LANGUAGE AND COGNITIVE PROCESSES, 1997, *12* (5/6), 613–656

# Integrating Form and Meaning: A Distributed Model of Speech Perception

M. Gareth Gaskell and William D. Marslen-Wilson

*Centre for Speech and Language, Birkbeck College, University of London, London, UK*

We present a new distributed connectionist model of the perception of spoken words. The model employs a representation of speech that combines lexical information with abstract phonological information, with lexical access modelled as a direct mapping onto this single distributed representation. We first examine the integration of partial cues to phonological identity, showing that the model provides a sound basis for simulating phonetic and lexical decision data from Marslen-Wilson and Warren (1994). We then investigate the time course of lexical access, and argue that the process of competition between word candidates during lexical access can be interpreted in terms of interference between distributed lexical representations. The relation between our model and other models of spoken word recognition is discussed.

## INTRODUCTION

The representation of information in a distributed manner is one of the key assumptions of connectionist theory (Hinton, McClelland, & Rumelhart, 1986; Smolensky, 1988). Here, we describe and evaluate a model that applies this assumption to the process of retrieval of lexical knowledge about spoken words. The model represents lexical knowledge using a set of features that encode information about the meaning and form of words.

Requests for reprints should be addressed to Gareth Gaskell, Centre for Speech and Language, Department of Psychology, Birkbeck College, Malet Street, London WC1E 7HX, UK. E-mail: g.gaskell@psyc.bbk.ac.uk

This research was supported by a UK MRC grant awarded to William Marslen-Wilson and Lorraine Tyler. We thank members of the Speech and Language Research Group at Birkbeck, participants at the 1995 Sperlonga Meeting, and members of the PDPNLP seminar group at the University of California, San Diego for much valuable discussion. We thank David Plaut for making his corpus of monosyllables available. Part of this research was presented at the Seventeenth Annual Conference of the Cognitive Science Society, Pittsburgh, PA.

Previous models of speech perception (e.g. McClelland & Elman, 1986) have generally employed an explicit ordering of information types, in which typically one or more phonological levels mediate between input representations and lexical entries. This allows an explanation of selective access to information in the lower levels, such as phonological information about nonwords. Such ordering is redundant when applied to fully distributed representations, since differences in the speed or success of retrieval of different forms of knowledge can instead be modelled by the partial activation of a distributed representation. Thus, the difference between the perception of a word and a nonword is in the types of information made available by the perceptual process. For a word, a complete lexical representation can be accessed, including stored syntactic, semantic and phonological information, whereas for a nonword, or for an unfamiliar word, only the phonological code is retrieved.

Our model eliminates explicit intermediate levels and treats lexical access as a direct mapping from fairly low-level information about the speech signal simultaneously onto a distributed substrate incorporating abstract representations of both the form and the meaning of words. The model is implemented using a simple recurrent network (Elman, 1990), for which lexical representations are distributed patterns of activity on a set of output nodes. Thus the network, rather than modelling the recognition of word forms, concentrates on the retrieval of lexical phonological and semantic information, with explicit recognition being a secondary product, relevant for psycholinguistic tasks such as lexical decision, rather than a goal of the model.

A number of connectionist models have demonstrated aspects of lexical processing using distributed semantic networks (e.g. Joordens & Besner, 1994; Kawamoto, 1993; Kawamoto, Farrar, & Kello, 1994; Masson, 1995; Plaut, 1995; Sharkey & Sharkey, 1992). However, these models have focused largely on processing *within* the lexicon rather than access to lexical information. Where lexical access has been addressed, it has been modelled as a mapping from static (orthographic) form representations onto meaning. This ignores the transient nature of speech, making competition effects during lexical access (e.g. Zwitserlood, 1989) impossible to incorporate.

The structure of the article is as follows. First, we discuss the assumptions underlying our model and describe how these can be implemented in the connectionist network. We then evaluate the performance of the network in two ways. We examine the influence of matching and mismatching speech on the retrieval of lexical knowledge by comparing the model's behaviour to the experimental study of Marslen-Wilson and Warren (1994) on the integration of partial cues in lexical access. These data proved difficult to model in the multilevel localist tradition, but can be accommodated by our model. We then turn to more standard phenomena in the literature on spoken word

recognition, showing how competition effects can be reformulated and modelled in terms of interference between fully distributed patterns of lexical representation.

## A DISTRIBUTED MODEL OF SPEECH PERCEPTION

We intend to model the process of speech perception as a direct mapping from low-level featural information onto a distributed representation of lexical knowledge and form. The principal assumptions of the model are as follows:

1. Lexical knowledge is represented in a fully distributed fashion.
2. Different forms of lexical knowledge (e.g. phonology, semantics) are represented in parallel and accessed simultaneously.
3. Speech input maps directly and continuously onto lexical knowledge.
4. The lexical access process operates with maximal efficiency.

The value of distributed representations in the modelling of cognitive functions is well documented (e.g. Hinton et al., 1986; Hinton & Shallice, 1991). We envisage the representation of a lexical item to be a distributed pattern encompassing its semantic, syntactic, morphological and phonological specification. The activation of a single complete lexical representation involves setting the correct values for all representational units. This view of the lexical access process differs radically from currently popular models of spoken word recognition such as TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994) and Cohort (Marslen-Wilson, 1987), which view the selection of word candidates as a parallel localist process of competition. Instead of mapping speech input onto many localist representations, we shall explore the possibility that lexical selection operates on a single distributed level of representation.[1]

Assumptions 2 and 3 are based on the proposals of Marslen-Wilson and Warren (1994), which we shall discuss in detail in the following section. Our intention is to produce a model of speech perception that treats all types of information derived from the speech signal as outputs of the system, with detail preserved in the mapping onto these forms of knowledge. These assumptions fit very easily into a distributed learning approach. Although it is possible to train connectionist networks to perform a multistage mapping (e.g. Plaut & Shallice, 1993), it is often simpler and more desirable to restrict

[1]The position we take is at the opposite extreme of the continuum of information distribution to the localist position, allowing us to highlight their contrasting predictions. However, we must not forget that intermediate positions are also plausible.

teacher input to the output level. This allows the network to develop whatever intermediate representations are necessary to perform the mapping effectively.

Our asssumption of maximal efficiency implies that at all points our model must derive the most informative output available from its analysis of incoming speech. Thus, if it is possible to isolate a single lexical match to the current input (i.e. at the word's uniqueness point), the relevant information about that word should be extracted. At other points, where more than one lexical entry matches the speech presented so far, the output of the model should reflect this ambiguity and activate the stored knowledge about these candidates. Thus, the network should entertain in parallel multiple hypotheses about the lexical identity of incoming speech, as do the majority of current models of speech perception. However, the distributed nature of the lexical representations used in our model places limitations on the effectiveness of parallel evaluation of multiple candidates. Our model assumes that speech is mapped more or less directly onto distributed representations of lexical knowledge, implying that multiple lexical candidates can only be evaluated by their influence on this level of representation, rather than at some independent stage of competiton (as assumed in models such as TRACE and Cohort). Since different lexical candidates will generally have different lexical representations, this suggests that they will interfere, producing a lexical "blend" of the various candidates (Smolensky, 1986). We will address the consequences of these proposals in some detail.

## Network Architecture

The above assumptions translate readily into a description of the starting state and processing environment of a connectionist network. The first three assumptions define the basic architecture of the model, while the maximal efficiency assumption is fulfilled by the application of an error-reducing learning rule to the mapping at all points during the processing of each word. To allow the network to generalise over patterns of phonetic features spread across time, our model is based on a simple recurrent network architecture (Elman, 1990; Jordan, 1986). These networks are an extension of the standard backpropagation procedure, having an extra set of input units which hold a copy of the hidden units at the previous time-step. This architecture has already proven valuable in the modelling of various aspects of speech perception (Elman, 1990, 1993; Norris, 1992, 1993).

We trained the network on the mapping between a stream of phonetic features and the internal representations of words (see Fig. 1). The featural input is passed through a set of 200 hidden units, which have access via recurrent links to the state of the hidden units at the previous time-step. The

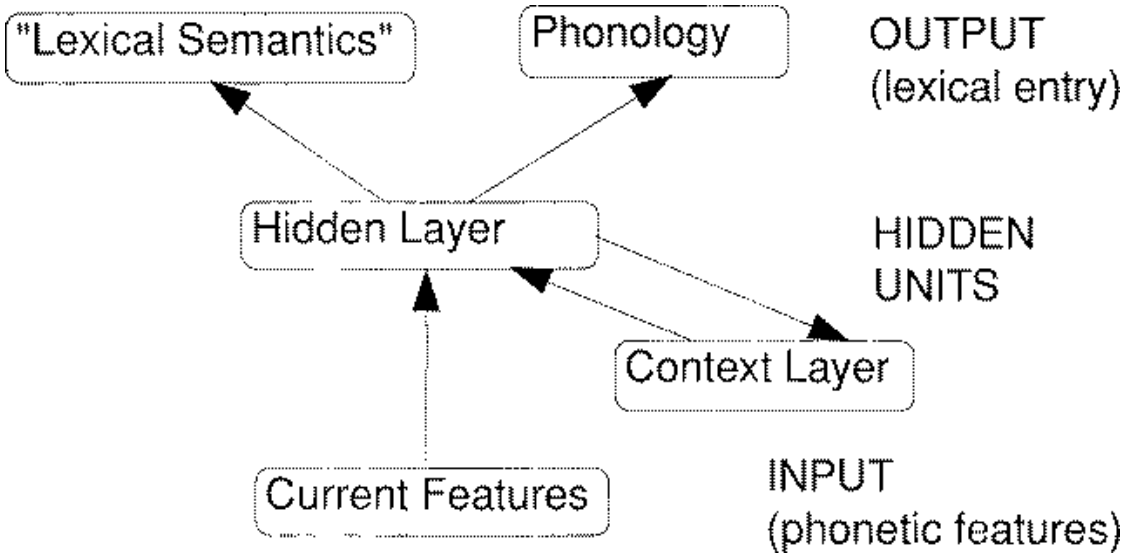


FIG. 1. A distributed model of speech perception.

hidden units are also connected to two sets of output units, representing the semantics and phonology of the words contained in the speech stream.

*Input Representation.* The model's input is a representation of the speech signal. Ideally, this would be a set of features or parameters derived directly from analysis of the speech wave (e.g. Stevens, Manuel, Shattuck-Hufnagel, & Liu, 1992). The structure imposed on the model by such a representation would be minimal, leaving the network free to develop the necessary generalisations from the speech stream. However, the computational cost of such an approach currently makes it infeasible. The representation we have adopted instead is a compromise, designed to be flexible and unintegrated enough to accommodate some subphonemic spread of information, but categorical enough to allow training to be carried out in a reasonable amount of time. Auditory input to the network was represented on a set of 11 binary input units. These encoded the phonetic features of the current input segment using the system of Jakobson, Fant and Halle (1952). To simulate the experiments of Marslen-Wilson and Warren (1994), which involved conflicting subphonemic cues to place of articulation, two additional input units were added. These allowed the representation of vowel transition information.

*Semantic Representation.* The output nodes we have labelled "semantic" encode a distributed representation of the stored non-phonological knowledge about a word. This is a provisional means of representing lexical information that has an arbitrary relationship with the incoming speech, such as semantic or syntactic knowledge. We make no

attempt to create a realistic representation of these forms of knowledge, partly because of the infeasibility of such a task and partly because we wish to focus on the basic properties of the lexical access process. The lexical knowledge for any word is instead represented by a random binary pattern of 0's and 1's.

This pattern could be thought of as a microfeatural representation of the meanings of the words such that each value represents the relevance of that feature to the word's representation (Hinton & Shallice, 1991; Plaut & Shallice, 1993). However, the representations we shall use differ from microfeatural representations in an important way. For most simulations we chose representations with 50% of units set to 1. These representations are much less sparse than microfeatural representations, which generally only have a small number of features on, with the remainder off. When we come to examine the time course of competition in our model, we shall see that variables such as sparseness have an important effect on the capacity of the network to represent multiple meanings in parallel.

*Phonological Representation.* The second set of output units encodes a representation of the underlying phonological structure of words. This is the representation on which judgements of form are based and is treated as a product of the perceptual mapping alongside semantic knowledge. This single level may well display properties of both lexical and non-lexical routes to phonological knowledge (cf. models of reading such as that of Seidenberg and McClelland, 1989) and is available for phonological judgements on both words *and* word-like nonwords. An important aspect of the choice of phonological representation is its capacity to promote generalisation to novel forms.

To satisfy this constraint, the phonological output was based on the representation of Plaut, McClelland, Seidenberg and Patterson (1996), originally designed for their model of word naming. This is a compact structured phonemic representation of monosyllabic words, divided into three groups of units corresponding to the syllable onset, nucleus and coda. Within each group, units generally represent single phonemes, with a small set of units representing phoneme clusters such as /ts/.

Because of incompatibility problems between the input representation and the phonemic representation of Plaut et al. (1996) (chiefly due to differences in the representation of long vowels and diphthongs), that of Plaut et al. was altered so that five vowels in the original representation were represented by combinations of two vowels, presented sequentially on the input units.

*Training.* Although the composition of the training sets varied from simulation to simulation, the overall procedure remained the same. The

corpus of words used in any particular simulation was translated into two forms. The first form was a continuous sequence of phonetic feature bundles representing incoming speech. The second was a sequence of the same length, but represented the semantics and phonology of the words, as described above. The second form was used for comparison with the network output, allowing connection weights to be altered by backpropagation of error. There were no gap markers between the words, nor was the context layer reset during training. At all points during the presentation of a word, the lexical knowledge about that word was available for adjustment of weights.

