the mapped ASL vocabulary consists of approximately 6000 signs and still expands, it is clear that large-scale applications would be so expensive computationally that they would become intractable, let alone the problem of collecting sufficient training data.

## Modeling Movement Epenthesis

The problems with context-dependent modeling outlined in the previous subsection, and the limited improvement over context-independent modeling, led us to the adoption of a radically different approach to tackling the coarticulation problem. This new approach models movement epenthesis explicitly, instead of letting the HMM modeling take care of it implicitly. The idea behind explicit modeling is that the movements introduced through epenthesis belong to a *finite, limited class of movements* only. They are largely determined by the starting and ending locations of the lexical forms of the signs, while the signs themselves are relatively unaffected, with the exception of hold deletion.

In order to obtain the necessary class of phonemes for modeling the movements between signs, we performed k-means clustering with a least-squares distance criterion on the start and endpoints of the signs’ lexical forms. The least-squares optimality criterion translates directly to optimal distances as far as Gaussian densities from the HMMs’ output probabilities are concerned.

The k-means algorithm yielded eight distinct clusters; thus, there were 64 different phonemes all in all, of which nine did not occur in the entire sentence database at all. The remaining 55 phonemes almost all had at least five training examples available. We merged the few phonemes that had fewer than five training examples available with the most closely related phonemes in the least-squares sense.

The bootstrap and training procedures for this approach were very similar to the procedures used for context-dependent modeling, except that they perform no parameter tying and do not introduce context-dependent HMMs after the first few training runs. The recognition network was specifically constrained such that for any transition between any two signs, only a path through the correct transition phoneme could be taken. The topology of the recognition network is depicted schematically in Figure 3.

This kind of modeling necessitates some changes in the HMM topology. Because of the elimination of context in the HMMs that represent the signs, they do not differ in length greatly anymore. This change requires the removal of the arcs skipping states. In addition, the transient movements between signs are usually short, but show some variety in their length.

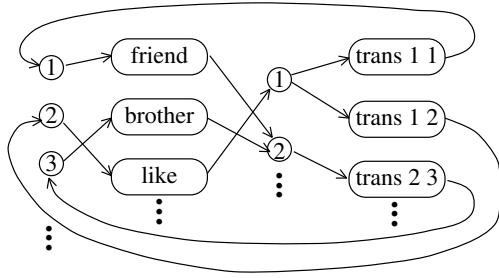


Figure 3: Epenthesis model recognition network. The numbered states correspond to the cluster numbers.

Thus, they require a topology distinct from the one used for the signs. The new topologies are depicted in Figure 4.

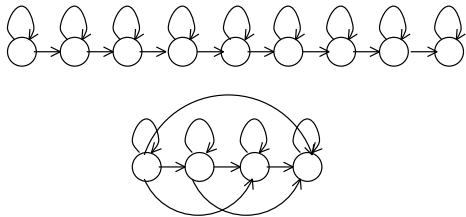


Figure 4: Topology of the epenthesis models. The top model is used for signs, and the bottom model is used for transient movements.

An advantage of movement epenthesis modeling over “traditional” context-dependent modeling is that it is more closely related to the actual phonological processes in ASL. It might be the first step toward a comprehensive incorporation of insights from ASL phonology into ASL recognition research. Another advantage is the reduced complexity of the approach; the recognition network requires only  $O(W + C^2)$  states, where  $W$  is the number of signs in the lexicon and  $C$  is the number of clusters found (hence  $C^2$  is the number of epenthesis phonemes). For large-scale applications, where  $C$  stays relatively small because of the limited number of different phonemes, these savings are substantial.

### Bigram Language Models

In the previous two subsections we described two approaches toward coping with the coarticulation problem. Regardless of the method employed, using a statistical language model that models the probabilities of one sign following another, can improve recognition accuracy. Bigram modeling, too, is inspired by speech recognition work, where it has been shown to have a significant effect on recognition accuracy [6]. This kind of modeling is important for task domains where using a constrained task grammar is impractical, due to the large number of different possible expressions. This is the case for the diverse sentence structure that we used in our experiments.

Bigram probabilities require a large corpus of labeled data to be effective. This is a problem for ASL recognition, because at present no such corpus exists. Collecting the corpus will be an important task for future research. Nevertheless, to test the feasibility of bigram modeling for ASL recognition, we computed bigram probabilities from our restricted data sets and used them in our recognition networks. The experimental results are given in the next section.

## 6. EXPERIMENTS

In the previous section we described several approaches to HMM modeling for ASL recognition and discussed the coarticulation problem. We performed several experiments to measure the relative merits of each approach against one another. These experiments consisted of a run that did not use any context-dependent models or bigram probabilities, two runs that used context-dependent models with and without bigram probabilities, and two runs that used epenthesis modeling with and without bigram modeling. We used Entropic’s Hidden Markov Model Toolkit Version 2.02 [15] for all experiments.

