# Modeling the Perception of Spoken Words

**M. Gareth Gaskell**
Centre for Speech and Language,
Psychology Department, Birkbeck College,
Malet Street, London WC1E 7HX
g.gaskell@psyc.bbk.ac.uk

**William D. Marslen-Wilson**
Centre for Speech and Language,
Psychology Department, Birkbeck College,
Malet Street, London WC1E 7HX
w.marslen-wilson@psyc.bbk.ac.uk

**Abstract**

We present a new distributed connectionist model of the perception of spoken words. The model employs an internal representation of speech that combines lexical information with abstract phonological information. We show how a single distributed representation of this type can form the basis for the perception of words and nonwords alike. The model is tested against lexical and phonetic decision data from Marslen-Wilson and Warren (1994). These experiments examined the integration of cues to place of articulation during lexical access and showed a pattern of results which proved difficult to accommodate in previous models. The use of a single, late, phonological representation allows this pattern of results to be simulated and has the potential to incorporate many other properties of the human system.

## Introduction

This paper describes a new approach to the perception and recognition of spoken words. This departs from previous approaches by postulating a fundamentally different relationship between speech input, lexical representations of meaning and form, and the listener's perceptual experience of speech. The conventional approach, standard across essentially all current theories and models, assumes a processing architecture where speech is first analyzed in terms of some form of pre-lexical phonological unit (such as strings of phonemes or syllables) constituting a separate level of perceptual and computational representation. This pre-lexical level forms the input to the mental lexicon and is the basis for the listener's perceptual experience of the speech stream. We argue, instead, that there is no such pre-lexical level, that the speech input, analyzed in featural terms, is mapped directly onto combined phonological and lexical representations, and that the listener's perceptual representation of speech is an abstract post-lexical product of the system.

Evidence for the abstractness of the phonological percept comes from a series of studies of phonological variation (Lahiri & Marslen-Wilson, 1991; Gaskell & Marslen-Wilson, 1994). These indicate that subjects have relatively little awareness of the surface form of speech. Instead, they base phonological judgments on the abstract representation of speech that underlies surface variations. Evidence that the phonological percept is a late product of the perceptual system comes from studies of the integration of phonological cues in speech perception. In particular, lexical and phonetic decision data in Marslen-Wilson and Warren (1994; henceforth MW94) argue against the pre-lexical integration of phonetic cues into segmental or similar units.

Here we describe a distributed connectionist model that operates on these premises. The model's behavior in simulations of lexical and phonetic decision closely mirrors human performance in the MW94 experiments. This success, contrasting with the failure of localist models to account for these data, is based on the use of a single, post-lexical level of phonological representation, providing the basis for the listener's perceptual experience of words and nonwords alike. This model, it also turns out, provides a successful framework for explaining a wide range of properties of human spoken-word recognition (Gaskell & Marslen-Wilson, in preparation).

## Experimental Background

Marslen-Wilson and Warren examined the integration of featural cues to segment identity in words and nonwords. They created cross-spliced 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*. The vowel transitions here point to a final velar consonant (the [g] from *jog*), which conflicts with the place information in the burst, indicating a labial consonant (the [b] from *job*). The purpose of the experiments was to examine the effects of these conflicts between cues as a function of the lexical status of the stimuli involved, and of the task the subjects were performing.

As summarized in Table 1, triplets of monosyllables containing either one word and two nonwords or two words and one nonword were cross-spliced to produce six types of stimulus. These varied in terms of the presence or absence of mismatching cues and in the lexical status of the pre- and post-splice components. In a lexical decision experiment, where subjects make a timed judgment as to whether the stimulus is a word or not, there were interference effects for all mismatch conditions except N3N1, where the stimulus as a whole formed a nonword, and where both pre- and post-splice components derived from a nonword. Surprisingly, a very similar pattern was found in a phonetic decision task, where subjects make a timed forced-choice judgment as to the identity of the final consonant of the stimulus (e.g., between "g" and "b"). There were again strong interference effects for all

mismatch conditions, but a greatly reduced effect for N3N1.

Table 1: MW94 experimental contrasts

| 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.