## General Performance

Before examining selected characteristics in detail, we shall review the basic behaviour of the trained network. At each time-slice, a new set of phonetic features is presented to the network, and the output vector is modified to accommodate the information provided by the input. Although the network was trained to output the phonological and semantic representations of each word at every point within that word, it cannot actually carry out that task when tested. This is because the identity of a word is often ambiguous early on in the word, and in some cases even the position of the onset of the word is ambiguous. What the network learns to do instead is to produce "cohort-like" behaviour, where the output of the network represents the set of word candidates compatible with the input so far (Content & Sternon, 1994; Norris, 1990). At certain times (on or after the uniqueness point of a word), this set has only one member, and the network can indeed output a vector very close to the training vector for that word. At other times, the network must entertain multiple hypotheses in parallel, until disambiguating information is encountered.

These hypotheses are represented in different ways for the two components of the output vector. The behaviour of the phonological output is fairly simple. Assuming that the network can identify the onset of a word, the network simply activates the phonemic nodes corresponding to the segments presented so far. For example, when the /æ/ of /kæt/ (*cat*) is presented, the /k/ node in the onset section will be activiated close to a maximum of 1.0, as will the /æ/ node in the vowel section. Other nodes may be activated more weakly, representing hypotheses about continuations of the speech. So in this case, perhaps, the /p/ and the /t/ in the coda section of the word would become slightly active, reflecting the fact that *cap* and *cat* are members of the network's training corpus.

The output at the semantic level is more difficult to interpret, because of the arbitrary relationship between the form of speech and word meanings. Generally, the meanings of words in a cohort set will be unrelated, implying

that their distributed semantic representations will have few similarities. In these circumstances, deterministic networks output a "blend" of the relevant representations (Smolensky, 1986), where the value of each element in the output vector is the arithmetic mean (weighted by frequency of presentation during training) of all the relevant training values for that element. For example, at some point during testing, the input might be compatible with just two words, *gear* (/giə/) and *geese* (/gis/). The first five elements of the random semantic vectors for these words are (1, 0, 0, 1, 1) and (0, 0, 1, 1, 0). The network at this point should output the frequency weighted blend of these two words. This will contain a 0 for the second element and a 1 for the fourth element (since the representations match for these elements), and will contain a value between 0 and 1 for the remainder, with a bias towards the more frequent word.

The dynamical process of modifying this blend as information comes in can be viewed in terms of movement through semantic space. The elements of the semantic vector can be thought of as axes in a many-dimensional space, in which word representations are fixed points. The changes in the state of the output units represent movement through this space. Assuming the elements of the semantic vector are binary, the network starts near the middle of the space and moves outwards as more information is presented. At all times, the network tries to find a point as close as possible to all the matching candidates. (Again, this process is modified by the relative frequencies of the words, so the words presented often during training will exert a greater "draw" than less frequent words.) As the number of candidates is reduced, the output moves to the midpoint of the remaining options and the distance from these representations decreases. Finally, at the uniqueness point of the word, the output can move onto the fixed point representing that word, remaining there until the onset of a new word is reached.

This view of lexical access lends itself to comparison with localist activation-based models (Marslen-Wilson & Welsh, 1978; Morton, 1969). The "activation" of a word in the distributed model is inversely related to the distance between the output of the network and the word representation in lexical space. A zero distance is equivalent to the maximum activation, whereas greater distances imply weaker activation. However, despite this overall similarity, there are important differences in the way distributed and localist systems model such an activation process. For example, in a localist system, there is in principle nothing to stop two or more words having the maximum activiation. The maximum activation in a distributed system is a zero distance from the target vector (i.e. a zero error score), where the output of the network exactly matches the representation of the word. But, by definition, if the network output matches the representation of one word, it must mismatch the representations of all words that do not share its

representation, and so no unrelated word can also be activated to the same extent. More generally, it is unclear whether the distributed blending approach can really accommodate the experimental data on parallel activation of meaning (Zwitserlood, 1989; Zwitserlood & Schriefers, 1995; Moss, McCormick, & Tyler, this issue). We shall investigate these issues in detail later in the paper.

## THE PROCESSING ENVIRONMENT FOR SPEECH PERCEPTION

In this section, we shall examine the competition dynamics of our model by simulating two experiments from Marslen-Wilson and Warren (1994). Their research takes a detailed look at the effects of mismatch on lexical access, and poses a challenge for current models of speech perception. We shall first review the main points of the experimental data and then go on to simulate their experiments using the network model.

### Experimental Data

Marslen-Wilson and Warren (1994) examined the integration of featural cues to word identity in words and nonwords. In particular, the integration of cues to place of articulation was examined, by cross-splicing monosyllabic words and nonwords that contained conflicting cues to the place of articulation of the final consonant. For example, subjects might hear a token consisting of the initial consonant and vowel of *jog*, followed by the final consonant burst of *job*. This stimulus contains information in the vowel to final consonant transition that accords with a velar consonant (i.e. the /g/ from *jog*). However, the burst information indicates a bilabial consonant (i.e. the /b/ from *job*). In cases like this, the burst information is dominant and the spliced stimulus is perceived as a token of *job*. However, previous research has shown that this featural mismatch can produce interference effects in tasks such as lexical decision (Streeter & Nigro, 1979; Whalen, 1982, 1984).

Marslen-Wilson and Warren (1994) manipulated the lexical status of the monosyllable from which the cross-spliced tokens were created. Triplets of monosyllables containing either two words and one nonword (e.g. *jog*, *job*, *jod*) or one word and two nonwords (e.g. *smog*, *smob*, *smod*) were cross-spliced to produce six types of stimulus (see Table 1). These stimuli varied in terms of the presence or absence of mismatching cues and in terms of the lexical status of the pre-splice and post-splice components. The baseline conditions were two tokens of the same word or nonword spliced together (e.g. *jo*b + *jo*b [W1W1] or *smo*b + *smo*b [N1N1]). For the remaining conditions, the post-splice token was held constant (either W1 or N1) and the pre-splice token was manipulated to create four conditions of

TABLE 1
The Experimental Contrasts of Marslen-Wilson and Warren (1994)

| *Lexical Status* | *Code* | *Example* |
|---|---|---|
| **Word sequences** | | |
| Word1 + Word1 | W1W1 | *job* + *job* |
| Word2 + Word1 | W2W1 | *jog* + *job* |
| Nonword3 + Word1 | N3W1 | *jod* + *job* |
| **Nonword sequences** | | |
| Nonword1 + Nonword1 | N1N1 | *smob* + *smob* |
| Word2 + Nonword1 | W2N1 | *smog* + *smob* |
| Nonword3 + Nonword1 | N3N1 | *smod* + *smob* |

*Note*: The underlined sections represent the segments spliced together to create the stimuli.

mismatch, which differed on the lexical status of the two components (e.g. *jog* + *job* [W2W1], *jod* + *job* [N3W1], *smog* + *smob* [W2N1] or *smod* + *smob* [N3N1]).

Three experiments, using lexical decision, gating and phonetic decision tasks, examined the perceptual consequences of these manipulations. Concentrating on the lexical and phonetic decision experiments, which illustrate the results most clearly, we find an interesting effect of the lexical status of the pre- and post-splice tokens (see Fig. 2).

In the lexical decision experiment, where subjects made a timed "yes/no" response to the cross-spliced stimuli, there were inhibitory effects of mismatching cues for all conditions made up from at least one word token. In other words, "yes" responses to W2W1 and N3W1 stimuli were slower than for the W1W1 baseline, and "no" responses to W2N1 stimuli were slower than for the N1N1 baseline. Crucially, however, the mismatching cues in the N3N1 condition did not slow responses significantly.[2] The mismatching cues to the word-final place of articulation had no effect when both pre- and post-splice cues were taken from nonword stimuli.

A surprisingly similar pattern was found when subjects were asked to make a timed, forced-choice phonetic decision on the final consonant of the stimuli. The alternatives were compatible with either the pre- or post-splice phonetic cues to the final consonant. Since burst information dominates in the perception of phonological form, subjects generally responded with the interpretation that agreed with this information. However, the speed with which this response was made depended in part on the information in the vowel transition. As in the lexical decision experiment, there were inhibitory

[2]These results are based only on the stimuli ending with a voiced stop. A fuller description of the results can be found in Marslen-Wilson and Warren (1994).

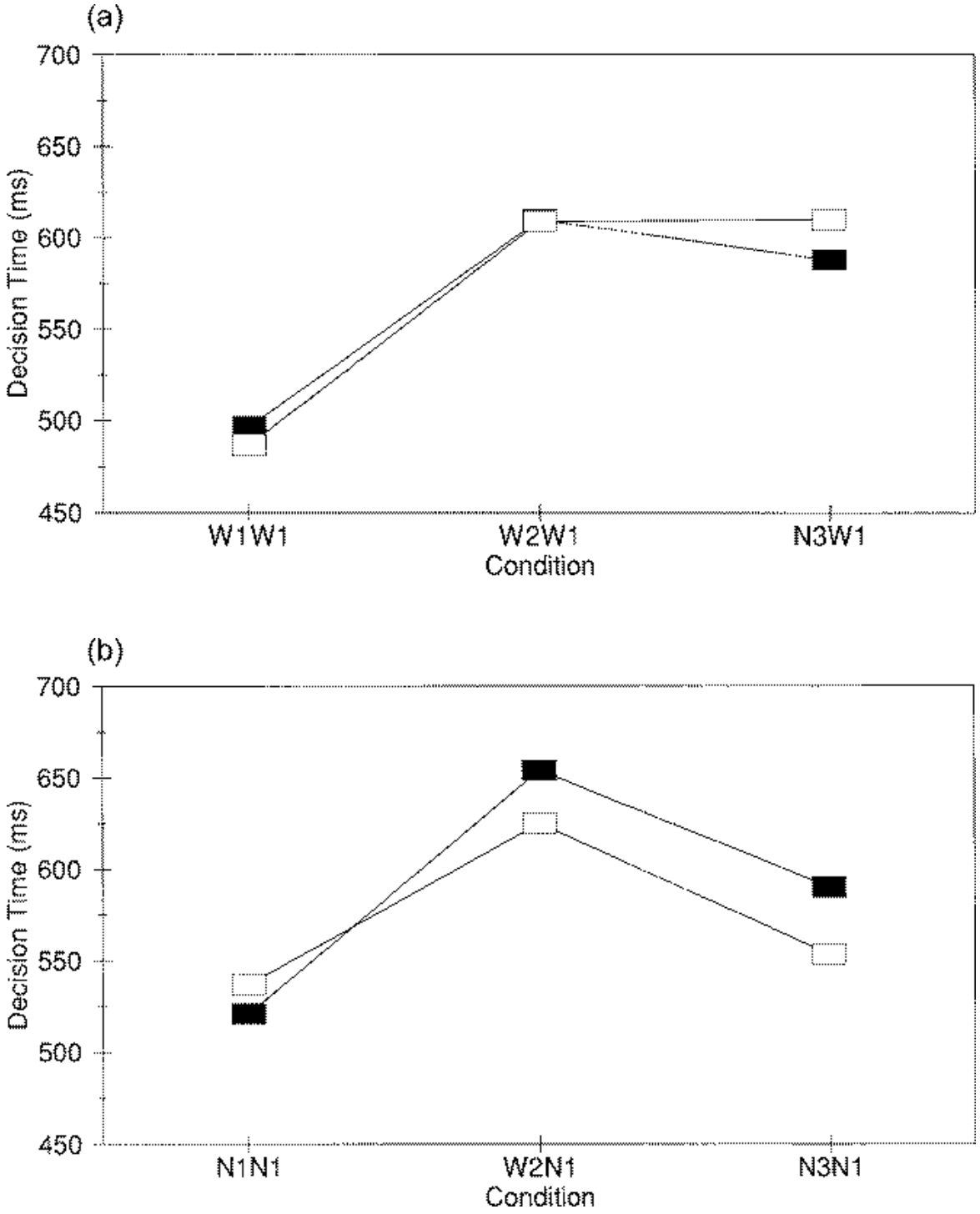


FIG. 2. Summary of the phonetic decision (■) and lexical decision (□) experiments of Marslen-Wilson and Warren (1994). Only the data for the voiced-stop stimuli are plotted. (a) The data for the word triplets; (b) the data for the nonword triplets. Response times were measured from the splice point.

effects for all conditions involving mismatch with at least one word. There was also some inhibition for the mismatching stimuli made up from two nonwords (N3N1), but this mismatch effect was significantly smaller than for the W2N1 condition.

In summary, effects of featural mismatch in both lexical decision and phonetic decision tasks are strongly influenced by lexical factors, with little or no mismatch effect for stimuli made up of two different nonword tokens. These results seem difficult to square with a view of lexical access which feeds off an autonomous segmental representation of incoming speech. Such a theory predicts that all mismatching conditions would have a similar disrupting influence on the word recognition process and would inhibit both phonetic and lexical decisions in a similar manner.

Marslen-Wilson and Warren (1994) argue that their results are best accommodated by a model of word recognition in which featural information is mapped directly and continuously onto lexical representations (Klatt, 1989; Stevens, 1986; Warren & Marslen-Wilson, 1987, 1988). The absence of a pre-lexical level of feature integration, such as a phoneme level, provides a simple explanation of the pattern of mismatch effects found in the lexical decision experiment. For the word sequences, all mismatching tokens delay responses because they reduce the goodness of fit to the target word. For the nonword sequences, response times depend on how well the tokens match the nearest lexical item, which in this case is W2. Thus the W2N1 tokens delay responses because they are more similar to a word than the N1N1 baseline. However, the N3N1 tokens do not delay responses because they are equally dissimilar to the nearest word as the N1N1 baseline. The ambiguity of the phonemic structure of the N3N1 tokens is irrelevant for this model because phonemic information is not integrated before contact with lexical representations.

