Is Recurrence Redundant? Revisiting Allen and Christiansen (1996)

C. Anton Rytting (rytting@ling.ohio-state.edu)
The Ohio State University
Department of Linguistics
222 Oxley Hall
1712 Neil Ave.
Columbus, OH 43210 USA

Abstract

Several influential connectionist models of the word segmentation task (e.g. Cairns, Shillcock, Chater, & Levy, 1997; Christiansen, Allen, & Seidenberg, 1998) follow Elman (1990) in using simple recurrent networks (SRNs). The use of SRNs in this context appears to be traditional rather than independently motivated.

This paper investigates whether alternatives to SRNs can achieve similar performance. Specifically, it reports a replication of part of Allen and Christiansen (1996), and a comparison with multi-layer perceptrons (MLPs). The major results are confirmed; however, the SRN topology is shown not to be necessary for the task addressed. MLPs perform equally well, even when they have access to less context.

Introduction

Word segmentation is the task of recovering word boundaries from an initially unsegmented stream of linguistic input. English writing marks these boundaries, making the task trivial, but Chinese writing does not, neither does any variety of natural speech.

If a language learner is to make any headway in acquiring the words in a language, that learner must learn how to pull these words out of running speech and hear them as separate words. By learning the indirect cues that signal word boundaries for a given language, the word learner solves the word segmentation task. This begins early in the language acquisition process: studies show that babies begin to learn word segmentation strategies around 7-9 months after birth (see e.g. Jusczyk, Houston, & Newsome, 1999, for a review).

Many of the initial experimental studies of word segmentation focused on individual cues: prosody (Jusczyk, Cutler, & Redanz, 1993), allophonic variations (Jusczyk, Hohne, & Bauman, 1999), distributional regularities (Saffran, Aslin, & Newport, 1996), and phonotactics (Friederici & Wessels, 1993). Similarly, early computational simulations of the word segmentation task focused on single cues. Brent and Cartwright (1996) used distributional regularities as a cue, combined with minimum description length as a language-independent heuristic. Cairns et al. (1997) also modeled the use of distributional regularities, but as a precursor to a later stage using stress (which was not explicitly modeled). Aslin et al. (1996) modeled the generalization of segmental cues at the ends of utterances to the ends of words.

Christiansen, Allen, and Seidenberg (1998) explicitly addresses the interaction among several cues: distributional regularities, stress, and utterance-boundary information. Their model reflects a particular conception of the word segmentation task within the larger context of language acquisition. The infant's primary task is the comprehension of the speech around it (as well as producing comprehensible speech in turn). As the child attends to the speech input to which it is exposed, it notices patterns in the signal (Saffran et al., 1996). Christiansen et al. (1998) conceive of a second, immediate task in language acquisition: namely, to continually update a statistical representation of these salient patterns.

This process is exemplified in a simple model presented in Allen and Christiansen (1996). In this scenario, a network receives input one phone at a time. Its immediate task is learn to predict the identity of the next phone.[1] The immediate task is "self-supervised" in that the next input provides immediate feedback as to what the preceding output should have been.

To model the contribution of distributional regularities at the phone level to the discovery of word boundaries, a neural network with an input unit and an output unit for each phone may be constructed, and the likelihood of a word boundary approximated by measuring points at which the error in predicting the next phone is high (see Cairns et al., 1997, for a model based on this idea). In order to model the interaction of utterance-boundary info and distributional regularity, an additional unit, symbolizing an utterance boundary or pause in speech, is added to the input layer, and corresponding unit to the output layer. The idea of having multiple output nodes learning multiple problems simultaneously is referred to as "hints" or "catalyst" nodes (see Suddarth & Kergosien, 1991). The intuition (following Aslin et al., 1996) is that since utterance boundaries are a subset of word boundaries with more or less representative properties, the utterance boundary output unit should be activated not only at utterance boundaries, but (to a lesser degree) at all word boundaries. Hence, above-average activation of the utterance boundary output unit is used to measure the network's performance on the derived task: identifying word boundaries. The other output units, corresponding to phones, can be thought of as catalyst units to the utterance boundary unit.

