How Babies Learn to Find Words

Michael Gasser (GASSER@CS.INDIANA.EDU)
Eliana Colunga (ECOLUNGA@CS.INDIANA.EDU)
Computer Science Department, Cognitive Science Program
Indiana University
Bloomington, IN 47405

Abstract

A recent study by Saffran and colleagues (1996) has demonstrated that young infants have a striking ability to rapidly learn the statistical relationships present in sequences of syllables. Eight-month-olds were able to distinguish three-syllable "words" that they had heard previously in a stream of syllables from those that they had not. Since the only evidence available in the input stream is the relative transition probabilities between syllables, the babies have apparently learned to distinguish the probabilities between syllable within words from those between words. In this paper we propose a neural network model which simulates the behavior of the babies in the experiment. The output layer of the network consists of units responding to sequences of syllables, and after training, these units are more highly activated following a word sequence than a non-word sequence. The model also offers an account of how this behavior relates to segmentation and to the learning of more abstract, grammatical regularities in the input.

Background

A recent study by Saffran and colleagues (1996) demonstrated that young infants have a striking ability to rapidly learn the statistical relationships present in sequences of syllables. Eight-month-olds were first presented a two-minute long sequence of syllables in which four three-syllable "words" were concatenated together in random order. Later the babies were tested on word and non-word sequences of three syllables, and the results indicated that they clearly distinguished the two categories. Since the only evidence available in the input stream was the relative transition probabilities between syllables, the babies had apparently learned to distinguish the probabilities between syllables within words from the probabilities between syllables across word boundaries. This ability to find words in the input using statistics will obviously come in handy as the babies are faced with learning real language.

What sort of mechanism could accomplish this task? In this paper we propose a neural network model which simulates the behavior of the babies in the experiment and which offers an account of how this behavior relates to segmentation and to the learning of more abstract, grammatical regularities in the input.

The Task

Sequences

In Saffran et al.'s experiment, the infants are sensitive to differences in transitional probabilities between syllables. A model with this sensitivity must have a means of dealing with patterns in time. Most neural networks which deal with sequential patterns accomplish this through the use of **time delays** on some of the connections in the network. In this way units can respond to the activations that other units had at times in the past, giving the network a form of short-term memory. We will assume that the elements of sequences are evenly spaced and that the delays are multiples of these primitive intervals, which we will refer to as "time steps." Time-delay connections permit a network to transform short sequences, those within the capacity of the short-term memory, into static patterns. For example, if a network has a layer of units which are connected by delays of 0, 1, and 2 time steps to the input units, that layer can represent sequences of 2 or 3 elements as static patterns because it has access to the state of the input units over 3 succeeding time steps. Longer sequences in such a network would be transformed into sequences of static subsequence pattern chunks, each the length of the short-term memory.

Segmentation

The speech perception system develops in infants in order to enable the segmentation of the auditory input stream into words (Jusczyk, 1997), and there is evidence that the basic units they have to work with are syllables (e.g., Jusczyk & Derrah, 1987). While the babies in Saffran et al.'s experiment were not performing any sort of explicit segmentation of the strings of syllables, sensitivity (and attention) to the syllable-to-syllable transition probabilities may contribute to the sort of segmentation which will be required later on for processing sentences. The boundaries between words would correspond to points where the transition probabilities between neighboring syllables are relatively low. Thus, we would argue, there should be a means of tying together the mechanism which learns the statistical properties of the input and the mechanism which segments the input.

Segmentation of an auditory or visual scene requires that different regions in the scene become associated with one another. That is, a mechanism that performs segmentation must have a way of solving the **binding problem**, the problem of tying together subgroups of features in short-term memory when more than one "object" is present in the scene. In a simple neural network, the activation of a collection of units representing object features represents only the presence of those features, not how they are grouped together as distinct objects. Some recent network models solve the problem through the use of some form of synchronization or alignment (Hummel & Biederman, 1992; Shastri & Ajjanagadde, 1993; Sporns, Gally, Reeke, & Edelman, 1989). Units in such a network are outfitted with a dimension of variability in addition to activation, and coincidence along this dimension represents "same object." We will refer to this additional dimension as the "binding dimension." For the segmentation of a sequence of syllables by such a network, the units representing each word would need to be aligned with one another and to be out of alignment with the units representing neighboring words.

