Towards a One-Way American Sign Language Translator

R. Martin McGuire[1], Jose Hernandez-Rebollar[2], Thad Starner[1],
Valerie Henderson[1], Helene Brashear[1], and Danielle S. Ross[3]

[1] GVU Center
Georgia Tech
Atlanta, GA 30332
haileris,thad,vlh,brashear
@cc.gatech.edu

[2] Engineering and Applied Science
George Washington University
Washington, DC 20052
jreboll@gwu.edu

[3] Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
psycholing@earthlink.net

Abstract

Inspired by the Defense Advanced Research Projects Agency's (DARPA) recent successes in speech recognition, we introduce a new task for sign language recognition research: a mobile one-way American Sign Language translator. We argue that such a device should be feasible in the next few years, may provide immediate practical benefits for the Deaf community, and leads to a sustainable program of research comparable to early speech recognition efforts. We ground our efforts in a particular scenario, that of a Deaf individual seeking an apartment and discuss the system requirements and our interface for this scenario. Finally, we describe initial recognition results of 94% accuracy on a 141 sign vocabulary signed in phrases of fours signs using a one-handed glove-based system and hidden Markov models (HMMs).

1. Introduction

Twenty–eight million Deaf and hard–of–hearing individuals form the largest disabled group in the United States. Everyday communication with the hearing population poses a major challenge to those with hearing loss. Most hearing people do not know sign language and know very little about deafness in general. For example, most hearing people do not know how to communicate in spoken language with a Deaf or hard–of–hearing person who can speak and read lips (e.g. that they should not turn their head or cover their mouths). Although many Deaf people lead successful and productive lives, overall, this communication barrier can have detrimental effects on many aspects of their lives. Not only can person–to–person communication barriers impede everyday life (e.g. at the bank, post office, or grocery store), but essential information about health, employment, and legal matters is often inaccessible.

Common current options for alternative communication modes include cochlear implants, writing, and interpreters. Cochlear implants are not a viable option for all Deaf people. In fact, only 5.3% of the deaf population in America has a cochlear implant, and of those, 10.1% of these individuals no longer use their implant (complaints cited are similar to those of hearing aides) [2]. The ambiguity of handwriting and slowness of writing makes it a very frustrating mode of communication. Conversational rates (both spoken and signed) range from between 175 to 225 WPM, while handwriting rates range from 15 to 25 WPM [5]. In addition, English is often the Deaf person's second language, American Sign Language (ASL) being their first. Although many Deaf people achieve a high level of proficiency in English, not all Deaf people can communicate well through written language. Since the average Deaf adult reads at approximately a fourth grade level [1, 9], communication through written English can be too slow and often is not preferred.

Interpreters are commonly used within the Deaf community, but interpreters can charge high hourly rates and be awkward in situations where privacy is of high concern, such as at a doctor or lawyer's office. Interpreters for Deaf people with specialized vocabularies, such as a PhD in Mechanical Engineering, can be difficult to find and very expensive. It can also be difficult to find an interpreter in unforeseen emergencies where timely communication is extremely important, such as car accidents.

2. The One-Way Translator

Our goal is to offer a sign recognition system as another choice of augmenting communication between the Deaf and hearing communities. We seek to implement a mobile, self–contained system that a Deaf user could use as a limited interpreter. This wearable system would capture and recognize the Deaf user's signing. The user could then cue the system to generate speech for the hearing listener. However, this idea is complicated with the problem of machine translation of ASL to English. To help constrain the problem, we assume the signer will use Contact Sign.

2.1. Language Modeling

American Sign Language (ASL) grammar is significantly different than English grammar, and many hearing students of ASL have difficulty with its complex features if they learn it after early childhood. Thus, native signers (someone who has learned from birth and is fully fluent) will often use contact signing, which uses many of the grammatical features of English and less of ASL, when encountering hearing signers [11]. By using Contact Sign, we reduce the complexity of the language set we are seeking to recognize, while maintaining a language set that is already familiar to the Deaf community as a tool for when communicating with the hearing.

We choose to further constrain the problem by leveraging the idea of "formulaic" language. Formulaic language is language that is ritualized or prefabricated. It includes routines, idioms, set phrases, rhymes, prayers and proverbs[16]. The DARPA one–way speech translation systems used by peace–keeping troops, maritime law enforcement, and doctors uses this idea to ask questions designed for specific responses. The system provides translations of predetermined phrases designed to provide information or elicit feedback. Informative phrases include sentences like "I am here to help you" and "The doctor will be here soon". Requests and questions include "Please raise your hand if you understand me", "Is anybody hurt?" and "Are you carrying a weapon?"[12]. Requests and questions are limited to those whose answers involve simple gestures, such as nodding yes/no, pointing, or raising a number of fingers (e.g. "How many children do you have?").