Marslen-Wilson and Warren (1994) also argue that the phonetic decisions are based largely on lexical knowledge of phonological form, which explains the close similarity between the results of the two experiments. However, this explanation is problematic in itself, since, in its simplest form, the model they advocate does not provide a basis for phonetic or phonological decisions based on nonwords. To solve this problem, they propose that phonological representations of nonwords are based on analogy to lexical forms, drawing on arguments used in connectionist models of word naming (Seidenberg & McClelland, 1989).

## Network Simulations

The data of Marslen-Wilson and Warren (1994) pose two challenges for our model of speech perception: to provide a basis for phonological decisions on words and nonwords that is not pre-lexical and to explain the pattern of

mismatch found in their lexical and phonetic decision experiments. In this section, we show that the assumption of a continuous mapping process from featural input onto parallel distributed lexical representations allows these challenges to be met.

*Training Corpus.* The network was trained to extract the phonological and semantic information for a set of 36 monosyllabic words drawn from Marslen-Wilson and Warren's (1994) test words. These comprised the unspliced words required to create 24 spliced triplets (12 word triplets and 12 nonword triplets) for testing. All words ended with a single consonant which was either a voiced (/d/, /b/ or /g/) or an unvoiced (/t/, /p/ or /k/) stop. These words were presented as input to the network in the form of bundles of phonetic features. All test words consisted of either four or five segments.

To maintain a more realistic competitor environment for the test items, a number of other words were added to the training corpus. Since we were predominantly interested in the network's evaluation of word-final consonants, a set of 71 close cohort competitors was added, which shared the initial onset and vowel segments with the target words but diverged on the final consonant cluster. These ensured that the test words had an average of 3.5 close competitors (range 0–10). This is not a realistic overall competitor environment, but it ensures that even at the ends of the test words, there remained strong competition between lexical candidates.

In addition, the token frequencies of the training set were manipulated. Corpus analyses of English (e.g. Johansson & Hofland, 1989) reveal a skewed distribution of token frequencies—a small number of words occur very frequently, whereas the vast majority of words have very low token frequencies. The training corpus was structured to reflect the gross statistics of this situation. The test words were all given a token frequency of 20 within the training corpus, preventing any frequency effects from obscuring the object of the simulations. The cohort competitors were then assigned random frequencies between 1 and 40, with a mean frequency of 20. A further 2998 monosyllables, taken from the simulations of Plaut et al. (1996), were added to the training corpus, with a token frequency of 1. The low-frequency words ensured that the network was exposed to a representative range of phonological forms, allowing it to generalise to novel (nonword) forms. However, we did not expect the capacity of the network to be great enough to learn the semantic representations of these words, because of the arbitrary nature of the phonology to semantics mapping. The corpus of just over 5000 tokens was presented to the network 60 times during training, implying that the network was trained on 300,000 monosyllabic tokens.

Five separate networks, with different initial weights, were trained using this corpus. The results of the simulation were highly consistent across

networks and the analyses we present are based on the average values over all five simulations.

*Simulating Coarticulation.* Successful simulation of the data of Marslen-Wilson and Warren (1994) relies on the presence of cues to place of articulation spread over a period of time (or processing cycles in the case of the network). To simulate this coarticulatory spread of place information between consonant and preceding vowel, we added two extra features to the 11 features standardly used by Jakobson et al. (1952). These features are set to zero for all segments except vowels immediately preceding nasal or stop consonants. For these vowels, the two features represent the place of the following consonant, mirroring the *diffuse* and *grave* feature values for that consonant.

This system does not allow the following consonant to be identified on presentation of a "coarticulated" vowel, but it does indicate its place of articulation: velar, coronal or bilabial. However, the completely deterministic use of these cues still overestimates the informativeness of coarticulatory place cues in vowels. The duplication of place feature-values from consonants to preceding vowels does not reflect the finding that vowel transition cues to place are weaker than the corresponding cues in the consonant burst. This finding is evident from the results of Marslen-Wilson and Warren (1994), as well as from many other studies (e.g. Martin & Bunnell, 1982; Streeter & Nigro, 1979; Warren & Marslen-Wilson, 1987, 1988). First, burst place cues nearly always dominate the resultant perceptual experience of cross-spliced tokens. Secondly, in gating studies of normal speech, subjects frequently mistake the place of articulation when asked to make judgements on tokens with burst information removed. In contrast, subjects almost never mistake the place of articulation of burst information, even when the preceding vowel transition cues mismatch. For example, in the gating study of Marslen-Wilson and Warren (1994), only 55% of responses to N3N1 stimuli at vowel offset were consistent with the cues to consonant place in the vowel. At the offset of these stimuli, almost all responses were compatible with the place of the burst.

To accommodate this difference in informational content, we made the place cues in the vowel probabilistic. These cues were correct (i.e. agreed with the place of the consonant) 70% of the time, with the remaining 30% of vowel cues consistent with either of the other two places of articulation used in the stimulus triplets.

*Training and Testing.* The network illustrated in Fig. 1 was trained on 50 sweeps through the corpus of roughly 5000 tokens described above. On each cycle, the 13 input nodes were activated with the phonetic pattern of one

segment of a word (modified as described above). The copy-back units, holding the activations of the hidden units at the previous cycle, also acted as input units. The activations of the 200 hidden units and the 102 output units (50 semantic units, 52 phonological units) were updated and compared to the training pattern, which consisted of the full semantic and phonological patterns for the current word. This means that right from the presentation of the first phoneme of a word, its full phonological representation was available as a teacher signal. Backpropagation of error was then used to adjust the weights of the connections in the network (Rumelhart, Hinton, & McClelland, 1986). The training output for a word remained constant throughout the presentation of each of its constituent segments. Each word followed on from the previous word without a gap and without resetting the context units.

The connection weights developed during training must allow the network to perform two operations. First, the network must act as a kind of buffer, collecting input over multiple time-slices to allow words to be recognised. Secondly, it must encode the phonological and semantic patterns for each word. Performing the semantic side of this mapping is particularly costly in terms of network resources because it is an arbitrary mapping. Our expectation was, therefore, that the number of hidden units in the network would be too few for it to isolate the correct output pattern for the vast majority of the low-frequency words, which were included in the training corpus to ensure that the network was exposed to a wide range of phonological forms.[3]

The basic performance of the trained network was tested by presenting the 107 test and competitor words to the network in a novel order. These words varied in frequency between 1 and 40 instances per epoch. To gauge the overall success of the training, the output of the network was recorded on presentation of the final phoneme of each word and compared to the training values. For 86% of the test set, the network output was closer (in terms of root mean-squared distance) to the full target vector (including both semantic and phonological representations) for that word than to any other target vector from this set. Of the 15 cases where this was not the case, seven involved stimuli that were still ambiguous at offset (e.g. *bell*, where *belt* was also part of the training set). A further five were low-frequency words (all had a corpus frequency of 5 or less per epoch). For these items, the network often extracted the phonological representation, but not the semantic representation. Finally, for three words there was no obvious pattern to the network's response.

[3]Experimentation with the number of hidden units confirmed that the degree of success at the semantic mapping depended strongly on the number of hidden units. However, increasing the number of hidden units also led to impractically long training times.

In summary, two factors interfered with the capacity of the network to perform the task. First, stimuli that were onset-embedded in other words could not be recognised properly at offset. This is a property of the dataset rather than the model and we shall return to this issue later. Secondly, words which had a low frequency were also not learned well, with the phonology of the words often extracted, but not the more arbitrary lexical representation. This is a consequence of limiting the resources available to the network—if there is a limit on the number of arbitrary mappings the network can learn, then it only learns the ones that are presented most often during training.

The network was then tested on a set of stimuli designed to simulate the test conditions of Marslen-Wilson and Warren (1994; see Table 1). The W1W1 and N1N1 baseline stimuli all contained vowel place cues that matched the place of the following segment. All other stimuli contained mismatching cues to the place of articulation of the final consonant. For example, a W2W1 stimulus contained place information in the vowel that was consistent with the W2 word combined with the final consonant of W1. Only the word tokens had been presented to the network during training. The test words were presented to the network in a random order, with each test item preceded by the same two filler words. This ensured that the activations of the context units were equivalent at the start of each test word. The phonological and semantic activations were recorded at each time-step.

*Lexical Decision.* When comparing the model with the experimental data, we shall assume that output error scores (the discrepancy or distance between the network's output and the target output) correlate with response times derived from an attractor-based settling system (see Plaut et al., 1996, for a test of this assumption), and that a lexical decision response depends predominantly on the semantic rather than phonological output of the model.[4]

In the interpretation of activation-based models of word recognition, we are faced with the question of whether decisions should be based on activations in absolute terms (e.g. Morton, 1969) or relative to the set of competitors (e.g. Luce, Pisoni, & Goldinger, 1990; Marslen-Wilson, 1987). For distributed models, the same choice occurs—absolute activations correspond to absolute distance values, whereas relative activations correspond to relative distances. We looked at both absolute and relative values in the lexical decision simulation and found highly similar patterns of results. For the sake of brevity, we will report only the relative values.

[4]Much the same pattern of results is found when lexical decision is based on the entire output vector. This is because the target words and their most active competitors are phonologically highly similar, and thus including the phonological dimensions of the output vector does not, in this case, help to discriminate between them.

Using the relative figures, we must assume that a "yes" response in a lexical decision task relies on the successful isolation of a single lexical item that matches the current speech input. More specifically, the output of the network must become sufficiencly closer to a single word representation than to all its competitors. This is analogous to reaching a threshold separation between the most active localist candidate and its more weakly activated competitors—at this point, we can say that the network has isolated a single word as matching the speech input.

The network's predictions for lexical decision "no" responses are less obvious. The simplest option is to assume some kind of deadline for reaching the threshold separation, after which a "no" response may be made. However, this cannot be correct, because many experiments (including the ones we wish to model) show variations between experimental conditions in the time taken to make a "no" response. The original Cohort model (Marslen-Wilson & Welsh, 1978) assumed instead that a token could be rejected as a token of a word as soon as the cohort of matching candidates was reduced to zero. Unfortunately, our current model does not allow such an explicit prediction. This is partly because the dichotomy between membership and non-membership of the Cohort is broken in our model (cf. Marslen-Wilson, 1987) and partly because the model is operating on continuous speech and may predict that the end of a nonword is in fact the beginning of a word, weakly activating a large number of words. These are both desirable properties of a model of speech perception, but make the modelling of lexical decision data more difficult. We shall nonetheless adhere to the spirit of the original Cohort model, by assuming that active words (which do not reach the threshold separation) will delay "no" responses. When the output of the network is close to a word in semantic space without ever getting close enough to generate a "yes" response, it will delay the rejection of the current input as a word.

Figure 3 illustrates the behaviour of the network's semantic output vector as the test items are presented. The x-axis represents the time course of presentation of a test word, with a single phoneme presented to the network at each time-step. Within each triplet, the stimuli are identical up to word position $-2$. At position $-1$, the vowel is presented to the network, along with the probabilistic coarticulatory information about the final consonant (equivalent to the pre-splice cues in the experimental stimuli). At position 0, the deterministic information about the final consonant is presented (equivalent to the post-splice cues in the experiment). The solid lines plot the separation of W1 from its competitors for the word triplets and the dashed lines plot the separation of W2 from its competitors for the nonword triplets. Separation values were calculated by subtracting the RMS distance between the output vector and the target from the distance to the nearest non-target word representation. A positive value for this measure indicates

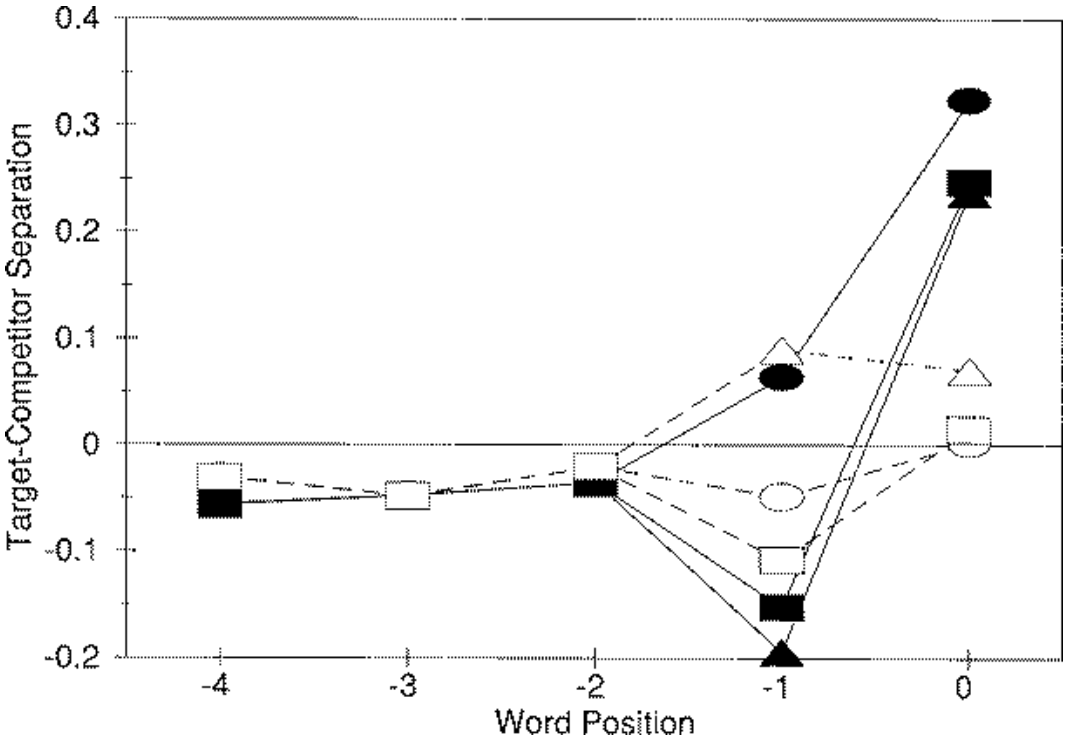


FIG. 3. Results of the lexical decision simulation. Each curve plots the mean distance from the target semantic representations relative to the nearest competitor for one experimental condition. The x-axis marks the time course of presentation of stimuli, with the final consonant (post-splice cues) presented to the network at position 0 and the vowel (pre-splice cues) presented at position −1. ●, W1W1; , W2W1; ■, N3W1; ○, N1N1; △, W2N1; □, N3N1.

that the network output is closer to the target than any competitor. We assume that word recognition would depend on a critical separation being reached.

