Coping with Variation in Speech Segmentation

Morten H. Christiansen & Joseph Allen

Program in Neural, Informational and Behavioral Sciences
University of Southern California
Los Angeles, CA 90089-2520, U.S.A.
morten@gizmo.usc.edu, joeallen@gizmo.usc.edu

Abstract

This paper presents results from word segmentation simulations in which a Simple Recurrent Network (SRN) was exposed to speech input incorporating high degrees of variation. In Experiment 1, a network trained on a speech corpus transcribed to include variation in terms of coarticulation was compared with a network trained on a citation form version of the same corpus. The results show that the network accommodates this variation without significant impairment to its performance on the segmentation task. Experiment 2 involved a novel approach to the modeling of segmental variation in which feature values were systematically varied according to a predetermined probability schedule. Results demonstrate that following training the networks were able to withstand a very high degree of segmental variation within words and still able to locate word boundaries in the input. Together the experiments indicate that the SRN provides a robust mechanism for the modeling of early speech segmentation.

1. Introduction

One of the first tasks that a child is faced with in language acquisition is the segmentation of the speech stream into words. The lack of the acoustic equivalents of white spaces in written text makes this a nontrivial task. Recent computational models of early speech segmentation have utilized the integration of multiple probabilistic cues to address this problem. Models by Aslin, Woodward, LaMendola & Bever (1996) and Brent & Cartwright (1996) achieved a good level of performance using a combination of phonology and utterance boundary information. Christiansen, Allen & Seidenberg (in press) showed that combining these two cues with information about lexical stress resulted in improved performance. However, the input to these models abstracted away from many important aspects of real language. The question remains as to how such computational models will fare when exposed to input more closely approximating the variation characteristic of actual speech.

Working from the insight that much of this variation is systematic, we present an investigation of the effects of variation on a connectionist model of infant speech segmentation. One type of variation is coarticulation, where segments vary on the basis of surrounding material. Previous computational models have used corpora in which every instance of a particular word always had the same phonological form. In contrast, we employ a phonetically transcribed corpus in which the phonological form of a word varies with its context.

Another way which the input signal varies is in the fact that individual segments of what is transcribed as the same phoneme actually vary considerably in their acoustic realization. Earlier models, such as Cairns, Shillcock, Chater & Levy (1997), modeled this variation by flipping random features with a certain probability. However, the variation in acoustic realization is not random; rather, for any segment certain features are more susceptible to change than others. Taking these ideas into account, we introduce a novel approach to modeling such segmental variation.

In the remainder of this paper, we present results from simulations involving SRNs trained on input consisting of segmental features, utterance boundary information, and lexical stress. First, we describe the model as well as the training and test materials employed in the simulations. In the following section, we present results from a simulation experiment in which performance is compared between nets trained on coarticulation and citation form versions of the same corpus. In the second experiment we test the same two networks under conditions of high degrees of segmental variation. Together the results show that our model performs well on the segmentation task—despite being faced with input characterized by considerable variation. This outcome is important because it shows that SRNs provide a robust mechanism for the integration of multiple cues even under less idealized conditions, and how such integration may form the basis of early speech segmentation. The conclusion considers further implications of these experiments for the modeling of speech segmentation.

2. Modeling Speech Segmentation

Previous work (Allen & Christiansen 1996, Christiansen 1997, Christiansen et al. in press) has established that SRNs constitute viable models of early speech segmentation. These models, like most other recent computational models of speech segmentation (Aslin et al. 1996, Brent & Cartwright 1996), were provided with input which abstracted away from many important aspects of real speech. This is in part due

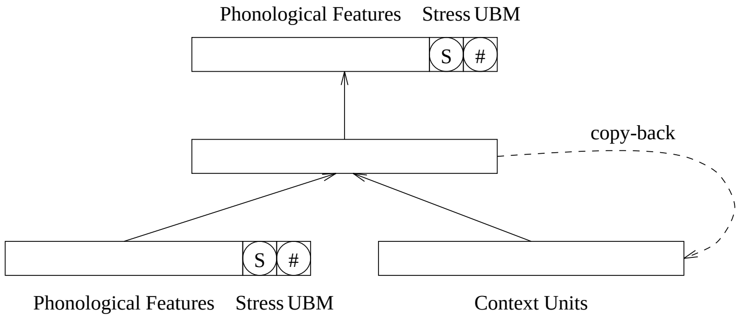

Figure 1: The architecture of the SRN used in the simulations. Arrows with solid lines denote trainable weights, whereas the arrow with the dashed line denotes copy-back connections. The SRN had 20 input/output units and 80 hidden/context units.

