A Connectionist Model of Phonological Representation in Speech Perception

M. Gareth Gaskell[1], Mary Hare[2] and William D. Marslen-Wilson[1]

[1]*Centre for Speech and Language,*
Birkbeck College,
University of London
and
[2]*Center for Research in Language*
University of California in San Diego

In press, *Cognitive Science*

Please address all correspondence to:
Gareth Gaskell
Psychology Department,
Birkbeck College,
University of London,
London WC1E 7HX
ENGLAND

E-mail: g.gaskell@psyc.bbk.ac.uk
Phone: 0171-631-6242
Fax: 0171-631-6587

Abstract

A number of recent studies have examined the effects of phonological variation on the perception of speech. These studies show that both the lexical representations of words and the mechanisms of lexical access are organized so that natural, systematic variation is tolerated by the perceptual system, while a general intolerance of random deviation is maintained. Lexical abstraction distinguishes between phonetic features that form the invariant core of a word and those that are susceptible to variation. Phonological inference relies on the context of surface changes to retrieve the underlying phonological form. In this paper we present a model of these processes in speech perception, based on connectionist learning techniques. A simple recurrent network was trained on the mapping from the variant surface form of speech to the underlying form. Once trained, the network exhibited features of both abstraction and inference in its processing of normal speech, and predicted that similar behavior will be found in the perception of nonsense words. This prediction was confirmed in subsequent research (Gaskell & Marslen-Wilson, 1994).

Introduction

Phonological variation is generally described in terms of a mapping from underlying to surface forms in the production of speech. For example the word *wicked*, which has the canonical form /wɪkɪd/, can be realized as [wɪkɪb] in the context of [wɪkɪbpræñk][1] (*wicked prank*). This is an example of place assimilation, in which the place of a syllable-final coronal segment (e.g., /t/, /d/, /n/) changes to become more like the following segment. In the above case, the place change involves closure of the lips, transforming the coronal /d/ into a labial [b]. A similar process operates on coronal segments when followed by segments with velar place, as in the assimilation of /d/ to [g] in [regkoœt] (*red coat*). These changes may be complete, producing tokens with no residual coronal information. In other cases, the result of the place change is partial, producing a segment with phonetic information relating to two different places of articulation (Browman & Goldstein, 1991; Nolan, 1992; Byrd, 1992). A similar assimilation process operates on oral vowels followed by nasal segments, as in the word *ban*. Here, the nasality of the consonant spreads back to the vowel, producing a phonetically nasal vowel. In English, this process is allophonic, since the product of the assimilation does not cross a phonemic boundary. However, in other languages, such as Bengali, the process *neutralizes* phonemic distinctions. This is because such languages contain vowels which are underlyingly nasal and hence indistinguishable from assimilated oral vowels.

Neutralizing processes, such as place assimilation in English or nasal assimilation in Bengali, appear problematic when we examine their effect on the perceptual system. Understanding speech involves a process of lexical access, in which a representation of the speech stream is compared to stored representations of words, which are used as the key to word meaning. The presence of fully assimilated segments of speech creates lexical ambiguity, in that the underlying identity of assimilated segments cannot be directly extracted from their phonetic properties. For example, a surface [b] could be an ordinary token of an underlying /b/, but it could also be an assimilated token of /d/, as in the *wicked prank* example above. Thus, an important challenge for any model of human speech perception is to explain how the system is organized to ensure that natural surface variants gain access to the stored knowledge about a word.

This picture is further complicated by the intolerance we find for random deviations in lexical access. The research reported in this paper is described with reference to activation models of lexical access such as Cohort (Marslen-Wilson, 1987) and TRACE (McClelland &

[1] The electronic form of this paper contains some nonstandard phonetic transcrptions.

Elman, 1986). In these models the state of the lexical access process is reflected in the activations of a lexical candidate. Candidates which closely match a featural representation of a section of the speech stream will be highly active, whereas candidates which match less well will have a lower activation. Access to stored information about a word is dependent on its activation, either in absolute terms or relative to other activations. Thus, the success of the lexical access process depends critically on the goodness-of-fit between sensory evidence and stored information about the form of words.

A number of recent studies using on-line priming techniques indicate that, unless the goodness of fit between speech and lexical forms is high, access to lexical information is disrupted. For example, Marslen-Wilson and Zwitserlood (1989) used cross-modal associative priming to study the effects of word-initial mismatch on lexical access. The experiments compared the priming effects of nonword rhyme primes to the effects of the source words (for example, comparing *noney*-BEE to *honey*-BEE), finding that rhyme primes were always less effective than the source words in the facilitation of the target words. Indeed, only when the competitor environment of the source word was particularly sparse was there any significant facilitation by the rhyme prime.

Similar studies (Connine, Blasko, & Titone, 1993; Marslen-Wilson, Moss, & van Halen, in press) show that even single phonetic feature deviations disrupt lexical access, although in these cases the disruption may not be complete. These findings suggest that natural phonological variation and random variation are treated in different ways by the perceptual system. Small random deviations, such as the phonetic changes used in the priming experiments, have a strong disruptive effect on the process of lexical access. In contrast, systematic phonetic changes, such as assimilatory change, are accommodated by the perceptual system.

This contrast reflects two aspects of speech perception, which can in classical terms be separated into the *process* of lexical access and the *representation* of phonological form. We argue that lexical form representations are abstract, containing only the invariant properties of the words. In addition, a process of context-dependent phonological inference assesses the phonological viability of surface changes in their segmental context.

Lexical Abstraction

Lahiri and Marslen-Wilson (1991) argue that phonological representations in the mental lexicon are structured in accordance with the theory of radical underspecification (Archangeli, 1988). This theory states that only certain aspects of the form of a word are lexically represented. Only phonetic features which are unpredictable[2] form part of the specification of a word, with all other information left unspecified. Generally, two types of argument are used to support underspecification (Keating, 1988). The first is based on transparency to phonological rules involving the spread of a harmonizing feature across intervening segments (see Hare, 1990, for a connectionist treatment of this phenomenon). The second, which we shall deal with in this paper, links underspecification to phonological variation.

Underspecification can be used as a basis for the explanation of both assimilatory change in speech production as well as tolerance of regular variation in perception. For example, in the representation of vowels, the feature [oral] is generally accepted as the default state and is therefore not present in an underspecified representation (Archangeli, 1984). Thus, the nasalization of vowels through assimilation is explained as a consequence of this unspecified state: the unspecified segment can gain the nasality of a following nasal, as in *ban*. However,

[2] Predictable features are those that can be derived by either context-insensitive (default) or context-dependent (redundancy) rules.

for languages which contain underlyingly nasal vowels, a nasal vowel cannot assimilate to a following oral consonant, because its underlying representation is already fully specified. It is the representational asymmetry which radical underspecification provides that is crucial to the current study. However, it is important to note that other varieties of underspecification, as well as other systems making use of privative features, exhibit similar asymmetries. Here, we take radical underspecification as our exemplar of these systems.

Stemberger (1991) found evidence for underspecification in the production of speech errors. Analyses of both naturally occurring and experimentally induced speech errors revealed strong asymmetries in the types of replacement errors made. This was consistent with the idea that some phonological feature values are left unspecified, and therefore prone to erroneous change during speech production, whereas other values are lexically specified and thus more resistant to change.

In the perceptual domain, the underspecified representation may act more like a filter. According to Lahiri and Marslen-Wilson (1991), successful lexical access depends predominantly on the degree of match between *specified* features of an underspecified lexical representation and the corresponding speech input. Therefore, if a vowel is specified in the lexicon as nasal, a matching segment of the speech stream must also be nasal. However, if a lexical entry contains an oral vowel, which is lexically unspecified for that feature, free variation is allowed on the value of the feature: the lexical entry will match both nasal and oral speech. This matching strategy tolerates systematic phonological variation, since it affects only unspecified features, which do not figure in the goodness-of-fit calculation. On the other hand, a general intolerance of random variation is maintained, since random variation will alter specified and unspecified features alike.

Evidence in support of this theory comes from a number of gating studies. Lahiri and Marslen-Wilson (1991) compared the effects of vowel nasalization on subjects' perceptions cross-linguistically, using Bengali and English. In English, vowels generally have an oral surface form, although vowel nasalization can occur as a product of assimilation, as in the *ban* example above. The same assimilation process occurs in Bengali, but the language also contains words which are *underlyingly* nasal (i.e., the difference between oral and nasal vowels is distinctive).

Lahiri and Marslen-Wilson presented Bengali and English subjects with three types of consonant-vowel-consonant (CVC) words. The first group were simple CVCs with an oral vowel and an oral final consonant (e.g., /bæd/, *bad*). The second type of item, denoted CVN, were words ending in a nasal consonant which contained an assimilated nasalized vowel (e.g., [tãn], *tan*). The Bengali subjects were also presented with CV_NC words containing an underlyingly nasal vowel followed by an oral consonant (e.g., /pãk/, 'slime'). As the stimuli were presented, in gradually increasing gated sections, subjects were instructed to predict the identity of the complete word.