We drew two main conclusions from these results. First, that lexical decisions and phonetic decisions were based on the same processing substrate, and second, that this substrate supported a complex pattern of interactions between the featural and lexical aspects of speech. A further simulation study using the localist TRACE model indicated that these results could not readily be modeled by a processing system of the classical sort, where lexical effects on phonetic decisions are accounted for in terms of top-down interactions from the lexical level to an independent pre-lexical phonemic level. We proposed instead a processing architecture of the type developed here, where the computational substrate for lexical and phonetic decisions is the same distributed representation, simultaneously encoding the mappings from speech input onto a phonological representation and from speech input onto a representation of lexical (or semantic) identity.

## Modeling Assumptions

Our model is based on a small number of assumptions about the perception of speech. These are partly drawn from previous models of speech perception (e.g., Morton, 1969; McClelland & Elman, 1986; Marslen-Wilson, 1987) and partly based on a functional analysis of the perceptual system. The principal assumptions are:

1) Lexical knowledge is represented in a fully distributed fashion.

2) Different forms of lexical knowledge (e.g., phonology, semantics) are represented at the same level and accessed simultaneously.

3) Speech input maps directly and continuously onto lexical representations.

4) The lexical access process operates with maximal efficiency by extracting the most informative lexical representation at all points during the perception of speech.

The value of distributed representations in the modeling of cognitive functions is well documented (e.g., Hinton, McClelland & Rumelhart, 1986; Hinton & Shallice, 1991). We envisage the lexical entry for a word to be a distributed pattern representing the semantic, syntactic, morphological and phonological specification of that word. These representations can be conveniently described in micro-featural terms (e.g., Plaut & Shallice, 1993). We also assume that units of lexical representation are not duplicated, in that the goal of lexical access—the activation of a single complete lexical representation—fills the representational space. This view of the lexical access process differs radically from currently popular models of word recognition such as TRACE and Cohort, which view the process of selection between 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.

Our model also differs from standard models in the ordering of different forms of information. Generally, phonological knowledge is seen as more "low-level" than semantic or syntactic knowledge. Almost all models suppose that a pre-lexical segmental (or similar) representation of speech is computed and that lexical access operates by matching this to phonological representations of words in an input lexicon. Thus, phonological representations are seen as the key to the lexical entries of words. Our model contains no such intrinsic ordering, with different forms of knowledge co-represented at the same level of the system. We are not trying to claim that the goodness of fit between the speech stream and stored representations does not rely on phonological information, but we propose that internal representations of phonological form are highly abstract and that no segmental representation of speech mediates in the lexical access process. By this view, the perception of a word and the perception of a word-like nonword (or an unfamiliar word) differ only in the degree to which different types of information are accessed. The perception of a word leads to the activation of all forms of lexical knowledge whereas the perception of a nonword leads only to the retrieval of phonological information, which is abstracted by the same process as that operating on words.

The assumption 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 simultaneously entertain 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 the 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 competition (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.

## Network Architecture

To allow the network to generalize over patterns of phonetic features spread across time, the model is based on a simple recurrent network architecture (Elman, 1990; Norris, 1990). The network is trained on the mapping between a stream of phonetic features and an internal representation of words. 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 hidden units are also connected to two sets of output units, representing the phonology and the lexical (or semantic) identity of the words contained in the speech stream (see Figure 1). From the perspective of the MW94 results, this is the crucial property of the model, since it means that the same substrate (the hidden unit weight space) is simultaneously coding both the phonological mapping and the lexical mapping. Our hypothesis is that this will allow the model to simulate the experimental results, in a situation where the same phonological output layer represents both words and nonwords.

For the purposes of this initial model, lexical/semantic identity was distributionally represented by an arbitrary vector of 50 zeros and ones. The phonological output was based on a slight adaptation of the Plaut & McClelland (1993) monosyllabic word representation. This is a compact phonemic representation of monosyllabic words, divided into 3 groups of units corresponding to syllable onset, rhyme and coda. Within each group, phonemes are represented by single units.