For the word triplets, the model should "recognise" W1 to predict a "yes" response in a lexical decision. For the nonword triplets, the network should reject the input as a token of a real word and predict a "no" response. For these stimuli, W2 is phonetically the most similar word to the input stimulus and so is the most difficult to reject.

As Fig. 3 illustrates, there is a clear partition of the words and the nonwords by word position 0 (the end of the word). For the word stimuli, the network always ends up much closer to W1 than any to other word, whereas for the nonword stimuli, no single word becomes sufficiently separated from its competitors (i.e. the relative distances remain close to the zero line). Provided we choose a suitable critical separation level for making a "yes" response (between 0.1 and 0.23), the word tokens should be accepted and the nonword tokens should be rejected. Furthermore, within each triplet, the pattern of results is similar to the experimental pattern. For the word triplets, the W1W1 condition is the baseline. Here, as the featural information is presented, the separation of W1 from its competitors

increases, reaching a maximum at the end of the word of 0.33 (that is, the output of the network is 0.33 distance units closer to W1 than to any other word). In contrast, both cross-spliced tokens result in reduced activation of the W1 target (a weaker separation of W1 from its competitors), mainly on presentation of the mismatching coarticulatory information in the vowel. The patterns for these two tokens are highly similar, with slightly more mismatch for the W2W1 than the N3W1 token. Thus the delay in reaching the critical value for a "yes" response should be roughly the same for both conditions of mismatch. This fits the pattern of human response times illustrated in Fig. 2.

For the nonword triplets, all activations hover around the zero line, suggesting that W2 never becomes sufficiently separated from its competitors to generate a "yes" response. Looking at the differences between the three conditions, W2 is best activated by the W2N1 token. The increased activation of W2 for this condition would predict "no" responses should be slower than for the baseline condition (N1N1) and the N3N1 condition, which show similar patterns of response. Again this matches the pattern of results found in the lexical decision experiment (Fig. 2). Crucially, the lack of coherence between the cues to place in the N3N1 condition has little effect on the network's response relative to the N1N1 baseline. In terms of their similarity to the closest word (W2), the N3N1 and N1N1 conditions are equivalent—they both mismatch in terms of the pre-splice and post-splice cues. Because these cues are not integrated before the mapping onto lexical information, their coherence or lack of coherence does not play a part.

*Phonetic Decision.* The translation from localist phonemic output values to predictions of phonetic decision responses is a comparatively straightforward matter: The network's predictions should depend on the relative activations of the word-final phoneme nodes involved. We shall assume that the network's response to word-final ambiguities in a forced-choice task depends only on the activations of the three segments in the coda output group that share the manner and voicing of the ambiguous segments, but vary in place of articulation. For example, the network's predicted response (both in terms of choice of response and time taken) to the stimulus token *jog*, constructed from the onset of *job* and the final burst of *jog*, would depend on the relative activations of the /b/, /g/ and /d/ nodes in the coda group of the phonological output units. The activations of other units within this group are minimal and unaffected by the experimental manipulations.

Figure 4 plots the difference between the activation of the target segment (the segment that agreed with the post-splice cues in the stimulus) and that of its nearest competitor from the three relevant nodes in the coda section of the phonological output (either /b/, /g/ and /d/ or /p/, /k/ and /t/). This relative

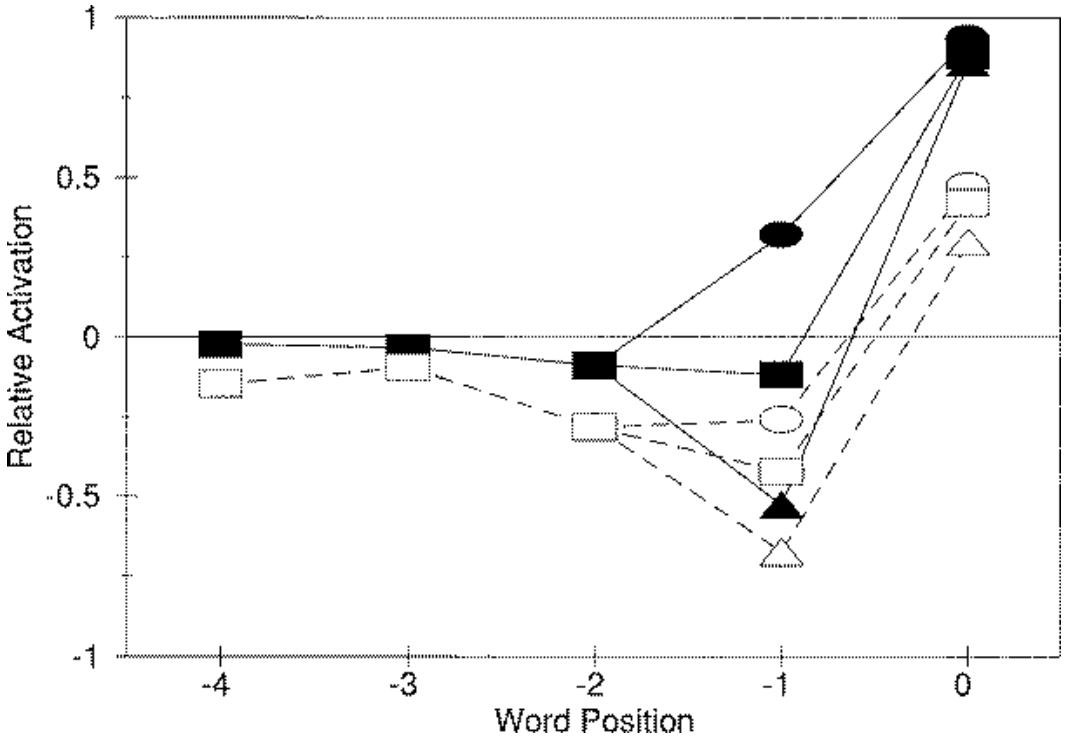


FIG. 4. Results of the phonetic decision simulation. Relative activation values were calculated by taking the activation of the target segment and subtracting the activation of its most active competitor from the bank of three word-final stop units sharing the target voicing (either /d/, /b/ and /g/ or /t/, /p/ and /k/). ●, W1W1; , W2W1; ■, N3W1; ○, N1N1; , W2N1; □, N3N1.

activation measure ranges from −1 (implying that the target has a zero activation and a competitor is fully activated) to 1 (implying that the target is fully activated and both competitors have zero activation). Comparison of the word and nonword triplets shows a strong effect of lexical status on the network response, which overshadows the effects within each triplet. For all word sequences the final relative activation is above 0.8, while for the nonword sequences the final activation varies between 0.32 and 0.49. This implies that responses should be slower for the nonword sequences than for the word sequences. A significant lexical effect was also found in the experiment of Marslen-Wilson and Warren (1994), but across conditions nonwords were only 23 msec slower than words, whereas the effects of mismatch within the word and nonword conditions were as much as 133 msec.

Aside from this overall difference between the responses to word and nonword sequences, the patterns within the two sequence types are reasonably similar. Compared to the baseline conditions, the patterns involving nonword onsets (i.e. N3W1 and N3N1) produce mismatch effects, but these mismatch effects are weaker than for the mismatching conditions with word onsets (i.e. W2W1 and W2N1). This pattern of results fits the response time data for the nonword sequences well (since the mismatching effect of the W2N1 stimuli was roughly twice that of the N3N1 stimuli),

although it slightly underestimates the effect of mismatch for the N3W1 condition (see Fig. 2). Even in the phonetic decision simulation, the effects of mismatch for two nonwords spliced together is weaker than for a word spliced onto a nonword. Here, the difference can be thought of as a lexical effect. The network has a strong preference for activating the phonological representations of words over nonwords (because all the training patterns are words). This means that when the vowel transition cues of a W2N1 stimulus are presented, there is a strong bias in the network towards a phonological response consistent with W2, which is inconsistent with the place of the final segment. For a N3N1 stimulus, the lexical bias is weaker or non-existent on presentation of the vowel transition, because the information received so far is already inconsistent with the phonological representation of any word. This makes it easier to switch to the place of N3 when the final consonant is presented, reducing the mismatch effect.

*Summary of Results.* Figures 5 and 6 summarise the network simulations (based on five separate training runs), illustrating a comparison between the experimental results of Marslen-Wilson and Warren (1994) and a transformation of the network output. For each graph, the experimental response times are plotted on a millisecond scale and the predictions derived from the network are plotted on an arbitrary interval scale (i.e. the zero value does not correspond to a zero response time). For 9 of the 12 experimental conditions (the 6 conditions in the phonetic decision experiment and the 3 word conditions in the lexical decision experiment), the network's predictions can be derived using threshold values, by interpolating the time course graphs in Figs 3 and 4. The threshold values for the lexical and phonetic decision simulations were chosen so that the predicted response times for the W2W1 conditions in the two tasks were equated (since the response times for these conditions were also the same). Each threshold value would be represented by a horizontal line in Figs 3 and 4. The correlate of response time in each condition can then be read off by finding the point on the x-axis at which the curve for that condition crosses the threshold (see Fig. 5).

For the remaining three conditions (the nonword conditions of the lexical decision simulation), a threshold model is not applicable. Instead, we simply plotted the separation values for W2, summed across the final two time-steps (the only points for which the stimuli diverge). The assumption underlying this comparison is that a larger separation value (implying a greater activation of W2) will inhibit a "no" response, resulting in a longer response time (see Fig. 6). Unfortunately, there is no obvious way to relate these values to the predictions derived from the word conditions.

Overall, there is a high level of agreement between the network's predictions and the experimental data. Within each triplet of conditions, for

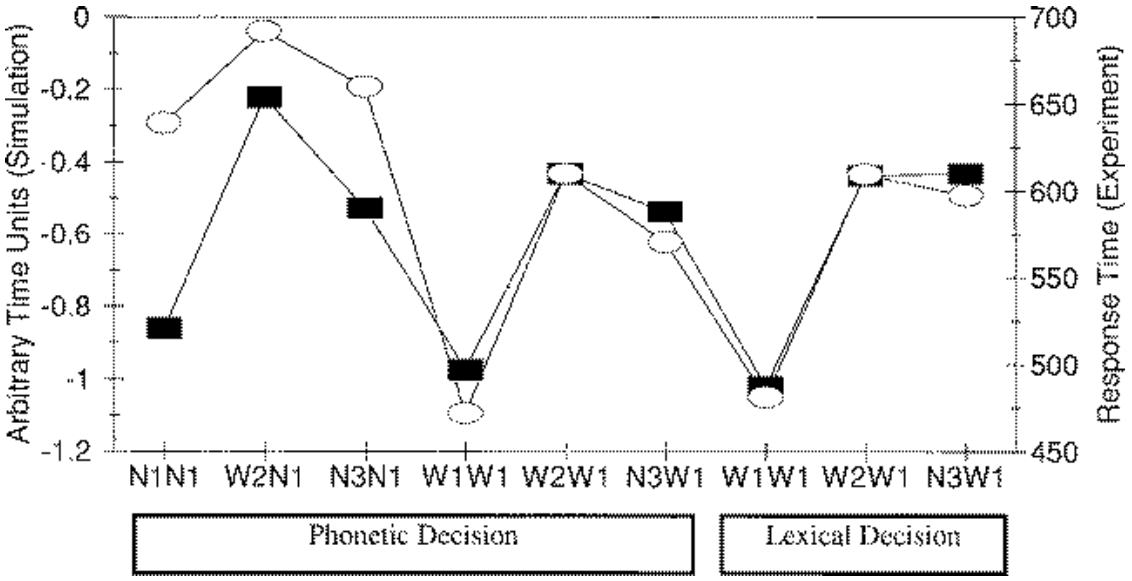


FIG. 5. Comparison between the experimental data (■) and the simulation results (○) for the phonetic decision and word lexical decision data. The experimental data use the millisecond scale on the right-hand side and the simulation data use the scale on the left-hand side (see text for an explanation of how these were calculated).

both lexical and phonetic decision simulations, the correct pattern of results is obtained. The only obvious discrepancy between the simulations and the experimental data is in the nonword conditions for the phonetic decision. Here, the pattern within the nonword conditions is correct (in that the effect

