Connectionist modelling of phonotactic constraints in word recognition

Joe Levy, University of Edinburgh Human Communication Research Centre,
2 Buccleuch Place, Edinburgh EH8 9LW, UK. e-mail joe@cogsci.ed.ac.uk

Richard Shillcock, University of Edinburgh Centre for Cognitive Science,
2 Buccleuch Place, Edinburgh EH8 9LW, UK. (e-mail: rcs@cogsci.ed.ac.uk)

Nick Chater, University of Edinburgh Department of Psychology,
7 George Square, Edinburgh UK. (e-mail:nicholas@cogsci.ed.ac.uk)

Abstract

Connectionist techniques for modeling the temporal statistics of phonemically transcribed spoken discourse are described. The aim is to investigate the limits of modeling psycholinguistic data at this pre-lexical level. The training data respect the frequency with which phoneme strings occur in conversational speech. The general model proposed uses a back propogation through time learning procedure to train a network that can predict the identitiy of the phoneme at the next time step, identify the current one and confirm the last five, after training on noisy data. The model eschews local representations of words and will have implications for current models of word recognition which employ such representations.

Introduction

Although most connectionist methods deal with static patterns, there has been much recent work on networks that can extract temporal structure from their input sequences (Elman 1990, Norris 1988; also see Hertz et al. 1991). The work reported here experiments with a class of these methods in an attempt to account for psycholinguistic data concerning word recognition and phoneme identification, using a network that can extract the temporal statistics from a continuous stream of (idealised) phonemically transcribed speech. In general we believe that there is a methodological imperative to investigate exhaustively the role of pre-lexical representations, and their low-level statistics of co-occurence, before higher level representations are invoked in modelling psycholinguistic data. The models presented here do not include lexical entries, i.e. local representations of the phonology of a particular word.

This paper will discuss some of the available algorithms and architectures and outline the type of data we seek to model. We will briefly describe some of the simulations of psycholinguistic data.

Connectionist approaches to learning temporal sequences

Of the many recent connectionist approaches to learning temporal structure, one of the easiest to implement and least computationally expensive is that described in Elman (1990) and Norris (1988). The back-propagation algorithm is given a limited ability to capture temporal sequence by "copying back" the activation of the hidden units on the last time step to a set of context units on the input layer. This method is a simplification of the more general method of "back propagation through time" (Rumelhart, Hinton and Williams 1986). One of our chief concerns is the extent to which this simplification weakens the ability of a network to extract high order statistical regularities (see Chater 1989). We are currently comparing the performance of the "copyback" method with full and truncated back propagation through time. Even the simplest copyback architecture has proved more adequate than we had predicted (Shillcock, Levy and Chater 1991).

CH 3065-0/91/0000-0101 $1.00 © IEEE

Models of lexical access

Speech sounds arrive over time and must be matched against some kind of stored representation. Psycholinguistic theories have employed various idealised levels of representation involving such entities as features, phonemes, morphemes and words. There has been a "lexicalist-localist" tradition in which the incoming signal is seen as directly contacting specific lexical representations: it is assumed that contacting the lexical entries for words makes available all of their associated information, in particular a complete phonological description. The contacted word(s) partly determine the sublexical processing of the input.

Within this approach, little has been said about the development of the representation which makes initial contact with the lexical entry. One computationally explicit account of the development of pre-lexical and lexical representations, the TRACE model (McClelland and Elman 1986), has captured many aspects of human spoken word recognition in a principled way and represents a coherent stance on issues such as constraining the activation of lexical representations and segmenting the continuous input. In contrast to TRACE, the model discussed in this paper employs the full range of phonemes in a description of spoken English, is capable of learning and does not involve local lexical representations.

The model advanced below does not distinguish at the input and output levels between representations of any frequent sequence, whether they are specifically morphemes, syllables, words or idioms. The model builds on a recent departure from the lexicalist-localist view of lexical access, involving a distributed connectionist model of word pronunciation (Seidenberg and McClelland 1989). Seidenberg and McClelland's model has achieved considerable coverage of the relevant psycholinguistic data. In the model of spoken word recognition described below, an analogous approach is taken within the auditory domain: feature-level representations are mapped onto phoneme-level representations. The training regime is taken from spoken discourse and reflects the frequency with which speech sounds corresponding to phonemes occur and co-occur in spoken language.

Our perspective does not rule out the possibility that explicit specific lexical representations might be necessary to account for certain data. Only after investigating exhaustively how much of the data can be accounted for by a model which does not possess such representations can the role of lexical representations in explaining psycholinguistic data be properly assessed.

A second computational model of spoken word recognition has been presented by Norris (1988), employing the copyback architecture mentioned above. In Norris' model feature-level representations of consecutive phonemes are mapped onto local representations of words via one layer of hidden units, a copyback mechanism giving the network the potential to respond to patterns of input across time.

