Finding the Gaps: Applying a Connectionist Model of Word Segmentation to Noisy Phone-Recognized Speech Data

C. Anton Rytting

Department of Linguistics
The Ohio State University, Columbus, Ohio, U.S.A.
rytting@ling.ohio-state.edu

Abstract

The Christiansen model of word segmentation [1] is a connectionist framework for modeling how infants combine multiple cues in learning and processing language. Most studies applying this model assume idealized input with adult-like representations of phonemes and features, with little or no degradation of the input signal. From these studies, it is difficult to tell if the model is robust to non-idealized, noisy input, which may correspond more closely to an infant language-learner's experience.

This study tests the robustness of the Christiansen model by providing input from a minimally-trained phone recognizer on infant-directed speech. Some degradation of performance is observed, but the model still performs above chance. This finding represents a first step in developing more realistic input representations for models of child language acquisition.

Index Terms: word segmentation, child-directed speech (CDS), computational models of language acquisition

1. Introduction

The acquisition of a language is an immense task. Not only are languages complex and greatly varied, but much of the structure underlying language competence cannot be observed directly, but must be inferred from indirect cues. An example of linguistic structure only indirectly observed is that of the organization of speech sounds into constituent units such as words or phrases. Unlike written English, where the boundaries between words are neatly marked by spaces, spoken language has no completely reliable cues for word boundaries. Even when an underlying vocabulary of possible words is known (as for adult native speakers or for most ASR systems), enough ambiguities arise to make the problem non-trivial. For infants who must find the boundaries in order to acquire the vocabulary in the first place, the task is much harder. However, human speech, and particularly infant-directed speech (IDS), contains a number of statistical cues which point to these boundaries. No single cue is perfect, but when taken together the cues reinforce each other and narrow the search space to manageable proportions. (See e.g. [2, 3] for reviews of experimental findings.)

A number of earlier models of the word segmentation task focused on understanding the role of single, specific cues (see [4] for a review). In contrast, the model proposed in [1] focuses on the interaction between different cues—specifically, distributional information over the segments, utterance boundaries, and lexical stress. This paper examines the Christiansen model's performance on input derived from actual IDS recordings, passed through a minimally-trained phone recognition system.

2. The Christiansen et al. (1998) model

Rather than approaching the detection of word boundaries directly, Christiansen et al. [1] differentiate between the *primary task* (learning the language), *immediate tasks* (updating one's statistical knowledge of the language's directly observable aspects), and *derived tasks* (inferring the non-observable structure of the language). Examples of immediate tasks include learning to predict what sound will come next (see e.g. [5]) or how soon the current utterance will end [6].

Christiansen et al. [1] (henceforth CAS98) treat word segmentation as a derived task, and do not train their network on it directly. Rather, they train an simple recurrent network (SRN) to learn three immediate prediction tasks: the next phone's identity, its level of stress, and whether the utterance is about to end. In order to train the network to learn the relationships between these tasks, the network is trained on the three tasks simultaneously. The network's word boundary predictions are derived from the activation level of the utterance boundary predictor. The results of this model on the Korman corpus [7] are given in Table 1, below:

Training	Boundary		Word	
Condition	Prec.	Rec.	Prec.	Rec.
phon-ubm-stress	70.16	73.71	42.71	44.87
phon-ubm	65.86	71.34	37.31	40.40
stress-ubm	40.91	87.69	8.41	18.02
utterances as words	100.00	32.95	30.79	10.15
pseudo-random	33.40	33.15	8.62	8.56

Table 1: Percent precision and recall for the three nets in CAS98, for an algorithm that treats utterances as words, and for a pseudo-random algorithm using mean word length (from [1], Table 3)

2.1. Limitations of the Christiansen et al. (1998) model

The CAS98 model articulates a plausible explanation for how children may combine cues of limited provenance in order to learn word boundaries with greater accuracy than they could with any single cue or heuristic. However, it has generally been tested with clean input, where observable cues used for detecting word boundaries (i.e., phone identities, level of stress, and utterance boundary locations) have been given to the system with a high degree of confidence and correctness. Their main study ([1]) assumes perfect accuracy and confidence. It transforms a word-level transcription of the mothers' utterances into string of phones derived from the pronunciations of the words listed in the MRC lexicon. Each phone is represented as a vector of binary phonological features, with each

feature presented with an activation of either 0 or 1. Hence, any variation in the actual speech signal below the level of the word is abstracted away from the input presented to the SRN.