to the use of corpora in which every instance of a word always has the same form. To more closely approximate the variability of the way words are phonologically realized in natural speech, we used a corpus that was phonetically transcribed such that the phonological form of a word varied with its context. More specifically, we gleaned the adult utterances from the Carterette & Jones (1974) corpus—a part of the CHILDES database (MacWhinney 1991). These utterances consist of informal speech among American college-aged adults.[1]

The SRN model employed in the current simulations experiments is illustrated in Figure 1. The SRN (Elman 1990) is feed-forward network with a crucial addition, the context units, to which the hidden unit activations from time step t is copied back to be paired with the next input at time $t+1$. This feedback loop allows information about previous hidden unit states to influence the processing of subsequent input, providing the SRN with limited ability to encode information spanning temporal sequences.

The network was provided with three probabilistic cues for possible integration in the segmentation task: (a) *phonology* represented in terms of an 18 value feature geometry, (b) *lexical stress* represented as a separate feature indicating the presence of primary vowel stress, and (c) *utterance boundary information* represented as a separate feature (UBM) which was only activated when pauses occurred in the input. In the simulations, the SRN was trained on the *immediate* task of predicting the next phonological feature set along with appropriate activations of the stress unit and the utterance boundary unit. In learning to perform this task it was expected that the network would also learn to integrate the cues such that it could carry out the *derived* task of segmenting the input into words.

Ultimately, any model of speech segmentation must be able to deal with the high degree of variation which characterizes natural fluent speech. The purpose of our simulations was therefore to investigate whether the success of the SRN model of early word segmentation (Allen & Christiansen 1996, Christiansen et al. in press) was dependent on the use of the simplified citation form input. Comparisons were made between networks exposed to a corpus incorporating coarticulation and networks exposed to a citation form version of the same corpus. In addition, the networks were tested on corpora involving high degrees of segmental variation. If the SRN is to remain a viable model of word segmentation, no significant difference in performance should arise from these comparisons.

The simulations involved two training conditions, depending on the nature of the training corpus. In the *coarticulation* condition the SRN was trained on the phonetically transcribed UNIBET version of the Carterette & Jones corpus. This transcription did not include lexical stress—a cue which contributed significantly to successful SRN segmentation performance in Christiansen et al. (in press). However, lexical stress was indirectly encoded by the use of the reduced vowel *schwa* (/6/ in UNIBET), so we chose to encode all vowels save the *schwa* as bearing primary stress.[2] Utterance boundaries were encoded whenever a pause was indicated in the transcript. In the *citation form* condition, the SRN was trained on a corpus generated by replacing each word in an orthographic version of the Carterette & Jones corpus with a phonological citation form derived via the Carnegie Mellon Pronouncing Dictionary (cmudict.0.4)—a machine-readable pronunciation dictionary for North American English which includes lexical stress information. This procedure was similar to the one used to gener-

[1] It would, of course, have been desirable to use child directed speech as in Christiansen et al. (in press), but it was not possible to find a corpus of phonetically transcribed child directed speech.

[2] This idealization seems reasonable because most monosyllabic words are stressed and because most of the weak syllables in the multisyllabic words from the corpus involved a *schwa*. Further support for this idealization comes from the fact that the addition of vowel stress implemented in this manner significantly improved performance compared to a training condition in which no stress information was provided.