Norris's model learns from its training set the frequency of the words it can recognize, and captures a range of human behaviour. In small scale simulations a spread of activation is generally assigned to all words congruent with the input up to and including the current phoneme. When the input word becomes unique the model generally opts overwhelmingly for that word and maintains its level of activation until the end of that word in the input. In contrast, the model described below instantiates a more comprehensive phoneme level description of spoken English, is potentially able to model data concerning infra-lexical processing (e.g. phoneme-monitoring) and, again, stops short of local lexical representations.

The general model

We aim to account for psycholinguistic data concerning word recognition using the temporal statistics of phoneme sequences derived from spoken discourse data. Connectionist nets provide a powerful way of extracting these statistics and implementing an explicit computational model. Such a model might be expected to have a degree of predictive ability as well as one of confirming past hypotheses about its input data. Our current model has seven groups of output units representing a temporal window of seven phonemes. The groups correspond to the prediction of the next phoneme, the identification of the current phoneme and the confirmation of the previous five phonemes. The confirmatory units constitute an "active memory" of the recent past input. Rather than act as a simple delay line we would like to see them display effects of "right context", perhaps correcting past mistaken hypotheses when more reliable current information is

received.

The input to our model is a vector representing the binary phonetic features of the current phoneme. At the moment we are using an encoding consisting of 11 features. Figure 1 illustrates the structure of our general model. The output of the model consists of sub-vectors specifiying the identity of the next phoneme, the current phoneme and the previous five phonemes. Each sub-vector contains a unit for each of the 36 phonemes we use. Thus, in the full case the output layer contains 252 units. In most of our simulations we have used 15 hidden units. In the simulations using the Elman/Norris architecture, this entails the use of 15 context units on the input layer.

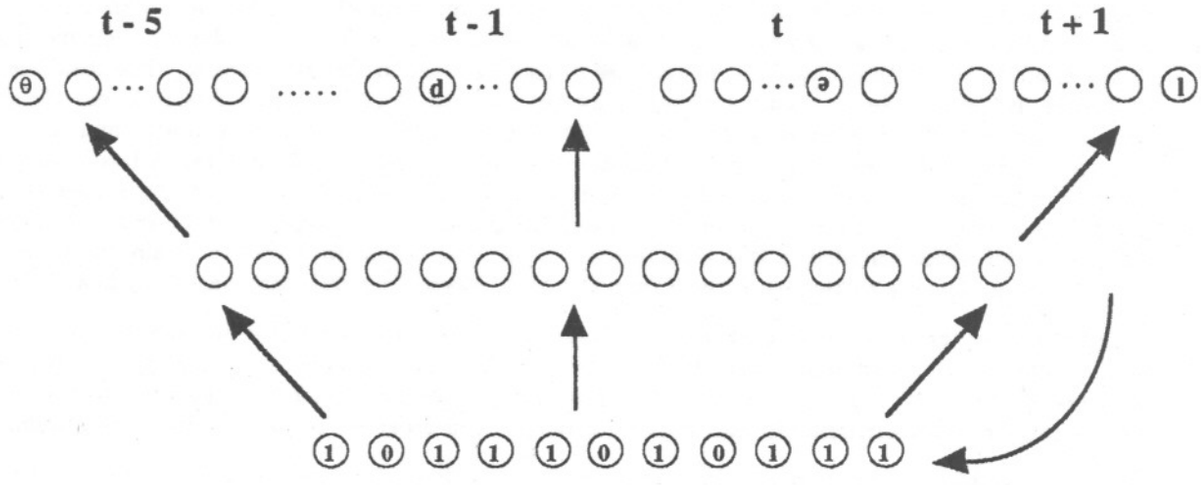

Figure 1: A schematic view of the general model. The input vector consists of 11 units representing the values of binary phonetic features specifying the current phoneme. Activation flows to a hidden layer consisting of 15 units. Recurrence is implemented by a "copy-back" or back propagation through time method. The output layer consists of seven groups of 36 units. Each group represents a separate time step from $t-5$ to $t+1$. The units within each group represent the evidence for the different phonemes at each time step. The example here displays the units that might be activated most strongly for the last seven phonemes from the phonetic transcription of "the model".

Previous work (Shillcock, Levy and Chater 1991) confirmed the promise of this approach using a cut-down version of our general model consisting of only three output slots (previous, current and next phoneme).

The full network (252 output units) is large and training it with the large amounts of data needed to expose it to a representative statistical sample of linguistic input is computationally very expensive. Using an algorithm such as back propagation through time effectively enlarges the network many times making training times even longer.

