Learning Distributed Representations for Syllables*

Michael Gasser
Departments of Computer Science and Linguistics
Indiana University
Bloomington, IN 47405
gasser@cs.indiana.edu

Abstract

This paper presents a connectionist model of how representations for syllables might be learned from sequences of phones. A simple recurrent network is trained to distinguish a set of words in an artificial language, which are presented to it as sequences of phonetic feature vectors. The distributed syllable representations that are learned as a side-effect of this task are used as input to other networks. It is shown that these representations encode syllable structure in a way which permits the regeneration of the phone sequences (for production) as well as systematic phonological operations on the representations.

Linguistic Structure and Distributed Representation

If the language sciences agree on one thing, it is the hierarchical nature of language. The importance of hierarchical, structured representations is now generally recognized for the phonological pole, where syllables and metrical units now play a major role (see, e.g., Frazier (1987) and Goldsmith (1990)), as well as for the syntactic/semantic pole of language and language processing. The major reason for believing in structured representations is the significance of structure-sensitive operations in language processing. A semantic inference rule may need to know where the subject of a clause is; a morphological reduplication rule may need to know where the coda (final consonant(s)) of a syllable is.

Traditional symbolic representations are based crucially on the simple notion of **concatenation** (van Gelder, 1990). A syllable representation, for example, is a (bracketed) string of concatenated phones. Recent connectionist work offers as an alternative to this widely accepted approach **distributed** representations, for which it is generally impossible to isolate which elements of the representation denote which of the lower-level units comprising the structure being represented.

What good are distributed representations? They certainly are harder to interpret directly, at least by external “users” of the system that creates them. And at first blush it seems cumbersome, if not impossible, to implement structure-sensitive operations on them, operations which present no particular difficulty for symbolic representations (Fodor & Pylyshyn, 1988). Clearly distributed representations would be useless for most purposes if they were not amenable to such operations. Recently, however, it has been shown that it is possible to arrive at a set of connection weights which implements structure-sensitive operations on distributed representations. Where the representations arise on hidden layers through training, the operations on them are also implemented through training (Chalmers, 1990). Where the representations arise as a result of the application of a set of primitive operations analogous to the filling of roles in symbolic models, the operations on them can be implemented more directly (Legendre, Miyata, & Smolensky, 1991).

There are three reasons to prefer distributed over symbolic representations for structured objects such as syllables and sentences.

1. Distributed representations do not necessarily increase in size as the complexity of the represented object increases. In the case of some types of representations, for example, those described in this paper, representations for objects of the same type are of fixed width (Pollack, 1990). This seems more important for syntax/semantics than for phonology, where there is apparently no recursive embedding, but in a learning context, it is a desirable feature for phonological representations too since a system cannot be expected to know beforehand how complex the representations will need to be and therefore how much memory to allot to them.
2. Complex transformations can be performed on distributed representations in a single parallel step, rather than through a series of symbolic `cons`es, `car`s, and `cdr`s (Legendre et al., 1991).
3. There are relatively simple algorithms for **learning** the structure in distributed representations (Elman, 1990; Pollack, 1990).

Most work concerned with distributed representations for structured objects has examined syntax or semantics. It remains to be shown whether it is possible to learn distributed syllable representations which embody the structure required for phonological operations of various sorts. This is in part what this study seeks to establish.

* In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (1992), 396–401

Linguistic Structure and Time

Language takes place in time: input to hearers and output from speakers is sequential. If linguistic knowledge is organized hierarchically, at least part of what hearers do in perceiving language must consist in taking in sequences of elements at one level and classifying them as belonging to a single unit at a higher level. Something temporal is turned into something static. In this sense a syllable is a static **summary** of a temporal sequence of phones. Speakers in turn carry out the reverse process: they turn static representations into temporal sequences. Given a syllable representation, they must unpack it into its component onset (initial consonant(s)) and rime (remaining segments). The sorts of syllable representations we seek should be accessible via the categorization that takes place during perception and should be expandable into their component elements during production.

The temporal nature of language is related intimately to the issue of short-term memory. The process by which a sequence of elements at one level is recognized as a single element at a higher level requires access to more than just a single element at a time; a context is necessary. The production of a sequence of elements, given a higher-level summary representation as input, requires as a context some representation of what has already been produced.

