A New Instrumented Approach For Translating American Sign Language Into Sound And Text

Jose L. Hernandez-Rebollar[1], Nicholas Kyriakopoulos[1], Robert W. Lindeman[2]

The George Washington University, Department of ECE[1], Department of CS[2]

[jreboll,kyriak,gogo]@gwu.edu

Abstract

This paper discusses a novel approach for capturing and translating isolated gestures of American Sign Language into spoken and written words. The instrumented part of the system combines an AcceleGlove and a two-link arm skeleton. Gestures of the American Sign Language are broken down into unique sequences of phonemes called Poses and Movements, recognized by software modules trained and tested independently on volunteers with different hand sizes and signing ability. Recognition rates of independent modules reached up to 100% for 42 postures, 6 orientations, 11 locations and 7 movements using linear classification. The overall sign recognizer was tested using a subset of the American Sign Language dictionary comprised by 30 one-handed signs, achieving 98% accuracy. The system proved to be scalable: when the lexicon was extended to 176 signs and tested without retraining, the accuracy was 95%. This represents an improvement over classification based on Hidden Markov Models and Neural Network.

1. Introduction

American Sign Language (ASL) is the native language of some 300,000 to 500,000 people in North America. It is estimated by Costello [3] that 13 million people, including members of both the deaf and hearing populations, can communicate to some extent in sign language just in the United States, representing the fourth most used language in this country. It is, therefore, appealing to direct efforts toward electronic sign language translators. In addition to the potential commercial application of such translators, sign linguists have interest in the use of automatic means to study signed languages, as Stokoe wrote [17]: "Looking back, it appears that linguistics was made possible by the invention of writing. Looking ahead, it appears that a science of language and communication, both optic (gestures) and acoustic (speech), will be enabled, in all probability, not by refinements in notational systems, but by increasing sophistication in techniques of recording, analyzing, and manipulating visible and auditory events electronically."

Researchers of Human-Computer Interaction (HCI) have proposed and tested some quantitative models for gesture recognition based on measurable parameters [15][4]. Yet, the use of models based on the linguistic structure of signs (Stokoe [17], Lidell [13]) that ease the task of automatic translation of sign language into text or speech is in its early stages. Linguists have proposed different models of gesture from different points of view, but they have not agreed on definitions and models that could help engineers design electronic translators. Existing definitions and models are qualitative and difficult to validate using electronic systems.

As with any other language, differences are common among signers depending on age, experience or geographic location, so the exact execution of a sign varies but the meaning remains. Therefore, any automatic system intended to recognize signs has to be able to classify signs accurately with different 'styles' or 'accents'. Another important challenge that has to be overcome is the fact that signs are already defined and cannot be changed at the researcher's convenience or because of sensor deficiencies. In any case, to balance complexity, training time, and error rate, a trade-off takes place between the signer's freedom and the device's restrictions.

2. Review of previous approaches

Previous approaches have focused on two objectives: the hand alphabet which is used to finger-spell words [5, 7, 10, 11, 18], and complete signs which are formed by dynamic hand movements [1, 16, 19, 20]. So far, body posture and face gesticulation have been left out.

The instruments used to capture hand gestures can be classified in two general groups: video-based and instrumented. The video-based approaches claim to allow the signer to move freely without any instrumentation attached to the body. Trajectory, hand shape and hand locations are tracked and detected by a camera (or an array of cameras). By doing so, the signer is constrained to sign in a closed, some-how controlled environment. The amount of data that has to be processed to extract and track hands in the image

0-7695-2122-3/04 $ 20.00 © 2004 IEEE

IEEE COMPUTER SOCIETY

also imposes a restriction on memory, speed and complexity on the computer equipment.

For instrumented approaches, all sensors are placed on the signer's limbs or joints. Although they might seem restrictive and cumbersome, the approaches based on gloves, such as the Data Entry Glove [5], the CyberGlove [10], the Data Glove [4], and The AcceleGlove [6], have been more successful in recognizing hand shapes than video-based approaches.

To capture the dynamic nature of hand gestures, it is necessary to know the position of the hand at certain intervals of time. For instrumented approaches, gloves are complemented with infra-red, ultrasonic or magnetic trackers to capture movement and hand location with a range of resolution that goes from centimeters (ultrasonic) to millimeters (magnetic). The drawback of these types of trackers is that they force the signer to remain close to the radiant source and inside a controlled environment free of interference (magnetic or luminescent) or interruptions of line of sight.

