LEARNING THE BASIC UNITS IN AMERICAN SIGN LANGUAGE USING DISCRIMINATIVE SEGMENTAL FEATURE SELECTION

Pei Yin, Thad Starner, Harley Hamilton, Irfan Essa, James M. Rehg

School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

The natural language for most deaf signers in the United States is American Sign Language (ASL). ASL has internal structure like spoken languages, and ASL linguists have introduced several phonemic models. The study of ASL phonemes is not only interesting to linguists, but also useful for scalability in recognition by machines. Since machine perception is different than human perception, this paper learns the basic units for ASL directly from data. Comparing with previous studies, our approach computes a set of data-driven units (fenemes) discriminatively from the results of segmental feature selection. The learning iterates the following two steps: first apply discriminative feature selection segmentally to the signs, and then tie the most similar temporal segments to re-train. Intuitively, the sign parts indistinguishable to machines are merged to form basic units, which we call ASL fenemes. Experiments on publicly available ASL recognition data show that the extracted data-driven fenemes are meaningful, and recognition using those fenemes achieves improved accuracy at reduced model complexity.

Index Terms— Machine Learning, American Sign Language, Feature Selection

1. INTRODUCTION AND RELATED WORKS

The natural language for most deaf signers in the United States is American Sign Language (ASL). ASL sentences are composed of signs, such as UNCLE, EAT, SAD, with a quite different grammar than English. In machine perception, American Sign Language Recognition (ASLR) algorithms infer those signs from the sensor readings, such as a video stream of the signer's hand movements. Comparing with general gesture recognition, ASLR is more structured and constrained. Since Stokoe's pioneering work [1] in 1960 demonstrated that ASL is compositional with an internal structure like spoken languages, various signal processing and machine learning techniques successful in speech recognition have been applied to ASLR. Although a decomposition of the language helps to achieve scalability in large vocabulary ASLR as in speech recognition, ASLR research has "not yet exploited the results of determining the appropriate basic units" [2].

ASL linguists have proposed several phonemic models since the 1960s, such as the Stokoe system [1], the Movement-Hold model [3], and the Hand-Tier model [4]. The Stokoe system describes a sign as one simultaneous bundle of three major formational categories: handshape, locations, and movements. This model is inadequate to capture sequential internal segments [5]. In order to represent both simultaneous and sequential phonemic contrast [6] in ASL, the Movement-Hold model describes ASL by two types of sequentially ordered segments: movement segments and hold (location) segments. Each segment is then defined by a simultaneous bundle of descriptors, such as handshape and location. The third model commonly used for ASL is the Hand Tier model [4], which organizes locations and movements sequentially as the Movement-Hold model and typically makes handshape characterize the entire sequence. This approach is designed to remove certain duplications in the Movement-Hold model, as well as to introduce benefits in the morphological representation. For ASLR, we adopt the idea of the Movement-Hold model, because its straightforward correspondence to the hidden Markov models (HMMs). In fact, most research groups [7, 8, 9, 10, 11, 12] use HMMs for sign language recognition.

While the Movement-Hold model provides a powerful linguistic tool for researchers analyzing ASL, the "conceptual" descriptors (such as handshape) used by the Movement-Hold model may not be available to the machines. One may suggest a two-step recognition method: first recognize those conceptual descriptors from sensor readings and then apply the phonological rules explicitly. However, such a method usually yields inferior performance in practice due to variance in signing, disfluencies [13], error accumulation, and many other factors. Alternatively, Vogler and Metaxas (VM) [9] have shown that low-level features can directly fit phonemic models of ASL. However, their manual transcription from the Movement-Hold model to HMMs will be infeasible for a large vocabulary size. In speech processing, an alternative to phonemes as basic units is called fenemes [14, 15], which are extracted directly from the acoustic features (such as cepstral coefficients) by clustering. Bauer and Kraiss (BK) [10] adopt such data-driven fenemic model based on k-means and HMMs. In applications such as ASLR, in which the "good" features are unknown, we believe that such generative models will be less accurate than discriminative models. We illustrate this phenomenon with a synthetic example in Section 2. The main differences between the aforementioned two papers and ours are summarized in Table 1. This paper learns a set of ASL fenemes from discriminative feature analysis. The intuition is that if two or more sign parts (a subsequence of a sign) are indistinguishable to machines, they can be merged to form a subword building block, which we call ASL fenemes. These data-driven fenemes are shared among the signs. The sharing not only reduces model complexity but also helps avoid over-training when certain discrimination is not necessary (explained in Section 2.2). To our knowledge, this is the first attempt to extract ASL fenemes from discriminative feature analysis. We describe our feneme extraction algorithm next.

