Spoken Word Recognition: A Combined Computational and Experimental Approach

Mark Gareth Gaskell

Birkbeck College

University of London

Thesis submitted for the degree of Doctor of Philosophy

March 1994

Abstract

The research reported in this thesis examines issues of word recognition in human speech perception. The main aim of the research is to assess the effect of regular variation in speech on lexical access. In particular, the effect of a type of neutralising phonological variation, assimilation of place of articulation, is examined. This variation occurs regressively across word boundaries in connected speech, altering the surface phonetic form of the underlying words. Two methods of investigation are used to explore this issue. Firstly, experiments using cross-modal priming and phoneme monitoring techniques are used to examine the effect of variation on the matching process between speech input and lexical form. Secondly, simulated experiments are performed using two computational models of speech recognition: TRACE (McClelland & Elman, 1986) and a simple recurrent network.

The priming experiments show that the mismatching effects of a phonological change on the word-recognition process depend on their *viability*, as defined by phonological constraints. This implies that speech perception involves a process of context-dependent inference, that recovers the abstract underlying representation of speech. Simulations of these and other experiments are then reported using a simple recurrent network model of speech perception. The model accommodates the results of the priming studies and predicts that similar phonological context effects will occur in non-words. Two phoneme monitoring studies support this prediction, but also show interaction between lexical status and viability, implying that phonological inference relies on both lexical and phonological constraints. A revision of the network model is proposed which learns the mapping from the surface form of speech to semantic and phonological representations.

Acknowledgements

I am extremely grateful to my supervisor, William Marslen-Wilson, for his support, advice, and interest throughout my studentship. Equally, I would like to thank Mary Hare, for her help and encouragement on the connectionist chapters of this thesis. I am also grateful to Jeff Elman, for making TRACE available for analysis, and to Richard Shillcock, for making his speech corpus and connectionist model available.

The friendship and support of the Birkbeck Speech and Language Group has greatly facilitated the course of my work, as have the pool and pinball skills of Nick Greenwood and Andy Nix. I am grateful to Catherine Gaskell for comments on an earlier version of this thesis and to Lou Jones for her patience and understanding.

This research was supported by a studentship from the Science and Engineering Research Council.

Contents

Abstract **2**

Contents **4**

Chapter 1 — Introduction **7**
1.1 Speech Perception 7
1.2 Variation and Speech Perception 7
1.3 Computational Modelling 8
1.4 Theoretical Assumptions 9
1.4.1 A Working Model of Word Recognition. 9
1.4.2 The Input Representation 10
1.4.3 Lexical Representation 10
1.4.4 The Matching Process 11
1.5 Thesis Structure 16

Chapter 2 — Mismatch and Lack of Match **17**
2.1 Introduction 17
2.1.1 The Cohort Model 17
2.1.2 The TRACE Model 18
2.2 Studies of Mismatch in Lexical Access 20
2.2.1 Shadowing 20
2.2.2 Mispronunciation Detection 21
2.2.3 Gating 21
2.2.4 Priming 22
2.2.5 Summary 23
2.3 TRACE Simulations 24
2.3.1 Experimental Data 24
2.3.2 TRACE Simulation 1 25
2.3.3 TRACE Simulation 2 31
2.4 Conclusions 34

Chapter 3 — Mismatch and Phonological Variation **36**
3.1 Introduction 36
3.2 Phonological Theory 36
3.3 Natural Variation in Speech 38
3.3.1 Allophonic Variation 38
3.3.2 Phonemic Variation 38
3.4 Psychological Studies of Phonological Variation 40
3.4.1 Place Assimilation 40
3.4.2 Assimilation of Nasality 42
3.5 Models of Variation in Speech 42
3.5.1 Phonological Variation as Noise 42
3.5.2 Representational Models 43
3.5.3 Inference Models 44
3.5.4 Summary 45
3.6 Experimental Considerations 45
3.7 Experimental Data 46
3.7.1 Pre-test 48
3.7.2 Experiment 1 50
3.7.3 Experiment 2 52
3.7.4 General Discussion 55
3.8 Conclusions 57

Chapter 4 — A Connectionist Model of Phonological Inference 58
4.1 Introduction 58
4.2 Connectionist Modelling 58
4.2.1 The Roots of Connectionism 58
4.2.2 Learning Algorithms 60
4.2.3 Properties of Connectionist Networks 64
4.3 Connectionist Models of Speech Perception 67
4.3.1 Waibel's Phoneme Recognition Model 67
4.3.2 Recurrent Network Models of Speech Perception 68
4.3.3 Recurrent Network Models of Phonology 72
4.4 Simulating Phonological Inference 73
4.4.1 A Model of Pre-lexical Compensation 74
4.4.2 Simulation 1 — Phoneme Monitoring 76
4.4.3 Simulation 2 — The Effect of Phonological Context 79
4.4.4 Simulation 3 — Lexical Effects in Pre-lexical Processing 81
4.5 General Discussion 83