In the following section we describe the relevant features of Playpen, a neural network architecture which handles sequential patterns using delay connections and which solves the binding problem with a separate binding dimension. We then discuss a simulation of Saffran et al.'s experiment within the Playpen framework. Finally we consider some implications of the model for the acquisition of the segmentation of speech by infants.

The Model

Units

Playpen (Colunga & Gasser, 1998; Gasser & Colunga, 1998) is a neural network architecture of the generalized Hopfield type (Hopfield, 1984; Movellan, 1990) which is designed to represent and learn relational knowledge and to deal with simple sequential patterns. Here we discuss only those features of Playpen which are relevant for the simulation of Saffran et al.'s experiment. To deal with the binding problem, some units in Playpen vary with respect to their **relative phase angle**, a quantity ranging from 0 to 2π. Relative phase angle plays the role of the binding dimension in the network. Each **micro-object unit** (**MOU**), representing an object feature, has a relative phase angle, and when a group of MOUs settles to a state in which they are all activated and have similar relative phase angles, the network has implicitly assigned the features represented by those units to a single object in the world. Similarly, when MOUs are out-of-phase with one another, the features represented by those units are implicitly treated as belonging to different objects.

Connections

As in other neural networks, the sign and magnitude of a weight on a connection have an effect on the activation of the receiving unit. Unlike most other neural networks, the sign and magnitude of the weight also have an effect on the relative phase angle of the receiving unit. Alongside its activation function, each unit has a coupling function which defines this effect. All else being equal, the sending unit *attracts* the phase angle of the receiving unit via a positive connection and *repels* the phase angle of the receiving unit via a negative connection.

In order to deal with patterns in time, each connection also has a delay associated with it. The network runs in discrete time, with one time step for each input event (one syllable in the case of the simulations reported here). During each time step the network is allowed to settle: the units in the network repeatedly update their activations and phase angles until the state of the network stabilizes.

Connections with delay 0 respond in the usual fashion. A connection with delay $d > 0$ causes the unit at the receiving end to respond to the activation that the unit at the sending end had d time steps before. As shown by Kleinfeld (1986) and others, a Hopfield network augmented with delay connections can learn to reproduce the sequences that it is trained on.

Learning

Learning in Playpen, as in most other neural networks, is Hebbian. Because a network may have hidden units, however, simple Hebbian learning often does not suffice; instead **contrastive Hebbian learning** (Movellan, 1990) is used.[1] Learning takes place in two phases. During the positive phase, the input units are clamped to a pattern sequence, the network is allowed to settle following each pattern element, and learning is Hebbian; that is, the change in weight on each connection for each pattern element is proportional to the product of the activations of the connected units. During the negative phase, no units are clamped, the network is allowed to repeatedly settle for the length of a typical pattern sequence, and learning is anti-Hebbian; that is, the change in weight on each connection is proportional to the *negative* of the product of the activations of the connected units. When the training patterns

[1] As originally formulated by Movellan (1990), contrastive Hebbian learning is a supervised algorithm. We have developed an unsupervised version of the algorithm, and we consider only that version in this paper.

have been learned, the two changes cancel each other out because the network's behavior in the two phases is identical.

