Simple Recurrent Networks and Competition Effects in Spoken Word Recognition

James S. Magnuson (magnuson@bcs.rochester.edu)
Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Richard N. Aslin (aslin@cvs.rochester.edu)
Department of Brain and Cognitive Sciences
University of Rochester, Meliora Hall, Rochester, NY 14627 USA

Abstract

Continuous mapping models of spoken word recognition such as TRACE (McClelland and Elman, 1986) make robust predictions about a wide variety of phenomena. However, most of these models are interactive activation models with preset weights, and do not provide an account of learning. Simple recurrent networks (SRNs, e.g., Elman, 1990) are continuous mapping models that can process sequential patterns and learn representations, and thus may provide an alternative to TRACE. However, it has been suggested that the features that allow SRNs to learn temporal dependencies lead them to work much like the Cohort model (e.g., Marslen-Wilson, 1987), such that items are activated by onset similarity to an input, but not by offset similarity (Norris, 1990). This would make them incompatible with TRACE and with recent results indicating that words that rhyme compete during spoken word recognition (Allopenna, Magnuson and Tanenhaus, 1998). We

between units are preset on the basis of theoretical assumptions. While TRACE, for example, can be criticized as unrealistic in several respects (see Norris, 1994), we find the largest draw-back of interactive activation models to be their obvious

Where models tend to differ is in the set of candidate words predicted to become active. One division that can be made is between alignment and continuous mapping (or continuous activation) models. Alignment models (e.g., Cohort: Marslen-Wilson, 1987; and Shortlist: Norris, 1994) postulate mechanisms which actively seek (or assume) word boundaries. In the Cohort model, candidates are evaluated as to how well they match an input word beginning from word onset. Activations are greatly reduced given mismatches between input and candidate.

Continuous mapping models give no special consideration to word onsets. Instead, items become active as a function of their moment-to-moment similarity to the input, with no explicit penalty for mismatches. The term continuous mapping is potentially confusing. It does not simply mean the model continuously provides an output. For example, TRACE is a continuous mapping model, but effectively becomes an alignment model when its explicit end-of-word ÒsilenceÓ phoneme is used to mark word boundaries.[1] Similarly, while the interactive activation and competition decision level of Shortlist provides continuous output, Shortlist is very much an alignment model, since mismatches are explicitly penalized based on aligning a candidate word with a known word boundary.

One might expect that explicitly searching for word boundaries would be an efficient or even optimal strategy. But consider the variability we experience in using spoken language. We recognize speech in countless circumstances where the acoustics of speech vary tremendously: outdoors, in stairwells, with different talkers, who might have different accents, or who might have just taken a bite out of a hamburger. A recognition mechanism optimized for clear speech (where word boundaries will still be difficult to find) may spend most of its time reanalyzing mis-segmented speech. A system which does not tie itself to word boundaries might prove more robust, since a wider range of possible matches to the input will be considered.

One result of the differences between continuous mapping and alignment models is a contrast in whether or not rhymes are predicted to compete. Both types of model predict that words sharing onsets will compete. Alignment models, because of the emphasis on mismatches, predict that candidates that mismatch at onset will compete weakly only if the initial mismatch is small (with evidence suggesting the mismatch can be no larger than one or two features; e.g., Connine, Blasko &

activated than words sharing onsets, since onset competitors will inhibit the rhymes before the input begins to overlap with the rhymes).

Until recently, the empirical record favored alignment models; evidence for rhyme activation was weak at best. However, Allopenna, Magnuson and Tanenhaus (1998) reported robust rhyme effects using the recently developed Òvisual world paradigmÓ (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In this paradigm, participants respond to spoken instructions to move objects in a visual display, and their eye fixations are measured continuously. Fixations turn out to be tightly time-locked to speech -- at least given a task in which visually guided movements are required (which avoids problems of interpretation raised by Viviani, 1990, since the eye movements have a functional interpretation;

rhyme competitors were present. TRACE activations were transformed into predicted fixation probabilities using a variant of the Luce choice rule (see Allopenna et al.). As TRACE predicts (TRACE accounts for greater than 90% of the variance in each of the critical items), the data indicate that items compete for recognition as a function of their similarity to a stimulus over time, and even substantial initial mismatches do not block rhyme activation (since all of the rhymes differed by more than two features).

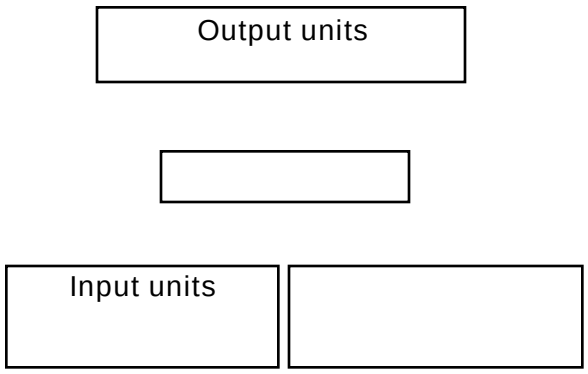
What do SRNs predict? Norris (1990) reported that the performance of SRNs

previous time step would have been influenced by its input from the context layer containing copies of hidden unit activations at time t-2, and so on.