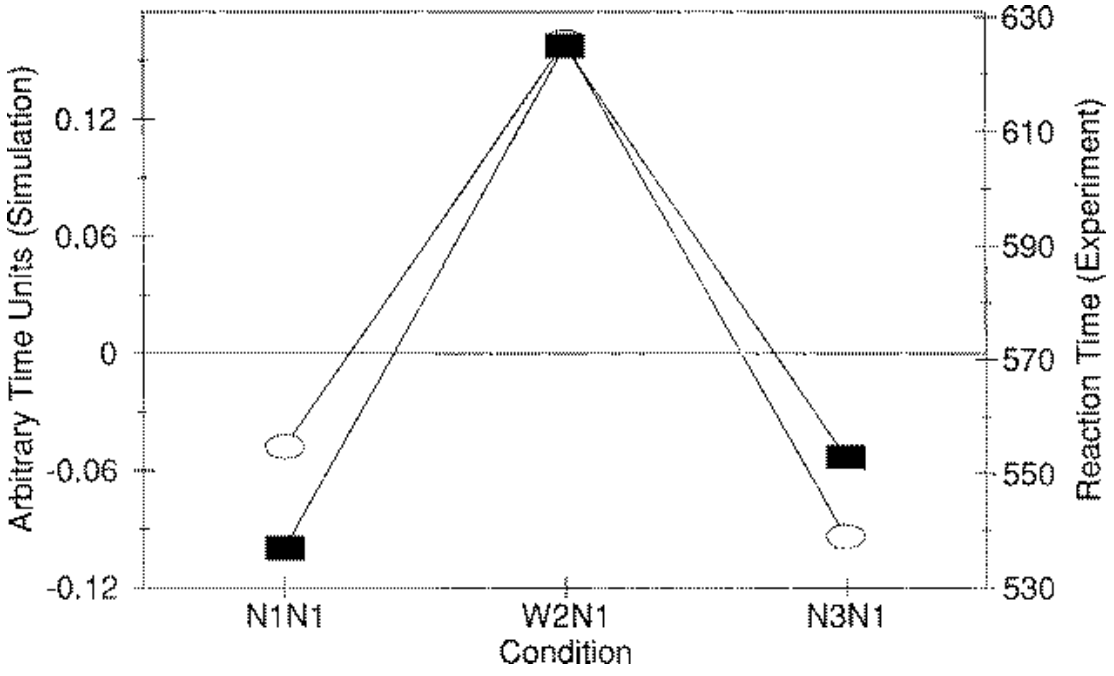


FIG. 6. Comparison between the experimental data (■) and the simulation results (○) for the nonword lexical decision data. The experimental data use the millisecond scale on the right-hand side and the simulation data use the scale on the left-hand side (see text for an explanation of how these were calculated).

of mismatch in the W2N1 condition, compared to the baseline N1N1, is roughly double the effect for N3N1, but the predicted response times are all much higher than the actual values. This reflects the fact that phonological activations of segments that combine to form nonwords are much lower than those that form words. This problem is similar to the problems highlighted by Besner, Twilley, McCann and Seergobin (1990) for Seidenberg and McClelland's (1989) model of reading (i.e. mapping from orthography to phonology), where performance on items used during training was good, but generalisation to nonwords was much worse than human performance.

Plaut et al. (1996) showed that better performance was possible on the reading task when a network was used which allowed a greater degree of interaction within and between levels. This allowed the network to develop attractors based on the subregularities involved in the mapping from orthography to phonology and was thus able to generalise better to other orthographic patterns. A similar modification applied to the feature-to-phoneme mapping could well produce better performance on the phonological layer for nonwords.

A second reason for the weaker activation of phonological output units for nonwords is unique to models of speech perception. Our network was trained on a continuous stream of speech, with no gaps or physical cues to the onsets or offsets of words. This reflects the situation that the human language learner is faced with, in which words are generally spoken in utterance context rather than in isolation. However, there is little doubt that to the fully developed perceptual system, silence at the beginning or end of a word (along with a corresponding lack of coarticulation) is valuable information. Because we had trained the network without gaps between words, we also had to test the network with a continuous stream of tokens (interspersed with filler words). In the evaluation of nonwords, this lack of word boundary cues makes the task of the network particularly difficult. For example, the /b/ in the token *smob* could, quite plausibly, be a word-initial /b/ following on from the nonword *smo*. Thus, there is ambiguity as to whether the network should activate the /b/ node in the onset or the coda section of the phonological output vector. For humans, this ambiguity is not present because they can make use of the silence before and after the speech to determine the syllabic position of the /b/. We expect, therefore, that if we introduced a small proportion of gaps between words in the training set and then tested the network on isolated words, the phonological performance on nonwords would be improved.

To summarise, our objectives in this section were to explain the pattern of match and mismatch found in the earlier experiments and to provide a single basis for representing both words and nonwords. Both these objectives were achieved. The network successfully simulated the pattern of mismatch found in the lexical and phonetic decision experiments of Marslen-Wilson and

Warren (1994). Compared to the baseline conditions, the amount of mismatch predicted by the network depended strongly on the lexical status of the pre-splice and post-splice components. In the case where both these components were spliced from nonwords, there was no effect of mismatch in the lexical decision simulation and a reduced effect of mismatch in the phonetic decision. The model is able to accommodate these data because it allows speech information to map directly onto lexical knowledge, without having to first integrate partial cues to phonemic identity.

The perception of phonological form is carried out in parallel with the access to semantic knowledge. This provides a single representational basis for the perception of words and nonwords alike. Nonwords access the same level of representation as words, but less completely. Lexical effects on nonword perception can thus be observed without recourse to interactive top-down flow of information (Norris, 1993). At present, the performance of the network on nonwords is not optimal, and predicts longer reaction times in the phonetic decision experiment than were actually found. However, we expect that improvements in the network architecture and training regime will lead to better nonword performance.

## THE TIME COURSE OF LEXICAL ACCESS

In this section, we investigate the basic properties of the model and compare them to our knowledge of the human system. There are now a number of properties of the lexical access system that are generally agreed on by researchers in the field. One of these is the assumption that as words are heard, the meanings of multiple candidates are activated (Marslen-Wilson & Zwitserlood, 1989; Zwitserlood, 1989). The state of these activations, and consequently the recognition point of a word, depends on the interaction between sensory input and factors such as competitor environment and word frequency (Luce et al., 1990; Marslen-Wilson, 1987, 1990).

### Multiple Lexical Candidates

At first sight, distributed representations seem inadequate for the simultaneous activation of multiple lexical candidates. In a localist system, such as the word level of TRACE, the activation of multiple candidates is simple, since each word in the model's lexicon has its own separate node. If during lexical access three word candidates match the speech input so far, this can be represented in a localist system by completely activating all three words.[5] In a distributed system, all words are represented on the same nodes, so only one word can be perfectly represented at any one time. Two or more

[5]TRACE incorporates lateral inhibition between word nodes to prohibit such a state, but this is an optional rather than a fundamental property of a localist system.

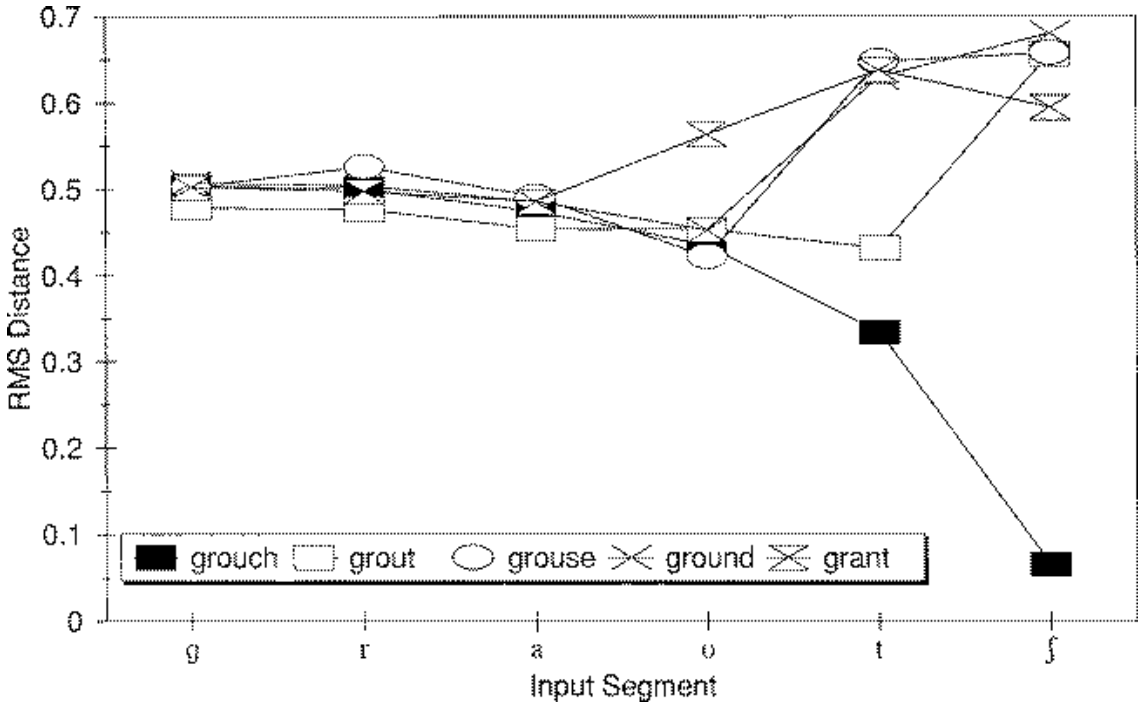


FIG. 7. Time course of semantic activation for the word *grouch*. Each line plots the distance from the network output vector for a selected word.

representations will interfere and can only be represented on the same nodes imperfectly. We will argue that this interference, far from being a problem, is an integral part of the word recognition process, since effects of factors such as competitor environment and word frequency fall naturally out of such a representational system.

Figure 7 illustrates the time course of activation at the semantic level as the sequence /graʊtʃ/ (grouch) is presented.[6] Each line shows the distance in semantic space between the output vector and a selected word representation. On presentation of the first segment, the network finds a vector that is roughly equidistant from all words (i.e. in the middle of the space). As more information is presented, the output vector moves closer to the matching words and away from the other words. This continues until, on presentation of the final segment, the network isolates the semantic representation of *grouch*.

This demonstrates a number of features of the model. First, the model shows evidence of the parallel activation of multiple candidates, with the distance from these candidates related to the number of candidates remaining active. When /ʊ/ is presented, the semantic vector is roughly 0.45 from four matching words. Later, the vector is slightly closer to the two

[6]The network was trained on a set of monosyllables designed to create a realistic competition environment for a single word-initial cohort (words beginning with /g/). The training set is described in the following section.

remaining matching words. However, only when the uniqueness point of the word is reached does the distance from the target word approach the minimum distance of 0. This behaviour differs from the old, localist Cohort model, where all word candidates matching the initial portion of the speech input (the word-initial cohort) are given a high activation value (equivalent to a low distance value) and the matching process involves the reduction of the activation of candidates that mismatch incoming speech.

The requirement that the network extracts the maximum lexical information at all points during word recognition causes the network to activate multiple lexical candidates, by constructing a blend of their representations (Kawamoto, 1993). Gaskell (1996) used statistical analyses of randomly populated vector spaces to examine the consequences of a distributed approach to parallel activation. The analyses showed that there is an upper limit to the number of words that can be informatively activated in this way. As the number of distributed patterns represented by the blend vector increases, their advantage over other randomly chosen patterns, in terms of similarity to the blend, decreases. In distance terms, there comes a point where the blend is closer to randomly chosen patterns than to some of the patterns it is supposed to represent. This is like having a version of the Cohort or TRACE models in which, at the start of a word, many cohort members have lower activations than non-cohort members. It does not mean that a distributed model cannot cope with a cohort-like selection process, but it does imply that the state of the output vector (e.g. the activation of the semantic units) will not portray the state of competition properly when many candidates match speech input early on in the processing of a word. Our model predicts that the degree of semantic activation engendered by a word-onset fragment should depend strongly on the number of words containing that onset. When there are many matching candidates, semantic activation will be weak or non-existent.

Gaskell (1996) also showed that this capacity was affected by both the structure and content of the lexical system. The number of dimensions in the representational space correlated positively with the capacity for multiple activation, so that a larger-dimensional space could cope with more patterns activated in parallel. However, increasing the sparseness of the representational system (e.g. by using a microfeatural semantic representation) acted in the opposite direction, reducing the limit on informative multiple activation.

The pattern of clustering of word representations in lexical space also affected performance. Initial simulations were carried out with randomly chosen distributed representations. This tends to produce an even distribution of representations within the lexical space. However, if this space encodes similarities between word meanings, it is unlikely to be evenly distributed. For example, the meaning of the word *apple* is likely to have

much in common with the meanings of other types of fruit, but little in common with the vast majority of word meanings. If this structure is preserved in the distribution in lexical space, then the problem of activating multiple meanings is compounded, since highly similar patterns of activation are more confusable than less similar ones. The exception to this rule is when the clustering encodes phonological information. In this case, the patterns that need to be activated in parallel (the representations of the cohort members) form part of the same cluster, which makes it easier to construct a vector that is close to those representations and relatively far from other representations.

## Competitor Effects in Lexical Access

To illustrate the results of the statistical analyses reported in Gaskell (1996), an investigation was carried out into the effect of candidate set size on lexical activations in the full model. We trained the network on a subset of English monosyllables designed to create a realistic competitor environment for a single word-initial cohort. The training set consisted of all the words in Plaut and co-workers' (1996) set of monosyllables that began with the phoneme /g/ ($n = 161$), plus all the words ($n = 115$) that were entirely embedded in one or more of the /g/ cohort set (e.g. *ape* in *gape* or *grape*).

This set of stimuli provide a reasonably realistic competitor environment for a single cohort of words, while remaining within the bounds of the monosyllabic phonological representation used at the output level. Since we wanted to look purely at the effect of the number of competitors, all training words were given an equal frequency of presentation during training. To reduce the co-occurrence strength between these words, 20 tokens of each word were assigned a random order in the training corpus, which was presented to the network 70 times during training.