A major point of Allen and Christiansen (1996) is that word boundaries can be learned in this way only for languages with particular statistical properties. It is straightforward to construct artificial languages that lack these properties. Specifically, for a language where each syllable is equally likely to follow any other (or to end a word or an utterance), the network can learn at best only syllable boundaries. But if certain segments or groups of segments (such as syllables) are more

[1] The assumption that the identity of the phone is unambiguously observable is itself an idealization. In real life, the intended segment is not observable with perfect accuracy, particularly for infants who may be presumed to be still learning the language's inventory of sounds. The automatic speech recognition community rightly treats phone identity as part of the hidden structure, not the observable. Fully addressing this problem would unduly complicate the model under discussion, although Christiansen and Allen (1997) addresses it in part.

likely to appear at the ends of words than others (as seems to be the case with most human languages), and a sufficiently large sample of the language is available to the learner, then the network can learn.

A summary of the Allen and Christiansen (1996) simulations

Method

To show this point, two artificial "mini-languages" were constructed to train the net: a "variable transition probability" (vtp) language, and a language with "flat" transitional probabilities between syllables. Each of these languages used the same twelve syllable types: 'b', 'd', 'p', or 't' followed by 'a', 'i', or 'u'. Each language consisted only of three-syllable, six-segment (CVCVCV) words. The "flat" language contained 12 such words constructed to maintain a word-internal transitional probability of 0.667 from one syllable to another; the "vtp" language used 15 words following different restrictions: for example, no word begins in 'b' or ends in 'u'.

Both languages were trained using a simple recurrent net (SRN) with 8 units each in the input and output layers and 30 each in the hidden and context layers. Word boundaries were not explicitly marked, but utterance boundaries (corresponding to the last input unit) were placed at intervals ranging from 2 to 6 words long. Training was done for seven iterations over a corpus with 120 instances of each word in the mini-language; testing was done on a corpus without marked utterance boundaries.

Results

For all experiments, the dependent measure was activation of the utterance boundary output unit. On the vtp language, the network predicts a significantly higher activation for word-boundary positions (0.204 on average) than for word internal positions (0.04 on average), including other syllable-boundary positions. The network trained and tested on the flat language showed higher activation at syllable- boundary positions, but these did not differ significantly between word-boundary and word-internal syllables (although the graph suggests that the activation after the first syllable of a word may be very slightly less). These results are seen to validate the points discussed above.

Issues of network structure

While the findings in Allen and Christiansen (1996) may seem of minor importance, their significance may be found in the basis of this same architecture in a number of studies involving actual language (Christiansen & Allen, 1997; Christiansen et al., 1998; Christiansen, Conway, & Curtin, 2005; Curtin, Mintz, & Christiansen, 2005). Before building on these works, it is useful to re-examine the initial assumptions made and methods used to distinguish the essential characteristics from the accidental.

For example, the notion of hints is essential to Allen and Christiansen's (1996) argument, and is an innovation over earlier studies such as Elman (1990); Cairns et al. (1997); Aslin et al. (1996). Where one might suppose that a network would find dealing with multiple prediction tasks more challenging, these results show that the additional tasks actually help the net learn the derived task of interest. Hence, it provides a plausible way of combining multiple cues in a computational model.

On the other hand, the specific type of network used (SRN versus time-windowed MLP) seems to be inherited from earlier studies such as Elman (1990). Elman (1990) developed his SRN topology to deal with sequential problems such as language, and SRNs have been used for modeling a number of language tasks where memory of indefinite length is needed. However, the usefulness of SRNs for modeling memory of longer time sequences has been shown to be somewhat limited, a weakness sometimes known as the "latching" problem (Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001; Tino, Cernansky, & Benuskova, 2004). Some have attempted to develop alternative ways of modeling memory, to overcome some of these limitations (Hochreiter & Schmidhuber, 1997). But for some language processing tasks, particularly those dealing with speech processing (rather than longer-range or syntactically complex tasks), arguably simpler and easier-to-train models such as time-windowed multilayer perceptrons (MLPs) are still used (e.g. Zhu, Chen, Morgan, & Stolcke, 2004; Morgan, Chen, Zhu, & Stolcke, 2004).