Mechanical skeletons achieve tracking that is immune to ambient noise by placing angle sensors directly on the signer's joints (wrist, elbow, shoulder). To the best of our knowledge the combination of gloves with skeleton trackers has not been used to capture gestures of ASL.

2.1. Phonetic structure

Selecting the right set of features is the decisive key to avoid ambiguity in a pattern recognition system. Ideally, these features are necessary and sufficient in number and nature to discriminate any pattern in the sample space as a member of one and only one class. Therefore it makes sense to base classification of ASL gestures on features that reflect the phonetic structure of the language.

By using traditional methods of linguistics to isolate segments of ASL, Stokoe [17] found that signs could be broken down into three fundamental constituent parts: the hand shape (*dez*), hand location with respect to the body (*tab*), and the movement of the hand with respect to the body (*sig*), and that these phonemes happen simultaneously. Lidell [13] proposed a model of movements and holds, Sandler [23] proposed movements and locations, and Perlmutter [14] proposed movements and positions, all of them happen sequentially. Under these sequential models, ASL follows the linear structure of spoken languages: phonemes make up words, words in turn make up sentences. It is interesting to note that these phonemes are based, in some degree, on the three simultaneous components of Stokoe, so the ASL structure is a sequential combination of simultaneous phonemes.

Some examples of automatic systems that have followed a model similar to Stokoe are described in [1, 12, 19, 21]. Vogler [20] followed Lidell's model. Starner [16] and Waleed [21] proposed ad-hoc set of features. Along with different models, these approaches also tested several recognition methods such as Hidden Markov Models (HMM) and Neural Networks (NN) to recognize either complete sentences [1, 16], isolated words [12, 20], or phonemes [19]. In these systems, the scalability promised by the phonetic model is compromised by the recognition method.

2.2. The Pose-Movement model

In this section we describe a phonetic model that treats each sign as a sequential execution of two measurable phonemes: one static, and one dynamic.

Definition 1: A ***pose*** is a static phoneme composed of three simultaneous and inseparable components represented by vector **P** = [hand shape, palm orientation, hand location]. The static phoneme occurs at the beginning and at the end of a gesture.

Definition 2: A ***posture*** is a vector of features **Ps** = [hand shape, palm orientation]. Twenty-four out of the 26 letters of the ASL alphabet are postures that keep their meaning regardless of location. The other two letters are not considered postures because they have movement.

Definition 3: ***Movement*** is a dynamic phoneme composed by the shape and direction of the trajectory described by hands when traveling between successive poses. **M**=[direction, trajectory].

Definition 4: A ***manual gesture*** is a sequence of poses and movements, P-M-P.

Definition 5: **L**, the set of purely manual gestures that convey meaning in ASL is called the ***lexicon***.

Definition 6: A manual gesture **s** is called a ***sign*** if s $\in$ L.

Definition 7: ***Signing space*** refers to the physical location where signs take place. This space is located in front of the signer and is limited by a cube bounding the head, back, shoulders and waist.

In this paper a Lexicon of one-handed signs of the type Pose-Movement-Pose is chosen for recognition based on the framework set by these definitions. By doing so, the recognition system is divided into smaller systems trained to recognize a finite number of phonemes, as opposed to training one to recognize an unlimited number of words. Since any word is merely a new combination of the same phonemes, the individual systems do not need to be re-trained when new words are added to the lexicon.

3. System implementation

The capturing system comprises two main elements: an AcceleGlove [6] and a two-link arm skeleton. Sensors and wires of the AcceleGlove were mounted on a leather glove to improve robustness without losing portability; the glove is able to detect hand shapes accurately for different hand sizes. The two-link arm skeleton comprises three components: one dual-axis accelerometer and two resistive angular sensors. One axis of the accelerometer detects arm elevation (θ_1), the second axis detect arm rotation (θ_2), one resistive angular sensor placed on the shoulder measures forearm rotation (θ_4) and the second angular sensor placed on the elbow measures forearm flexion (θ_3). In Figure 1, the shoulder and elbow are modeled as 2-degree of freedom revolute joints. Palm and finger are modeled as telescopic links whose lengths H and I are calculated as the projections of the hand and the index lengths onto the gravitational vector $\boldsymbol{g}$, based on the angle measured by the corresponding accelerometers on the AcceleGlove.