This representation provides a basis for decisions involving the form of speech. The use of phonemes here is a representational convenience. We assume that a segmental representation of speech emerges as a product of the interaction between orthographic and phonological knowledge in literate speakers of alphabetic languages (Marslen-Wilson & Warren, 1994; Morais, Bertelson, Cary & Alegria, 1986; Read, Zhang, Nie & Ding, 1986).

Auditory input to the network was represented segment by segment on a set of 13 binary input units. Eleven of these encoded the phonetic features of the current input segment using the Jakobson, Fant and Halle (1952) feature system. To simulate the coarticulatory spread of place information between consonant and preceding vowel, we added two further feature units. These were set to zero for all segments except vowels immediately preceding nasal or stop consonants. For these vowels, the two features represented the place of the following consonant, mirroring the *diffuse* and *grave* feature values for that consonant.

Consistent with evidence for the relative weakness of the vowel transition cues to place (e.g., Warren & Marslen-Wilson, 1987) these cues were made probabilistic during training: 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.

The network was trained to perform the joint phonological and semantic mapping for a set of 36 monosyllabic words drawn from the MW94 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 nasal (/n/, /m/, or /ng/), a voiced stop (/d/, /b/ or /g/) or an unvoiced stop (/t/, /p/ or /k/). These were presented as input to the network in the form of sequential bundles of phonetic features.

To maintain a more realistic learning environment for the network, a number of other words were added to the training corpus. Firstly, a set of 71 words were added to simulate the competitor environment for the test words. These were all close cohort competitors, sharing initial C(C)V segments with the target words but diverging on the final consonant cluster. This gave the test words an average of 3.5 close competitors (range 0-10).

In addition, the token frequencies of the training set were manipulated, reflecting the skewed distribution of word frequencies in English (Zipf, 1965). Test words were all given a token frequency of 20 within the training corpus. 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 and McClelland (1993), were added to the training corpus, with a token frequency of 1.

This corpus was presented to the network 50 times during training. On each cycle, the 13 input nodes were activated with the phonetic pattern of one segment of a word and the network was trained, using backpropagation, to produce the correct semantic and phonological patterns

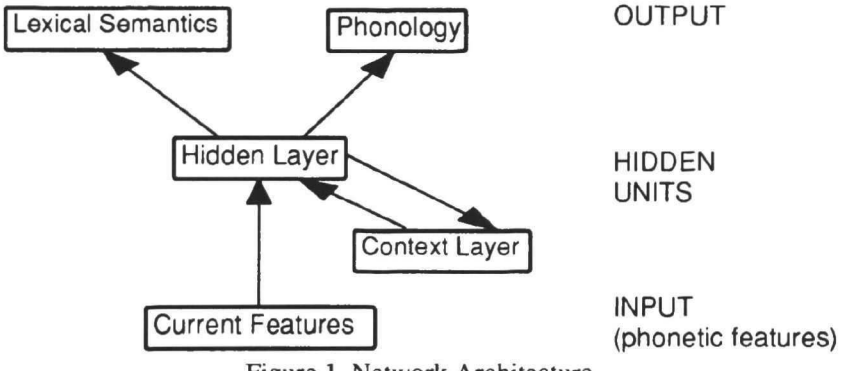


Figure 1. Network Architecture

for that word.

The network was then tested on a set of stimuli designed to simulate the test conditions of MW94 (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, the W2W1 stimuli contained place information in the vowel conforming 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 two filler words. The phonological and semantic activations were recorded at each time-step.

## Results

### Lexical Decision

Following other researchers we assume that output error scores correlate with response times in a cascaded processing system. We also assume that a lexical decision response depends predominantly on the lexical/semantic rather than phonological output of the model. Error scores at the semantic output can be transformed into word activation values using the function:

$$word\ activation = \frac{25 - \sum_{i=1}^{50} |t_i - o_i|}{25}$$

(where $t$ = training value and $o$ = output value for the $i$th unit).

This gives an activation value between -1 and 1, where 1 represents a perfect fit between the semantic output and the training pattern for that word and 0 is the expected activation value for an output pattern chosen at random. In a distributed representational system the competitor environment of a word can be directly reflected in this activation value. A semantic output which is similar to the training pattern for one word implies that it must also be dissimilar to the semantic patterns for all words which are not semantically related to that word. So, for example, no two unrelated words can have an activation of 0.9 at the same time. For this reason, there is no need to use relative activations to define a lexical decision criterion.