MLPs have also been used to model word segmentation. Aslin et al. (1996) decided to use a time-windowed MLPs rather than SRNs, explicitly stating their preference on grounds of simplicity. An additional complexity resulting from this choice is the need to explicitly vary the length of the time window, but this too had an upside, in that one knows explicitly what contribution each time step in the prior history is making.

Allen and Christiansen (1996) do not explicitly defend SRNs as an essential design choice, although they seem to prefer it implicitly to feed-forward networks (given that they incorporate the cue from Aslin et al. (1996) within their recurrent framework). Whether their use of SRNs rather than MLPs is crucial to their model is an open question. While this question is perhaps secondary to other considerations, it is nevertheless worth considering for two reasons:

First, the empirical clarity of a computational model depends in part on knowing and stating as precisely as possible which aspects of the model are essential to its success in simulating human behavior and which are accidental. A second and more practical reason is simple efficiency: if MLPs make just as good predictions, and are easier to work with and faster to train, then this allows for a greater number of explorations with less effort.

Revisiting Allen and Christiansen (1996)

Simulation 1: Simple Recurrent Networks

First, a replication of part of Allen and Christiansen (1996) was performed using the SRN package available within the **c**onx neural network toolkit (Blank, Kumar, Meeden, & Yanco, 2003).[2] The training parameters were matched to that work as nearly as possible. The network consists of an input layer and an output layer of eight units each (one for each symbol in the mini-language, including the utterance bound-

[2] available as part of the Python Robotics toolkit at pyro-robotics.org

ary marker), along with thirty units each in the hidden and context layers. "One-hot" encoding was used: each symbol was associated with a single input, for which the activation was 1 if the input was that symbol, and 0 otherwise. As in the original study, a learning rate of 0.1 and momentum of 0.95 was used, and weights were initialized with random values from a flat distribution ranging between -0.25 and 0.25. The activation function used was a logistic sigmoid $\sigma(x) = 1/(1+e^{-x})$. Standard backpropagation was used; error was measured with a squared error loss function.

The artificial languages used for the training and testing data were not available and had to be recreated according to the specifications in the text. These recreated languages are available from the author's web-site.[3] Separate training and test corpora are used: the training corpus for each language is marked with utterance boundaries; the test corpus is not. The level of activation of the utterance-boundary output node is predicted to be higher at word boundaries than at non-boundary positions. Since a certain amount of variation is to be expected from the random initial weights, the results of sixteen iterations, each with different starting weights, were averaged together for each condition tested.

The results from this replication were qualitatively similar to those reported in (Allen & Christiansen, 1996), although the difference in activation levels in the vtp condition was not so large. As predicted, activation levels in the vtp condition are higher at word boundaries (mean value of 0.136 across the sixteen runs, min 0.103, max 0.179) than at non-word boundaries (mean 0.018, min 0.014, max 0.023). There is also a difference between syllable boundaries and syllable-internal positions (0.076 vs. 0.0018). Crucially, word-internal syllable-boundaries show an average activation of 0.046—less than half that of the word boundaries. This shows that the network is learning word boundaries, not just syllable boundaries.

For the flat condition, activation is also higher at word boundaries (mean 0.069) than at non-word boundaries (0.025). However, this difference is fully accounted for in different activation levels between syllable boundary positions (0.069) and syllable-internal positions (0.0024). There is no appreciable difference between word boundary positions and word-internal syllable-boundary positions (both 0.069). Hence, the flat condition is only able to learn syllable boundaries, not word boundaries, just as Allen and Christiansen (1996) observed.

What is more important than the raw differences in average activation, however, is the degree of discriminability between true word boundaries and non-boundary positions from this activation level. This depends not only on the activation difference, but also on the variations between activation levels, which are not directly reported. A better measure of the overall discriminability is a receiver operating characteristic (ROC) curve. This measure plots the true positive rate (the probability of correctly detecting a word boundary) over the false positive rate (the probability of incorrectly positing a word boundary). Since any point along the curve corresponds to the performance at a given threshold of activation, the curve as a whole summarizes the performance of a binary decision procedure at any relevant threshold.