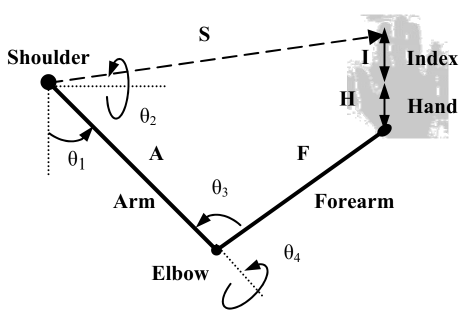


Figure 1. Four angles and four links make up the reduced arm model. H and I are telescopic.

The capturing system is augmented by two push buttons pressed by the user to indicate the beginning and ending of a gesture. Approximately one millisecond is needed to read each accelerometer's axis and resistive sensors by a micro controller PIC16F877 running at 20MHz. One byte per signal is sent via serial port at 9600 baud to a laptop think-pad IBM T-21 with a Pentium III running at 500 Mhz. The program to read the signals and extract the features, discriminate postures, locations, movements and search for the most likely sign, was written in Pascal 1.5 for Windows. The micro controller is connected to a speech synthesizer V8600 'DoubleTalk' from RC Systems which receives the ASCII string of the word corresponding to the recognized gesture.

4. Training and testing

Each module on the recognition system is linked to a part of the capturing hardware; they were trained and tested independently with help of 17 volunteers of different skill levels, from novice to native signer, which provided a range of accents and deviations with respect to the citation form. The complete recognition system was tested on 30 (later 176) one-hand gestures from one signer.

4.1. Palm Orientation

Two accelerometers placed perpendicularly to each other provide three axes of tilt to measure orientation of the palm. Since they react to gravity, only pitch and roll can be measured. The axis to measure 90^0 of pitch runs along the palm parallel to fingers. The other two axes measure 360^0 of roll. All seventeen signers were asked to hold the initial pose of FATHER, NICE, PROUD, PLEASE, THING and ASIDE to capture hand orientations: vertical, horizontal, vertical up-side down, horizontal tilted, horizontal palm up, and horizontal tilted counter clockwise.

The classification algorithm is a decision tree that starts finding vertical, horizontal and up-side down orientations based on hand pitch. The rest of the orientations are found based on hand roll. To test the classifier, all volunteers were asked to perform all the 53 static postures of the extended alphabet [8] fifteen times each.

In average, the orientation module accurately recognized 94.8% of the samples. The worst recognition rate corresponded to horizontal postures where the threshold is blurred by the deviations introduced by signers' accents, since they were asked to hold their poses, not to hold their hand in a certain position.

4.2. Postures

The posture module progressively discriminates postures based on the position of fingers on eight separate decision trees: five corresponding to each orientation plus three trees for the vertical postures divided into vertical-open, vertical-horizontal and vertical-closed based on the position of the index finger [7]. The decision trees are generated as follows:

For all eight trees **do**:

first node discriminates posture based on position of the pinky finger. Subsequent nodes based discrimination on the next finger.

If postures are not discriminated by finger flexion, **then** continue with finger abduction.

Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04)
0-7695-2122-3/04 $ 20.00 © 2004 IEEE

If postures are not different by individual finger flexions or abductions, **then** discriminate by the overall finger flexion and overall finger roll [7].
end.

To set the thresholds on each node, six novice signers were carefully instructed on how to perform the postures, so they are as close as possible to the citation form. Once thresholds were set, the algorithms were tested using new samples from seventeen signers including four of the initial six volunteers.

4.2.1. Aliases

Since accelerometers do not detect angular positions around the gravity vector, 10 postures were impossible to discriminate based on finger bending or spread around the gravity vector. These postures are called aliases. This aliasing reduced the number of recognizable postures from 53 to 43.

