Connectionist Models of Speech Processing

Nick Chater
Department of Psychology
University of Warwick

Morten H. Christiansen
Department of Psychology
Cornell University

Corresponding author: Nick Chater, Department of Psychology, University of Warwick, Coventry, CV4 7AL, U.K.

email: nick.chater@warwick.ac.uk.
Telephone: +44 1203 523537
Fax: +44 1203 524225

INTRODUCTION

Psycholinguistics refers to the empirical study of the human language processing system, typically using behavioral experiments. This chapter considers attempts to capture psycholinguistic data using connectionist models (Christiansen and Chater, 2001). We primarily focus on relatively 'early' aspects of speech processing--speech segmentation and word recognition.

This chapter has four sections. *Connectionist Modeling: A Bridge Between Psycholinguistics and Brain Theory?* outlines the gulf between theories of brain function and traditional account of language processing. Connectionist modeling promises to help span this gulf, by attempting to ground speech processing in a connectionist processing architecture, a type of architecture initially inspired by attempts to model the computational properties of the brain. The section *Segmentation and Recognition: Two processes or one?* asks how far the problem of segmenting speech into words occurs independently of word recognition—a critical question for computational modelling. *Competition and Interaction in Word Recognition* considers connectionist models of word recognition, and their interplay with empirical research and theory.

CONNECTIONIST MODELING: A BRIDGE FROM PSYCHOLINGUISTICS TO BRAIN THEORY?

Both theoretical and empirical aspects of the psycholinguistics of speech processing seem, at first sight, rather distant from brain theory.

Theoretically, the starting point in psycholinguistics has been to take ideas from linguistics, the study of the abstract structure of language. But a theoretical vocabulary of 'phonemes,' or 'nouns,' to say nothing of the subtle notions of modern linguistic theory, seems difficult to relate to neural mechanisms. We shall see, though, that neural network models of aspects of speech processing may be viewed as building a bridge between the

abstract domain of linguistic representation and processing and computational architectures that may capture some general properties of neural machinery.

Empirically, data on psycholinguistics also seems distant from brain theory at an experimental level, because relatively little is known about the detailed structure and function of the rather diverse brain areas involved in processing speech. Partly this is because, in contrast to the study of perception or motor control, it is not possible to gather relevant information from detailed neurobiological studies of non-human animals, because natural language appears to be unique to humans. Perhaps more important, it seems unlikely that brain structures underlying speech processing will have an neurophysiological basis as readily interpretable as the topographic maps in the visual and motor cortex. This is because the computational problems of speech processing problem have no apparent spatial structure that might be expected to map onto cortex in a spatially coherent way. In any case, at present, neurobiological considerations impose relatively coarse constraints on computational models of speech processing. Although data from neuropsychology and functional imaging are becoming increasingly important, the main empirical constraints on psycholinguistic models are derived from the vast body of sophisticated, but often highly equivocal, laboratory studies of human language processing.

Perhaps the strongest constraint on such models is that language processing must somehow be implemented in neural hardware, rather than on a conventional symbolic machine. Conventional symbolic models of language processing in cognitive science and artificial intelligence have typically ignored this constraint, often on the assumption that the brain must be as powerful as a universal Turing machine, and hence able to implement any computable procedure. But this argument ignores the fact that language processing operations must operate extremely rapidly, using large numbers of slow and simple neural components, thus requiring a highly parallel, co-operative style of computation. This style of computation is not easy to reconcile with the very complex chains of sequential symbolic operations involved in most conventional language processing models.

Connectionist modeling attempts to help bridge the gulf between psycholinguistics and neuroscience, by attempting to capture detailed psycholinguistic data using computational models that aim to embody at least some of the computational principles of the brain. It also has more general significance: As a crucial test case for the viability of neural network models of cognition. Because conventional linguistic theory describes the structure of language in terms of a highly complex set of symbolic rules, language processing appears to represent a very difficult challenge for neural network modelling (see CONSTITUENCY AND RECURSION IN CONNECTIONIST NETWORKS). At present, connectionist models of speech processing are only partially developed, but prospects are encouraging in a number of areas (Christiansen and Chater, 2001).

The problem of speech processing is, of course, extremely broad, ranging from acoustic processing to semantic analysis. Here, the focus is the middle-ground problem of understanding how the brain segments and recognizes individual words in continuous, fluent speech.

SEGMENTATION AND RECOGNITION: TWO PROCESSES OR ONE?