One approach to short-term memory is to give a system access to a buffer of some fixed width. This has several drawbacks, in particular the problem of how the system is to know beforehand how wide the buffer should be (Port, 1990). An alternative is an approach that permits a system to develop its own short-term memory. This is possible in connectionist networks with recurrent connections (Elman, 1990; Jordan, 1986; Port, 1990). It is this method that is utilized in the study described here.

The Learner's Task

Language acquisition begins with perception, so we expect the representations for syllables and other prosodic units to result from perceptual processes. There are several possibilities for how this might happen, though the most reasonable is probably some combination.

1. The hearer/learner may be learning phonology for its own sake, that is, either simply looking for regularity in the input, or looking for evidence that would allow the setting of some innate parameters (Dresher & Kaye, 1990).
2. The hearer/learner may be attempting to map perceptual features onto representations of articulatory gestures, as in various versions of the motor theory of perception (Liberman & Mattingly, 1986).
3. The hearer/learner may learn prosodic representations as a side-effect of word recognition.

It is the third possibility that is pursued here. The idea that phonology emerges as the child learns to recognize and produce words is an appealing idea, and an old one. It is based on the notion that phonology is not just arbitrary patterning, but rather a phenomenon with functions for the language processing system: to facilitate word recognition and to organize word production. According to the third view in the list above, the child acquires phonological representations in the context of using them.

Consider the relationship between the acquisition of word recognition and the acquisition of syllable structure. In learning to distinguish an initial subset of the words in the target language, a learner is provided with relatively direct information about the distinctiveness among a sizable subset of the possible syllables in the language. Because the syllables are contrastive units, the learner is forced to distinguish them in order to distinguish the words. The question addressed here is whether the word recognition task suffices to develop representations which support phonological operations.

A human learner/hearer is presented with unsegmented, continuous input. The task of the system studied here is a considerably simpler one: its input consists of sequences of phones, each in the form of a phonetic feature vector. The phones appear one at a time, and the internal state of the system on the previous time step provides the necessary context for recognition. The system's initial task is simply to assign sequences of phones (representing words in the language) to lexical categories. As a side effect of performing this task, it develops internal representations for various subsequences making up the words, in particular for the syllables in the language. These subsequence representations can then in turn be investigated by treating them as inputs to components with other tasks. Two further tasks are dealt with here: the transformation of a static sequence representation into the sequence of phones it represents (the production task), and the systematic mapping of one sequence representation onto another. In both cases, what is of interest is whether the sequence representations permit **generalizations** to be made. That is, trained on a subset of the sequence representations, does the system respond to others on which it was not trained?

The Approach

The networks used in the study described here are **simple recurrent networks** of the types first investigated by Jordan (1986) and Elman (1990). They consist of feedforward networks supplemented with recurrent connections from the hidden and/or output layers and are trained using the familiar back-propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986). Figure 1 shows the architectures of the networks used for the recognition and production tasks in the experiments described below. Earlier experiments indicated the superiority of these particular architectures over other variants of simple recurrent networks for these tasks.

The recognition network is presented with a sequence of phones, one at a time, each phone consisting of a vector of phonetic features. Among the features is **sonority**, which tends to correlate with proximity to the nucleus of a syllable. Each sequence ends with a boundary symbol, represented by an input pattern consisting entirely of zeros. The network is trained to auto-associate the input phone pattern (that is, simply to copy it to a set of output units), and to categorize the input sequence as belonging to one of a set of morphemes in the language. The auto-association task, while not