The 161 training words beginning with /g/ were split into cohort competitor sets based on their first and second segments. The 11 sets varied strongly in their size, from the two words beginning /gi/ (*gear* and *geese*) to the 59 words with the onset /gr/. One word was selected from each of these sets to examine the effects of cohort size on the time course of activation. All test words were three segments in length and nine of them became unique on their final segment, while the other two words (from the largest of the competitor groups) remained ambiguous at offset due to the presence of longer embedding words in the training set. Thus the 11 test words were controlled for the cohort competitor environment of their first and third segments and varied on the number of cohort competitors at the second segment.

These words were presented to the trained network in a random order interspersed with repeated filler words, which ensured that the state of the

context units was the same before each test word. The output of the network was recorded at each time-step, with semantic outputs translated into distance values. Our expectation, based on the earlier analyses, was that at the point where the competitor environments of the test words differed (i.e. on the presentation of the second segment of each word), the distance from the output of the network to the target word would be inversely related to the number of cohort competitors. In other words, when a test word is one of many matching candidates, the output of the network will be relatively far from that word. With this outcome in mind, we carried out a Pearson correlational analysis between the distances and the inverse of the cohort set size at each word position (see Fig. 8).

On presentation of the second segment, we found a strong negative correlation between the inverted cohort size and RMS distance from the target words ($r = -0.78$, $r^2 = 0.60$, $P < 0.01$). This confirms the findings of the earlier analyses, showing that before the uniqueness point of a word, its semantic activation depends strongly on the number of candidates that match the input so far. When the candidate set size is small, the network can construct an output vector which is fairly similar to all possible candidates' representations. As the set size increases, this becomes more difficult and the distance from the candidates also increases (in localist terms, their

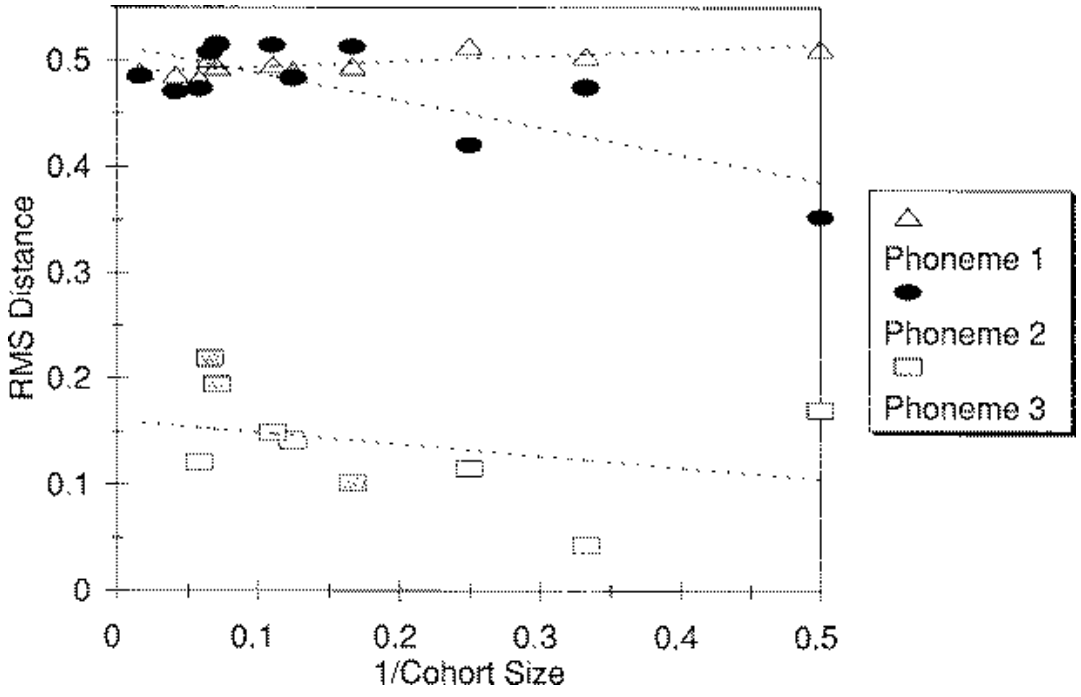


FIG. 8. Results of the competition simulation. The data for 11 three-segment words are plotted, with each word occupying a position on the x-axis corresponding to the inverse of the size of its cohort competitor set (the set of words matching the test word on its first two segments). For nine words, the RMS distance between the semantic output vector and the target representations of those words on presentation of all three segments are plotted. For the remaining two words, only the data for the first and second segments are plotted.

activations decrease). This aspect of the model diverges from the Cohort and NAM models, in which a word's competitor environment cannot *directly* affect its activation. Instead, the behaviour is closer to that predicted by interactive models such as TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994), which employ lateral inhibition between word candidates to reduce activations.

A closer examination of the data for phoneme 2 in Fig. 8 suggests that although the overall negative correlation between the inverse of cohort size and semantic distance is strong, it may be restricted to the smaller word cohorts (i.e. the points towards the right-hand side of the graph). This is a consequence of the limit on parallel activation of multiple distributed patterns discussed earlier (see also Gaskell, 1996). The ability of a network to activate multiple distributed patterns in parallel is severely disrupted for large numbers of patterns. Thus, for the larger cohort sets, it is unclear whether we can really say that the network is entertaining all the hypotheses in parallel.

On presentation of the third phoneme, most text words became uniquely identifiable and the network could isolate their full lexical representation. Excluding the data for the two non-unique items, there was no significant correlation on presentation of the final phoneme ($r = -0.31$, $r^2 = 0.10$, $P > 0.1$). This illustrates a further difference between our model and other models of competition. In a system of interactive activation such as TRACE, word activations build up gradually, since the activation of a word at any particular point in time depends on a function of both its current input and its activation at the previous time-step. This creates continuity and implies that the effects of competition will remain evident for some time, even after there is sufficient input to resolve any ambiguity. TRACE therefore predicts that activation will reflect the cohort competitor environment even after the uniqueness point has been reached.

The competitor effects observed here are more transient, and are eliminated as soon as there is sufficient bottom-up evidence to uniquely specify a word. This is because the network does not rely so much on residual activations in the calculation of its output. In effect, each time-step offers the opportunity for a complete reassessment of the words underlying the current input. This makes the lexical matching process more efficient, since competitors only affect word activations before the uniqueness point is reached.

The analysis of the lexical distances on presentation of the first segment showed an unexpected significant correlation in the opposite direction to the effect observed on the second segment ($r = 0.79$, $r^2 = 0.62$, $P < 0.01$). The first phoneme was the same for each word (/g/) and so the same number of words ($n = 161$) matched the speech input so far. Our expectation was that the network would construct a semantic vector which was (as far as possible)

equally close to all matching words. In fact, the output of the network was slightly closer to words with large cohort sets than to those with smaller cohort sets, reversing the normal effect of competition. Since this observation is based on the repeated presentation of a single phoneme, it should be interpreted with caution, but such a result does seem plausible when the interaction between the phonological and semantic tasks is considered.

Because all the words in the training set are presented with equal frequency, the network cannot make any useful prediction about the identity of a word on the basis of the initial /g/. At the same point in time, however, the phonological layer must reflect the relative probabilities of the available options. Although each word is equally frequent during training, different segments will occur in words starting with /g/ with different frequencies. Therefore, the state of the phonological level of output will reflect these frequencies, activating more frequent segments to a slightly greater extent than less frequent segments (as well as activating the word-initial /g/ node strongly). To some extent, these frequencies will correlate with the cohort competitor set size (for example, the greater cohort set size for /gr/ than /gl/ words implies that /r/ will occur more frequently than /l/ in the context of a word-initial /g/). There is an overriding tendency for the network to produce consistent semantic and phonological outputs. This interaction between the two sections of the output vector implies that the semantic layer will be biased towards an output coherent with the phonological output, meaning that the semantic vector should be slightly pushed towards the representations of words which contain the more frequent segments, namely those in the larger cohort sets. The upshot of this interaction is that there are very small, transient advantages for words in large cohort sets, similar to "gang effects" found in visual tasks (Andrews, 1989, 1992). However, these effects are quickly swamped by a much stronger bias in the opposite direction.

In summary, the main finding of these simulations is that our model predicts strong effects of competitor set size when the speech signal is ambiguous, but little or no effect once this ambiguity is resolved. The experimental data relating to this prediction are difficult to evaluate. Some evidence for effects of competitor set size comes from cross-modal priming studies of word recognition. Zwitserlood (1989) used prime words embedded in sentential context to examine the time course of word recognition. Subjects were presented with spoken sentences, and a visual target related to the prime word was used to probe activation levels at varying points during the presentation of the prime. Zwitserlood was primarily interested in the interaction of context during word recognition, but if we examine just the conditions where the context was neutral, we find that probes in early positions (where the speech is still ambiguous) elicit

20–30 msec of priming, compared with 40–50 msec effects at the recognition point (where the speech is now unambiguous). Similarly, Marslen-Wilson (1990) used cross-modal repetition priming to examine the activation of monosyllabic words. The prime words were either presented whole or with the final consonant cut off, and all primes had at least one cohort competitor with the same onset and nucleus. For the fragmented words, where more than one word candidate remained active, the priming of the visual target was modest (between 4 and 43 msec depending on frequency and competitor environment). The complete primes, however, evoked over 100 msec of priming in all conditions.

These results can be taken as evidence for an inverse relationship between the number of matching words and their activation. Early in a prime word there are multiple potential matches, and the priming of a target related to one of those is weak, compared to the amount of priming obtained by the full prime word. However, there are a number of reasons why we should be cautious about drawing such a conclusion. First, these studies use associative priming to assess lexical activations, which may reflect co-occurrence relationships between words rather than the activiation of their meanings (Moss, Hare, Day, & Tyler, 1994; Plaut, 1995). These studies also confound processing time (i.e. the time available to process the prime before the presentation of the target) with the availability of stimulus information. Zwitserlood and Schriefers (1995) have shown that both factors can influence the degree to which cross-modal priming occurs. Furthermore, a study by Marslen-Wilson and Gaskell (1992), which used a similar methodology to Marslen-Wilson (1990) to examine the time course of recognition of multisyllabic words, found no difference between the priming levels of complete bisyllabic words and the same words fragmented before their uniqueness point (the final consonant).

## Frequency and Competition Effects

The simulation described above employed a training set in which all words were presented with the same frequency. This is unrealistic, since there are large variations in the frequency of usage of words. It is well established that connectionist networks are sensitive to this frequency information (e.g. Kawamoto, 1993; Plaut et al., 1996; Seidenberg & McClelland, 1989). Learning in connectionist networks generally involves small adjustments of weights to reduce the difference between the network's output for any particular pattern and the desired output (the training output). Thus, the state of the network's weights depends predominantly on the patterns it is exposed to most often.

In terms of the current model, this means that the output of the network during word recognition reflects the lexical representations of the word

candidates still matching the speech input, but that this output is biased in favour of the higher-frequency word candidates. Again, this is a consequence of the training regime, in which the model must produce the maximum lexical information at all points during the presentation of a word. For example, if during word recognition two candidates remain active, one of which has been presented to the network 1000 times during training and the other only 500 times, the network's output should be closer to the lexical representation of the high-frequency word (i.e. its activation should be greater), since this is the more likely of the two candidates.

This behaviour was demonstrated using a simple variant of the competitor environment simulation in which half the words were presented during training 20 times per epoch and half 40 times. This gives a training frequency range which should produce a bias towards the more frequently presented training words during testing. Figure 9 shows the time course of semantic activation of selected words from the training corpus. The words form cohort competitor pairs, such as *glum* and *glove*, which diverge phonemically from each other on their final segment and from the rest of the training words on the penultimate segment (i.e. there are no other words in the training set beginning /glʌ/). One of each pair was a member of the high-frequency training group, and the other was a low-frequency word.

At word position −1 (the penultimate segment), only the two test words are consistent with the input so far. At this point, the network occupies a position in between these words in lexical space, but closer to the high-frequency word. The distance from the high-frequency word (0.26) is roughly half the distance from the low-frequency word (0.48), reflecting the 2:1 ratio of presentation frequency during training. When the ambiguity is resolved (position 0), the network moves towards the matching candidate and away from the mismatching one, although in this case a weaker frequency effect remains, suggesting that the high-frequency representation has been better learnt.