Cox describes a system, TESSA, that combines formulaic language with speech recognition and semantic phrase analysis to generate phrases in British Sign Language for Deaf customers at the post office [4]. A set of formulaic language phrases were compiled from observed interactions at the post office. These phrases were then translated into sign and recorded on video. The postal employee speaks to a system that performs speech recognition and uses semantic mapping to choose the most likely phrase. The clerk may say "Interest in the UK is tax free", and the system would cue the phrase "All interest is free of UK income tax" which would then reference the video of a signed translation for the Deaf customer to see.

The use of formulaic language allows for a reduction in vocabulary size and allows for better error handling. Cox showed a progressive decrease in error rates for the language processor, by allowing a user to select from larger N best lists: 1–best was 9.7%, 3–best was 3.8% and 5–best was 2.8% [4]. The application of the phrase selection options also resulted in a significant increase in user satisfaction with the system.

One of the reasons for TESSA's success was its limited domain. After consulting with members of the Deaf community, several scenarios were suggested where the one–way ASL to English translator may be beneficial: doctor's/lawyer's office, emergency situations such as car accidents, navigation in airports, and shopping for an apartment. We chose the last scenario due to its interactive nature and potentially limited vocabulary.

The apartment-hunting scenario is similar to the speech recognition community's Airline Travel Information Service (ATIS) scenario [7] where users would try to solve specific airline travel problems using speech access to a computerized database. Early versions of ATIS were actually "Wizard of Oz" studies where a human would be substituted for the computer to respond to the user's requests. In this way the experimenters could elicit "natural" speech from the subjects to determine what vocabulary should be included in the the actual speech recognition system. Thus, with a vocabulary of a few thousand words tuned to the specific scenario, the ATIS speech recognition system could give the user the impression of a unlimited vocabulary. We intend to perform similar studies with members of the Deaf community to determine the appropriate vocabulary for the apartment-hunting task.

2.2. Interface

In order to begin exploring the feasibility of a one-way translator, we are working on both the interface as well as the recognition components simultaneously. A preliminary interface is necessary to perform Wizard of Oz studies and elicit natural sign in the context of the apartment-hunting task. In addition, the preliminary interface generates useful feedback from the Deaf community.

Figure 1 shows an early prototype of the one-way translator. While the system shown is based on computer vision only (note the camera in the hat), the image demonstrates the head-up display used to provide a visual interface to the user while he signs. An early finding from interacting with the Deaf community is that the display should be mounted on the non-dominant-hand side of the signer to avoid collision during signs made around the face.

Figures 2-5 demonstrate a typical progression of the current interface during translation. Note that the interface is being designed for a hybrid computer vision and accelerometer approach where the signer wears a camera in a hat aimed at his or her hands, as in Figure 1. Thus, a video image from the camera is included in the interface so that the signer knows when the system is successfully tracking

Figure 1. Prototype one-way translator (vision system only shown). The head-up display provides a 640x480 color interface for the signer.

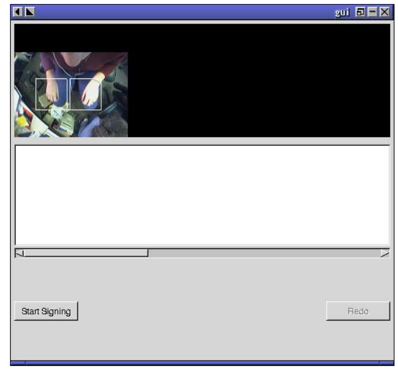

Figure 2. Initial screen for the translator. To start the system, the signer clicks a button mounted on the wrist.

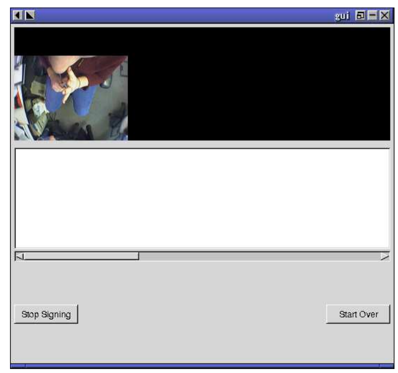

Figure 3. The system collects data as the user signs a phrase.

his hands. Figure 2 shows the initial screen for the translator. To start the system, the signer clicks a button mounted on his wrist. Such an interface may be implemented as part of a Bluetooth enabled wristwatch. At present, the interface is emulated with the buttons of a small optical mouse. As the user signs (Figure 3), the system collects data until the user clicks the wrist button again to indicate the end of the phrase. The user can also click a second button on the wrist to re-start the process. After clicking the stop signing button, the system recognizes the signed phrase, determines the most similar phrases in English from its phrase list, and allows the signer to select between them using a wrist mounted jog-dial (Figure 4). Note that these phrases could be displayed as a series of miniature sign language icons for signers completely unfamiliar with written English. Once the signer selects the closest phrase, the system speaks the phrase, showing its progress in bar as shown in Figure 5. The signer can interrupt the system or repeat the English phrase as desired.