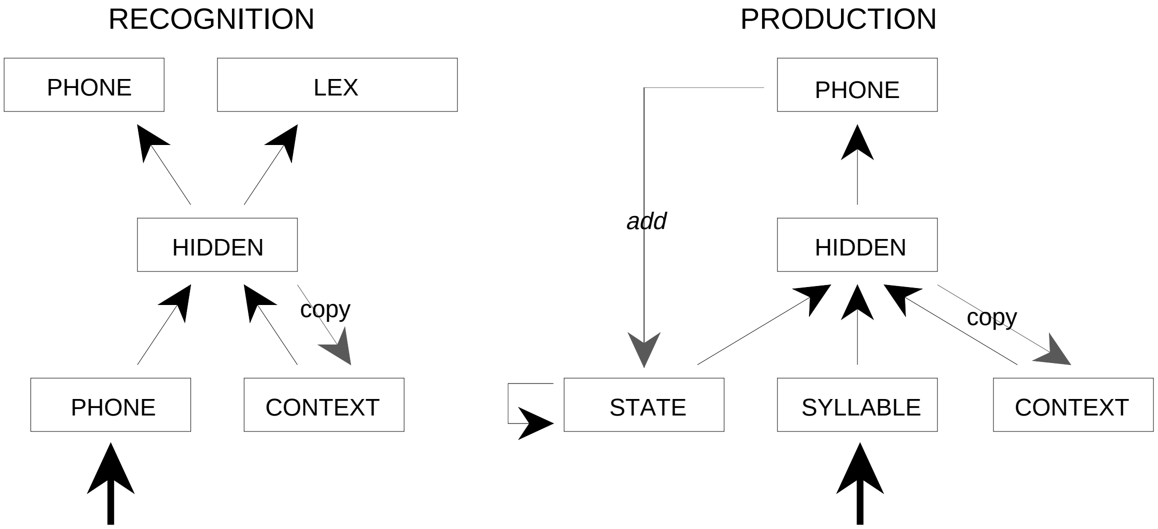

Figure 1: Network Architectures

directly related to word recognition, has the effect of forcing the network to distinguish the phones making up the sequences. The network is provided with targets for both the auto-association and recognition tasks. The lexical target remains constant throughout the presentation of the sequence. Via recurrent connections the network also has access to a copy of its hidden layer on the previous time step. A distributed syllable representation, to be used as input to other networks, is obtained by presenting the sequence of phones making up the syllable followed by a boundary symbol and saving the pattern which appears on the hidden layer at the end of this sequence.

The production network takes a distributed syllable representation (from the recognition network) as input. This remains constant throughout the production of the sequence. The network is trained to output, one at a time, the phones making up the sequence followed by a boundary pattern following the sequence. Each phone takes the form of a feature vector, identical to the pattern used as input to the recognition network. Targets are provided for each of the output phones. The production network has recurrent connections on both the hidden and output layers. The output pattern is added to a decayed version of the previous sequence of outputs and sent to the network as part of its input (on the STATE layer).

Experiments

Stimuli

Stimuli for the experiment consisted of phones and phone sequences in an artificial language. Phones were represented by vectors of 11 phonetic features. Possible syllables in the language are characterized as follows:

onset → {0,p,f,m,t,s,n,k,x}

nucleus → {i,e,a,o,u}

coda → {0,n,s}.

Thus there were 135 possible syllables in all.

Procedures

Each experiment began with the training of a recognition network to categorize a set of words in the artificial language. Each word consisted of two legal syllables in the language, and the set of words was generated by randomly combining pairs of syllables, with the restriction that no identical pairs were included. Once the recognition network had been trained on the words, representations for each of the 135 syllables in the language, consisting of hidden layer patterns following the presentation of the syllable sequences, were extracted from the network. These syllable representations were then used as inputs to other networks.

Experiment 1

First 100 two-syllable words were generated. This resulted in a set which contained 104 of the 135 possible syllables in the language. Next the recognition network was trained to identify the phone sequences representing the words. Previous experiments have shown that word recognition training on a relatively large set is more effective if the words are introduced gradually to the network rather than all at once, an idea inspired by the regimen used by Plunkett & Marchman (1991) to train a network to learn English past tense forms. Three new words were introduced to the training set each time the mean square error per pattern for the current training set dropped below 1.0. Training continued for 600 repetitions of the training set (43,048 words), by which time all 100 words had been introduced to the training set.

Performance on word recognition at this point was far from impressive. Only 17 of the 100 words were correctly identified at the point where the final word boundary was presented. Still it was felt that in attempting to learn to distinguish the words, the network might have developed distinct representations for the syllable sequences that made them up. Representations for all 135 possible syllables were set aside by presenting the network with the phone sequences and then saving the final pattern on the hidden layer. The hidden layer of the recognition network, and hence the width of the distributed syllable representations, was 25 units.

Next these syllable representations were used as inputs to a production network. 20% of the syllables were randomly selected to be set aside for testing the network for generalization. These included sequences which had been parts of the words in the original recognition training set and others which were not included in the set. The production network was trained to output each syllable sequence followed by a boundary symbol. Training continued for 110 repetitions of all patterns, at which point the network made errors on 7 of the 384 segments making up the training syllables. Errors were made on 7 of the 95 segments in the test sequences. Only one of these segments was one which did not lead to a legal syllable in the language.

These results indicate that the recognition network is able to generalize about syllable structure on words containing a subset of the possible syllables and that the distributed representations developed during training can be used for production as well. The fact that the errors made are reasonable ones indicates that the representations are encoding syllable structure in a systematic way.