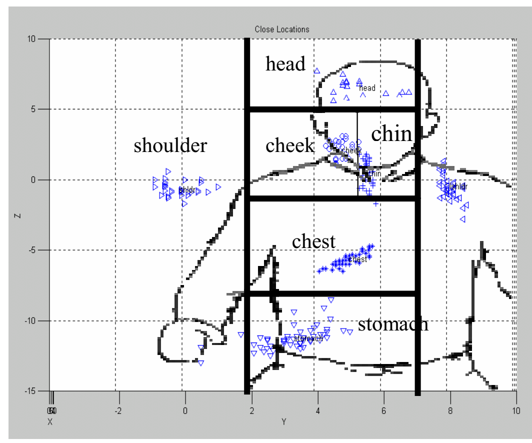
The highest accuracy (100%) corresponds to a vertical palm with knuckles pointing down, which is used to sign PROUD. The worst accuracy rate corresponded to postures C and E, with 68%. The total recognition average for all 43 postures is 84%.

4.3. Locations

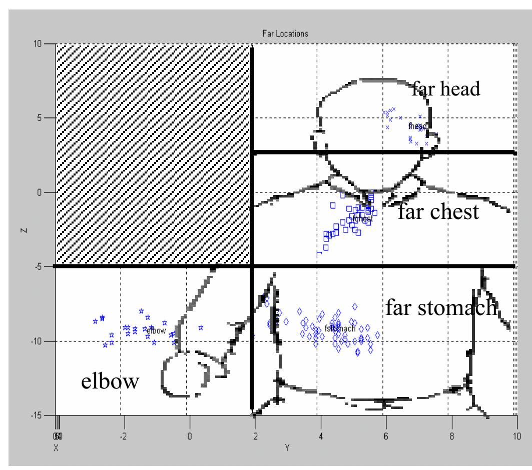
Eleven locations in the signing space were identified as starting and ending positions for the signs in the lexicon composed by one-handed signs: head, cheek, chin, right shoulder, chest, left shoulder, stomach, elbow, far head, far chest and far stomach. Four signers were asked to locate their hand at the initial poses of the following signs: FATHER, KNOW, TOMORROW, WINE, THANK YOU , NOTHING, WHERE, TOILET, PLEASE, SORRY, KING, QUEEN, COFFEE, PROUD, DRINK, GOD, YOU, FRENCH FRIES and THING. From all the signs starting or finishing at the eleven regions, these signs were selected randomly. The signers were selected because their heights represented the extremes and the average in the group of signers: 1.55, 1.82, 1.75 and 1.70 meters.

The coordinates of vector S, in Figure 1, were calculated using values of F=A=10, and H=I=3 that represent upper-arm, arm, hand and finger length's proportions. The sampled points in the signing space are plotted in Figure 2, as executed for the first volunteer who is 1.70 meters in height, Figure 2a corresponds to locations close to the body and Figure 2b corresponds to locations away from the body. A human silhouette is superimposed on the plane to show locations related to signer's body. The plane *y-z* is parallel to the signer's chest, with positive values of *y* running from the right shoulder to the left shoulder, and positive values of *z* above the right shoulder.

Similar to orientations and postures, locations are solved using a decision tree. The first node discriminates between close and far locations; subsequent nodes use thresholds on y and z that bound the eleven regions. Samples from the other three volunteers clustered with a similar distribution, but are shifted either to the right or to the left with respect to samples in Figure 2. On the female subject, a wider gap between chest and stomach was found. In all four subjects it was possible to set the thresholds on y and z at least 4σ around the mean, so that signers of different heights can use the skeleton system if a calibration routine is provided to set the proper thresholds.



(a)



(b)

Figure 2. a) Close locations b) Far locations.

The evaluation of the location module is based on the samples used to train the thresholds. On four signers, the accuracy rate averaged: head 98%, cheek 95.5%, chin 97.5%, shoulder 96.5%, chest 99.5%, left shoulder 98.5%, far chest 99.5%, elbow 94.5 %, stomach, far head and far stomach 100%. The overall accuracy was 98.1%. The advantage of the skeleton system is its portability (it does not need an external source), and its immunity to ambient noise.

4.4. Movements

Movements of the one-handed signs considered in this work are described by means of two movement primitives: shape and direction.

4.4.1. Shapes.

Shapes are classified based on the *curviness* defined by Bevilacqua in [2] as the relation of the total distance traveled divided by the direct distance between ending points. This metric is orientation and scale independent. As with the case of hand shapes and locations, the exact execution of a curve varies from signer to signer and from trial to trial. Thresholds to decide what is straight or circular were set experimentally by computing the mean over several trials performed by the same four signers. A curviness greater than 4 discriminated circles from straight lines with 100% accuracy.