While this interface is preliminary, it has been used for a simple demonstration recognizer combining computer vision and wrist-mounted accelerometers. Testing with native signers is necessary to determine if the system is acceptable to the community and if it can be used to reach conversational speeds. However, initial reaction has been positive.

3. Sign Language Recognition

In the past, we have demonstrated a HMM based sign language recognition system limited to a forty word vocabulary and a controlled lighting environment [13]. The user wore a hat–mounted camera to capture their signing. Data sets were taken in a controlled laboratory environment with standard lighting and background. The images were then processed on a desktop system and recognized in real–time. The system was trained on a 40 word vocabulary consisting of samples of verbs, nouns, adjectives, and pronouns and reached accuracy of 97.8% on an independent test set using a rule–based grammar.

However, this system was more appropriate to laboratory conditions than to a mobile environment. More recently, we have shown that combining accelerometer-based sensing with a computer vision hand-tracking system may lead to better results in the harsh situations typical of mobile sensing [3]. The systems are complementary in that the hat-based vision system tracks the hands in a plane parallel to the ground while the wrist-worn accelerometers, acting as tilt sensors due to the acceleration due to gravity, pro-

Figure 4. The signer selects among potential phrase translations.

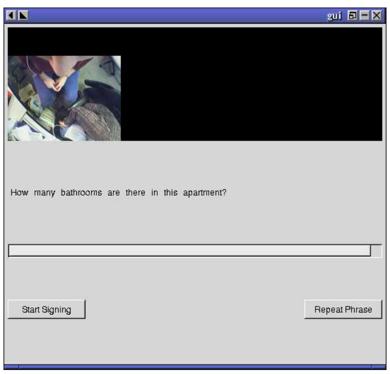

Figure 5. The translator speaks the selected English phrase.

vide information as to the angle of the hands in the vertical plane.

The Acceleglove (see Figure 6) provides another approach to mobile sign recognition. Accelerometers on the individual fingers, wrist, and upper arm provide orientation and acceleration information with respect to each other and potentiometers at the elbow and shoulder provide information as to the hand's absolute position with respect to the body. In previous work [8], the Acceleglove system was shown to recognize 176 signs in isolation using decision trees. Many signs are taught with a beginning hand shape, a movement, and an ending hand shape. With the Acceleglove system, the user makes the initial hand shape, and the recognizer shows which signs correspond to that hand shape. The system eliminates signs interactively as the user proceeds with the movement and end hand shape.

In this paper, we combine the Acceleglove hardware with the Georgia Tech Gesture Toolkit (GT2K) [15] to attempt

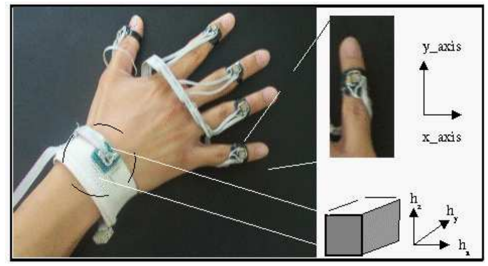

Figure 6. The Acceleglove. Five micro two-axis accelerometers mounted on rings read finger flexion. Two more in the back of the palm measure orientation. Not shown are two potentiometers which measure bend at the shoulder and elbow and another two-axis accelerometer which measures the upper arm angles.

phrase-level recognition with a 141 sign vocabulary. Our goal is to prove the feasibility of a phrase level ASL one-way translator using a mobile apparatus. A high word accuracy in a continuous sign recognition task with this system would suggest that a mobile phrase level translator is possible.

4. Recognition Experiment

Acquiring data with which to train our system began with choosing a set of signs to recognize. Since the Acceleglove was already part of an existing recognition system, a subset of signs was chosen from the list of signs understood by the original system. These signs were then organized into parts of speech groups of noun, pronoun, adjective, and verb for a total of 141 signs. Using a fairly rigid grammar of "noun/pronoun verb adjective noun", a list of 665 sentences was generated, ensuring that each sign appeared in the data at least 10 times.