Initially, all trainable weights are set to small, random values. The weights are then modified as each input is presented using backpropagation. Activation from one layer is passed through weighted, trainable connections to the next layer; input and context activations are passed to the hidden layer, and hidden unit activations are passed through weighted, trainable connections to the output units. Output error is computed for each output unit as the difference between a desired output and the actual output. Hidden-to-output weights are changed according to how much of the error was contributed by each weighted connection. Error is propagated back to the hidden layer by assigning each hidden unit a proportion of responsibility for the output error, and changing the incoming weights from the input and context layers accordingly.

For the current simulation using Norris' word list, we proceeded as follows. The network consisted of 18 input units (one for each phonetic feature), 20 hidden and context units, and 48 outputs (one for each lexical item, using a localist representation). Bias units were used for both the hidden and output units, and bias activation was always set to 1. The network was trained for many epochs, with a learning rate of .05. At each epoch, the list of 48 items was randomly ordered. Then each item was presented phoneme-by-phoneme. The network's task at each time step was to indicate the lexical item that was being presented by activating that word's localist output unit, and setting all other lexical units to zero. Context activations were not reset to 0 between words, as is sometimes done with SRNs. Resetting the context weights would effectively make the SRN an alignment model, since an explicit cue to word boundaries would be given.

As in Norris' simulation, we found little co-activation at offset between the reversed cohort pairs which overlapped only in one or two final phonemes. (Note



If we consider items with more complete rhymes, the results are quite different. Of the 48 items, there were 7 rhyme pairs (given in their orthographic forms here): baker/taker, renoroc/tenoroc (coroner/coronet reversed), reviled/timiled (delimit reversed), dish/finish, hsid/ksid (dish/disk reversed), hsinif/raef (finish/fear reversed), and flash/trash. We examined the performance of the network after 10,000 epochs. By this point, the most activated word unit was always the correct item by the last phoneme.

Strong rhyme co-activation was observed for three of the pairs after 10,000 epochs of training (baker/taker, renoroc/tenoroc, and hsid/ksid), and weak activation was observed for trash/flash, dish/finish. The two pairs which did not show even weak co-activation overlapped only slightly in the last syllable, and so the lack of activation is not surprising. Also, there were co-activation effects for these items earlier in training, with rhymes more active than unrelated items. However, prior to the 10,000 epoch mark, not all items were being Òcorrectly identifiedÓ by the last phoneme. We will return to this in the discussion section. The results for the three strong rhyme pairs after 10,000 epochs of training are presented in Figure 3.

There is an asymmetry in each of

the SRNÕs output was similar to that of our original simulation, albeit somewhat noisier. However, by 1000 epochs, rhyme effects disappeared, presumably because the model learned to give more weight to context activations for rhymes, and cohort co-activations nearly mirrored transitional probabilities. This means the SRN had learned the lexicon nearly perfectly, which we will argue later provides a poor analog to the human language processor.

Simulation 2: Allopenna et al. (1998)

Simulation 1 demonstrated that SRNs do predict rhyme activation under certain circumstances. We now turn to the question of how well those predictions match human data, specifically the data from Allopenna et al. (1998) shown in Figure 1.

We used an SRN similar to the one described for the previous simulation, except that it had 23 localist outputs (one for each possible response), 40 hidden and context units, and we used a learning rate of .1.[2] The items we used were phonemic transcriptions of the words beaker, beetle, speaker, carrot, carriage, parrot, candle, candy, handle, pickle, picture, nickel, casket, castle, basket, paddle, padlock, saddle, dollar, dolphin, collar, sandal and sandwich. The training procedure was identical to that for the previous simulation. For each epoch of training, the words were randomly ordered, and then presented phoneme-by-phoneme (using the 18-feature vector representation) to the SRN. The desired output was the current word, and context unit activations were not reset between words.

We chose to examine the model after 1500 epochs of training, because at that point, the correct output node was always the most highly active by the last phoneme of each word, but rhyme and cohort effects were still present. In order to compare the modelÕs performance to the data in Figure 1, we chose all of the target-cohort-rhyme sets in which the target was four phonemes in length (five of the eight sets, with the targets beaker, dollar, pickle, paddle, and sandal). Nearly identical effects were found for the other targets (carrot and candle, of length 5, and casket, of length 6), but we restricted our analyses to 4-phoneme targets because it is not clear how responses to phonemes of different lengths should be combined. We averaged cohort and rhyme conditions for each of the 4-phoneme targets. The average output is shown in the top panel of Figure 4. Target, cohort, and rhyme activations represent the averages across all 4-phoneme sets. The unrelated activation is the maximum value found at each phoneme from any set.