The critical data for underspecification theory are the responses during vowel presentation, where the nasality of the vowel is known but the final consonant is still unknown. For the nasal vowels, English subjects used the presence of vowel nasalization to predict CVN words, whereas Bengalis responded with a mixture of underlyingly oral and nasal words (roughly 60% of the Bengali responses were of type CV_NC, 6% were of type CVN and 28% were of type CVC). For stimuli with oral vowels, Bengali subjects almost never produced words with underlyingly nasal vowels (0.7% of responses), but responded with both oral CVC (80.3%) and nasalized CVN (13.4%) words.

These data reflect two aspects of the internal representation of phonological form. The fact that subjects are able to map surface oral vowels onto the lexical representations of CVN words, which have surface nasalized vowels, suggests that the representation on which their

responses are based is an abstract underlying representation, which does not encode the surface nasality of these vowels. Thus, assimilatory nasalization does not discriminate directly between lexical candidates because the representation on which the discrimination process is based does not encode such a surface change.[3]

In addition, the asymmetry found in the Bengali data, whereby surface nasal vowels can match lexical representations of words with oral vowels, but surface oral vowels mismatch specified nasal representations, indicates that this underlying representation is underspecified, with only nasal vowels actually represented in the lexicon.

A further study of neutralizing assimilation, comparable to the Bengali data, examined the representation of stop consonants in speech perception. Coronal place of articulation is generally regarded as the default place (cf. Paradis & Prunet, 1991) and is therefore absent from an underspecified representation of place. Again, this agrees with the susceptibility of coronals to place assimilation as described above and predicts a further asymmetry in the perceptual domain. This prediction was tested in a gating study using English (Marslen-Wilson & Nix, 1992; Nix, Gaskell & Marslen-Wilson, 1993). The study compared sentences such as (1) and (2) below, which contained words with coronal or non-coronal word-final stops.

(1) They thought the *lake cruise* was rather boring

(2) They thought the *late train* was rather boring

In (1) the critical word (*lake*) has a surface velar segment, but is underlyingly ambiguous: the underlying word could indeed be a token of *lake* but it could also be *late*, which has undergone a process of assimilation. However, there is no such ambiguity in (2), since assimilation from non-coronal to coronal place does not occur. For example, the /k/ in *lake* cannot assimilate to the /t/ in *train* to give *late train*. The only phonologically viable reading of the critical word is therefore *late*.

These sentences were presented to subjects in gradually increasing gated sections. At the offset of the critical word (i.e., *late*), the sentences of type (2) were fairly unambiguous, with 80% of subjects giving a response containing a word-final coronal. However, the type (1) words provoked a mixture of responses, with only 55% of responses conforming to the surface structure of the critical word.

This pattern of responses is similar to the nasality data of Lahiri and Marslen-Wilson. Although words like *late* or *lake* may be unambiguous when heard in isolation, they show a strong asymmetry in their perceptual quality when presented in normal utterance context. The lexical entries for words such as *lake*, with non-coronal final segments, are specified for place and so will not match any speech input (such as *late*) with a different place of articulation. This explains the lack of ambiguity for the sentences of type (2). On the other hand, lexical entries for words such as *late* are unspecified for place and so will match with both coronal and non-coronal segments of speech.

Taken as a whole, these results form strong evidence for an asymmetrical abstractness in the lexical representations of underlying form. In this paper we look at how this system might develop. We argue that such asymmetries develop as a response to the variability in the mapping between surface and underlying form. For speakers of Bengali, surface orality is a reliable indicator of an underlying oral vowel and is treated as such. On the other hand, surface nasality, since it can occur either through assimilation of an oral vowel or as the surface form of a nasal vowel, is treated more ambiguously. Similarly, for English speakers coronal surface segments are reliable indicators of an underlying coronal segment and are treated as such during speech perception. Labial and velar surface forms, since they can occur as the product

[3] Nasalization can be used *indirectly* to discriminate between candidates because it allows the listener to predict, in languages like English, that the following consonant is nasal.

of the assimilation of an underlying coronal segment, must be treated as more ambiguous, at least until their following context is known.

Phonological Inference

There is evidence that not only the form of lexical representations, but also the *process* of lexical access is organized to accommodate regular variation. Phonological inference is a process which analyses segments with reference to their phonological context in order to elucidate their underlying identity. This is essentially the opposite mapping to that assumed to take place in speech production. In order to carry out this process, the perceptual system must be sensitive to the rules or constraints under which regular phonological changes take place. Using again the example of place assimilation, phonological inference operates as a regressive process, checking the viability of a phonological change by comparison with its following context. Thus, the /d/ underlying the surface [b] in [wɪkɪbpræñk] (underlyingly *wicked prank*) can be inferred from the place of the following [p]: the segments both have a surface labial place of articulation and so conform to the phonological rule describing the assimilatory process.[4]

Experimental evidence for this inference process comes from a study of place assimilation (Gaskell & Marslen-Wilson, in press). The study examined the viability of phonological changes by comparing stimuli such as [wɪkɪbpræñk] with phonological changes which violate assimilation rules (e.g., [wɪkɪbgeɪm] — *wickib game* — where the labial place of the [b] conflicts with the velar [g], and thus the [b] cannot be an assimilated surface form of an underlying /d/). These stimuli were embedded in sentential context and were used as primes in a cross-modal repetition priming experiment (cf. Marslen-Wilson, Tyler, Waksler & Older, 1994). The visual target, which was the canonical form of the phonologically changed word (i.e., WICKED) was presented at the offset of the phonological change (i.e., at the offset of the [b]) and the time taken to make a lexical decision to the target word was measured. The response times were an indication of the degree to which the lexical entry of the target word was accessed by the phonologically changed prime. The results showed that an assimilated prime, in a context that validated the assimilation, primed responses as strongly as a canonical token of the prime (e.g., [wɪkɪdpræñk]). However, the same phonological change embedded in an unviable context for assimilation strongly reduced the priming effect.

This contextual effect shows that the process of lexical access is sensitive to viability in its assessment of regular variation. This effect, since it occurs between words, does not reflect an aspect of the lexical representation of words. Instead it indicates a process of inference which operates on the form representations, either pre-lexically or after initial contact with the lexicon. Phonological inference provides additional constraints on the mapping of speech input onto highly abstract lexical form representations, which under a simple mapping process would match a large number of surface forms.

Modeling Abstraction and Inference

In the remainder of this paper we shall present a model of abstraction and inference in speech perception. Our primary aim is to construct a model of speech perception which accommodates the existing experimental data in this area. We shall also show that the connectionist learning approach provides a plausible account of the formation of such a mapping, based on the statistical properties of speech. Finally, we shall examine the consequences of such an approach, showing that the perception of phonologically changed

[4] Such sequences do not always mark assimilation points (for example, the same sequence of segments occurs across the word boundary in *club player*). Often this kind of ambiguity can be resolved using lexical information (i.e. *club* is a word, but *clud* is not).

speech can be described in terms of the satisfaction of multiple lexical and phonological constraints. This, in turn, leads to testable — and in some cases tested — claims about the way phonologically changed speech is perceived.

The classic connectionist model of speech perception is the TRACE model of McClelland and Elman (1986). TRACE is a localist network employing interactive activation and competition between nodes to examine the matching process between speech input and lexical and phonemic candidates. However, TRACE is inappropriate for the purposes of this study for a number of reasons. First, we wish to examine how the perceptual system may develop given variable speech input, whereas the architecture of TRACE prohibits learning. Second, we have argued that TRACE is unable to model the results of the priming experiments examining the perception of place assimilated speech (Gaskell & Marslen-Wilson, in press). Specifically, the process of lateral inhibition, employed by TRACE to reduce activations is unable to model the swift, powerful effects of mismatch found in these and other priming studies (see also Marslen-Wilson, 1993).

Instead, the model we describe is a simple recurrent network (Elman, 1990; Norris, 1990), based on the backpropagation algorithm (Rumelhart, Hinton & Williams, 1986). Backpropagation allows the modification of links between nodes, by comparing the network's output to a set of training values. This enables the structure of the mapping to develop as a reflection of the network's experience. The simple recurrent network is an extension of this general method, incorporating time-delayed links between hidden units to allow generalizations to be made over temporal sequences of patterns. This is of particular value in the modeling of processes involving speech, where information is continuous and perceptual mechanisms operate on-line, often based on partial information.

Norris (1990, 1992) used a simple recurrent network to examine the perceptual mapping between a featural representation of speech and localist lexical units. The performance of the network broadly corresponded to the Cohort model (Marslen-Wilson, 1987). In the early stages of the presentation of a word, all the matching candidates became active to some extent; but as soon as the uniqueness point[5] of a word was reached, the activations of the mismatching candidates dropped sharply and the activation of the remaining matching candidate rose to nearly 1 (the maximum activation). Norris also found that the network, like humans, was intolerant of small deviations (Marslen-Wilson, 1993). For example, given the nonword input *horonet* (where *coronet* is a member of the training set), the activation of *coronet* never rose above 0.1. Simulations in our laboratory using a similar network architecture show intolerance of nonword deviation even at the end of a word (as in *coronef* for *coronet*).