We split the set of 486 sentences randomly into a training set with 389 sentences (containing 1818 signs) and a test set with 97 examples (containing 456 signs). We took special care to ensure that every sign in the vocabulary occurred at least once in the test set. The training and test sets were the same throughout all experiments to make results comparable. The experiments used a combination of three-dimensional wrist position coordinates (velocities by themselves turned out to be very unreliable features); polar coordinates in the  $x$ - $y$  plane, and the  $x$ - $z$  plane; wrist rotation angles; and velocities of the wrists as the features. The coordinate system was right-handed and relative to the signer’s spine, with the  $x$ -axis facing up, the  $y$ -axis facing to the right, and the  $z$ -axis facing forward. Table 2 gives the outcomes of the experiments.

We use word accuracy as the evaluation criterion. It is computed by subtracting the number of insertion errors from the number of correctly recognized signs. The results show that ignoring the coarticulation problem fares worse than tackling it with the two methods described previously. Probably the most remarkable result is that epenthesis modeling performed slightly better than context-dependent modeling even when the latter uses bigram probabilities and the former does not. Using bigram modeling yields a far bigger improvement for epenthesis modeling than for context-dependent modeling.

When bigram probabilities are not used, half of the substitution errors consist of confusions between “I, MY,” and “YOU, YOUR.” This is hardly surprising, as without hand configuration parameters there is no way to distinguish between the signs in each of the

| Type of experiment          | Word accuracy | Details                         |
|-----------------------------|---------------|---------------------------------|
| context independent         | 87.71%        | H=416, D=8, S=32<br>I=16, N=456 |
| context dependent           | 89.91%        | H=424, D=6, S=26<br>I=14, N=456 |
| bigram, context dependent   | 91.67%        | H=426, D=7, S=23<br>I=8, N=456  |
| epenthesis modeling         | 92.11%        | H=426, D=7, S=23<br>I=6, N=456  |
| bigram, epenthesis modeling | 95.83%        | H=438, D=11, S=7<br>I=1, N=456  |

Table 2: Results of 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.

two groups. Confusions between “INTERPRETER” and “INTERPRET-CAN” constitute the second large group of errors. The noun ends with a marker for persons, which is very similar to the verb “CAN” when hand configuration is ignored. We expect that adding hand configuration parameters will improve accuracy significantly.

These results provide evidence that modeling movement epenthesis is promising. Future research should expand on incorporating phonological features of ASL, make use of hand configuration data, and model the use of space. Followup on this research may also necessitate further research on data preprocessing, as the feature vector that we use is not invariant with respect to location and orientation. The current lack of invariance may complicate the modeling of the use of space.

## 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. Ioannis Kakadiaris helped in obtaining the computer vision samples.

## 7. REFERENCES

- [1] A. Braffort. ARGo: An architecture for sign language recognition and interpretation. In P. A. Harling, A. D. N. Edwards, editors, *Progress in gestural interaction. Proceedings of Gesture Workshop ’96*, pp. 17–30. Springer, Berlin, New York, 1997.
- [2] R. Erenshteyn and P. Laskov. A multi-stage approach to fingerspelling and gesture recognition. *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, pp. 185–194, Wilmington, DE, 1996.
- [3] I. A. Kakadiaris, D. Metaxas, and R. Bajcsy. Active part-decomposition, shape and motion estimation of articulated objects: A physics-based approach. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 980–984, Seattle, WA, 1994.
- [4] I. A. Kakadiaris and D. Metaxas. 3D human body model acquisition from multiple views. *Proceedings of the IEEE Fifth International Conference on Computer Vision*, pp. 618–623, Boston, MA, 1995.
- [5] I. A. Kakadiaris and D. Metaxas. Model based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 81–87, San Francisco, CA, 1996.
- [6] C. Lee, F. K. Soong, and K. K. Paliwal, editors. *Automatic Speech and Speaker Recognition, Advanced Topics*. Kluwer Academic Publishers, Boston, MA, 1996.
- [7] S. K. Liddell and R. E. Johnson. American Sign Language: The phonological base. *Sign Language Studies*, 64:195–277, 1989.
- [8] D. Metaxas. *Physics-based Deformable Models: Applications to Computer Vision, Graphics and Medical Imaging*. Kluwer Academic Publishers, November 1996.
- [9] Y. Nam and K. Y. Wohn. Recognition of space-time hand-gestures using Hidden Markov model. *ACM Symposium on Virtual Reality Software and Technology*, pp. 51–58, Hong Kong, 1996.
- [10] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.
- [11] T. Starner and A. Pentland. Real-time American Sign Language recognition from video using Hidden Markov Models. Technical Report 375, MIT Media Laboratory, 1996.
- [12] C. Valli and C. Lucas. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, Washington DC, 1995.
- [13] M. B. Waldron and S. Kim. Isolated ASL sign recognition system for deaf persons. *IEEE Transactions on Rehabilitation Engineering*, 3(3):261–71, September 1995.
- [14] M. W. Kadous. Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, pp. 165–174, Wilmington, DE, 1996.
- [15] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book (for HTK 2.0)*. Cambridge University, 1995.