One reason for this abstraction was due to necessity: at the time of their study, no phonetically-transcribed corpus of infant-directed speech was available, and the sound recordings of such corpora as the Korman corpus were too poor in quality to be usable. Nevertheless, a parallel study, [8] addresses the issue of natural variation in speech. This study uses as input the Carterette Corpus [9], which provides phone-level transcription of speech between adults. Moreover, some small amount of artificial noise is added to the activation levels of certain "peripheral" features of each phone (defined as those features whose change would not result in another phoneme found in the language). No phones suffered deletion, insertion, or substitution with another canonical phone except as recorded in the human-transcribed corpus.

Thus, [8] approximates the speech of adults as perceived by trained transcribers. By adding another (albeit controlled and artificial) layer of random noise, it also deals to a limited degree with the issue of variability in speech. It does not deal with the issue of a child's *perception* of speech. Rather, it assumes, as many studies have done before, that children perceive speech as adults do, hearing each phone in sequence as the speaker uttered it (or at least as an adult transcriber heard it). Some (e.g., [10, 3]) have questioned this assumption. Unfortunately, since it is difficult to know precisely what children do hear, the most that can be done in a simulation is another approximation. The current paper, while leaving some questions of representation to future work, reports on an approximation of input that preserves much more of the variability inherent to the speech signal that infants hear.

3. The Christiansen model on noisy data

In order to simplify comparisons with the original Christiansen model, the simulations reported here maintain the assumption that speech is represented as a string of segments drawn from the same phonemic inventory as adults, and encoded with the same phonological features. It does not, however, assume a uniform, canonical mapping from word to features as in [1], nor a mapping posited by adult transcribers as in [8]. Rather, it uses the output of an automatic phone recognizer. A similar approach was tested in [11], who showed that his algorithm could yield interesting results even very noisy input data. Since his simulations also were conducted before high-quality sound recordings of IDS were available, he had to conduct his studies on the TIMIT corpus, a commonly-used corpus of read speech from various American English dialects.

3.1. Materials

Since the time of the previous studies, Brent and Siskind [12] have developed and made available a new corpus to the CHILDES databank [13] that includes downloadable sound files with word-level transcriptions.[1]. The Brent corpus provides fourteen 90-minute sessions for each of eight American English-speaking mothers living in Baltimore who participated in the study, spaced at roughly two week intervals at 8-14 months of the baby's age. For each of the mothers, the middle 75 minutes of the earliest sessions (typically three or four) were transcribed at the word level.

The mothers' voices were recorded using portable DAT recorders and lapel-mounted microphones placed on the mother. The recordings took place in the families' home, in order to capture typical utterances of everyday life; in order to focus on speech directed at the child, the mothers were asked to avoid phone conversations with other adults. The mothers' utterances were defined as stretches of speech demarcated on either side by at least 300 ms of silence; time stamps for utterance boundaries are marked in the word-level transcriptions. (See [12] for more details.)

For the study reported here, three of the eight mothers' voices were used. In order to keep the input focused on as young as infants as possible, only the first few recordings from these mothers were used. The very first recording for each dyad was excluded, since the mother may have been more self-conscious due to the novelty of the microphone; recordings 2-5 from mothers "c1", "f1", and "f2" were used. The infants ranged between 9 and 10 months of age at the time of these recordings. These twelve sessions contain a total of 8285 utterances.

As with the Korman Corpus, no phone-based transcriptions were available, so these had to be created. Two versions were created: a canonical reference transcription created in the same manner as in [1], by replacing each word with the word's canonical pronunciation, and a noisy input created by a phone recognizer as described below.

3.2. Method

In order to create the input for this simulation, the SONIC Speech Recognizer [14, 15] was used as the basis for a phone recognition system. A version of the CMU dictionary adapted to the SONIC phone set was used as base dictionary, and words in the transcripts not found in this dictionary were added to it, with the pronunciation checked off the original sound files as necessary. To form the canonical input, each instance of a given word was transcribed using the first pronunciation of the dictionary. These pronunciations were then mapped from SONIC's phone set to the 36-phone MRC phoneset used in [1], to facilitate comparisons with that study.