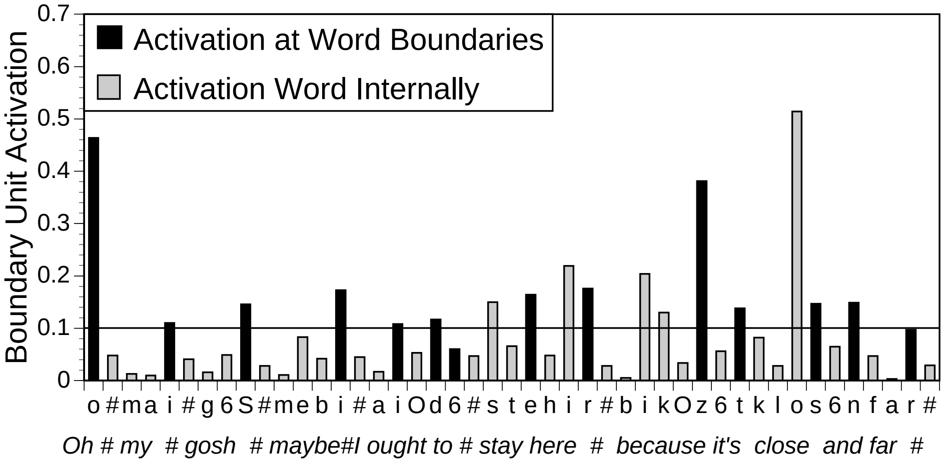

Figure 2: Boundary unit activation for the first 45 tokens in the coarticulation test corpus. A gloss of the input utterances is found beneath the input phoneme tokens (# = utterance boundary).

ate training corpora for the models reported in Christiansen et al. (in press). These pronunciations were subsequently translated into UNIBET format. Four vowels which were weakly stressed according to the dictionary were replaced with the UNIBET *schwa* and encoded as stressless, whereas the other vowels were encoded as stressed. Because the orthographic version of the Carterette & Jones corpus did not include indications of the pauses that occurred within a single turn in the phonetically transcribed version, the number of pauses that occurred on each of the phonologically transcribed lines were randomly inserted into the citation form version of the corpus. [3]

The overall corpus consisted of 1,597 utterances comprising 11,518 words. Test corpora were constructed by setting aside 10% of the utterances (the same utterances in both training conditions). Thus, the training corpora consisted of 1,438 utterances (10,371 words) and the test corpora to 159 utterances (1,147 words). In order to provide for more accurate test comparisons between the SRNs trained under the two conditions, utterance boundaries was inserted by hand in the citation form test corpus in the exact same places as found in the coarticulation test corpus. The networks in both training conditions were trained on two passes through their respective training corpora, corresponding to 74,746 sets of weight updates. Identical learning parameters were used in the two training conditions (learning rate: .1; momentum: .95) and the two nets were given the same initial weight randomization within the interval [-.2,.2].

Next in the first simulation experiment, we investigate whether the SRN model of early segmentation can perform as well in the coarticulation condition as in the citation form condition.

3. Experiment 1: Coping With Coarticulation

With respect to the networks, the logic behind the derived word segmentation task is that the end of an utterance is also the end of a word. If the network is able to integrate the provided cues in order to activate the boundary unit at the ends of words occurring at the end of an utterance, it should also be able to generalize this knowledge so as to activate the boundary unit at the ends of words which occur *inside* an utterance (Aslin et al. 1996). Figure 2 provides a snapshot of the segmentation performance of the coarticulation SRN on the first 45 phoneme tokens in the test corpus. Activation of the boundary unit at a particular position corresponds to the network's hypothesis that a boundary follows this phoneme. Black bars indicate the activation at lexical boundaries, whereas the grey bars correspond to activation at word internal positions. Activations above the mean (horizontal line) are interpreted as the postulation of a word boundary. As can be seen from the figure, the SRN performed well. It correctly activated the boundary unit before and after nine of the 14 words (i.e., *oh, my, gosh, maybe, I, ought, it's, and, far*), only missegmenting 5 words (i.e., *to, stay, here, because, close*).

In order to compare the performance of the two networks, the accuracy and completeness of their word predictions (Brent & Cartwright 1996) were calculated for the test corpora using:

[3] Note that the random insertion of utterance boundaries may lead to the occurrence of utterance boundaries were they often do not occur normally (not even as pauses), e.g., after determiners. Because the presence of pauses in the input is what leads the network to postulate boundaries between words, this random approach is more likely to improve rather than impair overall performance, and thus will not bias the results in the direction of the coarticulation training condition.