In speech processing, as in perception, a fundamental question concerns the relationship between segmenting the sensory input (e.g., the speech signal) into chunks, and recognising those chunks. Segmentation and recognition appear to stand in a chicken-and-egg relation—i.e., its simply not clear how one could precede the other. Unless the input is segmented, how do we know what chunks of speech we should even be attempted identify as specific words (or other linguistic units)? But, conversely, unless we know what linguistic unit has been said, how can we know where the boundaries between units lie?

One approach to resolving the paradox is to assume that segmentation and recognition are two aspects of a single process—that tentative hypotheses about each issue are developed and tested simultaneously, and mutually consistent hypotheses are reinforced. A second approach is to suppose that there are segmentation cues in the input that are used to give at least better-than-chance indications of what segments may correspond to identifiable words. So the question is: Does speech processing involve dedicated segmentation strategies, prior to word recognition?

Developmental considerations suggest that there may be specialised segmentation methods. The infant, initially knowing no words, seems constrained to segment speech input using some method not requiring word recognition. Moreover, infant studies have shown that pre-linguistic infants may use such methods, and are sensitive to a variety of information that is available in the speech stream and potentially useful for segmentation, such as phonotactics and lexical stress—probably before cues based on the possible meaning of what is being said can be used by the child (Jusczyk, 1997).

How can children learn to segment speech? Cairns et al. (1997) note that language is less predictable across, rather than between, words. They trained a recurrent network on a large corpus of phonologically transcribed conversational speech, represented as a sequence of bundles of binary phonetic features. The network was trained to predict the next bundle of features along with the previous and current feature bundles, based on the current input material. Where prediction error was large, it was assumed that a word boundary had been encountered. This model captured some aspects of human segmentation performance. For example, it spontaneously learned to pay attention to patterns of strong and weak syllables as a segmentation cue. However it was able to reliably predict only a relatively small proportion of word boundaries, indicating that cues need also be exploited. Christiansen, Allen and Seidenberg (1998) showed how multiple, partial constraints on segmentation could yield much better segmentation performance. They trained a simple recurrent network to integrate sets of phonetic features with information about lexical stress (strong or weak)

and utterance boundary information (encoded as a binary unit) derived from a corpus of child-directed speech. The network was trained to predict the appropriate values of these three cues for the next segment. After training, the network was able to integrate the input such that it would activate the boundary unit not only at utterance boundaries, but also at word boundaries inside utterances. The network was thus able to generalize patterns of cue information that occurred at the end of utterances to when the same patterns occurred *inside* an utterance. This model performed well on the word segmentation task while capturing additional aspects of infant segmentation, such as the bias toward the dominant trochaic (strong-weak) stress pattern in English, the ability to distinguish between phonotactically legal and illegal novel words, and having segmentation errors being constrained by English phonotactics.

Although it seems likely that segmentation cues are exploited to guide the process of word recognition, this can achieve only limited results. A definitive segmentation of speech can only occur after word recognition has occurred. Empirical evidence strongly indicates that, during word recognition in adulthood, multiple candidate words are activated, even if these correspond to different segmentation of the input. For example, Gow and Gordon (1995) found that adult listeners hearing sentences involving a sequence (e.g., *two lips*) which could also be a single word (*tulips*, in US pronounciation) showed speeded processing of an associate of the second word (*kiss*) and to an associated of the longer word (*flower*), indicating that the two conflicting segmentations were simultaneously entertained. This would not occur if a complete segmentation of the input occurred before word recognition was attempted. On the other hand, it is not clear how these data generalize to word segmentation and recognition in infancy before any comprehensive vocabulary has been established. How the segmentation and recognition develop into the kind of integrated system evidenced by the Gow and Gordon data remains a matter for future research.

COMPETITION AND INTERACTION IN WORD RECOGNITION

Gow and Gordon's (1995) result also suggests that word recognition itself may be a matter of competition between multiple activated word representations, where the activation of the word depends on the degree of match between the word and the speech input. Indeed, many studies point towards this conclusion, from a range of experimental paradigms. Such competition is typically implemented in neural networks by a localist code for words (the activation of a single unit represents the strength of evidence for that word (see LOCALIST AND DISTRIBUTED REPRESENTATIONS), with inhibitory connections between word-units. Thus, when an isolated word is identified, a 'cohort' of words consistent with that input is activated; as more of the word is heard, this cohort is rapidly reduced, perhaps to a single item.