	Phoneme/Feneme Extraction	Extraction Criterion	Sign Language
VM [9]	Manual	N/A	ASL
BK [10]	Data-driven	Generative	GSL
This paper	Data-driven	Discriminative	ASL

Table 1. Differences in sign language phoneme/feneme extraction between our approach and those in VM [9] and BK [10]. GSL: German Sign Language.

2. FROM DISCRIMINATIVE FEATURE SELECTION TO DISCRIMINATIVE FENEME EXTRACTION

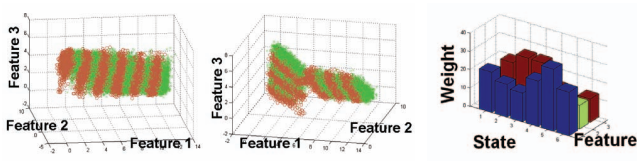
2.1. Segmentally-Boosted HMMs

The Segmentally-Boosted HMM (SBHMM) [12] is a discriminative feature selection algorithm for sequence classification problems [16] in which "good" features are unknown, such as ASLR and lip reading. SBHMMs first conduct standard HMM training on time sequences (signings). Inspired by the Segmental K-means Segmentation (SKS) [17], the time sequences are then segmented into states using the Viterbi algorithm. According to the Markovian assumption and the conditional independence assumption by HMMs, the samples of the same states are independent and identically distributed (i.i.d.). Therefore, feature selection algorithms, which usually require the data to be i.i.d., can be applied to those segments to compute state-dependent discriminative features. In SBHMMs, discriminative features that separate those states are extracted by multiclass boosting algorithm [18]. The outputs of the multiclass boosting ensembles span a discriminative feature space, in which traditional HMMs are then re-trained. SBHMMs have an advantage over previous discriminative feature analysis algorithms in two major aspects: (1) the features are evaluated *segmentally* (comparing with the "global" feature selection techniques [19, 20]); and (2) the new feature space is formed by a nonlinear projection computed by large margin classifiers (comparing with Tandem models [21]). Experiments in Yin *et al.* [12] have shown that SBHMMs reduce the test error of traditional HMMs by 17% to 70% in continuous ASLR, gait recognition, lip reading, and simple speech recognition tasks. SBHMMs also compare favorably to other discriminative learning techniques such as Conditional Maximum Likelihood (CML) [12].

2.2. Learning From Unsuccessful Separations

The multiclass boosting in SBHMMs searches for features that separate the samples of one state from those of any other states in the same sign and in the other signs. As indicated by the Movement-Hold model, many segments (states) are shared between different signs. For example, the sign WIFE [1] is composed of FEMALE, a transition move, and MARRIAGE; while HUSBAND is composed of MALE, a transition move, and MARRIAGE. If WIFE and HUSBAND are represented by three-state HMMs, their last states will be identical (in a noise-free situation). In such a case, multiclass boosting will not be able to find features that separate these two states. This "failure" in state separation does not necessarily affect the sequence recognition accuracy for SBHMMs, however [2]. In fact, we can use this failure as evidence that the two states should be tied. In this paper, we propose to take advantage of such tying to actively extract building blocks (fenemes) of signs.

Phonemes are the smallest contrastive units in a language, which can be illustrated in minimal pairs [22]. A minimal pair in ASL is defined as two signs bearing different meanings but are identical except for one formational aspect, such as location. In order to extract basic units for ASL, a learning model must be "sensitive" enough to represent the contrast in minimal pairs. In the synthetic example in Fig. 1, two types of three-dimensional time sequences moving from left to right are marked by orange and green. Assuming each dimension is a feature, both HMMs and SBHMMs with six states achieve 100% accuracy in the "easy"(clean) case of Fig. 1(a). However, when different states have different informative dimensions/features as in Fig. 1(b), HMMs, being generative, achieve a testing error of 10.1%. Meanwhile, SBHMMs, being segmentally discriminative, achieve an error rate of 2.3%. Fig. 1(c) shows the feature importance for each state are correctly estimated by SBHMMs.



(a) (b) (c)

Fig. 1. A synthetic example of segmental discriminative features. (a) "clean," (b) "twisted," and (c) SBHMMs assign highest weight to feature 3 in the first three states and feature 1 in the last three states for one HMM. This corresponds to the switch of the informative features. The feature weight for the other HMM is similarly assigned.