Shillcock and colleagues (Shillcock, Levy & Chater, 1991; Shillcock, Lindsey, Levy & Chater, 1992) examined the process of lexical access using a simple recurrent network trained to map from a segmented stream of phonemes to an output window, consisting of the network's evaluation of the current, the previous and the next phoneme. This simple architecture allowed the network to learn generalizations based on the statistics of the speech stream, using these generalizations to improve the network's performance in the prediction of upcoming phonemes. Effects normally explained as top-down or lexical, such as the Ganong effect (Ganong, 1980)[6], could thus be simulated in a bottom-up model with no explicit lexical level of representation (see also Norris, 1993).

---

[5] This is the point at which the set of lexical candidates matching a sequence of speech sounds reduces to one word.

[6] The Ganong effect refers to the experimental demonstration of word context influencing the perception of ambiguous segments by shifting the category boundary in voicing continua.

Following Shillcock et al. (1991, 1992), the approach we shall take is to impose on the network the minimum level of external structure necessary to model experimental findings about the perception of phonologically variant speech. For this reason, the network we describe does not incorporate explicit word or morpheme units in its output. Instead, we examine the extent to which phonological inference and abstraction in speech perception can be modeled simply by exploiting the surface properties of speech. The network is trained on the mapping from a systematically variant phonetic-feature representation of speech (i.e., the surface form of speech) to a canonical representation of the same speech (i.e., the underlying form). The model incorporates no explicit concept of a lexical entry, not because we believe phonological inference is independent of lexical access, but because we wish to determine the extent to which perceptual behavior can be explained as the result of a simple statistical learning mechanism, applied to the speech stream.[7]

Although the model may contain no explicit lexical entries, lexical form information is used in training the network. Training involves presentation of the underlying form of speech to the network as a standard by which the error of the network can be measured and reduced, and this information can only be gained if the canonical lexical forms of the words are available. As a developmental model of phonological processing in speech perception, this implies that lexical access must be successful as the phonologically variant speech is heard, so that the underlying form can be recovered and utilized. However, the research of Gaskell and Marslen-Wilson (in press) suggests that lexical access will only be completely successful if the phonological inference mechanism is already at work. A solution to this "boot-strapping" problem may be that, as perceptual mechanisms develop, the tolerance for error in the matching process is reduced. In other words, the process of learning to understand speech may involve a gradual tightening of the constraints involved in the goodness-of-fit computation in lexical access. This would allow access to lexical information for phonologically variant words early on in development and so enable phonological compensation processes to be learned. It is also possible that the kind of speech a young child is exposed to may contain fewer assimilated tokens, since a greater proportion of words will be spoken in isolation and utterances may be spoken more carefully (Kerswill, 1985, showed that the occurrence and extent of place assimilation depends strongly on factors such as speech rate and style).

A further implicit assumption of our model is that the unassimilated form of a word can be identified as the underlying form and then employed in the perception of assimilated forms. This identification may be the result of a frequency comparison between the different tokens of a word. Since place assimilation only occurs in a tightly constrained phonological context and even then only optionally, most coronal segments will be phonetically realized in their unassimilated form. Therefore the most frequent form of a word can be taken as the underlying form and used in the perception of assimilated forms. As pointed out above, these frequency proportions may be even more skewed in favor of unassimilated forms in the speech encountered by young children.

This view of the formation of phonological representations also has implications for the role of the perceptual system in the diachronic shaping of natural phonological changes. If the perceptual system for speech develops in the manner implied here, by gradual learning from experience, it follows that speech perception has a fairly passive role in the shaping of phonological rules. A connectionist network basically learns what it is trained to learn: there may be differences in the ease of learning different mappings (e.g., due to frequency

[7] It is possible that the structure of the speech on which the network is trained will influence the development of the network's internal representations. This may result in implicit lexical (i.e. word-form) representations being developed.

differences between the different types of mapping), but there is no reason why, for example, assimilations of coronals to non-coronals should be learned more easily than assimilations in the opposite direction. This contradicts theories (e.g., Kohler, 1990; Ohala, 1990) in which phonological change is seen as perceptually tolerated simplification.[8]

To facilitate comparison with experimental results, we examine the response of the network to place assimilated speech. Thus, the input to the network is a stream of feature bundles, corresponding to segments of speech, in which place assimilations are artificially introduced. During training, the network's output is modified by comparison to the canonical underlying form of this speech, again represented using phonetic features. We predict that the network will learn the phonological regularities underlying place assimilation and apply them to novel combinations of words, exhibiting a process of phonological inference. We also predict that the network will learn to develop different tolerance levels for different place feature values. Such behavior is functionally equivalent to an underspecified representation of place of articulation (cf. Seidenberg, 1994).

Method

Network Architecture

The network used in these simulations was a modified version of the Shillcock, Levy and Chater (1991) network illustrated in Figure 1. The main modification was in the input and output representations used: Shillcock et al. initially used a standard featural input (Jakobson, Fant & Halle, 1952), mapping onto localist phonemic output. In their later model they opted for Government Phonology elements (Kaye, Lowenstamm & Vergnaud, 1985, 1990) at both input and output. For these simulations we used the more traditional Jakobsonian features in both input and output representations. This gives the representation enough resolution to capture the phonological processes in speech, and in terms of these simulations seems to have no disadvantage compared to other phonetic systems.

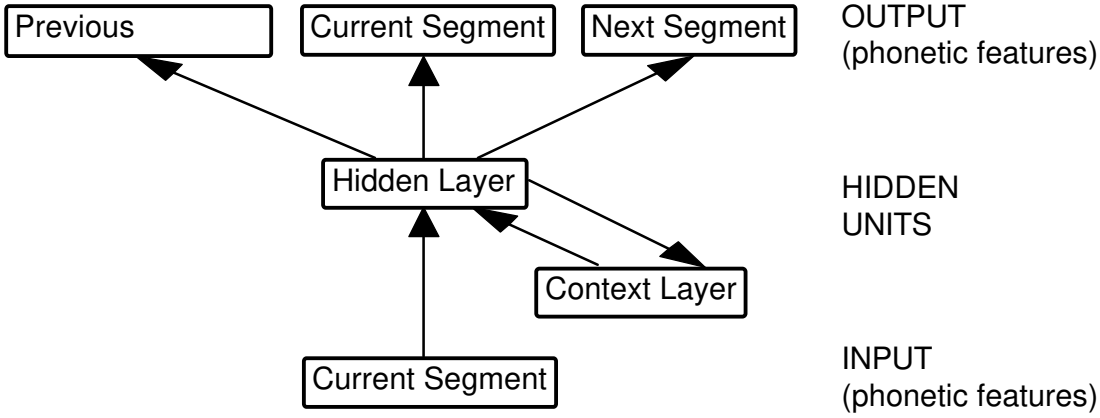

Figure 1. Network architecture. The connections from the hidden units to the context units are a direct copy of the hidden units at the previous time-step.

The network had 11 input nodes, corresponding to the 11 phonetic features used. These were connected to 100 hidden units, which had recurrent connections to 100 context units, allowing context-dependent processing to occur. The context units take their feedback from the hidden units (Elman, 1990) rather than the output units (Jordan, 1986) in order to allow

[8] However, true *acoustic* differences between speech sounds are beyond the scope of our model and do provide a means for perception to direct phonological change.

the network to develop representations based on regularities of unlimited length (cf. Servan-Schreiber, Cleeremans & McClelland, 1988; Dell, Juliano & Govindjee, 1993). The hidden units were fully connected to 33 output units, organized into three 11-feature output windows representing the network's hypotheses for the current segment, as well as the previous and following segments. The number of hidden units may seem excessive, since networks often employ fewer hidden units than input units, but here the network requires extra hidden units to allow contextual dependencies to be recognized. The exact number of hidden units was chosen fairly arbitrarily — a similar network with only 50 hidden units was tested, with only a slight reduction in performance.

Training Procedure

The network was trained, using standard backpropagation, to learn the mapping between a phonetic surface form of speech and its underlying representation. Rather than attempting to study phonological variation in general, we chose to model the process examined in earlier experimental work — place assimilation. In order to allow dependencies between place assimilated segments and their neighbors to be learned, a realistic corpus of speech was necessary, the larger the better. However, because most transcriptions of speech do not represent phonologically variant forms it was necessary to train the network using a phonetic representation of speech in which place assimilation was artificially introduced. The corpus was the speech data used in the Shillcock et al. studies, which was generously made available by the authors. This was a selection of conversational English, mostly taken from the LUND corpus of Svartvik and Quirk (1980), which was translated from orthographic to phonetic form using a translation program, with errors corrected by hand. The corpus was unfortunately rather small, containing only 3719 words by token (roughly 12000 segments), but was expected to be large enough for the coarser statistics of speech to be learned.