To capture the sign data, the original Acceleglove recognition program was altered to include user prompts and to log the data from the glove's sensors. The signer sat in front of the capturing computer at a fixed distance, wearing the glove on his right arm, and holding in his left a pair of buttons attached to the glove, with both arms on the armrests of the chair. The program displayed the sentence to be signed, and when the signer was ready, he would press one of the buttons to begin capture. At the end of the sentence, the signer would return his arms to the armrests, and press the other button to signify that the sentence had ended. The computer would then save the captured data to a numbered file, and increment to the next sentence in the list. This pro-

Grammar	Testing on training	Indep. test set
part–of–speech	98.05%	94.47%
unrestricted	94.19%	87.63%

Table 1. Sign accuracies based on a part–of–speech and an unrestricted grammar.

cess was repeated for all 665 sentences, with a camera filming the process to aid in the identifying incorrect signs.

Training of the HMM-based recognizer was done with GT2K [15]. After filtering the data to account for irregular frame–rates from the glove, the data was labeled using the sentence list. To minimize the impact of the signer's arms beginning and ending at the chair armrests, the "signs" start–sentence and end–sentence were added to the recognition list. A pair of grammars was created. The first follows the same parts of speech based form used to generate the sentence list, surrounded by the start–sentence and end-sentence signs. The second was a more unrestricted grammar, looking only for the start–sentence sign, followed by any number of any of the signs in the vocabulary, followed by the end–sentence sign. A set of training and testing sets were created using a randomly selected 90% of the data for training and the remaining 10% for model validation. The model was then trained with the automatic trainer. Sign boundaries were re-estimated over several iterations to ensure better training. After training, the models were tested against the remaining 10% of the sentences. Recognition accuracy was determined, with the standard penalties for substitutions, insertions, and deletions. This process of training and testing was repeated 21 times, yielding an overall accuracy based on the average of each of the 21 sets. The models created were each tested with both the strict and the unrestricted grammars, resulting in accuracy ratings of 94.47% average for the strict grammar and 87.63% average for the unrestricted. An additional model was created, using all of the data for training and all of the data again for testing. Accuracy ratings for this testing–on–training model were 94.19% for the unrestricted grammar, 98.05% for the strict (see Table 1).

Accuracy was determined following the standard speech recognition formula of

$$Acc = \frac{(N - D - S - I)}{N} \times 100\%$$

where N is the number of signs, D is the number of deletions, I is the number of insertions, and S is the number of substitutions. Note that only substitutions are possible with the strict part–of–speech grammar.

5. Discussion and Future Work

The results above are very promising. HMM recognition systems tend to scale logarithmically with the size of the vocabulary in both speed and accuracy. Since ASL has approximately 6000 commonly used signs, the pattern recognition framework of this project should be scalable to the larger task. In addition, there is a significant amount of work in the spoken language community for applying context, grammar, and natural language frameworks to HMM recognizers. Hopefully, this prior work will allow rapid adoption of such framework for ASL. Even so, ASL is significantly different from spoken language in that it allows spatial memory and spatial comparisons. In addition, face and body gestures communicate a significant amount of information in ASL. Thus, we expect this field to be a challenge for many years in the future.

The results suggest the feasibility of our goal of a phrase level translator. A phrase level translator with the interface described previously has significant tolerance to individual sign errors. The system simply needs to recognize enough of the signs so that the closest phrase is returned in the top few choices for the user to select. Hopefully, the planned Wizard of Oz studies will show that a relatively low number of phrases and signs are necessary to handle most situations in the apartment-hunting scenario. However, as soon as this sign and phrase lexicon is gathered, the system will be tested against these phrases as signed by multiple native signers. These experiments will help to further refine the models as well as to produce an initial automatic recognition system that can be used to test the translator *in situ.* As the system improves, more signers can experiment with the system, and a larger corpus develops. If the system is successful with the apartment-hunting scenario, we hope to expand the translator's scope to other scenarios of concern to the Deaf community. In this way we hope to copy the model of the speech recognition community which leveraged its success in the ATIS task to more difficult scenarios such as taking dictation.

We also expect to experiment with the translator apparatus. We wish to combine the Acceleglove system with a computer vision system to determine which features and signs each system can best recognize. In addition, we will discuss the various wearable hardware options with members of the Deaf community to determine which options are most desirable. Finally, we will improve the translator interface as we gain more experience in making practical systems with the community.