Although the frequency effects before and at the uniqueness point are both caused by the same overriding principle of error reduction, there are important differences between them. The frequency effect at the uniqueness point (position 0) simply reflects how well the patterns have been learnt and is heavily dependent on absolute frequency as well as implementational parameters, such as the number of hidden units and the learning rate and algorithm. It follows that given sufficient resources and training time, the effect could diminish or disappear altogether as the performance of the network reaches a ceiling level. On the other hand, the frequency effect before the uniqueness point (position −1) has a quite different cause and is more robust. In cases of ambiguity, the network weights the available options according to their relative likelihood, based on frequency of presentation during training. No amount of training can eliminate this

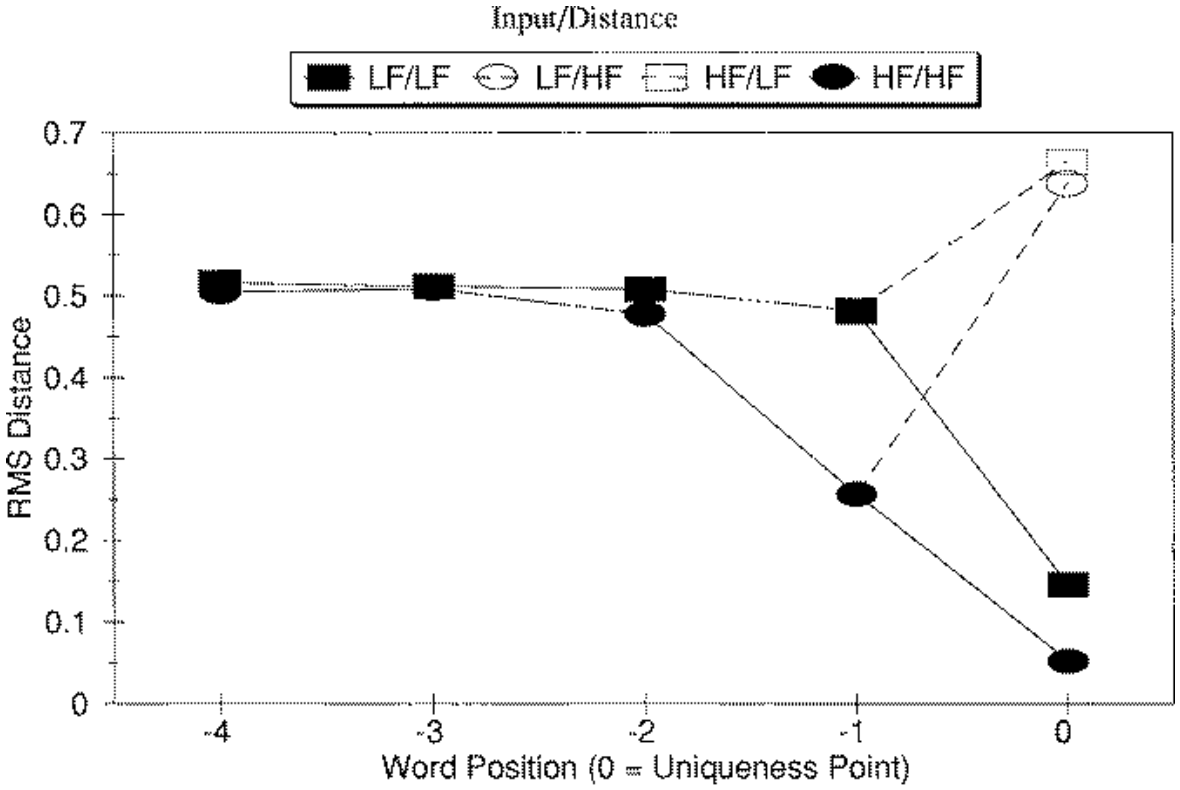


FIG. 9. The time course of activation of frequency variant word pairs (e.g. *glum* and *glove*) which diverge on their final segment. The squares plot the mean distance between the semantic output of the network and the low-frequency members of the pairs, and the circles plot the same measure for the high-frequency members. In each case, the solid line marks the activation pattern when the input matches the target and the dashed line marks the activation pattern when the input and target mismatch (e.g. the distance from *glove* when *glum* is presented).

temporary ambiguity and so, even when the network is performing perfectly, transient probabilistic frequency effects remain. These effects during the process of word recognition are similar to those predicted by the revised Cohort model (Marslen-Wilson, 1987) and are supported by data from cross-modal experiments (Marslen-Wilson, 1987, 1990) showing differences in the amount of priming engendered by high- and low-frequency words before their uniqueness points.

## DISCUSSION

We are now in a position to summarise the properties of the model we propose and evaluate it with respect to other models of spoken word recognition. We shall also extend our discussion to cover related topics that have not been explicitly simulated so far.

## The Processing Environment for Lexical Access

There is now a large body of evidence indicating that lexical access is highly sensitive to the phonetic details of speech (e.g. Andruski, Blumstein, & Burton, 1994; Connine, Blasko, & Titone, 1993; Marslen-Wilson et al., 1996; Warren & Marslen-Wilson, 1987, 1988). Although the perceptual system is highly successful at filtering out many types of noise and variation, phonetic and subphonetic deviations have a potent effect on the success with which access to lexical entries occurs. This is unsurprising, given the nature of the task that the perceptual system faces. A recognition system that has trouble distinguishing *sat* from *sap*, *sack* or *sad* will be inadequate for the task of understanding everyday speech. Even for words that have less crowded phonological neighbourhoods, the ability to distinguish new words from familiar ones is important. Indeed, lexical neighbourhood density seems to have little effect on the goodness of fit needed, since deviations at the ends of long words with no close cohort competitors (e.g. *apricod* for *apricot*) are just as disruptive to the lexical access process as shorter words with more competitors (Marslen-Wilson & Gaskell, 1992).

This seems an unlikely property to emerge from a connectionist simulation of such a process. Connectionist learning networks can be thought of as operating on a "need to know" basis. In a categorisation task, for example, they tend to encode the minimum information necessary to distinguish each stimulus from its neighbours.

However, our simulations of Marslen-Wilson and Warren's (1994) data show that our model, like humans, does not develop such tolerance. Tokens such as *smob*, for example, which match the closest word (*smog*) on everything but the place of articulation of the final segment (and in the case of the W2N1 stimuli, even contain vowel transition cues compatible with that place), do not strongly activate the lexical representation of that word. Why does the network require so much phonetic detail to access the meaning of the nearest word?

We suspect there are a number of reasons for this behaviour. First, there is the fact that the network is trained on a large number of different monosyllabic forms, which forces the network to retain a reasonably complete representation of phonological forms to distinguish them from each other. A second point is that the network is forced to pay attention to the full phonological representation of each word by the nature of the task. The network needs to encode the phonological representation of each word because it must output this information alongside the stored semantic knowledge. This prevents the network from developing poorly specified representations of word form. A final reason for the model's intolerance of deviation lies in the method of presentation of the input. Each word is embedded in a continuous stream of speech with no explicit cues to the

beginnings and ends of the words. This means that the network does not have the luxury of knowing that the /b/ in *smob* is part of that token. Instead, it could be the onset of a new word, meaning that a large new set of lexical representations must be considered, all of which will bias the output of the network away from the representation of *smog*.

The differences between the distributed model and TRACE have already been examined in some detail. TRACE employs a system of interaction between nodes which allows inhibition only between nodes in the same representational level and facilitation only between nodes in different levels. Our model is more similar to Shortlist (Norris, 1994), in that it makes less use of lateral inhibition and more of bottom-up inhibition. This similarity is not surprising, given that the initial stage of Norris's model (a lexical dictionary search) is intended to represent the behaviour of the recurrent networks examined in his earlier research (Norris, 1990, 1991).

## Distributed Representations

The use of a single distributed level of representation is a crucial feature of our current model. We have shown that competition between word nodes in a localist model can be re-described as interference between multiple distributed representations. The differences between the two accounts of competition can be subtle, and it is not easy to distinguish between them experimentally, even though the difference between the two representational systems is taken to be fundamental. Nonetheless, functional differences do exist and require experimental evaluation.

In the distributed model, direct competition is obligatory—the distance between a lexical blend and one of its component words depends strongly on the number of active words and their relative frequency. In effect, the model implements a form of the Universal Frequency Franchise argued for by Bard and Shillcock (1993), in which competitors are effective in proportion to their frequency. Localist models can also implement such a system, but this is an optional property. For example, TRACE uses inhibitory links between word nodes to effect a direct competition. On the other hand, Cohort—both in its original form (Marslen-Wilson & Welsh, 1978) and a more recent version (Marslen-Wilson, 1993)—and the NAM model (Luce et al., 1990) assume that word candidates do not directly affect each other's activations, although this factor can be taken into account at a decision level.

A second distinction between localist and distributed approaches is that the distributed blending model is inherently noisy. There is a limit on the number of representations that can be informatively activated in parallel. Beyond this limit, the system does not break down entirely, but it does become difficult to separate active from inactive words simply on the basis of

lexical distance. This limit depends on a number of factors, including the dimensionality of the representational space and the number of words in the lexicon, as well as the distribution of word representations within this space.

Marslen-Wilson (1987) describes the Cohort model in terms of three functions: access, selection and integration. *Access* refers to the initial mapping onto lexical representations, which in the Cohort model involves the access to multiple semantic information as well as phonological information. *Selection* describes the process of eliminating mismatching candidates and *integration* describes the mapping of the semantic and syntactic information about the recognised word onto the high-level utterance representation. The Cohort model proposed that multiple candidates were accessed and assessed in parallel. This was in direct contrast to serial search models, for which selection took place prior to access (Forster, 1976).

The distributed model refines and develops the stance taken by the Cohort model. The parallel assessment of multiple candidates is retained, but access to the lexical representations of these words becomes intimately tied in with the process of selection. As we have shown, the degree to which lexical knowledge about a word is accessed depends on the number of candidates remaining, with only partial and degraded information available for multiple candidates. It is only once selection is complete that the meaning of the remaining word can be fully isolated.

## Autonomy and Interaction in Lexical Access

For "box and arrow" models of perceptual processes, the direction of flow of information between the different boxes (levels of representation) is a critical issue. In the area of spoken word recognition, a wide range of views has been taken, ranging from a strict bottom-up flow of information (e.g. Forster, 1976; Norris, 1994) to fully interactive processing (McClelland & Elman, 1986).

Applied to connectionist learning models, the focus of this debate shifts slightly, since feedforward networks (once trained) operate in a strictly bottom-up manner, but show interaction between the various mappings they are required to perform (Norris, 1993). It still remains important, however, to ask what types of information interact, and under what circumstances (cf. Tabossi, 1993).

Our simulation of the data of Marslen-Wilson and Warren (1994) demonstrates strong effects of lexical status on the perception of word forms, which can be taken as evidence for interaction between lexical and phonological sources. However, our simulations so far have not addressed a more controversial form of interaction—effects of preceding sentential context on the recognition of words. In its current form, the model would no

doubt show little effect of sentence context on lexical activations because the dependencies and abstractions that must be learned for such behaviour to be observed far exceed the capacity of a simple recurrent network. However, we will assume that the capacity of the human perceptual system also exceeds that of a simple recurrent network, in which case we would expect contextual factors to create expectancies that affect the time course of lexical access.

A simple example of these expectancy effects comes from connectionist models of associative priming (Moss et al., 1994; Plaut, 1995). These models show sensitivity to word pairs that frequently co-occur in their training data (such as *cat* and *dog*), allowing one member of the pair to facilitate recognition of the other when tested. Word co-occurrence is a particularly obvious source of information, but more complex dependencies may also be learned and exploited in much the same way.

Even so, the effects of contextual factors on the word recognition process are unlikely to be strong. First, the information that can be gleaned from analysis of a word's preceding context is usually probabilistic, as opposed to the deterministic information available from analysis of the acoustic form of the word itself. Context will rarely provide firm enough evidence to rule out an incongruent candidate or accept a congruent one. All it can do is modify the output of the network moving it towards favourable candidates and away from unfavourable ones.

A second point to make is that there is a rather narrow time-window in which effects of context should be observable, as a consequence of the distributed blending approach to competition that we are proposing. Contextual effects can be likened to the transient frequency effect found in our current model. When the speech is ambiguous between two word candidates, such as *glum* and *glove*, there is a greater chance of the ambiguous token turning out to be the high-frequency word, and the network reflects these relative probabilities by moving the output vector closer to the representation of the high-frequency word (see Fig. 9). However, this effect is largely limited to the point in time at which the input could be one of these words, but no other. Earlier on in the processing of the word, the relative probabilities remain the same, but each word's *overall* probability is much smaller due to the presence of other matching word candidates. Thus, in terms of distance, the advantage enjoyed by the high-frequency word is much smaller.

A similar pattern of activity would be expected for contextually appropriate and inappropriate words—little advantage early on in the processing of a word, but stronger effects when few matching words remain. There is a slight difference between the effects of relative frequency and contextual appropriateness, since contextual appropriateness is a measure that can be applied to all words in a cohort, and may provide a consistent bias

in certain dimensions of lexical space. Nonetheless, this effect would still increase as the number of matching candidates diminishes.

This interpretation of context effects offers an alternative account of the effects of contextual appropriateness on the time course of spoken word recognition found by Zwitserlood (1989). She examined the effects of different types of preceding sentential context on the ability of a fragmented auditory prime word to facilitate recognition of a related visual target. She found that contextual appropriateness effects only emerged late in the word, but still before the sensory information was sufficient to entirely disambiguate the prime word.

Zwitserlood argued that this pattern of results was obtained because the appropriateness of the preceding context could only affect the selection stage of the recognition process and not the access stage (although Marslen-Wilson, 1989, argues that the same data imply that context affects only the integration stage). The distributed model makes similar predictions in terms of priming effects, but does so on the basis of the pattern of blending of distributed patterns that occurs during the processing of a word. Early on in the word, there can be little advantage of one word representation over a competitor, because there are too many other word representations that need to be represented by the single blend. Only later, when there is a small number of candidates, does the opportunity emerge for context to significantly affect the activations of the remaining candidates.

## The Role of Phonological Representation in Speech Perception

The incorporation of phonology alongside other forms of knowledge in a single distributed lexical space marks a departure from the standard model of lexical access in speech perception. Other models have assumed that phonology is either represented pre-lexically (e.g. McClelland & Elman, 1986), or both pre- and post-lexically (Cutler & Norris, 1979; Foss, Harwood, & Blank, 1980).

The model shares much in functional terms with these other models, since it shows lexical effects on the perception of phonological form and provides a basis for the perception of nonwords and words alike. However, it satisfies these constraints without integrating phonological knowledge pre-lexically. The model preserves relevant detail in the mapping process onto lexical representations, while still providing a compact and relatively abstract representation of the phonological form of both words and nonwords.