4.4.2. Direction.

Direction is defined as the relative location of the ending pose with respect to the initial pose (up, down, right, left, towards, and away) determined by the maximum displacement between starting and ending locations as follows:

$$\text{Direction} = \max (|\Delta x| ,\ |\Delta y| ,\ |\Delta z|) \qquad (1)$$

where $\Delta x = x_{final} - x_{initial}$, $\Delta y = y_{final} - y_{initial}$, $\Delta z = z_{final} - z_{initial}$; and x, y, z are the coordinates defining hand location.

To evaluate the movement module, the same four signers were asked to perform the six basic movements along the main axes ten times each. Only left and right (77% and 75%) were classified with less than 100% accuracy in all signers. The overall accuracy reached 92%.

4.5. Sign Classifier

To classify complete signs, we used c*onditional template matching*, a variation of template matching. Conditional template matching compares the incoming vector of components (captured with the instrument) with a template (in the lexicon) component by component and stops the comparison when a condition is met:

For the Lexicon **do**:
  extract a list of signs with same initial posture recognized by the corresponding module.
end. This is the first list of candidate signs.
For the list of candidates **do**:
  select the signs with same initial location recognized by the corresponding module.
end. This is the new list of candidate signs.
Repeat the selection and creation of new lists of
  candidates by using movement, final posture and final location.
Until all components have been used **OR** when there is only one sign on the list. That sign on the list is called 'the most likely'.

This search algorithm will stop after finding the initial pose if there is only one sign with such initial pose in the lexicon. In those cases, the probability of finding the sign is equal to P(ip|Xip)P(il|Xil), the product of the conditional probability of recognizing the **i**nitial **p**ose given the input X**ip** from sensors, times the probability of recognizing the **i**nitial **l**ocation given the input X**il**. In the worst-case scenario the accuracy of conditional template matching equals the accuracy of exact template matching when all conditional probabilities are multiplied:

$$P(sign) = P(ip|Xip)\, P(il|Xil)\, P(m|Xm)\, P(fp|Xfp)\, P(fl|Xfl) \qquad (2)$$

where P(m|Xm) is the probability of recognizing the movement given the input Xm, P(fp|Xfp) is the probability of recognizing the final posture, and P(fl|Xfl) is the probability of recognizing the final location given the input Xfl.

5. Evaluation.

To evaluate the search algorithm, a lexicon with only the one handed signs from Starner [16], Vogler [20], and Waldron [22] was created and tested, producing 30 signs: BEAUTIFUL, BLACK, BROWN, DINNER, DON'T LIKE, FATHER, FOOD, GOOD, HE, HUNGRY, I, LIE, LIKE, LOOK, MAN, MOTHER, PILL, RED, SEE, SORRY, STUPID, TAKE, TELEPHONE, THANK YOU, THEY, WATER, WE, WOMAN, YELLOW, and YOU.

To create the lexicon, the PMP sequences are extracted from the citation forms found in Costello [3] and in the Ultimate ASL Dictionary [9] and written in an ASCII file.

This reduced lexicon comprises eighteen postures, two trajectory shapes, and four directions. Almost all of them are identified immediately after recognizing the initial pose. The overall recognition rate was 98%.

5.1. Scalability

By using the conditional template matching to classify signs, the lexicon can be extended as long as the description of new signs is different from the signs already in the lexicon. To prove this statement, the lexicon was expanded to 176 one handed signs taken from Costello's and IDRT's [9] dictionaries, and one signer performed fifteen trials of each. The overall recognition rate on the 176 signs reached 94%.

6. Conclusions and Future Work

By breaking down the hand signs into their constituent phonemes, and facilitating their capture by a modular system, a syntactic classification algorithm to translate gestures of the American Sign Language in a straightforward manner was implemented. The work described in this paper leads to believe that this system is truly lexicon scalable, since retraining was not needed and accuracy was kept high when expanding the vocabulary, which represents the most valuable improvement over previous approaches for translating sign languages.

The combination of resistive and inertial sensors proved to be highly efficient in the two-link skeleton. This combination should be explored to recognize new classes of orientations and hand shapes around the gravity vector impossible to detect with accelerometers. The addition of proven low-cost, low-power wireless technologies will uncover new applications of the recognition system in many other areas of research beyond sign languages for instance, animation, virtual reality, tele-manipulation, rehabilitation, and gaming.