While competition at the word level has been widely assumed, considerable theoretical dispute has occurred over the nature of the interaction between different levels of mental representation. *Bottom-up* (or 'data-driven') models are those in which less abstract levels of linguistic representation feed into, but are not modified by, more abstract levels (e.g., the phoneme level feeds to the word level, but not the reverse). We note, however, that this does not prevent these models from taking advantage of supra-segmental information, such as in the inclusion of lexical stress in the Christiansen et al. segmentation model above, provided that this information is available in a purely bottom-up fashion (i.e., no lexical-level feedback). *Interactive* (also 'conceptually-driven' or 'top-down') models allow a two-way flow of information between levels of representation. Figure 1 provides an abstract illustration of the differences in information flow between the two types of models of word recognition. Note that bottom-up models allow information to flow through the network in one direction only, whereas an interactive model allows information to flow in both directions.

The bottom-up/interactive debate rages in all areas of language processing, and also in perception and motor control. Here we focus on putative interactions between

information at the phonemic and the lexical levels in word recognition (i.e., between phonemes and words), where experimental work and neural network modelling has been intense.

The most obvious rationale for presuming that there are top-down information flows from the lexical to the phoneme levels stems from the effects of lexical context on phoneme identification. For example, Ganong (1980) showed that the identification of a syllable-initial speech sound that was constructed to be between a /g/ and a /k/ was influenced by lexical knowledge. This intermediate sound was predominantly heard as a /k/ if the rest of the word was *iss* (*kiss* was favored over *giss*), but heard as /g/ if the rest of the word was *ift* (*gift* was favored over *kift*).

The TRACE model (McClelland and Elman, 1986) has an *interactive activation* architecture, with a sequence of layers of units. First layer units correspond to phonetic features, second layer units correspond to phonemes, and third layer units correspond to words. Within and between layers, there are fixed inhibitory bi-directional connections between units standing for incompatible states and fixed bi-directional excitatory connections between units standing for mutually compatible states. TRACE also deals with the temporal dimension of speech--there are many copies of the entire network, standing for different points in time, with appropriate connections between the units in each copy. TRACE captures effects of lexical context because lexical units influence phonemic input—McClelland and Elman modelled a wide range of data, and provided a model that has proved remarkably robust.

But 'context' effects on phoneme recognition can also be explained in purely bottom-up terms. If a person's decisions about phoneme identity depend on both the phonemic and lexical levels then phoneme identification will be lexically influenced, even though there need be no feedback from the lexical to the phoneme level. For example, the Ganong effect might be explained by assuming that the phoneme identification of an initial consonant that is ambiguous between /g/ and /k/ is directly influenced by the lexical level.

Thus, if *gift* is recognized at the lexical level, this will influence the participant to respond that the initial phoneme was a /g/; but if *kiss* is recognized, this will influence the participant to respond that the initial phoneme was a /k/.

A substantial experimental literature has attempted to distinguish TRACE from bottom-up models, indicating the importance of connectionist modelling in inspiring experimental research. One experimental result (Elman and McClelland, 1988), derived as a novel prediction TRACE, appeared to be particularly persuasive evidence against bottom-up connectionist models. In natural speech, the pronunciation of a phoneme will to some extent be altered by the phonemes that surround it, in part for articulatory reasons. This phenomenon is known as coarticulation. Listeners should therefore adjust their category boundaries depending on the phonemic context. Experiments confirm that people do indeed exhibit this 'compensation for coarticulation' (CFC; Mann and Repp, 1981). For example, given a series of synthetically produced tokens between /t/ and /k/, listeners move the category boundary towards the /t/ following a /s/ and towards the /k/ following a /sh/.

This phenomenon suggests a way of detecting whether lexical information really does feed back to the phoneme level. Elman and McClelland considered the case where compensation for coarticulation occurs across word boundaries. For example, a word-final /s/ influences a word-initial phoneme ambiguous between /t/ and /k/ to be heard as a /k/ (as in *Christmas capes*). If lexical-level representations feed back on to phoneme-level representations, the compensation of the /c/ should still occur when the /s/ relies on lexically driven phoneme restoration for its identity (i.e. in an experimental condition in which the identity of /s/ in *Christmas* is obscured, the /s/ should be restored and thus compensation for coarticulation should proceed as normal). Elman and McClelland confirmed TRACE's prediction experimentally. Recognition of the phoneme at the start of the second word was apparently influenced by CFC, as if the word-final phoneme in the first word had been 'restored' by lexical influence.