Chapter 5 — The Locus of Phonological Effects 85
5.1 Introduction 85
5.2 Phoneme Monitoring 85
5.3 Experimental Considerations 89
5.4 Experiment 3 90
5.4.1 Pre-test 91
5.4.2 Main Experiment 93
5.5 Experiment 4 98
5.5.1 Design and Materials 99
5.5.2 Pre-test 100
5.5.3 Main Experiment 101
5.6 General Discussion 103

Chapter 6 — In Search of Lexical Effects 106
6.1 Introduction 106
6.2 Simulation 4 — Memory Span 106
6.3 Simulation 5 — The Representation of Words 109
6.4 Speech Segmentation 112
6.4.1 Cues to Assimilation — A Re-analysis 112
6.4.2 A Connectionist Model of Speech Segmentation 113
6.4.3 Simulation 6 — The Segmentation Network 114
6.5 Simulation 7 — Word Frequency and Lexical Effects 119
6.6 General Discussion 124

Chapter 7 — Concluding Remarks 127
7.1 Models of Variation in Speech Perception 127
7.1.1 Variation as Noise 127
7.1.2 Lexical Representation of Phonological Change 127
7.1.3 Models of Phonological Inference 128
7.2 Linguistic Issues 130
7.3 Future Directions 130
7.4 Summary 131

Appendix A — Materials for Experiments 1 and 2 133

Appendix B — Materials for Experiment 3 136

Appendix C — Materials for Experiment 4 138

Appendix D — Simulation Materials 140
D.1 TRACE Simulations.....140
D.2 Simulations 1 and 2.....140
D.3 Simulation 3.....141
D.4 Simulations 5 and 6.....142
D.5 Simulation 7.....142

References 144

Chapter 1 — Introduction

1.1 Speech Perception

In this thesis I examine the role of variation in human spoken word recognition. This is a process that occupies a central role in the chain of events underlying speech perception. At the beginning of this chain, speech is transformed from a physical wave entering the ear to a pattern of neural signals. It is normally, but not generally, assumed that this process is followed by some form of feature extraction, allowing the perceptually relevant features of the speech to be isolated. Word recognition provides the key to the extraction of meaning, allowing access to the mental representations of selected word-forms. These representations, stored in the mental lexicon, consist of knowledge about the meanings and syntactic properties of the individual words, as well as their episodic and associative properties. Finally, the perception of speech involves a process of integration, allowing the individual semantic and syntactic properties of the words to be combined, generating the meaning of the sentence or utterance as a final product.

Earlier models of speech perception were based on existing knowledge of word recognition in the visual domain (e.g., Morton, 1969; Forster, 1976), largely because of the technical difficulties involved in the experimental manipulation of speech. But there are properties of speech that require special attention and suggest that visual and auditory word perception are distinct processes (but see Bradley & Forster, 1987).

Foremost among these properties is the fact that speech is temporally sequential and transient. Research suggests that very little speech can be retained in the auditory system in a fairly raw un-interpreted form (Crowder & Morton, 1969). This implies that the mapping from auditory information to word and sentence meaning must be both fast and effective. The sequential presentation of auditory information, as opposed to the more parallel information available in the visual domain, may also have implications for the way this information is processed.

Speech also lacks the boundary information that written and printed language contains. To a great extent speech is continuous: silent periods occur in speech waveforms, but these are as likely to be in the middle of words (e.g., before the aspiration of the /p/ in *rapid*) as they are to mark word boundaries. This continuum of information means that the units of representation in speech perception are opaque. Most alphabetic written languages have a rigid structure consisting of sentences made up of discrete words which are in turn made up of discrete letters. Although this external structure is not *necessarily* mirrored in the structure of the perceptual system, it is difficult to imagine that the visual word recognition system does not exploit these structural regularities. In contrast, no such structure is made obvious by the physical qualities of speech. Therefore the units of representation in the perceptual process, as well as the mechanism of the process itself must be elucidated experimentally.

The problem of the lack of boundaries in speech is compounded by the variance inherent in speech. This variation has all kinds of sources, from the simple background noise associated with the environment of the speech source to rule-governed variation found in many phonological processes. In all cases it affects the way speech is dealt with by the perceptual systems, often causing problems, but sometimes providing useful information about the environment of the speech being attended to.