Next, we examine the behavior of SBHMMs with ASL minimal pairs in a publicly available ASLR dataset [23] [3]. This dataset contains 665 phrases with 141 distinct signs (classes). The inputs are readings of 17 accelerometers mounted on hand, wrist, and shoulder in Table 2. We found that SBHMMs corrected three instances of misclassification made by HMMs with the minimum pair BROTHER and SISTER (as signed in the dataset). The sign BROTHER and the sign SISTER are illustrated in Fig. 2. The only difference in the two signs is the starting posture of the hand, which should relate to the accelerometer for the the wrist orientation (number 15) [4]. Indeed, the feature weight computed by the SBHMMs in Fig. 3(a) and (c) shows that the readings of the accelerometer for the wrist orientation are considered moderately important (the red dots at column 15) when we use all the 141 different signs for training.

To show that SBHMMs are sensitive to the impact of the minimum pairs, we have conducted training using only the samples of BROTHER and SISTER, and then using all the words including SISTER but excluding BROTHER. The feature weight are plotted in Fig. 3(b), (d), and (e). We can see that when using only the minimum pair of BROTHER and SISTER to train, the weights of the wrist orientation at the first state (the red dots at column 15) are significantly higher in Fig. 3(b) and (d) than those in Fig. 3(a) and (c). This difference illustrates that segmental feature selection is sensitive to such phonemic contrast. In addition, when training with the 140 types of signs without BROTHER, the weight of the wrist orientation at the first state (the red dot at column 15) for SISTER is reduced in Fig. 3(e) comparing with Fig. 3(a) and (c), because of the relaxed competition in classification. Note that the wrist orientation may help discriminate SISTER from other signs, so its weight does not necessarily vanish when BROTHER is absent. Another experiment with the minimal pair IGNORANT and MISUNDERSTANDING also proves that SBHMMs are able to disclose the phonemic contrast of the minimal pairs.

[1] In this paper, we use capitalized words to refer to ASL signing. Due to space limitations, we are unable to provide visual illustrations of all the signs mentioned in this paper. The video of most signs are available from online ASL dictionaries.

[2] For the samples from the identical states, their feature values are still similar in the new feature space. Therefore, those samples little side effect on the sequence recognition results.

[3] http://wiki.cc.gatech.edu/ccg/projects/asl/asl

[4] The location of the hand is actually a better feature to discriminate the two signs. However, such information is not available to accelerometers.

feature index	meaning
1 and 2	thumb outside
3 and 4	thumb top (on thumbnail)
5 and 6	index finger
7 and 8	middle finger
9 and 10	ring finger
11 and 12	pinkness
13	wrist perpendicular to bones
14	wrist parallel to fingers
15	wrist perpendicular to palm
16	shoulder elevation (forward)
17	shoulder (outward)

Table 2. The meaning of the 17 accelerometer readings

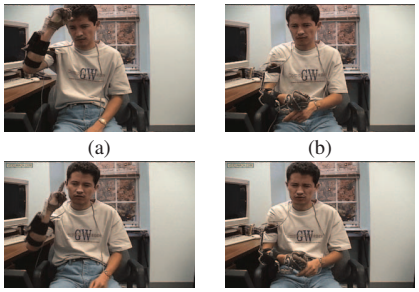


Fig. 2. Illustration of the formation of the minimal pair BROTHER and SISTER. (a)(b) BROTHER and (c)(d) SISTER.

2.3. The State Tying Paradigm

When two segments are indistinguishable to discriminative classifiers, they probably come from the same feneme. Based on this intuition, we apply state tying [15], which is successful in improving efficiency and scalability in speech recognition, to the inseparable states detected by SBHMMs. The state tying procedure can be executed in two ways: top-down or bottom-up. One major problem of the top-down approach is how to revise the state transition matrix to include the new states. The second way for state tying is the bottom-up approach. It starts with a fine-scale representation and sequentially clusters the states that are similar. Bottom-up state tying has been widely used in practical speech recognition systems due to its simplicity [15]. We adopt the bottom-up approach in this paper.

Algorithm 1 The DIST algorithm

1: randomly initialize the HMMs
2: run SBHMMs to select state-dependent discriminative features.
3: **while** NOT (a pre-set lower bound P number of fenemes are extracted OR no state pair exceeds θ OR a pre-set number of rounds I is reached) **do**
4: extract the confusion matrix in separating the states.
5: use the Houtgast algorithm to compute the similarity of each state pairs from the confusion matrix.
6: merge the top m most similar state pairs, or the state pairs whose similarity is above a threshold θ.
7: run SBHMMs to select state-dependent discriminative features in the new state space.
8: **end while**
9: train and test in the new feature space computed by SBHMMs.