Simulation

We simulated Saffran et al.'s task using a Playpen network structured as shown in Figure 1. The Syllables layer consisted of simple units (units with activation but no relative phase angle), one for each of the 12 syllables in the experiment. The Syllable units were connected to three Sequence layers of MOUs, one each for the most recent syllable and for the two syllables that occurred at one and two time steps in the past. That is, the Sequence layers had input connections with delays of 0, 1, and 2 time steps, giving the network an effective short-term memory of three time steps. Within each Sequence layer there was a separate unit for each of the 12 syllables. The weights joining the Syllable and Sequence layers were all positive and did not vary during training. There were trainable connections with delay 0 within and between all of the Sequence layers. These connections were initialized with weights of 0.0. Thus at the beginning of training, the presentation of a sequence of syllables resulted in a series of patterns in the Sequence layers, and for each of these patterns, a single unit was activated in each Sequence layer. The activated units represented the most recent syllable and the two syllables preceding it. The relative phase angles of these units showed no particular pattern because they had as yet no influence on each other.

The Sequence layers were completely connected by trainable, non-delay connections to the Words layer, consisting of 20 MOUs, with small initial random weights. These units were also completely connected to one another with trainable connections, initially with weights of 0.0. At the beginning of training, the activation of a pattern across the Sequence layers (a single activated unit in each layer) resulted in a weak pattern of activation across the Word units. At this point, the relative phase angles of the Word units showed no particular pattern.

As in Saffran et al.'s experiment, there were four distinct three-syllable "words." The words were composed of 12 distinct syllables; each syllable appeared in only one word. Inputs to the network consisted of sequences of four of these words. In these sequences each word followed each other word with equal probability. A sequence of syllables was presented to the network by clamping the units in the Syllables layer in sequence, with one network time step for each syllable.

During training, we expected the Sequence units representing words to be associated with each other by positive weights because they cooccurred frequently. This should lead these groups of three units to tend to activate each other and to align their relative phase angles. Furthermore, we expected each of these three-syllable sequences to be associated with a coherent pattern in the units in the Words layer, and the units in these patterns should also have synchronized relative phase angles. Presented with a 12-syllable sequence consisting of four words, the network should respond with more activation on the Words layer at the end of each word than within words. Presented with sequences of three syllables, it should respond with more activation on the Word units when the sequences constitute words than when they constitute non-words, and the relative phase angles of the Word units should align themselves as they are activated in response to a word. Figure 1 shows an example of what we expected at the end of a word sequence following training. The three units on the Sequence layers representing the last three syllables are in phase with one another and in phase with the activated units on the Words layer.

Training did result in the expected pattern of weights on the trainable connections and the expected response to sequences of words. The trained network was tested on three-syllable sequences, either words from the training set or non-words. Following each sequence, we recorded the total activation of the Word units. This was consistently higher following words than non-words. (The average total activation of the Word units following words was 1.05, while it was only 0.3323 following non-words.) In addition, for each of the words, the activated Word units consistently aligned their phase angles with one another. (The average standard deviation among the phase angles of units activated above 0.15 after the presentation of a word was only 0.0000244.)

Thus the network learned to respond differently to transitions within words than to transitions between words and to treat word sequences as different from non-word sequences. Word sequences not only resulted in greater activation at the Words layer. They resulted in synchronized patterns; that is, the network was treating sequences of syllables constituting words as units.

Conclusions and Future Work

We have shown how the model we described simulates the results of Saffran et al.'s task. In addition, it provides the basis for the segmentation that is necessary for more complex tasks. For segmentation to take place in a network that makes use of synchronization as a segmentation mechanism, the elements within each segment must take on the same value on the binding dimension. This is what we saw for the word sequences in our simulation. In addition, segmentation requires that between the segments the elements *differ* along the binding dimension. This does not take place in our simulation; the phase angles assigned to each group of Word units as a sequence of words is presented appear to be unrelated to one another. However, we believe that the pressure to