The upper graph in Figure 2 illustrates the activation of the semantic pattern of W1 for each of the members of the word triplets (averaged across all items). The stimuli for each condition are identical up to word position -1, where the coarticulatory information in the vowel is presented. At word position 0, the final consonant is presented.

The W1W1 condition is the baseline for comparison of effects of mismatch. Here, as featural information is presented, the activation of W1 rises, to a peak at the end of the word of 0.71. This figure does not represent perfect activation of the word, but implies that W1 is by far the most active candidate. Both cross-spliced tokens result in reduced activation of the W1 target, mainly on presentation of the mismatching coarticulatory information in the vowel. Furthermore, the patterns for the mismatching tokens are highly similar, with slightly more mismatch for W2W1 than N3W1 tokens. The model therefore predicts that both mismatching tokens should delay the recognition of the target to an equal extent.

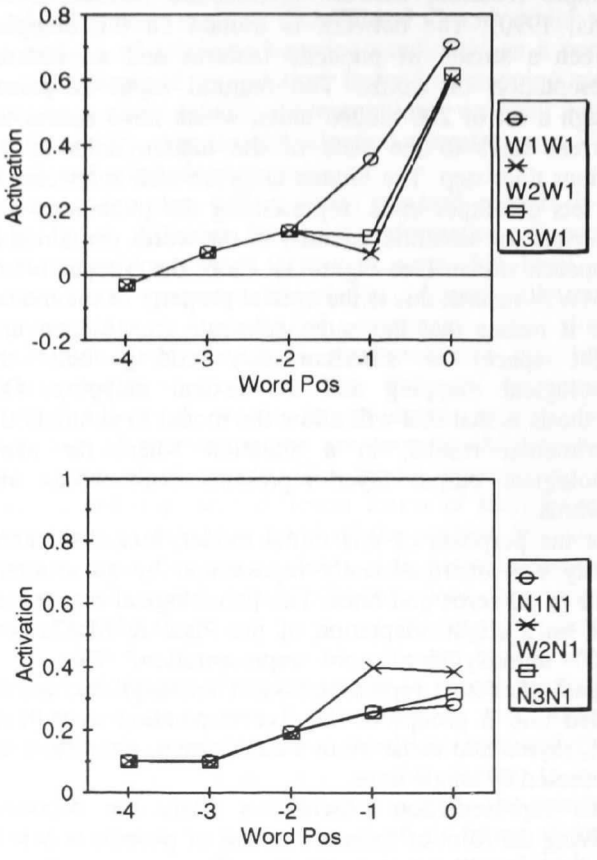


Figure 2. Semantic activations of W1 for the word triplets (above) and of W2 for the nonword triplets (below). The x-axis represents the position within the current stimulus (0 = final segment).

To examine the predictions for the nonword stimuli in a lexical decision task, we need to examine the activation of the semantic pattern for W2, the only word member of the stimulus triplets (lower graph, Figure 2). Unsurprisingly, W2 is best activated by the W2N1 token. However, the activation of W2 on presentation of this token does not rise above 0.4. Therefore, the model would predict a majority of "No" responses to these stimuli. In addition, the increased activation of W2 for this condition would predict "No" responses should be slower than for the baseline (N1N1) and N3N1 conditions.

The data from the lexical decision simulation were transformed to provide a comparison with the MW94 data (see Figure 3). In each case the simulation results were summed over the points during presentation of the word for which the conditions differed (i.e., on presentation of the final 2 segments). For the case where activation was assumed to be negatively correlated with response time (i.e., the "Yes" responses) these activations were negated. For both word and nonword conditions, the pattern of

activations produced by the network closely matches the pattern of response times found in MW94.