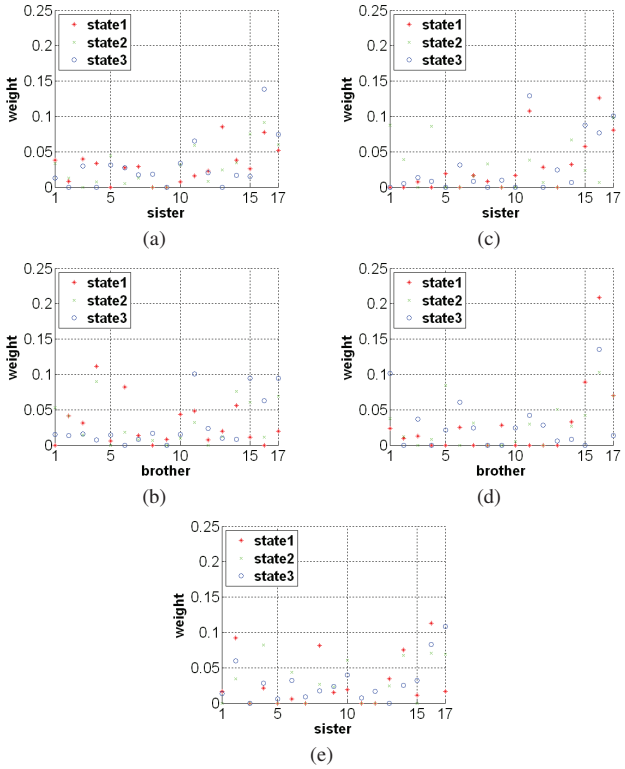


Fig. 3. The impact of minimum pairs on feature weighting obtained by SBHMMs. (a) and (b): the feature weighting for the signs SISTER and BROTHER, trained with all the 141 classes; (c) and (d): the feature weighting for the signs SISTER and BROTHER, trained with only those two signs; (e): the feature weighting for the sign SISTER, trained with all the classes except BROTHER.

2.4. DIscriminative State-space Tying (DIST)

In order to perform state tying, we need to compute the state similarities from the confusion matrix produced by SBHMMs. One conventional method for this conversion is the Houtgast algorithm [24] $s_{ij} = \sum_k^n min(c_{ik}, c_{jk})$, in which s_{ij} is the similarity score, and c_{ij} is the $(i,j)^{th}$ element in a $n \times n$ confusion matrix C. The flowchart of the DIscriminative State-space Tying (DIST) is in Algorithm 1.

3. EXPERIMENTAL VALIDATION

We have tested our DIST-SBHMMs algorithm on the accelerometer-based ASLR dataset [23]. Following the convention in Yin, *et al.* [12], we use one three-state HMMs to model each sign. We randomly split the ASL phrases, using 90% for training and the other 10% for testing. The resulting confusion matrix determines the state to be tied. The new reduced models (fewer states) are then re-trained. For simplification, DIST-SBHMMs only iterate once ($I = 1$) in our prototype, and only the top $m = 10$ most confusable state pairs are tied [5]. After state tying, the recognition error is reduced by 9% (from

[5]The value of m was arbitrarily set in this paper. We actually have manually examined the top 15 state pairs. Those state pairs are all correct according to ASL linguists.

Rank	Sign Name	State Number	Sign Name	State Number	Reason
01	DAUGHTER	3	SON	3	BABY
02	WIFE	3	HUSBAND	3	MARRIAGE
03	MORNING	1	THING	1	stretching arm
04	APOLOGIZE	1	APOLOGIZE	2	a repetitive pattern
05	WATER	1	WINE	1	W
06	ROOSTER	1	ROOSTER	2	a repetitive pattern
07	TOILET	1	TOILET	2	a repetitive pattern
08	SICK	1	SMART	2	folding middle finger
09	SUSPECT	2	PUZZLE	2	forehead touch
10	USE	1	USE	2	a repetitive pattern

Table 3. The top 10 similar states computed by DIST-SBHMMs in one of our tests

4.10% to 3.73%), which we believe is due to less overfitting.

In order to test the consistency of our feneme extraction algorithm, we have run the experiments four times, and the top 10 most confusable pairs have 60% overlap (24 out of 40). On the one hand, the 60% overlap illustrates that the perceptually meaningful fenemes are not obtained by chance. It proves that consistent fenemic representation for machine perception can be infered from data. On the other hand, the 40% difference suggests that more interesting patterns (fenemes) may be extracted with higher diversity in data. The top 10 most similar state pairs in one of our four runs are reported in Table 3. The discovered similarities and the extracted fenemes are considered perceptually meaningful by our sign linguist.

