Unsupervised Learning of Prosodic Boundaries in ASL

Joshua Falk and Diane Brentari

1 Introduction

In both spoken and sign languages, prosodic cues signal the ends of intonational phrases (Nespor and 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.

2 Background

Most work on prosodic boundary detection has focused on supervised learning, where the learning algorithm has access to information about where prosodic boundaries are located (Wightman and Ostendorf, 1994; Liu et al., 2008). A primary exception comes from Ananthakrishnan and Narayanan (2006), who attempted to automatically identify accents and phrase boundaries without labeled data. They used the Boston University Radio News Corpus (Ostendorf et al., 1995), an English language corpus which includes 7 speakers and approximately 3 hours of data. For learning, they used three techniques: k-means cluster, fuzzy k-means clustering, and a Gaussian mixture model, all with two categories (non-final and final). They used features based on intensity, F0, and timing, as well as part-of-speech tags. They report 87.3% accuracy under a supervised model, and between 77.6% accuracy and 81.1% accuracy for their unsupervised models. In this paper, we report the results of applying similar models to American Sign Language. To our knowledge, this represents the first case of unsupervised learning of prosody in a sign language.

3 Data

The data for this study comes from narratives by four adult native signers of ASL (Brentari et al., 2015). 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 and Sandler, 1999; Brentari et al., ~~2001~~).

Each cue was annotated by a separate coder: duration, head position, body position, blinks, and brow position. Their annotations were then checked by either a hearing native signer or a Deaf early learner of ASL. Annotations concerning prosodic constituency were completed independently by three proficient signers. Two had learned ASL at seven and eight years of age, respectively, and one was a full-time certified interpreter. The annotators achieved 90% agreement on I-phrase boundaries, and disagreements were resolved after discussion. For modeling purposes, durations were converted to z-scores (centered around the mean and set to have unit variance).

4 Mixture Model

The first model we built is a mixture model with two latent components. This is comparable to the Guassian mixture model used by Ananthakrishnan and Narayanan (2006). 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 model is defined by several parameters that generate the

distribution over cues observed in the data. First, there is the mixing probability, which represents how likely each category is. Each category then 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, where the probability of a each cue occuring depends on the cue.

The probability of the data given the parameters is called the likelihood, and a common technique for model fitting tries to find the parameters that maximize the likelihood. This method, the method of maximum likelihood, is strongly associated with the statistician Ronald Fisher, and has several theoretical justifications (Stigler, 2007). In many situations, the maximum likelihood estimate matches a "common-sense" estimator. For example, the maximum likelihood estimate for the mean is just the sample mean.

In order to find maximum likelihood parameters for the model, we used the Expectation-Maximization (EM) algorithm (Bishop, 2006). The EM algorithm is an iterative algorithm that provably converges to a local maximum of the likelihood, but not necessarily the global maximum. Running the EM algorithm starting from several random initializations of the parameters gave the same parameter estimates, suggesting that the recovered parameters represent a global maximum in the likelihood. The parameters for the supervised model are also maximum likelihood estimates, but with the categories known.

		Probability	Duration	Variance	Head	Body	Blink	Brow
unsuper.	non-final	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	non-final	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 (non-final), 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. We use a Bayes classifier for prediction, choosing the category that is more likely to have generated the observed data. 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. F1 is defined as $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$, and provides a way of combining precision and recall into an overall measure of predictive performance. Note that the predictive performance of the supervised model is likely subject to overfitting, as the model was trained and tested using the same data.

	Precision	Recall	F1
unsupervised	0.33	0.57	0.42
supervised	0.45	0.21	0.29

Table 2: Performance of mixture models.

Both models exhibit much lower performance than that reported in Ananthakrishnan and Narayanan (2006), with 80.5% accuracy for the Gaussian mixture model and 87.3% accuracy for the supervised model. In the discussion section below, we discuss possible reasons for this difference.

5 Markov Model

One limitation of the Gaussian mixture model is that it has no awareness of the temporal ordering of signs. It cannot capture the fact that, if a phrase-final sign has been observed, the following sign is very unlikely to be phrase-final. The second model overcomes this limitation by adding a dependence between adjacent states or categories: the probability of a category occurring depends on the category of the previous sign. This is known as a Markov model. Aside from this added structure, the model is identical to the mixture model discussed above. When the states or categories are unknown, the model is known as a hidden Markov model. For the hidden Markov model, parameters were estimated by the Baum-Welch algorithm (Bishop 2006), which is a type of EM algorithm like the one used to fit the Gaussian mixture model. The Baum-Welch algorithm also only guarantees that the likelihood of the parameters is a local maximum. Again, several random initializations resulted in the same parameter estimates, suggesting that the recovered parameters represent a global maximum in the likelihood.

	Duration	Variance	Head	Body	Blink	Brow
non-final	-0.38	0.54	0.19	0.13	0.26	0.09
final	1.05	1.23	0.60	0.40	0.61	0.36

Table 3: Parameters for unsupervised hidden Markov model.

Interestingly, the parameters recovered by the hidden Markov model are essentially identical to the parameters recovered by the unsupervised mixture model above. Looking at the transition probabilities explains this result. For the unsupervised model, a non-final sign is equally likely to occur after a non-final sign or a final sign (64% vs. 59%). The model has learned to essentially ignore the dependence we introduced. This differs from the supervised model, which knows that a final sign is very unlikely to occur after another final sign (1%). This occurs only when an entire intonational phrase consists of only one sign.