Surprisingly, bottom-up connectionist models can also capture these results. Norris (1993) provided a small-scale demonstration, training a simple recurrent network to map phonetic input onto phoneme output, for a small (12 word vocabulary) artificial language. When the net received phonetic input with an ambiguous first word-final phoneme and ambiguous initial segments of the second word, an analog of CFC was observed. The percentages of /t/ and /k/ responses to the first phoneme of the second word depended on the identity of the first word, as in Elman and McClelland (1998). But the explanation for this pattern of results cannot be top-down influence from word units, because there *are* no word units. Moreover, Cairns et al. (1995) scaled-up these results using a similar network trained on phonologically transcribed conversational English.

How can an autonomous computational model, where there is no lexical influence on phoneme processing, mimic the apparent influence of word recognition on coarticulation? Cairns et al. (1995) argued that sequential dependencies between the phoneme sequences in spoken English can often 'mimic' lexical influence. The idea is that the identification of the word-final ambiguous phoneme favored by the word level is also, typically, favored by transitional probability statistics across phonemes. Analysing statistical regularities in the phoneme sequences in a large corpus of conversational English, Cairns et al. showed that this explanation applies to Elman and McClelland's (1988) experimental stimuli. If these transitional probabilities have been learned by the speech processor, then previous *phonemic* context might support the 'restoration' of the ambiguous word final phoneme, with no reference to the word in which it is contained.

Pitt and McQueen (1998) tested between these two explanations experimentally. They carefully controlled for transitional probabilities across phonemes, and re-ran a version of Elman and McClelland's experiment: compensation for coarticulation was eliminated. Moreover, when transitional probabilities are manipulated in non-word contexts, compensation for coarticulation effects are observed. This pattern of results suggests that

compensation for coarticulation is not driven by top-down lexical influence, but by phoneme-level statistical regularities.

Against this, Samuel (1996) argues that the precise pattern of phoneme restoration does indicate the existence of small but discernible top-down effects. He conducted a statistical analysis of people's ability to discriminate whether a phoneme has been replaced by a noise in a word or non-word context, from the case where the phoneme and noise are both present. The logic is that to the extent that top-down factors 'restore' the missing phoneme, it should be difficult to tell whether or not the phoneme is actually present, and hence people's discrimination between the two cases should be poorer. Hence, phoneme present/absent discrimination should be poorer in word contexts than for non-word contexts, because top-down factors should be stronger. This prediction was confirmed experimentally (Samuel, 1996). Predictably, however, purely bottom-up explanations of this finding have since been proposed (Norris, McQueen and Cutler, 2000).

The theoretical debate concerning segmentation and word recognition has been profoundly influenced by connectionism. Connectionist models are now the dominant style of computational account, even for advocates of very different positions (as we have seen in relation to the bottom-up/interactive debate). Attempts to test between the predictions of competing models have generating experimental advances, which have in turn informed how models develop.

DISCUSSION

We have seen that connectionist models can provide a rich framework for modeling important aspects of human speech recognition, and is now central to the theoretical and empirical literature in the psychology of language. Moreover, connectionist methods can also be applied both to early processes in speech recognition, concerned with the early analysis of what is a highly complex and variable acoustic stimulus (see RECURRENT NETWORKS AND WORD RECOGNITION), and to later aspects of language

processing, where the main issues concern syntactic and semantic (see CONSTITUENCY AND RECURSION IN CONNECTIONIST NETWORKS). A critical issue for connectionist modeling is how or whether accounts of different aspects of the speech processing problem might ultimately be integrated into an overall model of speech processing. Presently, such an integration is a long way off—indeed, although progress has been substantial, connectionist and other models of speech perception are still some way from being able to identify words reliably in fluent continuous speech (and are not used, for example, in state-of-the-art automatic speech recognition), and work on syntactic and semantic analysis is still extremely preliminary.

Connectionist models appear, though, to provide a promising research direction, for a number of reasons. First, they provide a natural framework for modeling empirical psycholinguistic data. Second, learning is intrinsic to most connectionist networks, and hence the approach provides a natural source of developmental models (see COGNITIVE DEVELOPMENT). Third, connectionist models have provided a means of theoretical integration across different language processing domains. For example, interactive and bottom-up models of speech recognition as described here are closely analogous to interactive and bottom-up models of single word reading. Fourth, it is widely argued that connectionist networks capture some aspects of the computational 'style' of the brain---going at least some way to bridge between psycholinguistics and brain theory.