[3] http://www.ling.ohio-state.edu/~rytting/cogsci2006/AC96MiniLangs.txt

A useful summary statistic for an ROC is the area under the curve (AUC), which corresponds to the probability a randomly chosen pair of one positive item and one negative item will be correctly ranked (Hanley & McNeil, 1982). In this domain, this provides the probability, given a word-final symbol and a non-word-final symbol (both randomly selected), that the activation of the utterance-boundary unit at the word-final symbol will be higher than the activation at the the non-word-final symbol.

An AUC of 0.5 is no better than chance; perfect discrimination would have an AUC of 1. For each of the conditions here, an ROC curve was calculated using only the syllable boundary positions, rather than for every phone (since the syllable-internal positions are trivial to learn). As shown in Figure 1, the SRN in the vtp condition achieves a mean AUC of 0.85 (min 0.81, max 0.90), meaning that it performs better than chance at distinguishing word boundaries from other syllable boundaries. In the flat condition the SRN output has a mean AUC of 0.50 (min 0.43, max 0.54), showing it does no better than a syllable-boundary detector.

Simulation 2: Multi-layer Perceptrons

In the second experiment, the same simulation was replicated, substituting an MLP for the SRN with a context of 1, 2, and 3 preceding phones. For the MLP with 1 phone context, the network topology was identical (8 input, 30 hidden, and 8 output units) except for the removal of the 30 context units. The 2- and 3-phone context MLPs had 16 and 24 input units. All were fully connected with the hidden layer. The training procedure was the same.

1-phone context MLPs The average activations from the 1-phone condition were very similar to those reported above for the SRNs. Activation levels in the vtp condition are once again higher at word boundaries (0.12) than at non-word boundaries (0.024), syllable boundaries have higher activations than syllable-internal positions (0.079 vs. 0.00036), and activations at word-internal syllable-boundaries are again about half as strong actual word boundaries (0.060 vs. 0.12). This suggests that the MLP is more sensitive to word boundaries than to other syllable boundaries, even with only one phone of prior context. The discriminability of this network is not quite as good as the SRN: the area under the corresponding ROC curve (not shown) is 0.81. Since the MLP with only one phone of context converged to the same solution in all sixteen trials, the standard error is practically zero. This is worse than the SRN ($F(1,30) = 22.58$, $p < 0.001$), but clearly better than chance.

For the flat condition, in contrast, the average activation is equal for each of the syllable boundaries (0.0644) Just as in the SRN case, the network essentially fails to learn anything other than syllable boundaries (mean AUC = 0.515, min 0.501, max 0.517).

2-phone context MLPs The results for the 2-phone context MLP in the vtp condition are (unsurprisingly) better than the 1-phone context: The average activation for word boundaries (0.133) is once again higher than for word-internal syllable boundaries (0.659) and for syllable boundaries generally (0.71) The mean area under the ROC curve for the sixteen trials is 0.86 (min 0.845, max 0.869)—overlapping with the