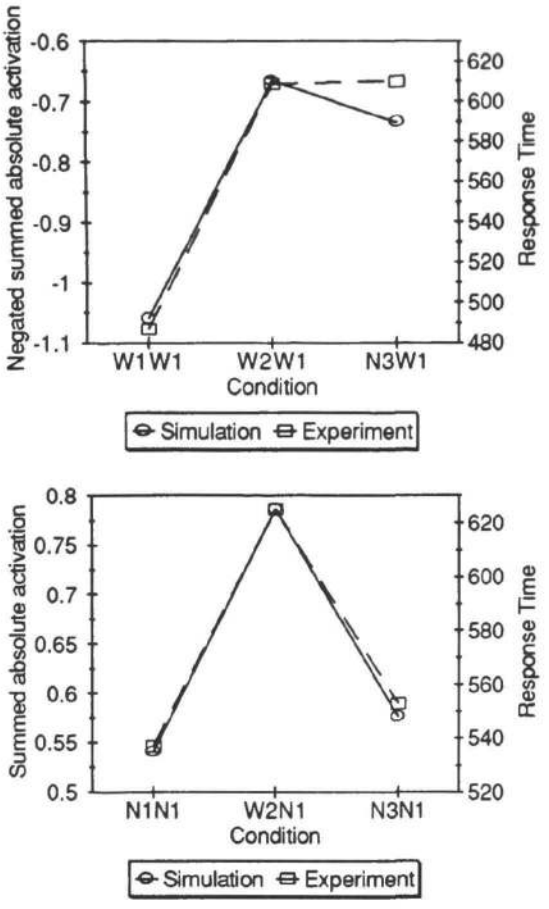


Figure 3. Comparison between lexical decision experimental data and network simulation for word (upper graph) and nonword (lower graph) stimuli.

Of particular interest is the finding that, like humans, the model shows no inhibitory effects when presented with two nonwords spliced together. This is because phonetic featural information is mapped directly onto lexical representations. The N1N1 and N3N1 conditions are equivalent in terms of the degree to which they match W2 phonologically: both contain information in the vowel transition and the following burst which deviates from the place of articulation of the W2 final consonant. Since there is no intermediate segmental level it does not matter that for the N3N1 condition the two sources of information conflict with each other segmentally—these cues are only integrated in the parallel mapping onto the phonological level. Thus, these conditions predict similar levels of inhibition of a lexical decision response.

## Phonetic Decision

The translation from localist phonemic output values to predictions of phonetic decision responses is straightforward: The network's predictions should depend on the relative activations of the word-final phoneme nodes involved. These are 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 response to the stimulus token *jog*, constructed from the onset of *job* and the final burst of *jog*, would depend on the activations of the /b/, /g/ and /d/ nodes in the coda group of the phonological output units. Therefore the difference between the activation of the target segment and its most active triplet competitor was used as a correlate of experimental response time (see Figure 4). As before, this measure is summed over the output for the final two segments of the input to provide a comparison with the experimental data.

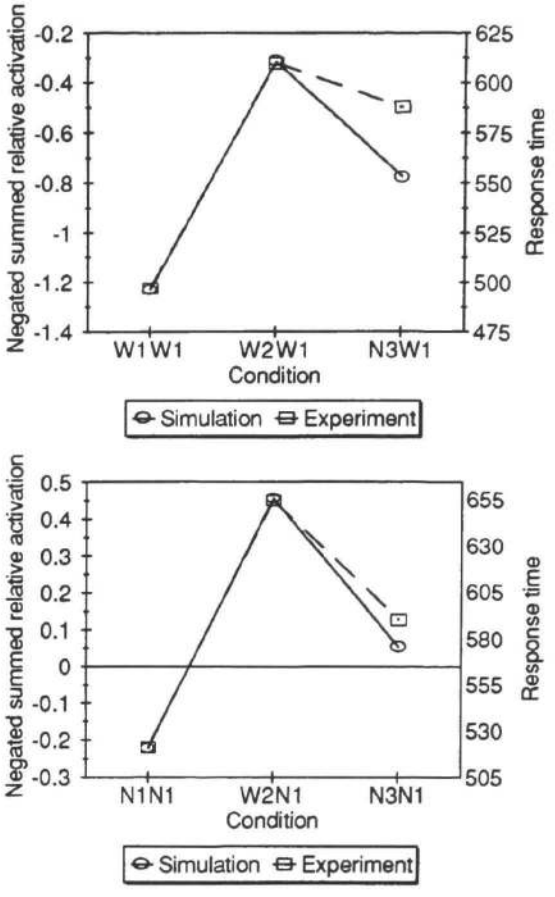


Figure 4. Comparison between phonetic decision experimental data and network simulation for word (above) and nonword stimuli (below).