6. Related Work

Following a similar path to early speech recognition, many early attempts at machine sign language recognition concentrated on isolated signs or fingerspelling. Space does not permit a thorough review, but, in general, most current systems rely on instrumented gloves or a desktop-based camera system. Before 1995 most systems employed a form of template matching or neural nets for recognition; however, many current systems now employ HMMs.

In 1998, Liang and Ouhyoung's work in Taiwanese Sign Language [10] showed very encouraging results with a glove-based recognizer. This HMM-based system recognizes 51 postures, 8 orientations, and 8 motion primitives. When combined, these constituents can form a lexicon of 250 words which can be continuously recognized in real-time with 90.5% accuracy. At ICCV'98, Vogler and Metaxas described a desk-based 3D camera system that achieves 89.9% word accuracy on a 53 word lexicon [14]. Since the vision process is computationally expensive in this implementation, an electromagnetic tracker is used interchangeably with the three mutually orthogonal calibrated cameras for collecting experimental data. Most recently at ICMI2003, Fang, Gao, and Zhao discussed a decision tree and SOFM/HMM recognizer that achieved approximately 91% accuracy on isolated Chinese Sign Language signs in a 5113-sign vocabulary [6]. This system is restricted to desk-based use as it employs two fully instrumented Cybergloves and three Polhemus six degree of freedom trackers.

7. Conclusion

We have introduced the concept of a one-way sign to spoken translator as a suitable task for sign language recognition research. We have demonstrated a preliminary system for phrase-level sign recognition with a per sign accuracy of 94%, suggesting that sufficiently large vocabularies and high enough accuracies are possible for a mobile one-way American Sign Language to English translator in a limited domain. Finally, we have discussed an interface design for the one-way translator that puts the signer-in-the-loop for selecting and verifying the appropriate phrase translation.

Acknowledgments

The authors would like to thank John Makhoul for his discussions which led to the one-way translator task.

References

[1] Stanford achievement test, 9th edition, form s. *Norms Booklet for Deaf and Hard-of-Hearing Students*, 1996. Gallaudet Research Institute.

[2] Regions regional and national summary report of data from the 1999-2000 annual survery of deaf and hard of hearing children and youth. Technical report, Washington, D. C., January 2001. GRI, Gallaudet University.

[3] H. Brashear, T. Starner, P. Lukowicz, and H. Junker. Using multiple sensors for mobile sign language recognition. In *ISWC*, pages 45–53, White Plains, NY, 2003.

[4] Stephen J. Cox. Speech and language processing for a constrained speech translation system. In *Proc. Int.. Conf.. on Spoken Language Processing*, Denver, Co., September 2002.

[5] J. Darragh and I. Witten. *The Reactive Keyboard*. Cambridge Series on Human-Computer Interaction. Cambridge University Press, Cambridge, 1992.

[6] Gaolin Fang, Wen Gao, and Debin Zhao. Large vocabulary sign language recognition based on hierarchical decision trees. In *Intl. Conf. on Multimodal Interfaces*, Vancouver, BC, November 2003.

[7] C. T. Hemphill, J. J. Godfrey, and Doddington G. R. The atis spoken language systems pilot corpus.

[8] J. Hernandez-Rebollar. *A New Instrumented Approach for Translating the American Sign Language into Sound and Text*. PhD thesis, George Washington University, Washington, DC, August 2003.

[9] Judith Holt, Carol B. Traxler, and Thomas E. Allen. Interpreting the scores: A user's guide to the 9th edition stanford achievement test for educators of deaf and hard-of-hearing students. *Gallaudet Research Institute Technical Report*, 91(1), 1997. Gallaudet Research Institute.

[10] R. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Third International Conference on Automatic Face and Gesture Recognition*, pages 558–565, Nara, Japan, 1998.

[11] C. Lucas and C. Valli. *Sign language research: Theoretical issues*, chapter ASL, English, and contact signing, pages 288–307. Gallaudet University Press, Washington DC, 1990.

[12] Ace Sarich. DARPA one-way phrase translation system (PTS). http://www.sarich.com/translator/.

[13] T. Starner, J. Weaver, and A. Pentland. Real-time American Sign Language recognition using desk and wearable computer-based video. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 20(12), December 1998.

[14] C. Vogler and D. Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *ICCV*, Bombay, 1998.

[15] T. Westeyn, H. Brashear, A. Atrash, and T. Starner. Georgia tech gesture toolkit: Supporting experiments in gesture recognition. In *Intl. Conf. on Multimodal Interfaces*, 2003.

[16] A. Wray and M. Perkins. The functions of formulaic language: an integrated model. *Language & Communication*, 1(20):1–28, 2000.