To create the "noisy" input, each of the sound files was segmented and resampled with a 8 kHz sampling rate. SONIC's default feature extraction system, Perceptual Minimum Variance Distortionless Response (PMVDR) cepstral coefficients [16], along with an off-the-shelf acoustic model for female speech was used. No adaptation or re-training of the acoustic model was done, except for on-line adaptation. A triphone language model was used, taken from 90% of the utterances used in the study; however relatively little weight was given to the model, under the assumption that the child would have only limited knowledge of transition frequencies between phones, and would have to rely more on the raw acoustics. Obviously, no dictionary was given to the recognizer, as that would defeat the purpose of the pre-lexical word segmentation task (which of course is to find the word boundaries so as to facilitate the acquisition of a lexicon). In place of a dictionary, SONIC was given the same 55-to-36 phone mapping that was used for the canonical phones, with each of the 36 phones treated as a word entry. Settings were adjusted to produce roughly as many phone insertions as phone deletions.

Due to occasional lapses in recording quality (most likely due to variation in the mother's volume, head position, background noise, and other factors incident to out-of-studio recordings), certain utterances failed to be recognized by SONIC (which gave either null output or recognized a single non-continuous, non-sonorant phone like [t]). Such utterances were excluded from both the canonical and noisy input sets. Of the 8285 utterances, 992

[1] Available from http://childes.psy.cmu.edu/data/Eng-USA/brent.zip (transcripts); http://childes.psy.cmu.edu/media/Eng-USA/Brent/ (sound).

were excluded, leaving 7293 utterances.

Even with these problematic utterances removed, recognition was, as may be expected, extremely noisy: the correctness measure was 43% and the accuracy was 23%. As a result of this noise, each of the 1576 word types in the canonical input had on average five different "noisy" realizations. Naturally, such noisy input makes for a very demanding test of any word segmentation model's abilities. It may serve more as a lower bound of what we might reasonably expect from a given model. We would expect the model's true performance to fall somewhere between the noisy and the canonical trials tested here.

Both the noisy and canonical input sets were divided 90%-10% into training and test sets, with 6564 utterances in the training set and 729 utterances in the test set. The canonical sets had 77952 and 8903 segments in the training and test sets, respectively; the noisy sets 77303 and 8866 segments. All utterance-internal pauses were deleted from both the canonical and noisy input strings, and a pause symbol was inserted if missing at the end of each utterance (symbolizing the 300ms pause after the end of the sound file).

In evaluating performance of the word segmentation model for the case of noisy input, one may wonder where word boundaries ought to be placed in the gold standard. For this study, the noisy transcriptions were previously aligned with the corresponding canonical transcriptions for those same utterances in order to evaluate the performance of the recognizer. These same alignments were used to assign the gold-standard word boundaries for the noisy transcriptions. In each noisy transcription, a word boundary was placed before each segment aligned with a word-initial segment in the canonical transcription of that utterance.

3.3. Training and testing the model

Two versions of the CAS98 model were trained and tested: the original 1998 version, which encoded the MRC phoneset into a vector of binary 11 phonological features, and a slightly modified version referenced in [17], using 17 features which are arguably more in line with those used in theoretical phonology. Stress information was not included in this study; methods of automatically detecting stress (or equivalent) information from the acoustic signal (and learning the role it plays) is left for future work. Hence, the most appropriate comparison to CAS98 is the *phon-ubm* condition rather than the *phon-ubm-stress* condition.

As with the models in [1, 17], the SRNs were trained to predict the upcoming phone's identity rather than its features; hence each SRN has 37 output units (one for each phone plus one for the utterance boundary symbol). Eighty hidden units and eighty context units were used. Since the utterance boundary symbol also has its own separate input node, the 11-feature model results in a 12-80-37 SRN, and the 17-feature model in an 18-80-37 SRN. Each of the two variant models was trained on one iteration of the training corpus, and then tested on the test corpus. Five iterations of training and testing were performed (with different initial starting weights), and the results of the iterations averaged, on each of the two transcriptions (canonical and noisy).