Comparison of the two graphs shows a strong effect of lexical status on the network response. The responses to the word sequences ranged between -1.2 and -0.3, compared to -0.2 to 0.4 for the nonword sequences. This is consistent with the finding in the experiment that responses were slower for the nonword than the word sequences.

The patterns within the two sequence types are quite 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 very well (since the

mismatching effect of the W2N1 stimuli was roughly twice that of the N3N1 stimuli), although it underestimates the effect of mismatch for the N3W1 condition.

Again, it is interesting that the mismatching tokens composed of two nonwords (N3N1) show less of an inhibitory effect than the W2N1 condition. Here the difference can be explained in terms of the interaction between lexical/semantic and phonological levels. As the lexical decision simulation shows, the W2N1 tokens activate the W2 representation more strongly than the N3N1 tokens. This biases the phonological activations in favor of an output which is coherent with this word and thus inhibits activation of the "correct" phonemic nodes.

## Discussion

In both lexical and phonetic decision simulations, the predictions of the network closely follow the pattern of responses found in the MW94 data. In the lexical decision simulation the network shows strong inhibitory effects for consonant place mismatches involving words (i.e., in the W2W1, N3W1, and W2N1 conditions) but little effect of mismatch involving only nonwords (in the N3N1 condition). In the phonetic decision simulation all mismatching stimuli show inhibitory effects on responses, but the strength of these effects depends on the lexical status of the components of the stimuli.

The model achieves our objective of providing a basis for phonological perception which is not pre-lexical and is strongly influenced by lexical activations, but still allows the form of nonwords to be identified and to be influenced by lexical factors. Indeed, the influence of lexical competitors is slightly too strong in the current model, although this is more likely to reflect properties of the training corpus than the choice of architecture.

The fact that these results can be accommodated by our model is an important validation of this approach. Although this research is at an early stage we expect to model many other properties of the speech perception system in a similar manner. In particular, the use of a distributed lexical representation provides a straightforward explanation of many time course effects in lexical access. Effects such as the multiple activation of lexical candidates, frequency and competition effects in lexical access, and priming of associatively and semantically related words can be explained in terms of the interference caused by semantic "blending" of outputs (cf. Joordens & Besner, 1994) in a fully distributed lexical representation (Gaskell & Marslen-Wilson, in preparation).

## Acknowledgments

This research was supported by UK MRC, SERC and ESRC grants awarded to Lorraine Tyler and William Marslen-Wilson. We thank Mary Hare for valuable advice.

## References

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

Gaskell, G., & Marslen-Wilson, W. (1994). Inference processes in speech perception. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society* Hillsdale, NJ: Erlbaum.

Gaskell, G., & Marslen-Wilson, W. (in preparation). *Integrating Form and Meaning: A Distributed Model of Speech Perception.*

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.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, **98**(1), 74-95.

Jakobson, R., Fant, G., & Halle, M. (1952). Preliminaries to Speech Analysis. Cambridge MA: MIT 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.

Lahiri, A., & Marslen-Wilson, W. D. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, **38**, 245-294.

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

Marslen-Wilson, W., & Warren, P. (1994). Levels of representation and process in lexical access. *Psychological Review*, **101**(4), 653-675.

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

Morais, J., Bertelson, P., Cary, L., & Alegria, J. (1986). Literacy training and speech segmentation. *Cognition*, **24**, 45-64.

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

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.

Plaut, D. C., & McClelland, J. L. (1993). Generalization with componential attractors: Word and non-word reading in an attractor network. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* Hillsdale, NJ: Lawrence Erlbaum.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: a case study of connectionist neuropsychology. *Cognitive Neuropsychology*, **10**(5), 377-500.

Read, C., Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition*, **24**, 31-44.

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

Zipf, G. K. (1965). *The psycho-biology of language: An introduction to dynamic philology.* Boston: Houghton-Mifflin.