Place assimilation was artificially introduced into the corpus by randomly selecting 50% of the coronal segments in viable context for assimilation and altering their place to become the same as the following context. Contextual viability was assessed according to the phonological rules described below.

t → p / -# (p, b, m)
d → b / -# (p, b, m)
n → m / -# (p, b, m)
t → k / -# (k, g)
d → g / -# (k, g)
n → ñ / -# (k, g)

Thus, for example, the /d/ in /gɷdbɔɪ/ was presented as input to the network in its assimilated form, [b], and the network was trained to map this onto the underlying /d/.

The result of this replacement strategy conforms to empirical estimates of the level of place assimilation found in normal speech (Barry, 1985; Kerswill, 1985). The introduction of assimilation in this way meant that only a small proportion of coronal segments were altered (since most did not occur in a viable context for assimilation, and only half of those that did were actually assimilated) and the overall proportion of changed segments was minimal (less than 0.5%). To simplify the procedure, no attempt was made to simulate a gradient of assimilatory change (Nolan, 1992; Holst & Nolan, in press). Instead, segments were presented to the network either unassimilated or fully assimilated.

In training, the assimilated transcription was presented to the network as a continuous stream of feature bundles and the output was trained, using standard backpropagation, to produce the unassimilated transcription. Because of the small proportion of assimilated segments this was mainly an auto-association task (or, in the case of the next segment

window, a simple prediction task), with compensation for phonological change only required for one segment in every two hundred. The network was trained on 100 sweeps through the entire corpus of speech, with the identity of the assimilated segments varying between sweep. The software used both in training and testing was the TLearn program developed by Jeff Elman.

The root mean square (RMS) error of the network is a comparison between the output of the network and the training data, and is therefore a measure of the success of the mapping learnt by the network. After 100 cycles through the training set, the RMS error was dropping, implying that the network was still learning, but at a tiny rate. Presentation of samples of unassimilated speech to the network produced error-free performance in both the current segment and the previous segment output windows (i.e., the network output was always more similar to the correct segment than to any other).

Simulation 1

Simulation 1 examines the representation of place of articulation formed by the network while learning the mapping between surface and underlying forms of assimilated speech. This is accomplished by examining the response of the trained network when presented with tokens of speech in either assimilated (e.g., *reg car*) or unassimilated (e.g., *red car*) surface form, allowing comparison of the model's treatment of coronal and non-coronal segments. As discussed earlier, coronal place of articulation is generally regarded as the default. Thus, according to Lahiri and Marslen-Wilson (1991), successful lexical access depends critically on matching speech input with lexical representations containing non-coronal segments, such as /b/ or /g/, but for lexical representations of words containing coronal segments, such as /d/ or /t/, more or less free variation across place of articulation is tolerated, since the lexical representations are unspecified in this respect. Looking at this underspecified representation the other way round (i.e., examining the degree to which variation is tolerated for any particular segment forming part of the speech input), this implies that there should be little variation in the way surface coronals are mapped onto underlying forms, since they cannot map onto specified non-coronal segments. However, there should be more variation in the way surface non-coronal segments are treated, since they can map onto non-coronal and coronal segments alike.

Therefore, for the network to show behavior in concordance with an underspecified representation of place, it should exhibit greater variation in its treatment of surface labial and surface velar segments than for coronal segments. This kind of result would suggest that the network's treatment of place of articulation conforms to an underspecified description, and that this asymmetry has developed as a consequence of the regular variation present in the speech input during training (i.e., as a consequence of its experience of regular variation).

The contrasts employed also allow comparison with the gating studies of place assimilation described above (Marslen-Wilson & Nix, 1992; Nix et al., 1993), showing a difference in ambiguity between assimilated and unassimilated segments.

Design

The simulation involved the presentation of a set of two-word stimuli to the trained network, segment by segment. The first word ended with a coronal segment, which was presented either unassimilated (e.g., [kwaɪt], *quite*) or with velar (e.g., [kwaɪk]) or labial (e.g., [kwaɪp]) place change, such as would occur as a result of place assimilation. The following word always provided a viable phonological context for assimilation, as defined by the phonological rules described above. The output of the network was examined at two points for each stimulus pair. At the presentation of the final segment of the first word, the output of the current segment window was recorded, allowing the network's initial response to that

segment to be examined. The second measurement was taken on presentation of the following segment, which formed the phonological context for assimilation. Here, the activations of the previous-segment window were recorded, which were again directed to the identification of the final segment of the first word. In other words, two measurements were made of the network's response to the word-final segment; one as the segment was presented, and one when the following context was known. The design of the simulation thus involved two variables, surface place (unassimilated coronal vs assimilated labial vs assimilated velar) and phonological context (unknown vs known).

Since the output of the network is a vector representing activations of features, a further transformation is required in order to model human data, such as gating or phoneme monitoring responses. It is plausible to assume that the model's prediction of the response to a word-final segment is proportional to the goodness of fit between the output of the network in one or more of the output windows and the ideal activation pattern of that segment (cf. Seidenberg & McClelland, 1989). In this simulation, the target segments differ only in their place of articulation. We therefore require a dependent variable that directly measures the place of articulation of the network's output. Coronal and labial places are distinguished in the Jakobson, Fant and Halle (1952) feature system by the value of the *grave* feature: a value of 1 represents labial place and 0 represents coronal place. Coronal and velar places, however, differ on two features: *grave* again and *diffuse*. To control for these representational differences, the measure we chose as the dependent variable was the mean deviation from coronal (the underlying place) on the distinguishing features. For labial segments this was the deviation for the *grave* feature, and for velar segments it was the average of the deviations found on the *grave* and *diffuse* features. To simplify matters, we shall describe this *coronal score* in percentage terms, such that a value of 100% corresponds to a coronal place of articulation and a 0% score represents a non-coronal (either labial or velar) place of articulation.

Materials

The 18 items in this simulation consisted of single words from the training corpus (i.e., words already encountered by the network at least 100 times), followed by one of 5 simple context words (*pay, by, may, go* and *'cos*[9]). The test words were at least three segments long and contained a word-final coronal segment (six each of /t/, /d/ and /n/). Half the items had been presented in assimilated form at some point during training whilst the other half had always been presented in their canonical form.[10] The five context words were, in the representation system used, three segments long (some consisted of one consonant followed by a diphthong, some had a consonant-vowel-consonant structure) and began with a labial or velar segment, one each from /g/, /k/, /m/, /p/ and /b/.

Each item was represented in each condition. The place variable was manipulated by altering the final segment of the word plus the context word if necessary (e.g., *quite go* vs *quike go* vs *quipe by*).

Procedure

The stimuli were presented as a continuous list to the trained network. Because of the possibility of the initial state of the hidden units, recycled through the context units, affecting responses, the list was presented twice with the output of the first presentation ignored. On the second presentation, the measurements described above were made.

---

[9] The abbreviated form of *because*.

[10] This factor had no effect on responses and is not discussed further.

Results and Discussion

The performance of the network was generally accurate, with minimal error on the current and previous windows to the non-test segments. For the experimentally manipulated segments, the coronal scores for the six conditions are summarized in Table 1.

Table 1.
Results of Simulation 1.

Surface place	*Without phonological context*	*With phonological context*
Coronal (unassimilated)	99 (1.5)	98 (9.6)
Labial (assimilated)	5 (9.3)	49 (37.0)
Velar (assimilated)	15 (21.7)	55 (29.0)

Note. The figures quoted are the coronal scores (in percentage terms) for each condition (standard deviations in parentheses).

A two-way item ANOVA was carried out on the data using the variables surface place and context. The main effect of surface place ($F[2,34] = 84.7$, $p < 0.001$) showed that surface labial and velar segments are treated as underlyingly less coronal than surface coronal segments. The main effect of context ($F[1,17] = 52.4$, $p < 0.001$) showed that the network is more likely to treat a segment as coronal once its following context is known. However, the interaction between these variables ($F[2,34] = 20.2$, $p < 0.001$) indicates that this context effect is restricted to the surface labial and velar conditions.

To ensure that this pattern of results was not simply an artifact of the initial weights used in training, the network was retrained four times using different starting weights. The replications showed almost exactly the same error curve during learning and were again tested after 100 sweeps through the training corpus. Each replication was analyzed separately and showed exactly the same pattern of results as the original network. Because of the close similarity found between the responses of the original network and the replications, only the original network was tested in subsequent simulations.