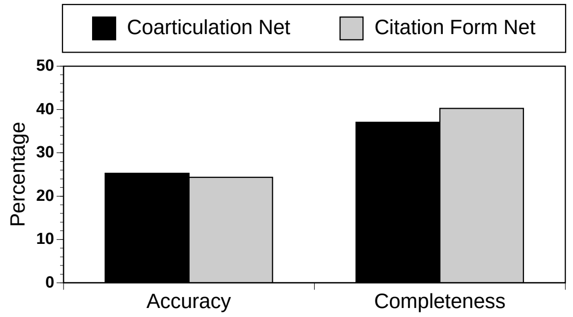

Figure 3: Word accuracy and completeness for the coarticulation net (black bars) and the citation form net (grey bars).

$$\text{Accuracy} = \frac{\text{Hits}}{\text{Hits} + \text{False Alarms}}$$

$$\text{Completeness} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}}$$

Accuracy provides a measure of how many of the words that the network postulated were actual words, whereas completeness provides a measure of how many words, of those which were there to be found, the net actually discovered. A hit requires that the network correctly predicts both the beginning and the end of a word (without any false positives). A miss occurs if the net does not segment a word at its appropriate boundaries, and a false alarm stems from segmenting a word incorrectly. Consider the following hypothetical example:

#the#dog#s#chase#thec#at#

where # corresponds to the postulation of a word boundary. Here the hypothetical learner correctly segmented out two words, *the* and *chase*, but also falsely segmented out *dog*, *s*, *thec*, and *at*, thus missing the words *dogs*, *the*, and *cat*. This results in an accuracy of $\frac{2}{2+4}$ = 33.3% and a completeness of $\frac{2}{2+3}$ = 40.0%.

With these measures in hand, we can compare the performance of the SRNs trained in the coarticulation and citation form conditions. Figure 3 shows the accuracy and completeness scores for the two networks. The coarticulation SRN obtained an accuracy of 25.27% and a completeness of 37.05%. The citation form SRN reached an accuracy of 24.33% and a completeness of 40.24%. There were no significant differences between the accuracy scores ($\chi^2 = 0.42, p > .9$) or the completeness scores ($\chi^2 = 2.46, p > .19$). The SRN model of word segmentation thus was able to cope successfully with variation in the form of coarticulation, suggesting that it provides a good basis for discovering word boundaries in input that is closer to natural speech than the input used in previous computational models . The next simulation experiment investigates how the model fares when exposed to additional segmental variation.

4. Experiment 2: Coping With Segmental Variation

Individual segments of what is transcribed as the same phoneme in reality vary considerably in their acoustic realization. Earlier models have attempted to model such segmental variation by flipping random features with a certain probability (Cairns et al. 1997). However, the variation in acoustic realization does not occur randomly; rather, for any segment certain features are more susceptible to change than others.

Taking this observation into account, we divided the features for each phoneme into a set of *core features* and a set of *peripheral features*. Whereas peripheral features are prone to change, core features tend to be more resistant because a change to the these would alter the basic nature of the phoneme. For example, in our feature encoding scheme, changing the *voiced* feature of a /p/ phoneme would result in a /b/ phoneme, however, changing the continuant feature would only result in unorthodox instance of /p/, rather than another phoneme altogether. The peripheral features for each phoneme were chosen such that a change to these features for a given phoneme would not dramatically impair its recognition in the dialect of American English transcribed in the Carterette & Jones (1974) corpus. To capture segmental variation we created test sets in which the peripheral features would be subject to change given a certain probability. Results are presented from three sets of simulations in which the probability of peripheral feature change was .01, 05, and .1, respectively. A phoneme on average has 1.9 peripheral features. Thus the chances for a given phoneme to undergo segmental change is higher than indicated by the above probabilities. The same is true with respect to the chance of getting a segmental change in a particular word. Thus, in the three test sets approximately 6%, 26% and 41% of the words underwent segmental change.[4]