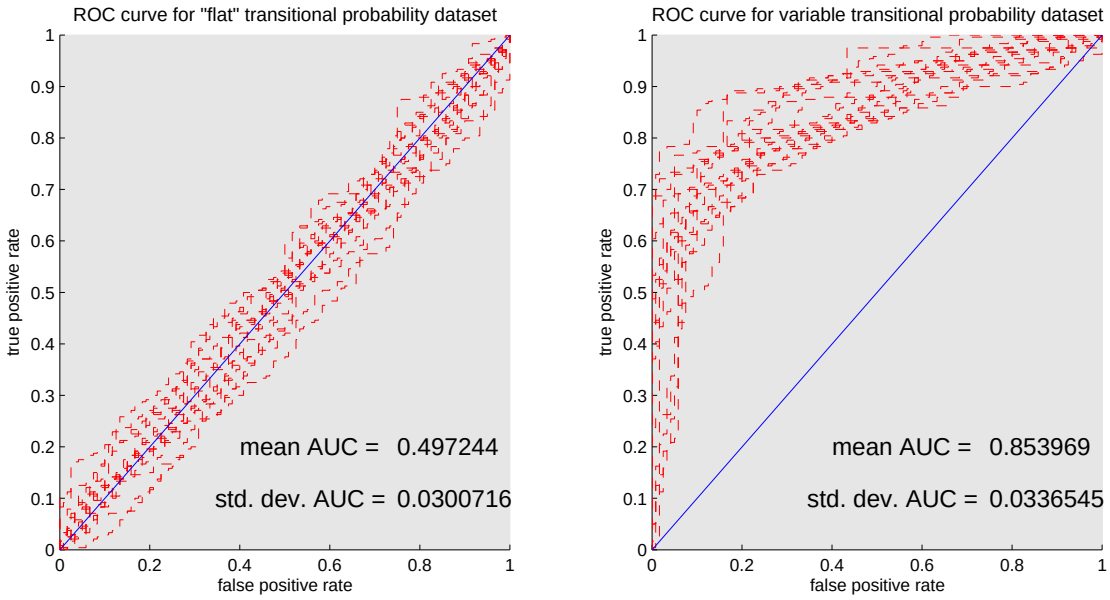

Figure 1: Receiver operating characteristic (ROC) curves for 16 runs of the SRN in Simulation 1. Area under the curve (AUC) shows the SRN's discrimination between word boundaries and other syllable boundaries. Syllable-internal positions are not included.

range of scores seen for the SRN, though with less variation between trials. The difference between the 2-phone MLP and the SRN is not significantly significant ($F(1,30) = 0.69, p = 0.41$).

For the flat condition, syllable boundaries are once again higher than non-syllable boundaries (0.065 vs. 0.0008). However, word boundaries are no higher than other syllable boundaries, just as in the 1-phone case. The mean AUC is 0.472, slightly below chance (min 0.471, max 0.476).

3-phone context MLPs The average activations in the vtp condition follow the same pattern as the other MLPs for this condition, but with somewhat greater differences between the word boundary activations (mean 0.148, min 0.121, max 0.170) and the word internal averages (0.015). As with the other conditions, there is a large, consistent difference between word boundaries and syllable boundaries generally (mean 0.073, min 0.064, max 0.083). However, the difference between the 3-phone MLP and the other nets is most clearly seen in the areas under the ROC curves. As shown in Figure 2, the AUC for distinguishing word boundaries from other syllable boundaries is 0.92 (min 0.893, max 0.932). The performance as measured by the AUC is significantly better than that on the SRN condition ($F(1,30) = 48.21$, $p < 0.0001$).

For the flat condition, syllable boundaries are once again higher than non-syllable boundaries (0.067 vs. 0.0004). However, this time word boundaries are higher than other syllable boundaries (0.074 vs. 0.062). This is not because word boundaries themselves are being learned directly, but rather because the net is learning that the first syllable boundary **c**annot be a word boundary: the average activation for the end of the first syllable is 0.049, markedly lower than the second (0.077) or the third (0.074). This effect also may be seen in the ROC curve, which is no longer merely at chance (AUC = 0.596, min 0.553, max 0.646). Although there are no cues in the syllabic transitional probabilities, by looking at more than one syllable of context, the net observes longer-range regularities that arise in the data. This effect may need further study to be completely understood, however.

Discussion

We may see in the results above that MLPs also account as well as SRNs for the effects of multiple-cue integration in a restricted domain. Generally speaking, the more context, the more successful the net is at differentiating word boundaries from other syllable boundaries, though even a single phone of context, without recurrence of any sort, is sufficient for better-than-chance performance. This is due largely to the design of the artificial language used. One of the most obvious cues to an upcoming word boundary in the vtp language: 'a' or 'i' vowel (as opposed an 'u') was of course just as visible to the 1-phone MLP as to any other net. With two phones of context, it was possible to pick up also on a second major cue in the language: 'b' (not being allowed at the beginning of a word) is slightly more likely to start a word-final syllable than the other three consonants. This may explain why the two-phone MLP did just as well as the SRN: because the most salient, relevant cues in the language occurred just two phones from the end of the word.[4]

[4]Tino et al. (2004) and others mention another potential difference between SRNs and MLPs: namely, an *architectural bias* that may result from the net's structure combined with its initial random weights before training. To test for this bias, pretests were run on the testing data before any training took place. No evidence of such a bias, or any appreciable difference between the MLP and the SRN prior to training, was found for this task as described above.