4. Results

4.1. Activation levels

As mentioned above, the CAS98 model works by training the net on clearly observable utterance boundaries while simultaneously performing a phoneme-prediction task, and relies on it to generalize the activation for the utterance boundary output unit to word boundaries as well. Hence, a higher average activation on word boundaries (both utterance medial and utterance final) relative to word-internal positions indicates the SRN's ability to learn this generalization. As in [1], the SRNs successfully learned both to predict upcoming utterance boundaries and to generalize this prediction to utterance-internal word boundaries. The average activations for word-medial, word-final, and utterance-final positions are shown in Table 2.

Training Condition	Avg. Activation		
	Word-medial	Word-final	Utt-final
canon-11	0.024	0.080	0.123
canon-17	0.027	0.093	0.125
noisy-11	0.032	0.082	0.120
noisy-17	0.035	0.083	0.119

Table 2: Average activations for the four nets trained with the utterance boundary cue

One-tailed t-tests on the four sets of five iterations show the average activation of the utterance boundary unit is significantly higher in word word-boundary positions than in word-medial positions ($p < 0.001$).

4.2. Boundary detection and word extraction

In order to translate these continuous activation levels into boolean boundary predictions, it is necessary to set a threshold on the activation. In this section, we follow [1], in positing the mean activation level over all segments as the threshold. Any activation above this mean is treated as a posited word boundary.

A more crucial measure of a word-boundary detector's ability to help with vocabulary acquisition is its ability to find words within the running speech. This requires a higher level of accuracy than mere boundary detection: in order to successfully segment an instance of a word, the learner must find both the initial and final boundaries of the word, and refrain from any false-positive boundaries within the word. Precision and recall measures for both of boundaries and words are shown in Table 3.

Training Condition	Boundary		Word	
	Prec.	Rec.	Prec.	Rec.
canon-11	53.28	57.56	20.18	21.92
canon-17	57.66	66.34	25.20	29.00
noisy-11	46.36	54.30	14.56	17.04
noisy-17	44.16	62.30	13.48	18.98

Table 3: Percent precision and recall for the four nets trained with the utterance boundary cue (as in Table 1) over the subset of the Brent Corpus

Clearly, the models trained on noisy transcriptions do not perform as well as those with the canonical transcriptions. For the canonical transcriptions, the larger, linguistically better-motivated 17-feature set performs better, but for the noisy condition, representation seems to make little if any difference.

As Table 3 shows, the models here (and particularly those trained on the noisy transcriptions) suffer from over-segmentation due to too-low thresholds. Accordingly, the results are supplemented with percentage figure for the AUROC (area under the ROC curve). For these measures, an AUROC of 0.5 would be expected for at-chance performance.

As with the figures above, once again the 17-feature encoding of the canonical transcription performs the best, with an AUROC of 0.832. The 11-feature canonical transcription had an AUROC of 0.778, and the two noisy inputs had AUROCs of 0.7338 (11-feature) and 0.7376 (17-feature). The differences between the canonical and noisy representations are significant (two-tailed t-test: $p < 0.005$), as is the difference between the 11- and 17-feature canonical inputs ($p = 0.00506$).

5. Discussion

As stated above, the noisy output from the SONIC-based phone recognizer is exceedingly noisy, and as such may be said to represent a lower bound on what children would be expected to recognize, insofar as the assumption holds that they have access to the same phonological representations as adults (or linguists). Hence, in spite of the significantly degraded performance of the SRNs on noisy input, the fact that it is still able to learn at all is encouraging. It is quite possible that the cue of lexical stress (or rather, its acoustic counterparts), which was not examined here, may prove even more of a crucial aid to word segmentation in when the phonological input is noisy than when it is clean (as in [1]). This question will have to remain for future study.

6. Conclusions and future work

The CAS98 model provides one example of a connectionist approach to the word segmentation task faced by infants at early prelexical stages of language acquisition. It addresses an interesting question relevant to psycholinguistic studies of the word segmentation task: the integration of multiple cues such as lexical stress, transitional probabilities over segments, and word boundary information, all known from psycholinguistic experimentation to play a role in infants' detection of word boundaries.