7. References

[1] Bauer, B., Hienz, H., and Kraiss, K., "Video-Based Continuous Sign Language Recognition Using Statistical Methods". IEEE 2000, pp 463-466.

[2] Bevilacqua F., Naugle L., and Valverde I., " Virtual Dance and Music Environment Using Motion Capture". Proc. of the IEEE-Multimedia Technology and Applications Conference, Irvine, CA 2001.

[3] Costelo, Eleine. Random House Webster"s Concise American Sign Language Dictionary. Random House Inc. NY. 1999

[4] Fels, Sidney S., and Hinton, Geoffrey E., "Glove Talk -A Neural-Network Interface Between a Data-Glove and a Speech Synthesizer". IEEE Transactions on Neural Networks, vol. 4, No. 1. January 1993.

[5] Grimes, G.,. US Patent 4,414,537. November 1983.

[6] Hernandez Jose L., Kyriakopoulos, N., Lindeman, R.. "The AcceleGlove a Hole-Hand Input Device for Virtual Reality". ACM SIGGRAPH Conference Abstracts and Applications 2002. pp 259.

[7] Hernandez, Jose L., Kyriakopoulos, N., Lindeman R., "A Multi-Class Pattern Recognition of Practical Finger Spelling Translation", IEEE International Conference on Multimodal Interfaces ICMI'02. October 2002, pp 185-190.

[8] J. L. Hernandez-Rebollar, 'A New Instrumented Approach for translating the American Sign Language into Sound and Text'. D. Sc. Dissertation appendix A, The George Washington University, August 2003.

[9] IDRT Inc. The Ultimate American Sign Language Dictionary. The Institute for Disabilities Research and Training Inc. Copyright 2001.

[10] Kramer, J., and Leifer, L., "The Talking Glove: An Expressive and Receptive "Verbal" Communication Aid for the Deaf, Deaf-Blind, and Nonvocal", SIGCAPH 39, pp.12-15 (spring 1988).

[11] Lamar, Marcus V., Bhuiyan, S., and Iwata, A., "Hand Alphabet Recognition Using Morphological PCA and Neural Networks". Proceedings of the International Joint conference

[12] Liang, R., and Ouhyoung M. "A Real-time Continuous Gesture Recognition System for Sign Language", Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. 558-567.

[13] Lidell S., and Johnson, R., American Sign Language: The phonological base. Sign Language Studies, 64: 195-277, 1989.

[14] Perlmutter, D., "A moisac theory of American Sign Language syllable structure", paper presented at Second Conference on Theoretical Issues in Sign Language Research. Gallaudet University, Washington, DC 1988.

[15] Rubine, The Automatic Recognition of Gestures. Ph.D. Thesis C. Cs. Dept. Carnegie Mellon U. 1991

[16] Starner, T., Weaver, J., and Pentland, A., A Wearable Computer Based American Sign Language Recognizer, MIT Media Lab. Technical Report 425. 1998

[17] Stokoe, William C., Armstrong, David F., Wilcox, Sherman E. Gesture and the Nature of Language. Cambridge University Press, 1995

[18] Uras, C., and Verri, A., "On the Recognition of The Alphabet of the Sign Language through Size Functions". Dipartimento di Fisica, Universitá di Genova. Proceedings of the 12th IAPR Int. Conf. On Pattern Recognition. Conference B: Computer Vision and Image Processing Vol 2,1994. pp 334-338.

[19] Vamplew, P., Recognition of Sign Language Using Neural Networks", Ph.D. Thesis, Department of Computer Science, University of Tasmania, 1996.

[20] Vogler, C. and Metaxas, D., "Toward Scalability in ASL Recognition: Breaking Down Signs into Fonemes". Gesture Workshop"99, Gif-sur-Yvette, France, March 17-19, 1999.

[21] Waleed, M., Grasp: Recognition of Australian sign language using instrumented gloves. Master"s thesis, University of New South Wales, Oct. 1995

[22] Waldron Majula B., "Isolated ASL Sign Recognition System for Deaf Persons", IEEE Trans. On Rehabilitation Engineering vol 3 No. 3 September 1995.

[23] W. Sandler, 'The Spreading hand autosegment of American Sign Language'. Sign Language Studies 50: 1-28. 1986.