Input coding

The transcribed words constituting the training data were converted to idealized phoneme-level descriptions using a text-to-phoneme program, developed by the University of Edinburgh Centre for Speech Technology Research (CSTR), and employing 36 different phonemes based on those of the CSTR Machine Readable Phonetic Alphabet. The eight dipthongs were each converted to sequences of two phonemes. This phonemic transcription was then converted to an idealized feature-level representation, consisting of the following 11 features based on those of Jakobson, Fant and Halle (1952): vocalic/non-vocalic, consonantal/non-consonantal, voiced/unvoiced, discontinuous/continuant, strident/mellow, nasal/oral, diffuse/non- diffuse, compact/non-compact, tense/non-tense, grave/acute, flat/plain. Thus the phonemes schwa and /I/ were represented respectively as below.

schwa	1	0	1	1	0	0	0	0	0	1	0
/I/	1	0	1	1	0	0	1	0	0	0	0

All 36 phonemes were given a value of 1 or 0 for each of the 11 features. The final form of the training set was a continuous stream of feature-level descriptions of segments, with no word boundary information.

Training regimes

One of our main concerns is that the data presented to the networks during their training reflects the statistical nature of real spoken language. As far as possible the frequency of different phoneme sequences in the training set should match the frequencies of those strings in spoken dialogue. The ideal method of achieving this is to take a corpus of spoken language (e.g. the LUND corpus, Svartvik and Quirk 1980) and transcribe it into the appropriate input coding for the network. This is laborious and it is difficult to transcribe enough material to represent relatively infrequent phoneme sequences. A less ideal but more practical alternative is the approach taken by Seidenberg and McClelland (1989). They sampled words from a dictionary with a probability proportional to their frequency in the language. Each epoch consisted of a different set of words sampled in this way. In some of our simulations we have used a similar method using a 33,000 word phonetic dictionary derived from the MRC Psycholinguistic Database (Coltheart 1981).

In order to force the networks to capture as much temporal structure as possible we have usually added noise to the training set in our simulations. Usually this took the form of randomly changing a feature value of, on average, one of the binary features in the input vector. This should tend to make the network consider the temporal context surrounding each particular time-slot in order to reduce its error during training.

In order to avoid the network overfitting the data, we set aside a portion of the original transcribed data to use for cross-validation. We stop training when the error on the cross-validation set seems to have stopped decreasing. This decision can sometimes be difficult to make, especially when noise has been added to the training set. The method is convenient however, since it allows us to estimate the appropriate number of hidden units without having to find the ideal minimum number for generalisation empirically.

A further concern is the use of an appropriate learning algorithm. Although we have achieved some success with the Elman/Norris approximation to back-propagation through time, we are concerned that such approaches may be limited in the richness and extent of temporal structure they can capture. Our current simulations use a truncated version of back propagation through time where the current time step and the four previous ones are used during training. The closer a training regime approaches full back propogation in time the more computationally expensive it becomes. We plan to compare the degree of structure captured by different regimes such as the Elman/Norris architecture and full and truncated back propagation through time.

Modelling psycholinguistic data

Using a network architecture where the output layer represented the last, current and next phoneme, Shillcock et al. (1991) reported two successful simulations of psycholinguistic data from phoneme restoration and phoneme monitoring tasks. Here we extend this work to the seven time-slot architecture.

Capturing phonotactic constraints

Simulations using an Elman/Norris algorithm have shown that the expansion of the previous work to a model with a five-slot buffer is feasible. When this model is given an input corresponding to the phonemically transcribed phrase "this is a test of the model", the mean activations of the predictive unit, the current unit and the five confirmatory units are all significantly greater than when the same phonemes are presented in random order. The network is sensitive to the phonemic context both before and after a specific phoneme; the five confirmation units were not simply functioning as delay lines.

Phoneme restoration

Listeners' perception of degraded individual speech sounds in words is often restored: if the /s/ in legislature is replaced by white noise, the word is still heard as intact, but with a cough perceptually displaced from the word (Warren 1970). This was modelled by putting test words in the carrier sentence "and the next word is ... and the next word is" and observing activation levels. Restoration effects are limited at this stage of the development of the model. Even for frequent words like *got*, *this*, and *yes*, the phonotactic knowledge encoded in the network was insufficient to overturn clear feature-level descriptions of one phoneme into a different one: *thif* is not interpreted as *this* — the unambiguous feature-level description of /f/ is not interpreted as /s/ simply because of its phonemic context. Thus the model respects the input and does not hallucinate phonemes on the basis of word frequency. This captures the effect more accurately than TRACE, for instance, in which a local lexical level representation of *vocabulary*, for instance, overturns bottom-up evidence and converts the erroneous /t/ in *vocabulaty* to /r/. Limited phoneme restoration does occur, however, when feature-level descriptions are ambiguous between two phonemes. For instance, when the input was /* e s/, in which * was ambiguous between /y/ and /r/, the model restored the /y/. In principle, the model is capable of recruiting both right and left context in the identification of a particular phoneme.