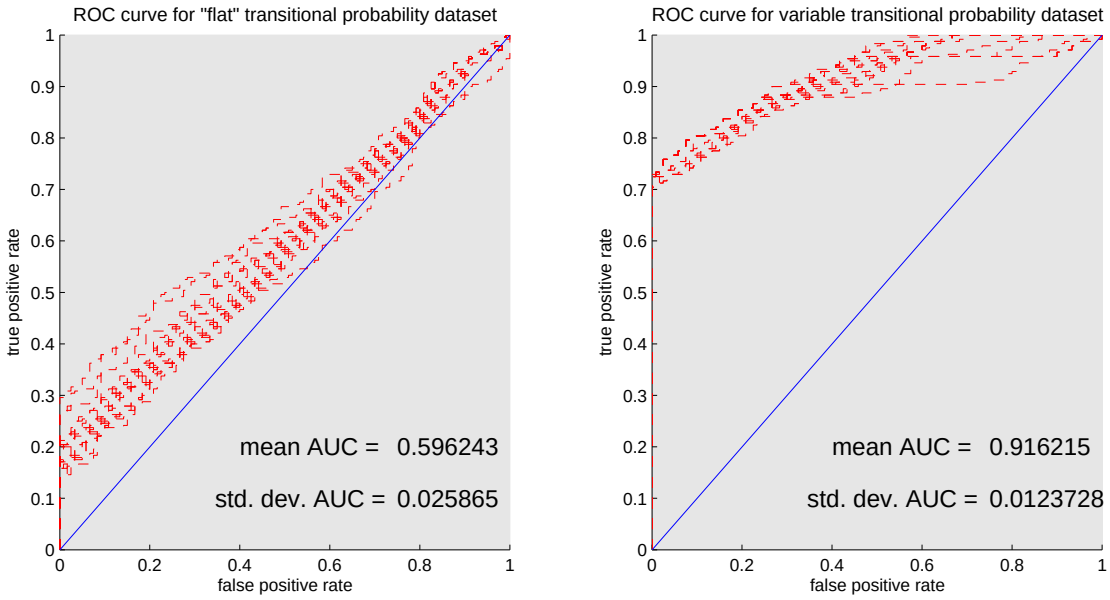

Figure 2: Receiver operating characteristic (ROC) curves for 16 runs of the MLP with three phones of context in Simulation 2. Area under the curve (AUC) shows the MLP's discrimination between word boundaries and other syllable boundaries. Syllable-internal positions are not included.

While the toy language may in hindsight look like a ridiculously easy problem for the net, it is worth noting that such patterns are not unheard of in human languages. Many languages (e.g., Spanish, Italian, Modern Greek) have a restricted set of phonemes word finally, compared with other positions in the word (even the final position of word-internal syllables), and a great many languages have certain phonemes (particularly those commonly found in word-final inflections, like English /-s/) that are extremely frequent at the end of words compared with other positions. Thus, word segmentation may not be a problem for which recurrence is particularly necessary, although more context does of course help. In contrast, no such claim is made about problems of language acquisition involving higher-level structures, such as syntax, where unbounded dependencies pose problems that connectionist models quite likely need some recourse to recurrence in order to solve.

Conclusion

The notion of hints and multiple tasks within neural-net frameworks is a useful paradigm for modeling and investigating the exploration of multiple cue interaction in problems of language acquisition, such as the word segmentation task. However, the use of hints is not dependent on any particular topology of neural network; these results suggest that the procedure works equally well for two different network topologies, SRNs and time-windowed MLPs, at least for a constrained task on artificial data. Furthermore, the exact size of the time-window is not always crucial.