Next the trained recognition network was presented a representative set of 142 bogus syllables, sequences which did not conform to the language the network had been trained on. These included sequences with phones not among the phoneme inventory of the language (e.g., *b* and *d*), sequences with illegal codas (e.g., *fap*), sequences with long nuclei (e.g., *mua*), sequences with cluster onsets, and sequences with no nuclei. The hidden-layer representations for each of these sequences were saved and presented to the trained production network. The output of the production network was then examined to determine whether the networks would in effect correct the representations. The production network responded to 97 of the 142 sequences (68%) by replacing the original sequence with a legal syllable in the language. Typical responses included the following: *kn* → *ken*, *kfe* → *ke*, *xou* → *xu*, *pik* → *pi*, *zan* → *nan*.

These results are further evidence that the recognition and production networks have learned about the structure of syllables in the language. They also indicate that the representations are robust.

Experiment 2

Finally, the syllable representations from the recognition network were used as inputs to simple feedforward networks which were designed to determine whether the representations could be used for phonological transformations. Each feedforward network took as input a syllable representation and yielded as output the syllable representation that resulted when applying a particular rule to the input syllable. Three rules (and three networks) were used: a rule which replaced the vowel in a syllable with *u*, a rule which made the coda of the syllable -*s*, and a rule which replaced the onset of the syllable with the fricative in the same place of articulation as the onset of the original syllable (or by *s* if there was no onset).

Each network was trained on 80% of the syllables until there were no errors, then tested on the remaining 20%. Training required about 25 repetitions of all of the patterns. The network's response was taken to be that syllable (of the 135 possible) whose distributed representation was closest (in Euclidian distance) to the network's output pattern. For each rule, over 95% of the test syllables were generated correctly. In all cases errors resulted in syllables which satisfied the basic constraint imposed by the rule in question (*u* nucleus, *s* coda, fricative onset).

These results indicate that the syllable representations learned by the recognition network encode syllable structure in a way which makes it accessible to the sorts of operations which are common in the phonological systems of natural languages.

Discussion

The experiments reported on here demonstrate that simple recurrent networks can be trained to develop representations of syllables which encode information about structure in a distributed form. These representations present a viable alternative to traditional concatenative types of representations. Like their symbolic counterparts, the distributed representations can be unpacked into the sequences they represent and can be transformed in systematic ways. Unlike their symbolic counterparts, the distributed syllable representations are learned; are of fixed width; and permit parallel, single-step operations.

There are at least two other connectionist approaches to the acquisition of syllables. Goldsmith & Larson (1990) model the syllabification of words in a variety of languages using a constraint satisfaction network in which units represent segments in the word and activations represent the "derived sonority" of the segments, an indication of their role in the syllabic structure of the word. Two simple parameters characterize syllabification in each language. The model provides an elegant account of a range of phenomena, but it is not clear what it has to do with processing since what is modeled is abstract, atemporal derivation. It is also not specified how a language learner might have access to the derived sonorities needed to learn the parameters.

More in the spirit of the present approach is an experiment by Corina (1991), in which a simple recurrent network was presented with sequences of phonetic segments from a database of spoken English utterances. Trained simply to predict the next segment, the network showed clear evidence of having discovered the statistical regularities that characterize the structure of the English syllable. That is, its output predictions corresponded closely to the actual probabilities of particular segment classes in particular positions. This is evidence that a network can also learn about syllable structure from training on an unsupervised task. It remains to be seen whether the hidden layer patterns from Corina's network are suited for recognition and production or whether there is anything to be gained by combining the supervised recognition and unsupervised prediction tasks.[1]

How might the syllable representations learned in the network fit into to a more complete model of word recognition and production? I noted above that the recognition network was not especially successful in learning to distinguish the 100 words it was trained on. As the number of words to be recognized increases to more plausible ranges, we can expect very serious degradation in this capacity, though increasing the hidden layer size would offset the degradation to some extent. Yet the problem might go away in a hierarchically organized system with simple recurrent networks operating with different units as inputs. Word recognition might then be a process of assigning sequences of syllables and/or larger metrical units to word or morpheme units. Thus the syllable representations learned in the network described here would provide the input to a syllable-level network. See Gasser (1991) for more on this proposal.