Monitoring for word-medial phonemes

Listeners are faster to monitor for word-medial phonemes like /p/ in prefixed words like *repel* compared with monomorphemic words like *lapel*, reflecting the fact that strings of phonemes beginning with prefix-like phonemes (/rIp/) are more frequent (regardless of position in words) than the comparable strings beginning with un-prefix-like strings (Shillcock *submitted*). The stimulus materials from the experiment were transcribed and embedded in the context "and the next word is ... and the next word is" and presented to the network. Activations for the critical phoneme in each word (e.g. /p/ in *repel*) were recorded. Mean activations were higher for the prefix cases, compared with the monomorphemic cases, at all positions in the net's memory (including the prediction position) except for the last one, where the two were precisely equal. The difference was only significant for the second position, however, ($t = 2.26, df = 14, p < .02$) possibly reflecting the poor sampling of polymorphemic words in the training corpus. Althought the training corpus was small, the network still reflected the bigram and trigram information available from a large phonemic dictionary which had proved to be a good predictor of this particular human data.

Conclusions and future work

The architecture discussed above promises to be useful in modelling the pre-lexical phonotactic constraints which affect word perception.

Current work is involving changes to the characterization of the input; specifically we are replacing the Jakobsonian features with those elements recognized by Government Phonology (Kaye et al. 1985). This work begins to converge with the connectionist approach to phonology, particularly with the processing approach espoused by Gasser and Lee (1989), who suggest an architecture similar to the model described above, but without the confirmatory output units. Extending the corpus, so as to ensure better sampling of the open-class vocabulary is important, as is the enrichment of the corpus to ensure that the input contains information currently excluded, such as phonological reduction and coarticulation. We expect the back propagation through time algorithm to extract richer temporal statistics than the copyback algorithms used so far.

References

Almeida, L. B. 1987. A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In Proceedings of the IEEE First International Conference on Neural Networks, San Diego, California, June, 1987.

Chater, N. 1989. Learning to respond to structures in time, RIPRREP 1000/62/89 Research Initiative in Pattern Recognition, St Andrews Road, Malvern, Worcs., U.K.

Coltheart, M. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33A**:497–505.

Elman, J. L. 1988. Finding structure in time. Technical Report, CRL TR 8801, Centre for Research in Language, UCSD.

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

Gasser, M. & Lee, C. (1989). Networks that learn phonology. Computer Science technical report number 300, Indiana University.

Hertz, J., Krogh, K., Palmer, R. G. 1991. *Introduction to the theory of neural computation* Addison-Wesley.

Jakobson, R., G. Fant and M. Halle 1952. Preliminaries to speech analysis. Technical Report **13**, MIT. Acoustics Laboratory, MIT Press.

Jordan, M. 1986. Serial order: a parallel distributed approach. Institute for Cognitive Science Report, 8604, University of California, San Diego.

Kaye, J., Lowenstamm, J. and Vergnaud, J-R. 1985. The internal structure of phonological elements: a theory of charm and government. *Phonology Yearbook* **2**:305–328

McClelland, J. L. & Elman J. L. 1986. Interactive processes in speech perception: the TRACE model. In D. E. Rumelhart & J. L. McClelland (eds.) *Parallel Distributed Processing*, Vol. 2., 58-121, Cambridge, Mass: MIT Press.

McClelland, J. L. & Rumelhart, D. E. 1988. *Explorations in Parallel Distributed Processing: Models, Programs and Exercises.* Cambridge, Mass: MIT Press.

Minsky, M. & Papert, S. 1969. *Perceptrons.* Cambridge, Mass: MIT Press.

Norris, D. G. 1988. A dynamic-net model of human speech recognition. Talk given at the Cognitive Models of Speech Processing Workshop, Sperlonga 1988. See G. Altmann (ed.) 1990, Cognitive Models of Speech Processing, MIT Press.

Pineda, F. 1987. Generalization of back-propagation to recurrent and higher order neural networks. *Neural Information Processing Systems*, New York.

Seidenberg, M. S. & McClelland, J. L. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**, 523-568.

Svartvik, J, & Quirk, R. 1980. *A corpus of English conversation.* Lund: Gleerup.

Shillcock, R. C. 1990. The processing of prefixed words: a connectionist account. Submitted to Memory & Cognition

Shillcock, R., Levy, J., and Chater, N. 1991. A connectionist model of auditory word perception in continuous speech. Proceedings of the 1991 meeting of the Cognitive Science Society, University of Chicago.

Warren, R. M. 1970. Perceptual restoration of missing speech sounds. *Science*, **167**, 392-393.