This finding is useful in that it frees the researcher to consider the best design for the problem at hand, independent of the cue interaction issues. While SRNs are certainly an intuitively attractive topology for modeling the prediction of sequential data (including unsupervised discovery of sequential structures), particularly when the exact extent of the relevant context is unknown, there may be situations when other models are preferable: e.g., allowing for the use of more efficient training techniques and faster simulation of larger-size problems.

Finally, keeping in mind what parts of the model are essential, and which are incidental, may be helpful in relating these and future models to issues of biological plausibility. Naturally, all ANNs are to some degree biologically implausible, and there is much that is still unknown about the design of the neural system actually used in language processing. If it should someday be shown that recurrent models of the Elman type are incompatible with the actual mechanism for speech processing, the multiple cues model need not be rejected automatically, inasmuch as it has been shown not to depend crucially on recurrence. By abstracting away from non-essential elements such as recurrence, the model not only gains some flexibility, but may better highlight its essential element: the hints themselves.

Future directions

This work is a preliminary step toward extending connectionist models of the word segmentation task to other types of input, including automatically pre-processed (noisy) audio data, and input from languages besides English. Future steps include the replication of larger-scale studies using transcriptions of both adult- and child-directed English (Christiansen & Allen, 1997; Christiansen et al., 1998). Later steps include moving away from human transcriptions to acoustic input from speech recordings and finally to exploring the in-

tegration of automatic, concurrent, unsupervised acquisition of a phonemic inventory with the word segmentation task.

References

Allen, J., & Christiansen, M. H. (1996). Integrating multiple cues in word segmentation: A connectionist model using hints. In *Proceedings of the eighteenth annual cognitive science society conference* (pp. 370–375). Mahwah, NJ: Lawrence Erlbaum Associates.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), (pp. 117–134). Mahwah, NJ: Lawrence Erlbaum Associates.

Blank, D., Kumar, D., Meeden, L., & Yanco, H. (2003). Pyro: A Python-based versatile programming environment for teaching robotics. *Journal of Educational Resources in Computing (JERIC)*, *3*(4), 1–15.

Brent, M. R., & Cartwright, T. A. (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.

Christiansen, M. H., & Allen, J. (1997). Coping with variation in speech segmentation. In A. Sorace, C. Heycock, & R. Shillcock (Eds.), *Proceedings of gala*.

Christiansen, M. H., Allen, J., & Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13 (2/3)*, 221–268.

Christiansen, M. H., Conway, C. M., & Curtin, S. (2005). Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language acquisition, change and emergence: Essay in evolutionary linguistics.* Hong Kong: City University of Hong Kong Press.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, *96*, 233-262.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179-211.

Friederici, A., & Wessels, J. (1993). Phonotactic knowledge and its use in infant speech perception. Perception and Psychophysics, *54*, 287–295.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. Kremer & J. Kolen (Eds.), *A field guide to dynamical recurrent neural networks.* IEEE Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Jusczyk, P. W., Cutler, A., & Redanz, N. (1993, June). Infants preference for the predominant stress patterns of English words. *Child Development*, *64*(3), 675–687.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, *61*, 1465–1476.

Jusczyk, P. W., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207.

Morgan, N., Chen, B., Zhu, Q., & Stolcke, A. (2004, May 17–21). TRAPping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition. In *Proceedings icassp-2004.* Montreal: IEEE.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Suddarth, S., & Kergosien, Y. (1991). Rule-injection hints as a means of improving network performance and learning time. In L. Almeida & C. Wellekens (Eds.), *Neural networks/eurasip workshop 1990* (pp. 120–129). Berlin.

Tino, P., Cernansky, M., & Benuskova, L. (2004, January). Markovian architectural bias of recurrent neural networks. *IEEE Trans. Neural Netw.*, *15*(1), 6–15.

Zhu, Q., Chen, B., Morgan, N., & Stolcke, A. (2004). On using mlp features in lvcsr. In *Interspeech 2004 - icslp, 8th international conference on spoken language processing* (pp. 921–924). Jeju Island, Korea: ISCA.