From the perspective of its plausibility as a model of phonological acquisition, the present model has a number of inadequacies and gaps. First, I have only scratched the surface in terms of what might be required of such a model. How, for example, might this approach account for learning how to assign stress to novel words (in a language which does this in a non-arbitrary way)? Recently, Gupta & Touretzky (1991) have shown that perceptrons can learn to assign stress to syllable sequences from 19 natural languages (apparently encompassing the range of possible stress systems). The present approach would attempt to achieve this in the context of the hierarchical architecture referred to above, by training a sequential network which takes distributed syllable representations (one at a time) as input to recognize words involving one or more metrical units (sequences of stressed and unstressed syllables). The hope would be that distributed representations for these units, and eventually for the entire words, would arise, and that these would provide the input to the word production process, where stress assignment takes place. While considerably more involved than the approach of Gupta & Touretzky (1991), this would respect the sequential nature of language and maintain the relationship between word recognition and phonological learning.

A further weakness of the framework in its current state involves the learning of production. While the learning of syllable representations as a side-effect of the process of word (or morpheme) recognition seems reasonable, the learning of the reverse process is another matter. The network trained on the production task was provided with targets for each output phone, a degree of supervision that clearly does not correspond to anything in the experience of the human language learner. For now it may be best to view this task as nothing more than an existence proof that the representations can be unpacked for production or alternatively a technique for analyzing the distributed representations, which, unlike their symbolic counterparts, are not directly interpretable. Of course, the issue of how children learn to produce, as well as perceive, linguistic forms, when they are not provided with targets, is one facing any approach to language acquisition.

Finally, the present approach presupposes some mechanism for segmentation, first, at the level of the phones that are the inputs to the recognition process, and second, at the level of the syllables (or words) themselves. Again, segmentation is a problem for all sorts of acquisition models. Recently Doutriaux & Zipser (1990) have had some success in training simple recurrent networks to discover segments in speech. Thus this seems to be a problem that can be approached within the framework outlined in this paper.

References

Chalmers, D. (1990). Syntactic transformations on distributed representations. *Connection Science*, *2*, 53–62.

Corina, D. P. (1991). *Towards an Understanding of the Syllable: Evidence from Linguistic, Psychological, and Connectionist Investigations of Syllable Structure*. Ph.D. thesis, University of California, San Diego.

Doutriaux, A. & Zipser, D. (1990). Unsupervised discovery of speech segments using recurrent networks. In Touretzky, D., Elman, J., Sejnowski, T., & Hinton, G. (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*, pp. 303–309. Morgan Kaufmann, San Mateo, CA.

Dresher, B. E. & Kaye, J. D. (1990). A computational learning model for metrical phonology. *Cognition*, *34*, 137–195.

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

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *28*, 3–71.

Frazier, L. (1987). Structure in auditory word recognition. *Cognition*, *25*, 157–187.

Gasser, M. (1991). Sequence comparison and simple recurrent networks. *Center for Research in Language Newsletter*.

Goldsmith, J. & Larson, G. (1990). Local modeling and syllabification. In Deaton, K., Noske, M., & Ziolkowski, M. (Eds.), *Papers from the 26th Annual Regional Meeting of the Chicago Linguistics Society: Parasession on the Syllable in Phonetics and Phonology*. Chicago Linguistics Society.

Goldsmith, J. (1990). *Autosegmental and Metrical Phonology*. Basil Blackwell, Cambridge, MA.

Gupta, P. & Touretzky, D. S. (1991). Connectionist networks and linguistic theory: investigations of stress systems in language.. Unpublished report, Carnegie-Mellon University.

[1] In some preliminary experiments, I have not found better performance on word recognition from networks which are also expected to predict their next sequence.

Jordan, M. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pp. 531–546 Hillsdale, New Jersey. Lawrence Erlbaum Associates.

Legendre, G., Miyata, Y., & Smolensky, P. (1991). Distributed recursive structure processing. In Lippmann, R. P., Moody, J. E., & Touretzky, D. S. (Eds.), *Advances in Neural Information Processing Systems 3*, pp. 591–597. Morgan Kaufmann, San Mateo, CA.

Liberman, A. M. & Mattingly, I. G. (1986). The motor theory of speech revised. *Cognition*, *21*, 1–36.

Plunkett, K. & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition*, *38*, 1–60.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*, 77–105.

Port, R. (1990). Representation and recognition of temporal patterns. *Connection Science*, *2*, 151–176.

Rumelhart, D. E., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. E. & McClelland, J. L. (Eds.), *Parallel Distributed Processing, Volume 1*, pp. 318–364. MIT Press, Cambridge, MA.

van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cognitive Science*, *14*, 355–384.