Unsupervised Learning of Prosodic Boundaries in ASL

Joshua Falk and Diane Brentari, University of Chicago

In both spoken and sign languages, prosodic cues signal the ends of intonational phrases (Nespor & Sandler 1999). Children must somehow learn to associate these cues with phrase boundaries without explicitly being told where those boundaries are. In this paper, we present two unsupervised statistical models that learn to identify the ends of intonational phrases (I-phrases) in American Sign Language (ASL) based on prosodic cues: a mixture model, and a hidden Markov model. Although neither model is presented with labeled phrase boundaries, both achieve performance comparable to models that are trained with labeled boundaries. The success of these models sheds light on how infants might learn the prosodic system without explicit instruction.

The data for this study comes from narratives by four adult native signers of ASL. Signers were asked to describe the Sylvester and Tweety cartoon 'Canary Row,' and their narratives were recorded. Previous work has shown that phrase-final signs in ASL are generally longer, and that they are more likely to co-occur with non-manual cues, such as eye blinks, changes in brow position, and recalibrations of the head and body (Wilbur 1994, Nespor & Sandler 1999, Brentari et al. 2011). For each narrative, each sign was annotated in ELAN for these prosodic cues, and I-phrase boundaries were annotated by three proficient signers (90% agreement, with disagreements resolved by discussion). Durations were converted to z-scores.

The first model is a mixture model with two latent components. This model assumes that signs come from one of two categories, which we hope will ultimately reflect the presence or absence of an I-phrase boundary. The mixing probability represents how likely each category is. Each category has its own distribution over the prosodic cues. Sign duration is drawn from a normal distribution, with separate mean and variance for each category. The presence of each non-manual cue is drawn from a Bernoulli distribution that also depends on the category. The parameters for the unsupervised model were estimated using the Expectation-Maximization (EM) algorithm (Bishop 2006). The parameters for the supervised model are maximum likelihood estimates.

		Probability	Duration	Variance	Head	Body	Blink	Brow
unsuper.	medial	0.63	-0.37	0.54	0.19	0.13	0.27	0.09
	final	0.37	1.07	1.23	0.60	0.40	0.61	0.36
supervised	medial	0.81	-0.14	0.75	0.27	0.18	0.32	0.14
	final	0.19	0.58	1.62	0.43	0.27	0.49	0.24

Table 1: Parameters for mixture models

Comparing the parameter estimates, we see broad agreement between the unsupervised and supervised values. The unsupervised model correctly identified one higher frequency category with shorter signs and fewer non-manual markers (medial), and a lower frequency category with longer signs and more non-manual markers (final). However, the unsupervised model finds more extreme differences between the two categories.

We also compare how well the two models predict phrase boundaries. In building a system to predict boundaries, we want to simultaneously maximize precision (the percentage of predicted boundaries that are actual boundaries) and recall (the percentage of actual boundaries that the model predicts). Interestingly, while the supervised model achieves higher precision, the unsupervised model achieves significantly higher recall.

	Precision	Recall	F1
unsupervised	0.33	0.57	0.42
supervised	0.45	0.21	0.29

Table 2: Performance of mixture models

The second model is a hidden Markov model with two latent states. This model adds a dependence between adjacent states: the probability of observing one state depends on the identity of the previous state. For example, if a phrase-final sign has been observed, the following sign is very unlikely to be phrase-final. Aside from this added structure, the model is identical to the mixture model discussed above. For the hidden Markov model, parameters were estimated by the Baum-Welch algorithm (Bishop 2006).

Interestingly, the parameters recovered by the hidden Markov model are identical to the parameters recovered by the unsupervised mixture model above. Looking at the transition probabilities explains this result. For the unsupervised model, a medial sign is equally likely to occur after a medial sign or a final sign (64% vs. 59%). The model has learned to essentially ignore the dependence. This differs from the supervised model, which knows that a final sign is very unlikely to occur after another final sign (1%).

	m $\rightarrow$ m	m $\rightarrow$ f	f $\rightarrow$ m	f$\rightarrow$f
unsupervised	0.64	0.34	0.59	0.41
supervised	0.77	0.23	0.99	0.01

Table 3: Transition probabilities for Markov models

We also compare the predictive performance of the unsupervised and supervised Markov models. For both models, the predictions are given by the most probable sequence of states for the observations. This was computed by the Viterbi algorithm (Bishop 2006). Both models have very similar predictive performance. The supervised model still has higher precision and lower recall, but the difference is much smaller.

	Precision	Recall	F1
unsupervised	0.33	0.57	0.42
supervised	0.37	0.49	0.42

Table 4: Performance of hidden Markov models

We have offered the first unsupervised systems to learn the prosodic phrasing of a sign language. Both unsupervised systems recover reasonable parameters and have predictive performance on par with supervised systems, but adding a dependency between adjacent states does not improve performance. However, the overall predictive performance of even the supervised system is surprisingly low. One possibile explanation is the high rate of constructed action in the narratives, which may show similar cues to phrase-final signs. Another possibility is that additional cues or more complex combinations of known cues contribute to prosodic phrasing. While further work on I-phrase final cues in ASL is necessary, the success of such simple systems at determining cue distributions sheds light on how infants might learn the prosodic system without explicit instruction.

Bishop, Christopher. 2007 *Pattern Recognition and Machine Learning*. New York: Springer.

Brentari, D., C. González, A. Seidl, & R. Wilbur. 2011. Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech* 54(1). 49–72.

Nespor, M., & W. Sandler. 1999. Prosody in Israeli Sign Language. *Language and Speech* 42. 143–176.

Wilbur, R. 1994. Eyeblinks and ASL phrase structure. *Sign Language Studies* 84. 221–240.