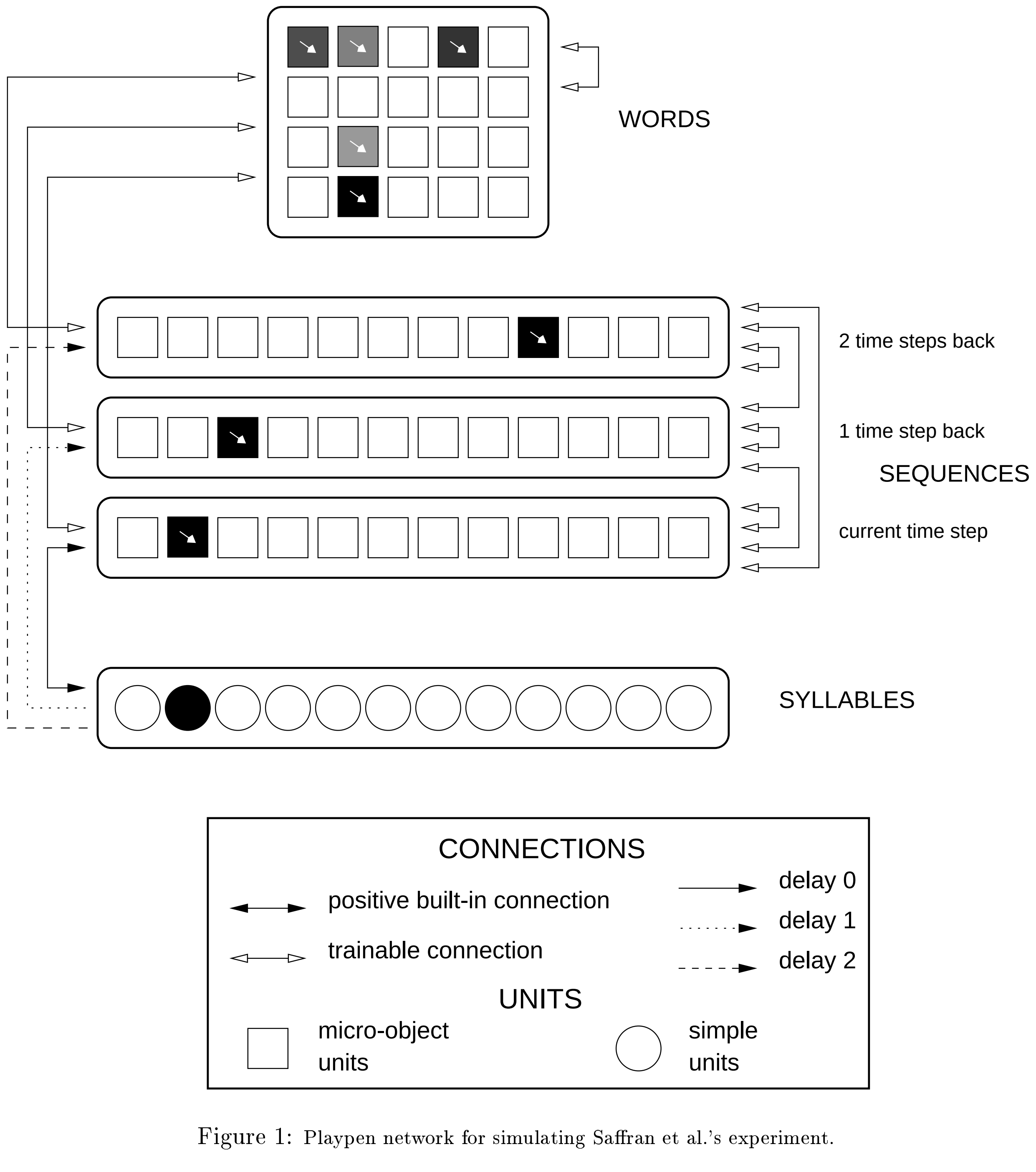

Figure 1: Playpen network for simulating Saffran et al.'s experiment.

perform this second aspect of segmentation would come from another, higher-level task.