As Table 1 shows, the coronal scores for unassimilated conditions are roughly 100%: the network treats coronal segments as completely unambiguous. The assimilated segments are treated as much more ambiguous: before the phonological contexts of the segments are known, the network tends to treat assimilated segments at face value, giving a non-coronal (low) response. Once the following context is known, the coronal scores rise sharply, to 49% for the surface labial segments and 55% for the surface velars. The high standard deviation values for the assimilated conditions with phonological context (37% and 29%) suggest that the output of the network fluctuates strongly between highly coronal and highly non-coronal responses. This conclusion is supported by the distribution of responses for these conditions (Figure 2), which shows a roughly bimodal pattern, with peaks at both ends of the response range. This contrasts with the pattern of responses one would expect if the mapping from surface to underlying place was random. In such a case, a deterministic network would produce a unimodal distribution, peaking towards the middle of the range of responses (Movellan & McClelland, 1993). Instead, the network makes use of the available contextual cues to make a best guess at the underlying place. This aspect of the network's performance can be argued to mirror the categorical performance found in human perception of phonological form.

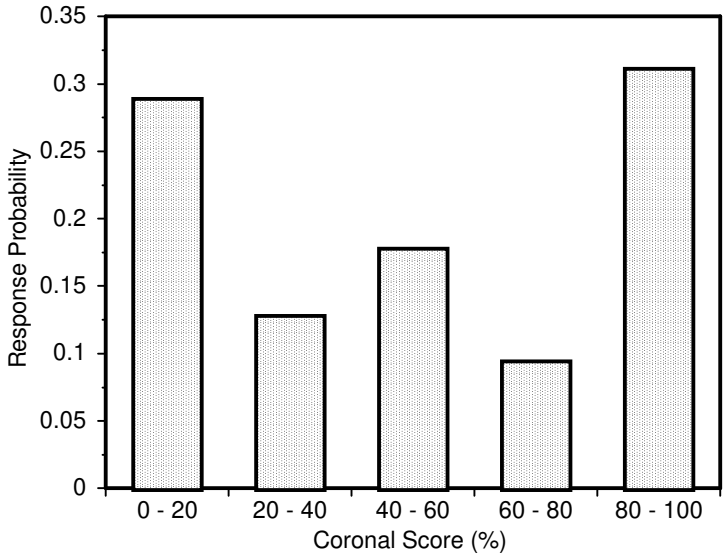

Figure 2. Distribution of the responses (in terms of coronal score) to the labial and velar surface place conditions on presentation of the following context (based on the data from the initial simulation plus the 4 replications).

The variability found in this simulation provides some support for the notion that the behavior of the network is equivalent to a simple mapping based on an underspecified representation of place. The high coronal score and low standard deviation for unassimilated conditions show that surface coronal segments are consistently mapped onto underlying coronal representations. In contrast, the surface labial and velar segments provoke a very variable response, ranging from highly coronal to highly non-coronal. Similar asymmetries were found in the gating studies cited in the introduction. Lahiri and Marslen-Wilson (1991) showed that speech tokens with oral vowels were consistently mapped onto lexical entries containing oral vowels, whereas tokens with nasal vowels were sometimes mapped onto lexical entries with underlyingly nasal vowels and sometimes onto lexical entries containing oral vowels. Similarly, Marslen-Wilson and Nix (1992) found that coronal segments were consistently perceived as coronal whereas labial and velar segments provoked a more variable response. Since the representations used to train the network were fully specified, this suggests that this *functional underspecification* has developed as a reflection of the asymmetry of the training corpus, in which surface coronal segments are reliable cues to an underlying coronal segment, whereas surface non-coronals can be underlyingly either non-coronal or coronal.

However, this difference in variability is mainly restricted to the conditions in which the following context of the target segment is known. Before this point, both coronal and non-coronal segments are treated by the network at face value (although there remains a small amount of variation in the response to velar segments at this point). This lack of variation is because the vast majority of surface non-coronal segments in the training corpus are also underlyingly non-coronal. The network is sensitive to these statistics and therefore opts for the surface form of non-coronal segments during testing. Once the following context of these segments is known to license assimilation there is a much better chance of the place of the underlying segment being coronal and so the network shows more compensation in these circumstances. Even here, however, the underlying identity is not certain. This is because, although the conjunctions of consonants used here will sometimes occur as a result of place

assimilation, they can also occur naturally, both within words like *ambulance* as well as across word boundaries. This ambiguity can only be resolved with the use of lexical information.

The importance of lexical information in the resolution of surface ambiguity is emphasized by the gating study of Marslen-Wilson and Nix (1992; see also Nix et al., 1993). For assimilated sentences such as 'They thought the *lake cruise* was rather boring', where both derivations of the surface ambiguity are words (i.e., *late* and *lake*), subjects' responses are correspondingly mixed: when presented with sentences with speech following the assimilated segments removed, 55% of subjects responded with the coronal form and 45% with the non-coronal form. However, when there was a clear lexical bias towards either the underlying coronal derivation (as in *red/reg*) or the underlying non-coronal derivation (as in *blat/black*), subjects almost never opted for the nonword. Simulation 3 investigates the incorporation of lexical effects such as these into the network model.

Simulation 2

Simulation 2 examines the second aspect of the perception of systematically variant speech: phonological inference. Phonological inference involves the application of phonological rules or constraints to sequences of surface speech in order to recover the underlying canonical form. This process has been demonstrated in a cross-modal priming study (Gaskell & Marslen-Wilson, in press), by examining the effect of *phonological viability* on lexical access. Here, we replicate these manipulations, for example, comparing the network's response to a target surface /k/ embedded in a context which either renders assimilation from an underlying coronal segment viable (e.g., [kwaɪkgoω], where the [k] may be an assimilated /t/, as in *quite go*) or makes assimilation phonologically unviable (e.g., [kwaɪkbaɪ] — *quike by* — where the [k] and [b] mismatch in terms of place of articulation).

For the network to demonstrate a context effect of this type, it must make use of its recurrent links to learn the phonological regularities involved in place assimilation. Specifically, it must learn that for a surface labial or velar segment to be underlyingly coronal, it must be followed by a segment with the same place of articulation. During testing, therefore, a viable context should increase the amount of compensation for assimilation the network exhibits, pushing responses to the target towards an underlying coronal place. An unviable context, on the other hand, should push responses the other way, towards a non-coronal reading of the target segment.

As well as contrasting viable and unviable phonological contexts, Simulation 2 examines variation within the general class of phonologically unviable changes. This is done by comparing following contexts which are either minimally or maximally unviable, as defined by their global featural similarity to a viable context.

Design and Materials

The trained network was tested using the same carrier words as Simulation 1. However, the following contexts of these words were altered so that each item was presented in four ways. In the baseline condition, each item was presented in its unassimilated form, with a word-final coronal segment (e.g., [kwaɪtpeɪ], *quite pay*). In the remaining conditions, the place of this segment was altered, either to labial or velar. In one condition, this change was made viable by its following context (e.g., [kwaɪppeɪ]). The other two conditions used following contexts which invalidated the phonological change as a natural assimilatory process. In the minimally unviable condition, the following context was featurally similar to a validating context, differing only on the place of articulation (e.g., [kwaɪpgoω], *quipe go*). In the maximally unviable condition, the following context was a vowel, which mismatched the

conditions for a viable phonological change on a number of phonetic features (e.g., [kwaɪpoʊn], *quipe own*).

For the conditions in which a place change occurred, each item was presented twice, once with a labial resultant place and once with a velar resultant place, and the mean coronal score across these two presentations was used in the analysis. Where necessary (i.e., in the viable and minimally unviable conditions) the context words were alternated to maintain viability or lack of viability. Again the network's evaluation of the word-final segment of the carrier word was used as the dependent variable, this time only measuring when the following context was presented. All other aspects of the procedure remained constant.

Results And Discussion

The data (see Table 2) were subject to a one-way item ANOVA with 4 levels (unassimilated, viable assimilated, minimally unviable and maximally unviable), which showed a highly significant main effect ($F[3,51] = 71.2$, $p < 0.001$). A post hoc Newman-Keuls test on the data showed that all conditions differed significantly at the 5% level. Most critically, the responses to the viable condition were significantly more coronal than the responses to both minimally and maximally unviable conditions.

Table 2.
Mean coronal scores (%) and standard deviations (in parentheses) for Simulation 2.

Condition	***Coronal score***
Unchanged	98 (6.0)
Viable	51 (23.0)
Minimally Unviable	39 (24.0)
Maximally Unviable	11 (25.0)

These results suggest that cues to the presence of assimilation provided by the phonological context of assimilated segments are used by the network in its evaluations of the speech. The presence of a viable context for assimilation significantly increases the coronality of the response of the network to place-changed segments, as compared to either unviable context. This implies that the network has learned to identify the phonetic context in which assimilation occurs. The network uses this contextual knowledge to extract the underlying identity of assimilated segments.

The similarity between the viable and minimally unviable contexts explains why the minimally unviable context still provoked a certain amount of compensation for assimilation, reflected in the higher coronal score for that condition (39%) compared to the maximally unviable condition (11%). This is another example of the graded behavior found in many connectionist networks — the viability of the phonological context of an assimilation is measured by the similarity between that context and the ideal (i.e., correct) phonological context. So a phonological context which matches the viable context on 10 out of 11 features is, to the network, quite a good context for assimilation and so provokes a fair amount of compensation. The vowel contexts, on the other hand, mismatch viable contexts on many other features, such as *vocalic*, *consonantal*, and *continuant*. This stronger violation of phonological constraints prevents the network from compensating for assimilation, producing responses similar to the no-context condition of Simulation 1. The prediction of a continuum of phonological viability is as yet untested. However the strong effects of minimally unviable context found in cross-modal priming experiments (Gaskell & Marslen-Wilson, in press) would seem to fit better in an all-or-nothing model of phonological viability.