[2] Note that the a wide range of parameters (number of hidden and context units, learning rate, and training epochs) lead to the same result (as evidenced by our successful replication of Simulation 1 with a larger learning rate and smaller number of trianing epochs). For training sets like the one used for this simulation, increasing the number of hidden units allows a smaller learning rate to arrive at the desired performance threshold (Òrecognitionó of targets, i.e., target units having the highest activation at word offset).

A weakness of the current input representations is that entire phonemes are presented in a single time-step. An input representation which allowed more continuous input presentation would clearly provide a better comparison to the human data. In order to compare the current simulation output to the data, we used linear interpolation and extrapolation to fit the simulation output to the data.

There were 30 frames of human data (each frame corresponding to a 33 msec video frame). In order to stretch our four simulation data points, we aligned point 1 to the fifth frame of the human data, which was the frame before any of the fixation probabilities were greater than .01. Then, 5 frames were inserted between each simulation data point. This took us intentionally to frame 20. At the last simulation data point, the rhyme activation has decreased from its peak value. Frame 20 corresponds to a similar point in the human data. From frame 21 to frame 30, we assumed the target probability should rise to 1, and the other values should decrease to 0, as is true for the human data. We then computed correlations between the interpolated simulation predictions and the human data. The r^2 values

clicked on one object, feedback was given by removing all the incorrect choices from the display and repeating the target name).

The lexicon could be divided into four sets of four words. For example, one set was /pibu/, /pibo/, /dibu/, and /dibo/. Each item has one cohort (/pibu/ and /pibo/, /dibu/ and /dibo/) and one rhyme (/pibu/ and /dibu/, /pibo/ and /dibo/). The real advantage of these subsets was the frequency manipulation they allowed. For example, if /pibu/ and /dibo/ (which were not predicted to compete significantly) were presented with high frequency in the learning phase, and /pibo/ and /dibu/ were low-frequency, we would have two different frequency conditions: high-frequency items with low-frequency competitors, and vice-versa. In Magnuson et al. (1998), items were either high- or low-frequency, such that there were four target/competitor frequency conditions: high/high, low/low, high/low and low/high. In Magnuson et al. (1999), a third, Òmedium Ó level of frequency was added, which allowed a crucial test. On some trials, high-frequency targets which had either high- or low-frequency competitors were presented among three unrelated, medium-frequency distractors. If competition effects in the paradigm were driven only by the characteristics of items displayed on a given trial, there should have been no difference in the fixation probabilities to these items, since

would provide good fits). The simulation results are shown in Figure 6. For all four panels, activations are based on simulations using the entire lexicon. Rather than using a variant of the Luce choice rule, as Allopenna et al. (1998) did, to capture the constraints of the subjectsô task, we present these results as evidence that the SRN provides a basis for the major trends of the artificial lexicon studies.

Discussion

The simulations described here show that rhyme effects are

important one would appear simply to be the amount of training; we replicated Simulation 1 even after increasing the learning rate by an order of magnitude. Whether or not we observed rhyme effects depended on when training was stopped. If an SRN is trained until virtually no changes occur in connection weights, and if it has a sufficient number of hidden and context units to represent the temporal dependencies of the input, its outputs will mirror the statistics of the lexicon perfectly. This does not provide a good analog to the human language processor.

There are obviously many differences between the learning situations of our SRNs and a human learner. One is that our SRN always received an input of perfect fidelity (with the exception of context activations at word onsets, which will contain arbitrary information about the ending of the preceding,

lexicon. Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society, 331-336.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. Cognition, 25, 71-102.

Matin, E., Shao, K. C., and Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. Perception & Psychophysics, 53, 372-380.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1-86.

McQueen, J. M., Cutler, A., Briscoe, T. & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. Language and Cognitive Processes, 10, 309-331.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G.T.M. Altmann (Ed.), Cognitive Models of Speech Processing: Psycholinguistic and Computational Persepectives, 87-104. Cambridge: MIT.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, 52, 189-234.

O̒Grady, W., Dobrovolsky, M., & Aronoff, M. (1989). Contemporary Linguistics. New York: St. Martin̒s.

Plaut, D. C., and Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist model. In B. MacWhinney (Ed.), The Emergence of Language (pp. 381-415). Mahwah, NJ: Erlbaum.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. C. (1995).

SciEmeritionVicogni, i Pthe 9(f)-18(33)]TJ0.833 Tc0.1[30.0(pe(Tc62.4)17 I)-w[In(190d) Eyepe
Cogni ., tive Procus,8, &Reviewsge