It is worth considering the possibility that although the network is not trained explicitly to develop pre-lexical phoneme-like units, it nevertheless develops distributed segmental representations to carry out its task.

Abstraction is common in connectionist simulations of linguistic tasks (e.g. Elman, 1990, 1993), and may also be occurring in our network, but subject to the constraint that relevant phonetic detail is not lost as a result. Indeed, such weak abstraction is functionally equivalent to the direct access from features approach.

## Competition Across Word Boundaries and Segmentation

Our discussion of competition effects so far has concentrated on the activation of onset-aligned words (e.g. *glum* and *glove*). However, McQueen, Norris and Cutler (1994) have demonstrated effects of competition between both onset-aligned and non-aligned words using a word-spotting task. Such effects are not necessarily problematic for our model provided the competing words remain the "focus" of the network output. In other words, the network should resolve ambiguity correctly—regardless of the alignment of the competing words—provided the critical discriminating input arises by the offset of the target word. However, when competition cannot be resolved by this time, the model will fail because it has only been trained to input information about the word containing the current speech input. The model lacks the ability to "look back" at previous ambiguities to resolve the conflicts on the basis of new evidence.

The lexical networks of TRACE and Shortlist offer one solution to this problem. However, it would be preferable, if possible, to develop a complete model of speech perception within the distributed learning approach. One way to encourage our current model to show these effects is simply to train the network to delay its output. By shifting the input one segment to the left, so that the task of the network is to extract the lexical and phonological information about the word that the previous segment belongs to, a greater proportion of ambiguities can be resolved. For example, the word *glue* cannot be fully identified by the current model because of the presence of the embedding word *gloom*. However, the same architecture trained with a lag of one segment allows this ambiguity to be resolved in the case of phrases such as *glue sniffing*, where the segment following glue (/s/) mismatches the embedding word.

Unfortunately, forcing a fixed delay on the network eliminates some of its attractive properties, particularly the ability to respond to incoming speech swiftly, as humans seem to do (Marslen-Wilson, 1973). Also, it is difficult to decide how long such a lag should be, since some ambiguities may continue several segments into the following word (and it is also unclear at what point the human perceptual system breaks down given such ambiguities). What is required is a more flexible output mechanism, which delays output only where necessary.

Content and Sternon (1994) modelled such effects using a simple recurrent network architecture by adding a second output window which was trained to output the identity of the previous word.[7] This allowed following context to affect the identification of a previous word, even for novel combinations of words. This approach, although still somewhat restrictive, appears a promising avenue of inquiry. It is important to recognise that the goal of our current model—access to word knowledge—is not the ultimate goal of the speech perception system. The problems of embedding may be more tractable for a distributed learning model when the target of the mapping is the meaning of the utterance, or at least the phrase, rather than the meaning of a single word.

A related issue is the segmentation of the speech continuum into word units. Models of segmentation have generally either emphasised the use of cues contained in the speech signal, so as to segment speech pre-lexically (e.g. Cutler & Norris, 1988), or advocated a process of lexical competition that operates in parallel with word recognition (McClelland & Elman, 1986; Norris, 1994). Experimental evidence for both these approaches is available (e.g. Cutler & Carter, 1987; Cutler & Norris, 1988; McQueen et al., 1994) and it is likely that the human system makes use of both types of process to segment speech efficiently. The statistical learning approach employed in our network is in a strong position to make use of all available cues to segment speech as the process of lexical access unfolds. We have shown that our network exhibits competition between lexical candidates, both within and across word boundaries. Simple recurrent network models have also displayed the capacity to pick up low-level cues to likely word boundaries, based on the statistical properties of the language (Cairns, Shillcock, Chater, & Levy, 1995; Gaskell, 1994). These cues can be employed flexibly and probabilistically by the network, allowing it to adapt its segmentation strategy according to the strength of information available.

## Predictions

The most prominent predictions of our model stem from its treatment of competition during lexical access. The distributed model predicts that the frequency and number of word candidates matching the speech input at any point directly affect their activations (as measured by distance in lexical space). This makes the model falsifiable, but unfortunately, a positive result would not rule out all localist models, since those that employ lateral inhibition (e.g. TRACE) make similar predictions. Evidence that multiple semantic representations can be activated without interference would be

[7] In fact, the model had only one output window, but the network was trained to output the identity of either the current word or the previous word, depending on the activation of an input node.

difficult to accommodate not only by our model, but also by models that treat semantic representations as fully distributed patterns of activation. We have begun to examine this prediction, looking at the effects of competitor environment on lexical activation using spoken word fragments in a priming task. The experiments show strong effects of the number of words matching the prime fragments, as predicted by our model (Gaskell & Marslen-Wilson, 1997).

The distributed model also predicts that the content of lexical representations should affect competition during lexical access. Gaskell (1996) has shown that sparseness affects the blending of distributed representations, and thus we might expect lexical variables such as imageability to interact with the competition between word candidates during lexical access. Preliminary data on this topic suggest that imageability does affect response times in naming and lexical decision tasks, particularly for spoken words in dense cohort neighbourhoods (Tyler, Voice, & Moss, 1996).

Finally, there is a possibility of determining the limit on the number of distributed representations that can be activated in parallel. This seems a more difficult question to test, because it relies on being able to tell the difference between very small amounts of activation and no activation at all (e.g. by using priming techniques). On the other hand, this would be a particularly interesting question to address, because it might give us clues about the dimensionality or richness of lexical representations (and of mental representations in general).

## REFERENCES

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 802–814.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 234–254.

Andruski, J.E., Blumstein, S.E., & Burton, M. (1994). The effects of subphonetic differences on lexical access. *Cognition*, *52*, 163–187.

Bard, E.G., & Shillcock, R.C. (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. Altmann & R. Shillcock (Eds), *Cognitive models of language processes: The Second Sperlonga Meeting*. Hove: Lawrence Erlbaum Associates Ltd.

Besner, D., Twilley, L., McCann, R.S., & Seergobin, K. (1990). On the connection between connectionism and data: Are a few words necessary? *Psychological Review*, *97*, 432–446.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1995). Bottom-up connectionist modelling of speech. In J.P. Levy, D. Bairaktaris, J.A. Bullinaria, & P. Cairns (Eds), *Connectionist models of memory and language*, pp. 289–310. London: UCL Press.

Connine, C.M., Blasko, D.G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition. *Journal of Memory and Language*, *32*, 193–210.

Content, A., & Sternon, P. (1994). Modelling retroactive context effects in spoken word recognition with a simple recurrent network. In A. Ram & K. Eiselt (Eds), *Proceedings of*

the Sixteenth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Cutler, A., & Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133–142.

Cutler, A., & Norris, D. (1979). Monitoring sentence comprehension. In W.E. Cooper & E.C.T. Walker (Eds), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 113–121.

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

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Forster, K.I. (1976). Accessing the mental lexicon. In R.J. Wales & E.W. Walker (Eds), *New approaches to language mechanisms*. Amsterdam: North-Holland.

Foss, D.J., Harwood, D.A., & Blank, M.A. (1980). Deciphering decoding decisions: Data and devices. In R.A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Gaskell, M.G. (1994). *Spoken word recognition: A combined computational and experimental approach*. PhD thesis, Birkbeck College, University of London.

Gaskell, M.G. (1996). Parallel activation of distributed concepts: Who put the P in the PDP? In G.W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pp. 284–289. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Gaskell, M.G., & Marslen-Wilson, W.D. (submitted). Discriminating local and distributed models of competition in spoken word recognition. In M.G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pp. 247–252. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Hinton, G.E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Hinton, G.E., McClelland, J.L., & Rumelhart, D.E. (1986). Distributed representations. In D.E. Rumelhart & J.L. McClelland (Eds), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. Cambridge, MA: MIT Press/Bradford Books.

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis*. Cambridge, MA: MIT Press.

Johansson, S., & Hofland, K. (1989). *Frequency analysis of English vocabulary and grammar*. Oxford: Clarendon Press.

Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank: Explorations in connectionist modelling. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1051–1062.

Jordan, M.I. (1986). Attractor dynamics and parallelism in a connectionist sequential network. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Kawamoto, A.H. (1993). Non-linear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*, 474–516.

Kawamoto, A.H., Farrar, W.T., & Kello, C. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1233–1247.

Klatt, D.H. (1989). Review of selected models of speech perception. In W.D. Marslen-Wilson (Ed.), *Lexical representation and process*. Cambridge, MA: MIT Press.

Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighbourhoods of spoken words. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing*. Cambridge, MA: MIT Press.

Marslen-Wilson, W.D. (1973). *Speech shadowing and speech perception.* PhD thesis, MIT, Boston, MA.

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

Marslen-Wilson, W.D. (1989). Access and integration: Projecting sound onto meaning. In W.D. Marslen-Wilson (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press.

Marslen-Wilson, W.D. (1990). Activation, competition, and frequency in lexical access. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

Marslen-Wilson, W. (1993). Issues of process and representation in lexical access. In G. Altmann & R. Shillcock (Eds), *Cognitive models of language processes: Second Sperlonga Meeting.* Hove: Lawrence Erlbaum Associates Ltd.

Marslen-Wilson, W.D., Moss, H.E., & Halen, S. van. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1376–1392.

Marslen-Wilson, W.D., & Gaskell, G. (1992). Match and mismatch in lexical access. *International Journal of Psychology*, *27*, 61.

Marslen-Wilson, W., & Warren, P. (1994). Levels of representation and process in lexical access. *Psychological Review*, *101*, 653–675.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29–63.

Marslen-Wilson, W.D., & Zwitserlood, P. (1989). Accessing spoken words: On the importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 576–585.

Martin, J.G., & Bunnell, H.T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant–vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 473–488.

Mason, M.E.J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 3–23.

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

McQueen, J.M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition—spotting words in other words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 621–638.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Moss, H.E., Hare, M.L., Day, P., & Tyler, L.K. (1994). A distributed memory model of the associated boost in semantic priming. *Connection Science*, *6*, 413–427.

Moss, H.E., McCormick, S.F., & Tyler, L.K. (this issue). The time course of activation of semantic information during spoken word recognition: function precedes form. *Language and Cognitive Processes.*

Norris, D. (1990). A dynamic-net model of human speech recognition. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing.* Cambridge, MA: MIT Press.

Norris, D. (1991). Rewiring lexical networks on the fly. In *Proceedings of the 2nd European Conference on Speech Communication.* Berlin: ESCA.

Norris, D. (1992). Connectionism: A new breed of bottom-up model. In R.G. Reilly & N.E. Sharkey (Eds), *Connectionist approaches to natural language processing.* Hove: Lawrence Erlbaum Associates Ltd.

Norris, D. (1993). Bottom-up connectionist models of "interaction". In G. Altmann & R. Shillcock (Eds), *Cognitive models of language processes: The Second Sperlonga Meeting.* Hove: Lawrence Erlbaum Associates Ltd.

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

Plaut, D.C. (1995). Semantic and associative priming in a distributed attractor network. In J.D. Moore & J.F. Lehman (Eds), *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pp. 37–42. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Plaut, D.C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., & Patterson, K.E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Rumelhart, D.E., Hinton, G.E., & McClelland, J.L. (1986). A general framework for parallel distributed processing. In D.E. Rumelhart & J.L. McClelland (Eds), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. Cambridge, MA: MIT Press/Bradford Books.

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

Sharkey, A.J.C., & Sharkey, N.E. (1992). Weak contextual constraints in text and word priming. *Journal of Memory and Language*, *31*, 543–572.

Smolensky, P. (1986). Neural and conceptual interpretation of PDP models. In D.E. Rumelhart & J.L. McClelland (Eds), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models*. Cambridge, MA: MIT Press/Bradford Books.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–74.

Stevens, K.N. (1986). Models of phonetic recognition II: A feature based model of speech recognition. In P. Mermelstein (Ed.), *Proceedings of the Montreal Satellite Symposium on Speech Recognition*. Montreal.

Stevens, K.N., Manuel, S.Y., Shattuck-Hufnagel, S., & Liu, S. (1992). Implementation of a model of lexical access based on features. In J.J. Ohala, T.M. Nearey, B.L. Derwing, M.M. Hodge, & G.E. Wiebe (Eds), *Proceedings of the 1992 International Conference on Spoken Language Processing*. Edmonton: University of Alberta.

Streeter, L.A., & Nigro, G.N. (1979). The role of medial consonant transitions in word perception. *Journal of the Acoustical Society of America*, *65*, 1533–1541.

Tabossi, P. (1993). Connections, competitions and cohorts: Comments on the chapters by Marslen-Wilson; Norris; and Bard and Shillcock. In G. Altmann & R. Shillcock (Eds), *Cognitive models of speech processing: The Second Sperlonga Meeting*. Hove: Lawrence Erlbaum Associates Ltd.

Tyler, L.K., Voice, J.K., & Moss, H.E. (1996). The interaction of semantic and phonological processing. In G.W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pp. 219–222. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Warren, P., & Marslen-Wilson, W.D. (1987). Continuous uptake of acoustic cues in spoken word-recognition. *Perception and Psychophysics*, *41*, 262–275.

Warren, P., & Marslen-Wilson, W.D. (1988). Cues to lexical choice: Discriminating place and voice. *Perception and Psychophysics*, *43*, 21–30.

Whalen, D.H. (1982). *Perceptual effects of phonetic mismatches*. PhD thesis, Yale University, New Haven, CT.

Whalen, D.H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception and Psychophysics*, *35*, 49–64.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*, 25–64.

Zwitserlood, P., & Schriefers, H. (1995). Effects of sensory information and processing time in spoken-word recognition. *Language and Cognitive Processes*, *10*, 121–136.