Simulation 3

So far, we have found behavior in accordance with both abstraction and inference in the network's treatment of phonologically variant speech. However, the analogy to underspecification of representation was weak, since the network only treated non-coronal segments in a variable manner once the following context was known to license the change. This is a quite different process to the lexical matching strategy proposed by Lahiri and Marslen-Wilson (1991), which allowed systematic variants to map onto underspecified forms irrespective of context. This strategy is able to explain the gating results of Marslen-Wilson and Nix (1992) in which place assimilated segments were perceptually ambiguous even when their following context was absent (i.e., spliced out).

We have argued that the reluctance of the network to compensate for assimilation before its following context is known is a reflection of the statistics of the training corpus: the vast majority of surface labial and velar segments presented to the network during training were also underlyingly non-coronal. The network can overcome this statistical bias by utilizing cues to the presence of assimilation embedded in the environment of the segment in question. As Simulation 2 illustrated, the network is able to pick up cues to the viability of a phonological change from the place of its following context. But if the network is to compensate for assimilation before following context is known, it must exploit cues that precede the assimilated segment. That is, it must exploit cues to the lexical status of the underlying words.

For example, a cue to the place of the /n/ underlying [m] in [limbeɪkən] (*lean bacon*) is the preceding [li]. Of the two underlying possibilities, /n/ and /m/, only the /n/ completes an English word, and so a process which has access to lexical information can exploit this, deducing the underlying coronal place of the [m] before its context is known to validate assimilation. Of course the model, as it stands, cannot use lexical information directly, as it has not been trained to isolate lexical items. However, it may still show a bias towards underlying responses that complete words over responses that complete nonwords, since the word sequence a more familiar sequence of segments. Thus, to the network, /lin/ is a familiar sequence of segments at the output layer, since it has been presented repeatedly during training, but /lim/ is not.[11] Since networks of this type operate by extracting statistical regularities from their input, this would imply that there would be a *pseudo-lexical* bias in the network's response to surface labial or coronal segments. Such an effect has been found by Shillcock et al. (1991, 1992) in the evaluation by a recurrent network of phonetic ambiguities (Ganong, 1980). It is also possible that the network may be able to extract more than simple co-occurrence statistics from the training data. Elman (1993) has shown that, under certain circumstances, networks of this type develop representations which reflect the higher level structure of the temporal sequences on which they are trained (in Elman's case, the syntactic structure of a stream of words) rather than acting merely as statistical analyzers.

Effects of familiarity are likely to be strongest at the ends of high frequency words, since these are the ones in which the information gained from the analysis of statistical regularities is most reliable. For this reason, Simulation 3 examines the effects of preceding context on the resolution of surface ambiguity using only high frequency words. Because of the lack of high frequency words with word-final coronal segments in the original corpus, a modified corpus is used, in which the frequencies of selected words are artificially increased. These high frequency words are used in testing, firstly to check that biasing effects occur, and secondly in

[11] This analysis ignores the potential for sequences such as /lim/ to occur across word boundaries, as in *He thought the flea might die*. Such occurrences, however, will generally be less frequent and should be more or less balanced out by similar cross-boundary sequences involving coronal segments (e.g., *flea never*). Cross-boundary occurrences may also be phonetically different to the same sequence of segments occurring within a word.

a replication of Simulation 1, to examine the representation formed of place of articulation for these words.

Design and Materials

Simulation 3(a).

The test items consisted of pairs of words, of which the first was the ambiguous test word and the second provided viable phonological context for assimilation (see Table 3). Twelve of these were items which contained a phonologically ambiguous surface labial or velar segment and were biased by the preceding lexical context to an underlying coronal reading of that segment (e.g., [deɪvɪbpeɪ]; underlyingly *David pay*). The remaining items contained a bias towards an underlying labial or velar segment (e.g., [jobpeɪ], *job pay*). In half the items, the ambiguous segment had a labial surface form and in the other half the critical segment had a surface velar place. Thus all test items contained a labial or velar word-final segment in surface form, but the preceding segments contained biasing information towards either the +coronal or -coronal resolution of the ambiguity. No items contained lexical biases towards both the +coronal and -coronal resolution (as in *line/lime*). The three independent variables were bias (+coronal or -coronal), surface place (labial or velar) and monitoring point (before at or after the ambiguous segment)

Table 3.
Example stimuli for Simulation 3.

Surface place	***+Coronal bias***	***-Coronal bias***
Labial	* davib pay	job pay
	(*David pay*)	(* *jod pay*)
Velar	* wickig go	along go
	(*wicked go*)	(* *alon go*)

Note. The italicized phrases are the underlying words assuming a place assimilated coronal. Phrases marked with an asterisk contain a nonword.

In order to boost the frequencies of the biasing words, the training corpus for the network was altered so that these words had token frequencies of 30 (out of a total of just over 3000 words). This was done, for each biasing word (e.g., *David*, *job*), by replacing all tokens of two or three other words by that word. The substituted words were chosen so that their combined corpus frequency, added to the original corpus frequency of the replacement word, was 30. In terms of the proportion of the corpus containing these words, a corpus frequency of 30 corresponds to a word frequency of roughly 10000 in a million, and so makes these words as frequent as fairly high frequency function words such as *is* and *it* (Johansson & Hofland, 1989). This adapted speech corpus was used to train a network identical in architecture to the original network (see Figure 1). As before, the network was trained on 100 sweeps through the entire corpus.

The network's evaluation of the surface ambiguity was measured in all three output windows, allowing any changes in prediction over time to be assessed. Details of procedure were the same as Simulation 1.

Simulation 3(b).

A replication of Simulation 1 was also carried out, using the 12 +coronal biased high frequency stimuli. This was to examine differences in the treatment of coronal and non-coronal surface forms for these words. Apart from the change of stimulus words, all aspects of design and procedure remained constant.

Results

The results of Simulation 3(a) are, as before, in the form of a coronal score. A high value of this measure indicates that the network is treating the input as being underlyingly coronal. The mean coronal scores for each condition are presented in Table 4.

Table 4.
Results of Simulation 3(a)

		Monitoring point		
Lexical bias	***Surface place***	***Before ambiguity***	***At ambiguity***	***After ambiguity***
+Coronal	Labial	96	47	57
+Coronal	Velar	91	41	59
-Coronal	Labial	21	7	18
-Coronal	Velar	29	5	20

Note. The figures quoted are the percentage coronal scores.

A three-way ANOVA was carried out on the data, using the variables bias (the lexical bias), surface place (the surface place of articulation of the ambiguity) and monitoring point (before, at or after the ambiguous segment).[12] There was a highly significant main effect of bias ($F[1,20] = 21.11$, $p < 0.01$), showing that the network was more likely to compensate for assimilation when the underlying coronal completed a real word. The main effect of monitoring point ($F[2,40] = 26.97$, $p < 0.01$) showed that the pattern of compensation shifted as further information was presented. There was a significant interaction between bias and monitoring point ($F[2,40] = 3.39$, $p < 0.05$), suggesting that the effect of lexical bias was strongest before the ambiguity was presented. No other effects approached significance.

The data at the three monitoring points were also analyzed separately in two-way ANOVAs. At each monitoring point the effect of the bias variable was significant (before: $F[1,20] = 75.22$, $p < 0.01$; at: $F[1,20] = 9.97$, $p < 0.01$; after: $F[1,20] = 8.78$, $p < 0.01$). Again, no effects involving surface place were significant. These biasing effects are illustrated in Figure 3.

[12] The 'Before' measurements were taken from the next-segment window, the 'At' from the current-segment window and the 'After' from the previous-segment window.

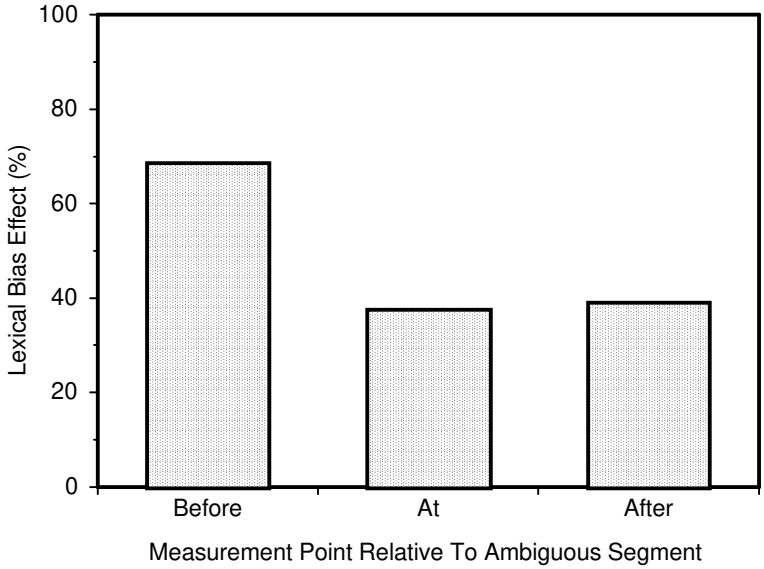

Figure 3. Difference scores ([+coronal] bias – [-coronal] bias) for Simulation 3(a). This shows the strength of the lexical bias effect at each of the three monitoring points.