Such a task, for example, is that in the related experiment by Marcus, Vijayan, Bandi Rao, and Vishton (1999). In this experiment, seventh-month-old infants were presented with sequences of three syllables separated by gaps. The sequences heard by each baby followed a particular pattern of similarity within the syllables, either AAB, ABB, or ABA. For example, AAB sequences were sequences such as *le le we* and *wi wi je*. Tested later on sequences of *novel* syllables, the babies clearly distinguished those which obeyed the pattern they had been trained on from those which did not. In describing their experiment, Marcus et al. (1999) speak of each of the syllables in their input sequences as "words," the sequences themselves as "sentences," and the task as a "grammatical" task, and they argue that the mechanism required to solve this task is completely different from that required to learn statistical reguarities, as in Saffran et al.'s experiment. In another paper (Gasser & Colunga, 1999), we have described a simulation of Marcus et al.'s experiment with a Playpen network. In addition to the mechanisms in Playpen described here, this simulation requires **micro-relation units**, a means of explicitly representing relational knowledge in a neural network.[2] In this network, the input consists of one-syllable Word units, each of which takes on a particular phase angle in response to built-in connections reflecting similarity between syllables. When the same syllable (word) appears twice in a sentence sequence, the activated Word unit has the same phase angle each time, and the unit for the syllable which differs from the other two is out-of-phase with them. Much as the network used in the present simulation learns to associate the three-syllable word patterns with patterns on the Words layer, the network used to simulate Marcus et al.'s results learns to associate the three-syllable sentence patterns with patterns on a layer of Grammar units.

We believe that these two Playpen networks, the one described in this paper and the one used in simulating Marcus et al.'s experiments, represent two different *levels* of language processing and acquisition. At the lower level, it is statistical correlations between specific syllable types which define the task. On the basis of these correlations, it is possible to distinguish frequently recurring sequences (words) from patterns that either do not occur or are less frequent. At the Words layer, this gives the system units which can provide the input to a higher level. At this higher level these units can be treated as different "objects" when they appear in sequences (the different "words" in Marcus et al.'s task). At the higher level it is gross similarities between objects (words) within sequences (sentences) which complete the segmentation of the sequences and guide learning.

Much remains to be done in tying together the two mechanisms. Most importantly we will need to show how the statistical properties of the task guide the system in treating it as one or the other type of problem; sequences of the type used by Saffran et al. should lead to learning of one type, while sequences of the type used by Marcus et al. should lead to learning of the other type. What we have described in this paper is a beginning, however. Within a single general associationist framework — a simple settling network augmented with delay connections, a binding dimension, and a means of representing relations explicitly — we have modeled what appear to be two very different types of language acquisition tasks.

References

Colunga, E. & Gasser, M. (1998). Linguistic relativity and word acquisition: a computational approach. *Annual Conference of the Cognitive Science Society*, *20*, 244–249.

Gasser, M. & Colunga, E. (1998). Where do relations come from?. Tech. rep. 221, Indiana University, Cognitive Science Program, Bloomington, IN.

Gasser, M. & Colunga, E. (1999). Grammatical rule learning in a connectionist network. Submitted to Boston University Conference on Language Acquisition, 1999.

Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*, 3088–3092.

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480–517.

Jusczyk, P. W. & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*, 648–654.

Jusczyk, P. W. (1997). *The Discovery of Spoken Language.* MIT Press, Cambridge, MA.

Kleinfeld, D. (1986). Sequential state generation by model neural networks. *Proceedings of the National Academy of Science*, *83*, 9469–9473.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77–80.

Movellan, J. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In Touretzky, D., Elman, J., Sejnowski, T., & Hinton, G. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*, pp. 10–17. Morgan Kaufmann, San Mateo, CA.

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

Shastri, L. & Ajjanagadde, V. (1993). From simple associations so systematic reasoning: a connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, *16*, 417–494.

Sporns, O., Gally, J. A., Reeke, G. N., & Edelman, G. M. (1989). Reentrant signaling among simulated neuronal groups leads to coherency in their oscillatory activity. *Proceedings of the National Academy of Sciences*, *86*, 7265–7269.

[2] We have motivated micro-relation units independently elsewhere (Gasser & Colunga, 1998).