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

Table 4: 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 5: Performance of hidden Markov models.

6 Discussion

Overall, the models trained without labels learn reasonable parameters for the distributions over cues. This shows that there is a strong enough correlation to support learning even in the absence of labeling. However, the hidden Markov model does not recover good parameters for transitions between states, harming the performance of the model. This suggets that, at least for these cues, there is not much evidence supporting the lack of adjacent final signs.

The accuracy and recall of even the supervised systems are surprisingly low relative to the performance reported in Ananthakrishnan and Narayanan (2006). There are several factors that could contribute to this poor performance. There are significant differences in the data used in each study. Ananthakrishnan and Narayanan (2006) use radio data, which is more carefully and deliberately produced. This may enhance phonetic cues for prosodic phrasing. In contrast, the data in our study is a spontaneous narrative, which may have more disfluencies, as well as less careful production. This may mask some of the phonetic cues for prosodic phrasing. The 'Canary Row' also video elicits a large amount of constructed action, which may exhibit similar phonetic properties to phrase-final signs, such as lengthening.

Additionally, Ananthakrishnan and Narayanan (2006) also include part-of-speech tags in their classifier. Their classifier has to use the prosodic cues to in effect bootstrap the information contained in the part-of-speech tags, but this still represents a significant additional source of information lacking in our study. Unfortunately, Ananthakrishnan and Narayanan (2006) do not report the performance of the model without POS information, making it difficult to assess the contribution of this difference.

Finally, the cues used by Ananthakrishnan and Narayanan (2006) are not the same as the cues used in this study, largely because of modality differences. It is possible that additional cues, including cues that have previously gone unnoticed in the literature, play an important role in ASL phrasing. Furthermore, even within the cues used here, there may be additional information being lost. For example, the duration or intensity of eyeblinks could vary depending on whether a blink occurs on a phrase-final sign.

7 Future Steps

From a modeling perspective, there are several steps that could be taken to improve the performance of the models. Hierarchical models can better capture individual variation. Instead of drawing cues directly from a distribution over cues, the model could draw parameters for cue distributions for each signer from a distribution over parameters (Gelman et al., 2013). This type of hierarchical modeling captures similarities between signers while also capturing consistent individual variation. The end result is improved parameter estimates and a better sense of the scope of individual variation. With better accounting for individual variation, the low probability of a final-final sequence might become apparent.

In the supervised setting, one option is to train a more sophisticated classifier that can assign different weights to different cues or account for interactions. Previous research on identifying prosodic boundaries has shown success with these types of models (Wightman and Ostendorf, 1994). A downside to this approach is that it is hard to compare the supervised and unsupervised models when their structures diverge.

Given the data from Ananthakrishnan and Narayanan (2006), it would not be difficult to run their model without part-of-speech tags. This would make it possible to eliminate one of the confounds discussed above in comparing unsupervised learning of prosody in English and ASL. It would also be interesting to introduce the between state dependence of the hidden Markov model and see how it affects the performance on their data set.

If the low probability of final-final sequences is discovered and the supervised models still outperform the ASL models, a potential future study would be to create comparable data sets in English and ASL that control for content and style. If even on these data sets the differences persist, this would suggest that richer ASL cues than the ones coded in this study are needed. If instead the differences disappear, it would suggest that the register and style differences discussed previously are responsible for the difference in performance.

Finally, using a similar paradigm to Brentari et al. (~~2001~~), it would be possible to test native ASL signers to determine their performance at identifying phrase-final versus non-final signs from the data set used in our study. Such an experiment would provide a better point of comparison for evaluating the predictive performance of any statistical model.

8 Conclusion

We have offered the first unsupervised systems to learn aspects of prosodic phrasing in 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. Furthermore, the overall predictive performance of even the supervised system is lower than what has previously been reported for similar models applied to English. One possibile explanation is differences between the data sets, such as the more formal production of radio speech or the high rate of constructed action in the narratives, which may show similar cues to phrase-final signs. Another confound comes from the additional part-of-speech information available in previous research on English but unavailable in our study. Perhaps the most exciting possibility is that additional cues or more complex combinations of known cues can mark 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.

References

Ananthakrishnan, Sankaranarayanan, and Shrikanth Narayanan. 2006. Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling. In *INTERSPEECH*, 829–832.

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Brentari, Diane, Joshua Falk, and George Wolford. 2015. The acquisition of prosody in American Sign Language. *Language* 91:e144–e168.

Brentari, Diane, Carolina González, Amanda Seidl, and Ronnie Wilbur. ~~2001.~~ Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech* 54:49–72.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis, Third Edition*. Taylor & Francis.

Liu, Fangzhou, Huibin Jia, and Jianhua Tao. 2008. A maximum entropy based hierarchical model for automatic prosodic boundary labeling in Mandarin. In *ISCSLP 08. 6th International Symposium on Chinese Spoken Language Processing.*, 1–4. IEEE.

Nespor, Marina, and Wendy Sandler. 1999. Prosody in Israeli Sign Language. *Language and Speech* 42:143–176.

Ostendorf, Mari, Patti J. Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University Radio News Corpus. *Linguistic Data Consortium* 1–19.

Stigler, Stephen M. 2007. The epic story of maximum likelihood. *Statistical Science* 598–620.

Wightman, Colin W., and Mari Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on speech and audio processing* 2:469–481.

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