The results of Simulation 3(b) are summarized in Table 5. A two-way ANOVA on these data showed a significant main effect of place of articulation ($F[2,22] = 16.3$, $p < 0.001$), indicating that surface coronals evoked a more coronal underlying response than surface labial or velar segments. There was also a main effect of the output window used ($F[1,11] = 6.7$, $p < 0.05$), showing that responses when the following context was known to be viable provoked more compensation for assimilation than responses before this context was known. The interaction between these two variables was marginally significant ($F[2,22] = 3.3$, $p = 0.058$).

Table 5.
Results of Simulation 3(b).

Surface place	***Without phonological context***	***With phonological context***
Coronal (unassimilated)	100 (0)	100 (0)
Labial (assimilated)	43 (38)	60 (33)
Velar (assimilated)	35 (33)	61 (36)

Note. The figures quoted are the percentage coronal scores (standard deviations in parentheses).

Discussion

The results of Simulation 3, examining the effects of lexical context on the disambiguation of phonological changes in high frequency words, form a clear pattern. In Simulation 3(a), each monitoring position shows a strong biasing effect of preceding context on the network's response to surface ambiguity. These effects are strongest before the ambiguous segment is presented, showing the extent to which the network has learned to use the reliable co-occurrence information in the frequent words. On presentation of the ambiguous segment, the network's responses maintain the strong lexical bias but are shifted towards the non-coronal response, reflecting the network's preference for a surface reading of a non-coronal segment. This behavior is due to the massive predominance of unassimilated non-coronal segments in

the network's training set. Finally, as the viable context for assimilation is encountered, the responses shift back somewhat towards a coronal response. This shows the sensitivity of the network to the phonological conditions in which place assimilation occurs. However, at all points the network displays a strong bias in its responses towards a real-word resolution of surface ambiguities.

Of course, given the high frequency of these test words in the training corpus, this strong frequency effect is not surprising.[13] What is more interesting is the effect these lexical biases have on the model's representation of place of articulation. Simulation 3(b) shows that these lexical biasing effects allow the network to construct a representation of place of articulation more compatible with an underspecified description. In both output windows, surface coronal segments are treated as unambiguously coronal, whereas labial and velar segment provoke much more variable responses (with standard deviation values ranging from 0.33 to 0.38). This asymmetry in the treatment of place suggests that under certain conditions (i.e., when a surface non-coronal segment is preceded by a context that provides strong cues to the underlying coronal nature of the segment), the network's behavior is functionally equivalent to an underspecified representation of place in the input lexicon.

It is important to note that despite the strong lexical biasing effects found here, the model remains intolerant of random deviation. To demonstrate this, we presented the network with the -coronal biasing words with a place change on the final segment (e.g., *job* presented as *jod*). This kind of deviation is not phonologically viable in any kind of context. Despite the fact that in these cases the preceding context provides strong cues towards the -coronal completion, the network treated these segments at face value: The mean coronal score on presentation of these segments was 99.5%. Thus, the network cannot be described as having a general tolerance of deviation towards the end of a word. Instead, compensation for phonological variation only occurs under tightly constrained conditions.

General Discussion

In this paper, we have examined a necessary component of human speech perception: the mapping from the phonetically variant surface form of speech to its underlying representation. Our model applied connectionist learning principles to this mapping, finding that the contextual constraints that govern assimilation can be acquired and exploited in a simple manner. The trained network shows effects of phonological inference in its processing, using the context of phonological changes to deduce their underlying identity. The network also exhibits behavior in concordance with an underspecified description of its underlying representations, despite the fact that all training and testing is carried out using fully specified phonetic transcriptions. In this final section we shall examine both these qualities, examining the implications of a connectionist model of such behavior.

Learned Phonological Inference

All three simulations showed that the trained network is sensitive to the phonological conditions in which place assimilation occurs. The network will compensate for assimilation when presented with a surface labial or velar segment, but not for a coronal segment. The viability of a phonological change is exploited by the network, provoking more compensation for assimilation when a phonological change is followed by a segment of the same place than otherwise. Finally, Simulation 3 showed that, for high frequency words, the lexical status of the preceding context of surface labial and velar segments affects the outcome of the mapping

[13] A pretest of these materials showed that, when tested on the original (unadjusted) training corpus, there was no significant biasing effect of preceding context (see Gaskell, 1994).

process. The network learns to perform what could be described as the reverse of the mapping occurring in speech production, deriving "rules" of inference from perceptual experience.

In fact, however, the performance of the network is better described in terms of the satisfaction of soft or weak constraints (cf. McClelland & Rumelhart, 1988). Contextual effects, such as the effect of preceding context in Simulation 3 and the effect of following context in Simulation 2 are graded, depending on measures of global similarity rather than all-or-nothing rules. What evidence is there that the human perceptual system behaves in this manner?

We have recently examined this question in a set of phoneme monitoring experiments (Gaskell & Marslen-Wilson, 1994). The experiments used carrier sentences such as "The city got two awards for its *clean parks*" in which the critical word (*clean*) was presented either in unassimilated form or with a word-final place change (*cleam*). Contextual effects on the perception of this change were examined by manipulation of the lexical status of the word containing the assimilation (e.g., comparing *clean* with *threan*) and by alteration of the segment following the place change to manipulate its phonological viability (e.g., comparing the viable assimilation, *cleam parks*, with the unviable change, *cleam guesthouses*).

These stimuli were used in two experiments, both of which showed a similar pattern of results. In the first experiment, subjects were required to monitor for the word-final coronal segment of the critical word (e.g., the /n/ in *clean* or the /n/ underlying *cleam*). For the conditions with place-changed segments, the target was not present in the surface form of speech and a response could only be made on the basis of a more abstract underlying representation. Unsurprisingly, the presence or absence of a place change had a strong effect on the proportion of responses made to each condition. However, within the phonologically changed conditions, there were also effects of preceding context (lexical status) and following context (phonological viability) on the responses. Although these effects interacted, there were still effects of both factors in the absence of the other. Thus, there was an effect of phonological viability for the nonword stimuli, with more responses when the phonological change was viable than unviable, and there was an effect of lexical status for the unviable phonological changes, with more responses for the word carriers than the nonword carriers.

In the second experiment, the target was the word-initial segment following the critical word (e.g., the /p/ of *clean parks*), which was held constant across conditions. Although effects were weaker in this experiment, and were based on response times rather than response proportions, much the same pattern of results was found. Across conditions of lexical status there was a significant inhibitory effect of the unviable phonological changes on responses. For words, this effect was a robust 51 ms, whereas for the nonword conditions there remained a tendency for the unviable context conditions to evoke longer responses than the viable changes, although the 25 ms effect did not reach significance.

This pattern of results falls naturally out of a model in which independent contextual cues to the presence of assimilation (such as lexical bias or phonological viability) are treated as weak constraints. In particular, the fact that compensation can occur in the perception of nonwords rules out a theory based purely on representational abstractness at the lexical level.

Functional Underspecification

We have argued that aspects of the network's behavior in Simulations 1 and 3 are consistent with an underspecified description of the network's underlying phonological representation. In Simulation 1, the network developed an asymmetry in its treatment of coronal and non-coronal places of articulation. Coronal segments were uniformly mapped onto underlying coronal segments whereas surface labial and velar segments were treated more variably, provoking both strong coronal and strong non-coronal responses. In the original

network, this behavior was delayed; being confined to the network's responses on presentation of the following context of the segments in question. However, Simulation 3 showed that for segments embedded in high frequency words that do not form minimal pairs, this pattern of behavior was exhibited on presentation of the target. Thus, in certain circumstances, the network treats surface non-coronal segments as ambiguous; sometimes mapping them onto underlyingly coronal segments and sometimes keeping the surface place in the underlying representation.

This asymmetric variability, which reflects a similar variability in the network's training regime, is similar to experimental studies of underspecification. For example, Lahiri and Marslen-Wilson (1991) show that surface oral vowels (which are analogous to English surface coronal consonants) are consistently perceived by Bengali subjects as underlyingly oral (only 0.7% of responses contained an underlying nasal vowel). On the other hand, surface nasal vowels (which are analogous to English surface labial or velar consonants) provoked a more mixed response, with over 30% of responses being underlyingly oral. A similar pattern of results was found by Marslen-Wilson and Nix (1992) in a study of English place assimilation.