The potential implications of the connectionist approach to language processing are enormous, raising the possibility of a radical rethinking not just of language processing, but of language structure itself. Perhaps the ultimate description of language resides in the structure of complex networks, and can only be approximately expressed in terms of rigid, grammatical rules. Or perhaps connectionist models can only succeed to by building in standard linguistic constructs; or perhaps connectionist learning methods do not scale up at all (see Seidenberg's and Smolensky's contributions to Christiansen, Seidenberg and Chater, 1999 for opposing perspectives). The future of connectionist models of language

processing may therefore have important implications for the theory of language processing and language structure, and the neural machinery underlying speech processing.

REFERENCES

Cairns, P., R. Shillcock, N. Chater, and J. Levy, 1995. Bottom-up connectionist modelling of speech. In Connectionist Models of Memory and Language, (J.P. Levy, D. Bairaktaris, J.A. Bullinaria and P. Cairns, eds.), London: UCL Press, pp. 289-310.

Cairns, P., R. Shillcock, N. Chater, and J. Levy, 1997. Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation, Cog. Psych., 33:111-153.

Christiansen, M.H., J. Allen, and M. S. Seidenberg, 1998. Learning to segment speech using multiple cues: A connectionist model, Lang. & Cog. Proc., 13:221-268.

Christiansen, M.H., and N. Chater, 2001. Connectionist psycholinguistics, Trends in Cog Sci, 5, 82-88.

Christiansen, M.H., N. Chater, and M. S. Seidenberg (Eds.) 1999. Connectionist models of human language processing: Progress and prospects, Special Issue of Cog. Sci., 23: 415-634.

Elman, J.L., and J. L. McClelland, 1988, Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes, J. Mem. & Lang., 27:143-165.

Ganong, W.F., 1980, Phonetic categorization in auditory word perception, J. Exp. Psych.:HPP, 6:110-125.

Gow, D. W., and P. C. Gordon, 1995. Lexical and pre-lexical influences on word segmentation: Evidence from priming, J. Exp. Psych.:HPP, 21:344-359.

*Jusczyk, P. W., 1997. The discovery of spoken language, Cambridge, MA: MIT Press.

Mann, V.A., and B. H. Repp, 1981. Influence of preceding fricative on stop consonant perception, J. Acoust. Soc. Am., 69:548-558.

McClelland, J.L., and J. L. Elman, 1986. The TRACE model of speech perception, Cog. Psych., 18: 1-86.

Norris, D., 1993 Bottom-up connectionist models of "interaction." In Cognitive models of speech processing (G.T.M. Altmann, and R. Shillcock, eds.), Hillsdale, NJ: Erlbaum, pp. 211-234.

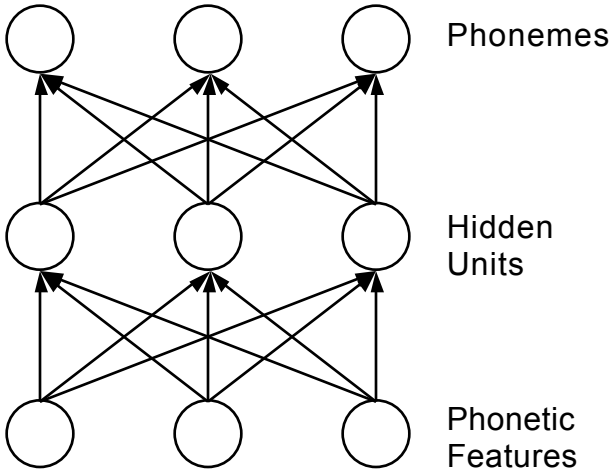
Norris D., J. M. McQueen, and A. Cutler, 2000. Merging information in speech recognition: Feedback is never necessary, Beh. & Brain Sci., 23.

Pitt, M. A., and J. M. McQueen, 1998. Is compensation for coarticulation mediated by the lexicon? J. Mem. & Lang., 39:347-370.

Samuel, A. C., 1996. Does lexical information influence the perceptual restoration of phonemes? J. Exp. Psych.: Gen., 125: 28-51.

Figure 1: Illustrations of a bottom-up model (top) and an interactive model (bottom). The links in the bottom-up model can be either excitatory or inhibitory and only allow for information to flow upwards from the phonetic features through the hidden units to the phonemes on the output. In the interactive activation model, the links are bi-directional and allow information to flow both bottom-up from the phonetic features through the letter units to the word units and top-down. Arrows denote excitatory links whereas filled circles denote inhibitory links.

Bottom-Up Model



Interactive Activation Model