4. CONCLUSION AND FUTURE WORK

In this paper, we have presented preliminary results for DIST-SBHMMs, which extract data-driven ASL fenemes from segmental discriminative feature selection. Experimental results suggest that DIST decompose ASL into perceptual meaningful pieces (fenemes) that can be used to reduce model complexity without sacrificing model accuracy. However, DIST-HMMs may produce unnecessary tying due to lack of "good" features to distinguish certain signing segments. Our discriminative model is also sensitive to "bad signing." In the future, we plan to execute more tests of DIST-SBHMMs to observe how ASL fenemes are extracted iteratively and whether such data driven method can discover fenemes that are unknown to ASL linguistics. We also plan to extend this discriminative temporal structure learning technique to less-constrained gesture recognition applications.

5. REFERENCES

[1] W. C. Stokoe, "Sign language structure: An outline of the visual communication systems of the American deaf," *Studies in Lingusitics*, vol. 8, 1960.

[2] G. T. Holt, P. Hendriks, and T. Andringa, "Why don't you see what I mean? Prospects and limitations of current automatic sign recognition research," *Sign Language Studies*, vol. 6, pp. 416–437, 2006.

[3] S. K. Liddell and R. E. Johnson, "American Sign Language: The phonological base," *Sign Language Studies*, vol. 64, pp. 195–277, 1989.

[4] W. Sandler, "The spreading hand autosegment of American Sign Language," *Sign Langauge Studies*, vol. 50, pp. 1–28, 1986.

[5] T. Supalla and E. Newport, "How many seats in a chair? The derivation of nouns and verbs in American Sign Language," in *Understanding Language Through Sign Language Research*, Siple, Ed., pp. 91–132. New York:Academic Press, 1978.

[6] S. K. Liddell, "Think and believe: Sequentiality in American Sign Language," *Language*, vol. 60, no. 2, pp. 372—399, 1984.

[7] T. Starner, J. Weaver, and A. Pentland, "Real-time American Sign Language recognition using desk and wearable computer based video," *IEEE Trans. on PAMI*, vol. 20, no. 12, pp. 1371–1375, 1998.

[8] R. H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in *Proc. of the International Conference on Face & Gesture Recognition*, 1998, pp. 558–567.

[9] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of American Sign Language," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 358–384, 2001.

[10] B. Bauer and K. Kraiss, "Towards an automatic sign language recognition system using subunits," *LNAI*, vol. 2298, pp. 64–75, 2002.

[11] Y. Chen, "Chinese sign language recognition and synthesis," in *IEEE International workshop on AMFG*, 2003.

[12] P. Yin, I. Essa, T. Starner, and J. M. Rehg, "Discriminative feature selection for hidden Markov models using segmental boosting," in *Proc. of ICASSP*, 2008.

[13] W. Sandler and D. Lillo-Martin, *Sign Language and Linguistic Universals*, Cambridge University Press, 2006.

[14] Frederick Jelinek, *Statistical methods for speech recognition*, MIT Press, Cambridge, MA, USA, 1997.

[15] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, Printice Hall, 1993.

[16] Thomas G. Dietterich, "Machine learning for sequential data: A review," *LNCS*, vol. 2396, pp. 15–30, 2002.

[17] B. Juang and L. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust. Sp. Sign. Process*, vol. 38, pp. 1639–1641, 1990.

[18] Y. Freund and R. Schapire, "A decision theoretic generalization of on-line learning and application to boosting," *Journal of Computer and System Science*, vol. 55(1), pp. 119–139, 1995.

[19] P. Yin, I. Essa, and J. Rehg, "Asymmetrically boosted HMM for speech reading," in *Proc. of CVPR*, 2004, pp. II:755–761.

[20] P. Smith and M. Shah N. Lobo, "Temporalboost for event recognition," in *Proc. of ICCV*, 2005, pp. 733–740.

[21] H. Hermansky, D. P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of ICASSP*, 2000, pp. 1635–1638.

[22] C. Valli and C. Lucas, *Linguistics of American Sign Language: An introduction*, Gallaudet University Press, third edition, 2000.

[23] R. M. McGuire, J. Hernandez-Rebollar, T. Starner, V. Henderson, H. Brashear, and D.S. Ross, "Towards a one-way american sign language translator," in *Proc. of the International Conference on Face & Gesture Recognition*, 2004, pp. 620–625.

[24] O. Anderson, P. Dalsgaard, and W. Barry, "On the use of data-driven clustering technique for identification of poly and mono-phonemes for four European languages," in *Proc. of ICASSP*, 1994, pp. 121–124.