The experimental results have been explained in terms of an underspecified lexical representation. Here, however, the behavior can be seen as a special case of the mapping between surface and underlying forms of speech in which cues to the underlying nature of surface non-coronal segments are strong enough to provoke compensation for place assimilation. In the absence of information about phonological viability, there must be sufficient lexical cues to underlying place present in the *preceding* context of a surface ambiguity.

In the network model, this effect of preceding context only appeared for the high frequency words in the training corpus, since these were the only words for which the statistics of the speech stream were reliable enough to allow the underlying nature of the surface ambiguities to be predicted. For less frequent words, however, it appears that such effects can only be modeled by a network which has developed word-like or morpheme-like internal representations. Whether these representations can be developed in the absence of specific lexical training is, as yet, unclear (Elman, 1993), and for the purposes of this research is unimportant. No matter how these lexical representations are developed, they are used in the disambiguation of surface form in a graded manner, interacting with other contextual weak constraints on the underlying identity of surface speech.

Our discussion of abstraction and inference has treated them as separable components of, respectively, representation and process. However, the connectionist approach we have taken blurs this distinction, transforming both into related aspects of a single mapping process. Phonological inference, in this paradigm, is a contextual sensitivity in the processing of non-coronal surface segments, and abstraction is an asymmetry in the tolerance for variation of place of articulation.

The Locus of Phonological Processing

The connectionist description of the inference process involved in speech perception places the locus of phonological inference neither entirely pre-lexically nor entirely lexically. Our model of phonological inference in speech perception makes the best use of the available cues to underlying identity. We have argued that for the behavior of the model to be consistent with data on underspecification, the network must develop word-like internal representations during training. Thus, the output of the network will depend strongly on lexical information when such cues exist. However, when these cues are weaker or non-existent, as in the perception of nonwords (or of minimal pairs), phonological inference may still occur,

depending on other cues such as the contextual viability of a phonological change. Here, phonological inference proceeds in a non-lexical manner.

This description of the locus of phonological inference bears a surface resemblance to dual route models of phoneme monitoring (Cutler & Norris, 1979; Foss, Harwood & Blank, 1980), in which phonemic knowledge can be tapped either lexically or pre-lexically. Here, however, the behavior falls naturally out of the application of a simple learning strategy to the mapping involved in the perception of normal connected speech. Again, this description agrees with the phoneme monitoring results of Gaskell and Marslen-Wilson (1994) in which compensation for assimilation was found for both words and, to a lesser extent, for nonword stimuli.

Conclusions

Our network simulations provide further evidence that the application of a connectionist learning strategy to the mappings involved in cognitive processing is a valid and fruitful approach. The model demonstrates varying degrees of both abstractness of representation and rule-like transformations depending on the relevance of each to the situation in hand. These aspects of linguistic structure are developed from an initially neutral processing environment by exposure to the language itself (or at least a simplified representation of the language). We cannot (and do not wish to) claim that our model is a general model of linguistic structure, since there are many issues relevant to underspecification that have not been touched on in our simulations. Instead, our aim has been to match our model with the range of psycholinguistic data on the perception of phonological change.

Our research provides particular support for an expanding group of studies describing phonological structure and psycholinguistic performance in terms of connectionist systems (e.g., Jordan, 1986; Hare, 1990; Dell, Juliano & Govindjee, 1993). It is also consistent with a broader set of studies, which argue that other areas of psycholinguistic behavior are best described in connectionist learning terms (e.g., Plunkett & Marchman, 1993; Plaut, McClelland, Seidenberg & Patterson, in press; Moss, Day, Hare & Tyler, in press; Hare & Elman, in press).

References

Archangeli, D. (1984). *Underspecification in Yawelmani phonology and morphology*. Unpublished doctoral dissertation, MIT, MA.

Archangeli, D. (1988). Aspects of underspecification theory. *Phonology*, **5**, 183-207.

Barry, M. C. (1985). A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics*, **4**.

Browman, C. P., & Goldstein, L. (1991). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Between the grammar and physics of speech* Cambridge, England: Cambridge University Press.

Byrd, D. (1992). Perception of assimilation in consonantal clusters: a gestural model. *Phonetica*, **49**(1), 1-24.

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.

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: Erlbaum.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, **17**, 149-195.

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.

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: Erlbaum.

Ganong, W. F. I. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, **6**, 110-125.

Gaskell, M. G. (1994) Spoken word recognition: a combined computational and experimental approach. Unpublished doctoral dissertation, Birkbeck College, University of London.

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

Gaskell, M. G., & Marslen-Wilson, W. D. (in press). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*.

Hare, M. (1990). The role of similarity in Hungarian vowel harmony: a connectionist account. *Connection Science*, **2**, 123-150.

Hare, M., & Elman, J. L. (in press). Learning and morphological change. *Cognition*.

Holst, T., & Nolan, F. (in press). The influence of syntactic structure on [s] to [sh] assimilation. *Laboratory Phonology 4.*

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, England: Clarendon Press.

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: Erlbaum.

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

Kaye, J. D., Lowenstamm, J., & Vergnaud, J. (1990). Constituent structure and phonological government. *Phonology*, **7**.

Keating, P. (1988). Underspecification in phonetics. *Phonology*, **5**, 275-292.

Kerswill, P. E. (1985). A sociophonetic study of connected speech processes in Cambridge English: an outline and some results. *Cambridge Papers in Phonetics and Experimental Linguistics*, **4**.

Kohler, K. (1990). Segmental reduction in connected speech: Phonological facts and phonetic explanations. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modeling* Dordrecht: Kluwer Publications.

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. (1993). Issues of process and representation in lexical access. In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Language Processes: The Second Sperlonga Meeting* Hove: Erlbaum.

Marslen-Wilson, W., Moss, H. E., & van Halen, S. (in press). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, **101**(1), 3-33.

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

Marslen-Wilson, W. D., & Nix (1992). *Phonological variation in lexical access.* Paper presented at the meeting of the Experimental Psychology Society, Oxford, England.

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.

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

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs and Exercises*. Cambridge, MA: MIT Press/Bradford Books.

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

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science*, **17**, 463-496.

Nix, A., Gaskell, G., & Marslen-Wilson, W. D. (1993). Phonological variation and mismatch in lexical access. *Proceedings of the 3rd European Conference on Speech Communication and Technology*, Berlin, Germany: ESCA.

Nolan, F. (1992). The descriptive role of segments: Evidence from assimilation. In D. R. Ladd & G. Docherty (Eds.), *Laboratory Phonology II,* Cambridge, England: Cambridge University Press.

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. (1992). Connectionism: A new breed of bottom up model. In R. G. Reilly & N. E. Sharkey (Eds.), *Connectionist Approaches to Natural Language Processing* Hove, England: Erlbaum.

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, England: Erlbaum.

Ohala, J. J. (1990). The phonetics and phonology of aspects of assimilation. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology 1: Between the Grammar and Physics of Speech,* Cambridge, England: Cambridge University Press.

Paradis, C., & Prunet, J. F. (1991). *Phonetics and Phonology Volume 2: The Special Status of Coronals.* San Diego, CA: Academic Press.

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

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, **48**, 21-69.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. 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. (1994). Language and connectionism: the developing interface. *Cognition*, **50**, 385-401.

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

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1988). *Encoding sequential structure in simple recurrent networks* (Technical Report No. CMU-CS-88-183). Carnegie Mellon University, Computer Science Department, PA.

Shillcock, R., Levy, J., & Chater, N. (1991). A connectionist model of auditory word recognition in continuous speech. In 13th Annual Conference of the Cognitive Science Society, (pp. 340-345). Chicago, IL: Erlbaum.

Shillcock, R., Lindsey, G., Levy, J., & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. In 14th Annual Conference of the Cognitive Science Society, . Bloomington, IN: Erlbaum.

Stemberger, J. P. (1991). Apparent anti-frequency effects in language production: the addition bias and phonological underspecification. *Journal of Memory and Language*, **30**(2), 161-185.

Svartvik, J., & Quirk, R. (1980). *A Corpus of English Conversation*. Lund: Gleerup.

Author Note

M. Gareth Gaskell, Centre for Speech and Language, Birkbeck College, University of London; Mary Hare, Center for Research in Language, University of California in San Diego; William D. Marslen-Wilson, Centre for Speech and Language, Birkbeck College, University of London. This research was supported by a UK SERC studentship awarded to Gareth Gaskell and by MRC, ESRC, SERC and JCI grants awarded to William Marslen-Wilson and Lorraine Tyler. We are grateful to Richard Shillcock for the use of speech data in the network training and to Jeff Elman for the use of the TLearn software.

Correspondence concerning this article should be addressed to Gareth Gaskell at the Centre for Speech and Language, Department of Psychology, Birkbeck College, Malet Street, London WC1E 7HX, England. Electronic mail may be sent to g.gaskell@psyc.bbk.ac.uk.