Figure 4 displays the results from testing the coarticulation net and the citation form net on versions of their respective test corpora with increasing degrees of segmental variation. Both networks performed well when exposed to these test corpora, only experiencing nonsignificant decrements in performance as the degree of variation increased. For the coarticulation network accuracy scores decreased from 25.27% to 23.99 ($\chi^2 = .73, p = 1$) and completeness scores from 37.07% to 34.48% ($\chi^2 = 1.62, p = 1$); and

[4] The probability of segmental change in a word was calculated as: $1 - (1 - p)^N$ where p is the (independent) probability of a phoneme undergoing a feature change (calculated by counting the actual number of changed phonemes) and N is the number of phonemes in a word (given an average of 3.22 phonemes per word we used $N = 3$).

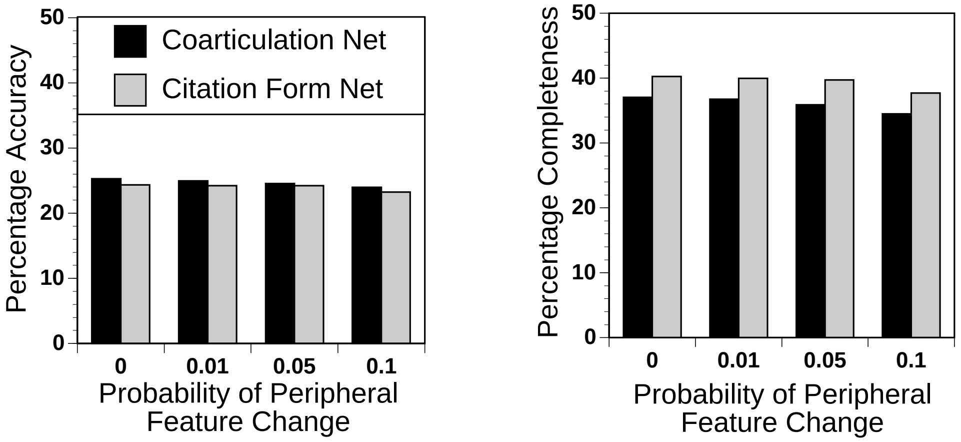

Figure 4: Accuracy (left panel) and completeness (right panel) obtained by the coarticulation net (black bars) and the citation form net (grey bars) under conditions of increasing segmental variation.

for the citation form network accuracy decreased from 24.33% to 23.25% ($\chi^2 = .60, p = 1$), and completeness from 40.24% to 37.72% ($\chi^2 = 1.54, p = 1$). There was no significant difference at any point between the accuracy ($p's > .9$) or the completeness ($p's > .1$) of the two networks. The coarticulation net was able to cope successfully with a high degree of segmental variation—even when as many as 41% of the words underwent segmental change. These results suggest that the model of speech segmentation introduced in Christiansen et al. (in press) is not only able to deal with coarticulation, but also with a high degree of subsequent segmental variation.

5. Conclusion

The results from the two simulation experiments show that our model performs well on the segmentation task—despite being faced with input characterized by considerable variation. This outcome is important because it demonstrates that the model provides a robust mechanism for the integration of multiple cues, whereas previous models have not been shown to be able to cope satisfactorily with coarticulation and segmental variation. For example, although the connectionist model by Cairns et al. (1997) was trained on a corpus of conversational speech in which assimilation and vowel reduction had been introduced into the citation forms using a set of rewrite rules, it performed poorly in comparison with the present model (e.g., when pauses were included, their model discovered 32% of the *lexical* boundaries whereas our model discovered 79% of the lexical boundaries). Both sets of results suggest that connectionist networks provide a useful framework for investigating speech segmentation under less than ideal circumstances. In contrast, it is not clear that other computational frameworks can readily provide the basis for such investigations. For example, a statistical model such as Brent & Cartwright (1996), which uses stored representations of familiar lexical items to discover novel items for subsequent memorization, would as a consequence of coarticulation include several different phonological versions of the same word into the lexicon. Given that such statistical models tend to use phonemes as the basic representational unit it is not clear how to investigate segmental variation save by a probabilistic process of phoneme replacement. This would, however, exacerbate the problem of creating multiple phonological forms for the same underlying word.