However, until now it, along with many studies, have finessed to a greater or lesser degree the issue of the lower-level input on which the model rests. This present study constitutes one step in bridging the gap between raw acoustic input and the input representation usually assumed. The results for the noisy input condition are noticeably worse than those for canonical, idealized input. However, they still manages to learn to generalize utterance boundary information to word boundaries (as shown by the difference in mean activation level and the AUROC measures), suggesting that the SRN does not degrade catastrophically even in the face of exceptionally noisy input.

This work represents a stage in on-going work to relax the current assumptions governing connectionist models of the word segmentation task and extending them to qualitatively different models of the acoustic input available to infants. Future steps include developing methods of automatically extracting acoustic correlates of lexical stress, as well as phonological features and utterance boundaries from these audio files. Later steps include combining the CAS98 model with unsupervised methods of acquiring phonemic inventories from the sound signal in order to have a model of language acquisition truly from the ground up.

7. References

[1] Morten Christiansen, Joseph Allen, and Mark Seidenberg, "Learning to segment speech using multiple cues: A connectionist model," *Language and Cognitive Processes*, vol. 13 (2/3), pp. 221–268, 1998.

[2] Peter W. Jusczyk, "How infants begin to extract words from speech," *Trends in Cognitive Sciences*, vol. 3, no. 9, pp. 323–328, September 1999.

[3] Janet F. Werker and Suzanne Curtin, "Primir: A developmental framework of infant speech processing," *Language Learning and Development*, vol. 1(2), pp. 197–234, 2005.

[4] Michael R. Brent, "Speech segmentation and word discovery: A computational perspective," *Trends in Cognitive Sciences*, vol. 3 (8), pp. 294–301, 1999.

[5] Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport, "Statistical cues in language acquisition: Word segmentation by infants," in *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, G.W. Cottrell, Ed., Hillsdale, NJ, 1996, pp. 376–380, Lawrence Erlbaum Associates.

[6] Richard N. Aslin, Julide Z. Woodward, Nicholas P. LaMendola, and Thomas G. Bever, "Models of word segmentation in fluent maternal speech to infants," pp. 117–134. Lawrence Erlbaum Associates, Mahwah, NJ, 1996.

[7] Myron Korman, "Adaptive aspects of maternal vocalizations in differing contexts at ten weeks," *First Language*, vol. 5, pp. 44–45, 1984.

[8] Morten H. Christiansen and Joseph Allen, "Coping with variation in speech segmentation," in *Proceedings of GALA*, A. Sorace, C. Heycock, and R. Shillcock, Eds., 1997.

[9] E. C. Carterette and M. H. Jones, *Informal Speech: Alphabetic and Phonemic texts with statistical analyses and tables*, University of California Press, Berkeley, CA, 1974.

[10] Mary E. Beckman and Jan Edwards, "The ontogeny of phonological categories and the primacy of lexical learning in linguistic development," *Child Development*, vol. 71, no. 1, pp. 240–249, 2000.

[11] Carl G. de Marcken, *Unsupervised language acquisition*, Ph.D. thesis, MIT, Cambridge, MA., 1996.

[12] Michael R. Brent and Jeffrey M. Siskind, "The role of exposure to isolated words in early vocabulary development," *Cognition*, vol. 81, pp. 31–44, 2001.

[13] Brian MacWhinney, *The CHILDES project: Tools for analyzing talk*, Erlbaum, Mahwah, NJ, 2000.

[14] Bryan Pellom, "SONIC: The University of Colorado continuous speech recognizer: Technical report TR-CSLR-2001-01," Tech. Rep., University of Colorado, March 2001.

[15] Bryan Pellom and Kadri Hacioglu, "Recent improvements in the CU SONIC ASR system noisy speech: The SPINE task," in *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, April 2003.

[16] Umit H. Yapanel and John H.L. Hansen, "A new perspective on feature extraction for robust in-vehicle speech recognition," in *Proceedings of Eurospeech'03*, Geneva, 2003.

[17] Morten H. Christiansen, Christopher M. Conway, and Suzanne Curtin, *Multiple-Cue Integration in Language Acquisition: A Connectionist Model of Speech Segmentation and Rule-like Behavior*, Hong Kong: City University of Hong Kong Press, 2005.