Of course, there is much more to the variation in the speech stream than we have addressed here. For example, the input to our coarticulation nets varied in terms of the individual phonemes making up a word in different contexts, but in real speech coarticulation also often results in featural changes across several segments (e.g., the nasalization of the vowel segment in *can*). Similarly, segmental variation also includes combined changes among clusters of segments in addition to the independent changes in individual segments implemented in Experiment 2. Future work must seek to bring the input to segmentation models closer to the actual variations found in fluent speech, and we have sought to take the first steps here.

Together, the results from the two experiments show that after a initial period of exposure to input with moderate variation (coarticulation), the SRN model is subsequently able to cope with additional high degrees of segmental variation. The ability of the network trained under relatively noisy conditions to perform under increasingly noisy conditions after training is not only an important demonstration of fault tolerance. Developmental data also indicates that early motherese is characterized by less variation than speech directed to adults (Morgan, Shi & Allopenna 1996, Ratner 1996). Our model shows how relatively careful (but far from perfect) speech characteristic of motherese may provide a bootstrap to a more robust segmentation mechanism which is then able to deal with higher degrees of variation in subsequent development.

The model upon which these experiments were based also demonstrates how the language specific, rule-like knowledge which we believe underlies both the capacity to segment continuous speech and, in the adult, to detect the naturalness of novel lexical items (phonotactics) might emerge as a consequence of the integration of multiple probabilistic sources of information. The idea that such knowledge consists of integrated low level probability distributions is consistent with a view of linguistic cognition in which representations are a consequence of the interaction among the mechanism in which they are instantiated and the statistics of the language signal to which the speaker is exposed (Seidenberg, Allen & Christiansen, this volume). The model's ability to make predictions based on novel sequences, for example, is a reflection of the similarity metric among such novel sequences and the set of sequences on which the model was trained. The model's sensitivity to these similarity metrics, and exactly how they are constituted are of course a consequence of the architecture used in the simulations.

6. Acknowledgments

We would like to thank Mark Seidenberg for helpful discussions of the issues addressed in this paper.

References

Allen, J. & Christiansen, M. H. (1996), Integrating multiple cues in word segmentation: A connectionist model using hints, *in* 'Proceedings of the 18th Annual Conference of the Cognitive Science Society', Lawrence Erlbaum, Mahwah, NJ, pp. 370–375.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P. & Bever, T. G. (1996), Models of word segmentation in fluent maternal speech to infants, *in* Morgan & Demuth (1996), chapter 8, pp. 117–134.

Brent, M. & Cartwright, T. (1996), 'Distributional regularity and phonotactic constraints are useful for segmentation', *Cognition* **61**, 93–125.

Cairns, P., Shillcock, R., Chater, N. & Levy, J. (1997), 'Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation', *Cognitive Psychology* **33**, 111–153.

Carterette, E. & Jones, M. (1974), *Informal Speech: Alphabetic and Phonemic texts with statistical analyses and tables*, University of California Press, Berkeley, CA.

Christiansen, M. H. (1997), Improving learning and generalization in neural networks through the acquisition of multiple related functions, *in* J. A. Bullinaria, D. W. Glasspool & G. Houghton, eds, 'Proceedings of the Fourth Neural Computation and Psychology Workshop: Connectionist Representations', Springer-Verlag, London, U.K.

Christiansen, M. H., Allen, J. & Seidenberg, M. S. (in press), 'Learning to segment speech using multiple cues: A connectionist model', *Language and Cognitive Processes*.

Elman, J. (1990), 'Finding structure in time', *Cognitive Science* **14**, 179–211.

MacWhinney, B. (1991), *The CHILDES Project*, Lawrence Erlbaum, Hillsdale, NJ.

Morgan, J. L. & Demuth, K., eds (1996), *Signal to Syntax*, Lawrence Erlbaum, Mahwah, NJ.

Morgan, J., Shi, R. & Allopenna, P. (1996), Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping, *in* Morgan & Demuth (1996), chapter 16, pp. 263–281.

Ratner, N. (1996), From "signal to syntax": But what is the nature of the signal?, *in* Morgan & Demuth (1996), chapter 9, pp. 135–150.