1.2 Variation and Speech Perception

The phenomenon of variation in speech perception is perhaps the most neglected of all the issues currently facing researchers of human speech perception, and it is this issue that I address in the research reported here. In particular, I examine the implications of variation at the phonological level on models of *word recognition*: the process that extracts the meanings of the words we hear. The initial research I report examines the implications of earlier research on mispronunciations for these models. The remainder of my thesis is devoted to variation described by linguists as rule-governed or

lawful variation, and in particular to an example of phonological variation occurring across word and morpheme boundaries in connected speech.

As an illustration of the extent of the problems caused by phonological variation, consider the recognition of the word *stand*. In isolation, as well as in many utterance contexts this will be realised in its canonical form, [stænd][1]. However, the same word can also surface as [stæn] in the context of *stand down*, [stæN] in the context of *stand close* and [stæm] in the context of *stand back*. These regular phonemic changes result from the phonological processes of segment deletion and assimilation of place of articulation.

Changes such as these pose problems for theories of lexical access and word recognition in speech perception. My research on these changes provides empirical evidence relating to a number of issues of word recognition. The way the speech system responds to phonological variation is of relevance to both the representations used in lexical access and the way these representations are manipulated. Do phonological changes, for example, imply some abstractness in lexical or pre-lexical forms? Or does lexical access involve the application of phonological rules or generalisations, similar to those assumed to apply in speech production?

This thesis also examines how these representations and processes may develop, using connectionist techniques to model the development of phonological processes in speech perception. The work is also of relevance to the interface between psychology and phonology (Kohler, 1990; Hura, Lindblom & Diehl, 1992), contrasting theories in which phonological change is dependent on restrictions of speech production to those in which speech perception dominates language change.

1.3 Computational Modelling

Computational modelling involves the implementation of psychological models on a computer system, and is therefore an extension of standard psychological theory. The research reported here contains roughly equal proportions of experimental work and theoretical, computational modelling using connectionist structures. What, then, does computer implementation bring to the modelling of psychological processes?

One advantage of computer modelling is the explicitness it brings to the theory in question. The value of this kind of explicitness depends to a great extent on the theory itself: many psychological theories are more simple to explain and understand if described in terms of rules or heuristics, but two types of theory have obvious advantages when implemented on computers. In models such as SOAR (Laird, Rosenbloom & Newell, 1986) or GPS (Newell & Simon, 1963) the theory is the computer program itself: the theory can be described in simpler terms but the most explicit description is the actual coding used to design the program. Connectionist models also lend themselves to computer implementation, but for a different and more interesting reason. The way connectionist systems work, and in particular the way distributed learning systems work, is difficult to describe in the symbolic terms of many theories. Connectionist learning involves a simple mapping between input and output, but the behaviour of the trained system depends strongly on the characteristics of its input during learning (i.e. its experience). Because of this, it is often easiest to study the behaviour of the model by implementing it using a simplified representation of human experience.

Computer implementations of models of this kind render them testable. This allows the researcher to see whether or not the model does what he or she expects, and allows flaws in the theory to be exposed by examining the model's performance on existing psychological data. It also allows concrete predictions to be made about human performance in as yet untested areas. The majority of connectionist models have been tested in the former sense, by comparison to existing data, but there have been some recent examples of psychological studies motivated by connectionist models which have confirmed their predictions (e.g., Seidenberg & Bruck, 1990).

The role of connectionism in language processing remains highly controversial. Pinker & Prince (1988) argue that, at best, a connectionist model of a linguistic process is merely an implementation of

---

[1]In the postscript version of this thesis, the phonetic representations are somewhat non-standard. hopefully, this will not cause too much confusion.

a rule-based system. However, there is growing evidence that in some areas of linguistics the rule-based system itself is the approximation to the graded behaviour exhibited by humans (see Chapter 4). This debate has been conducted largely using inflectional morphology as a test case. In the field of phonology, the dominant theories are rule-based theories (but see Prince & Smolensky, in press). However, in this thesis I describe a model which applies connectionist learning principles to the perception of phonological change. A consequence of this approach is that phonological inference in speech perception is seen as a graded constraint satisfaction process rather than a tightly defined rule-based change. This view receives support from the phoneme monitoring studies described in Chapter 5.

1.4 Theoretical Assumptions

Although my research is not intended to be support for any specific model of speech perception, there is a paradigm of word recognition that underlies this work. In this section I describe the assumptions of this paradigm and then review the experimental evidence relating to the features I describe.

1.4.1 A Working Model of Word Recognition.

In this thesis I shall use the term *lexical access* to denote the process that allows lexical information about a word to be retrieved, and *word recognition* to be the more general process of isolating the correct lexical interpretation of incoming speech. Thus lexical access is a necessary component of word recognition, but there may be an additional selectional component if more than one lexical meaning is accessed (Marslen-Wilson, 1987; Pisoni & Luce, 1987). The basic model I assume is illustrated in Figure 1.1.

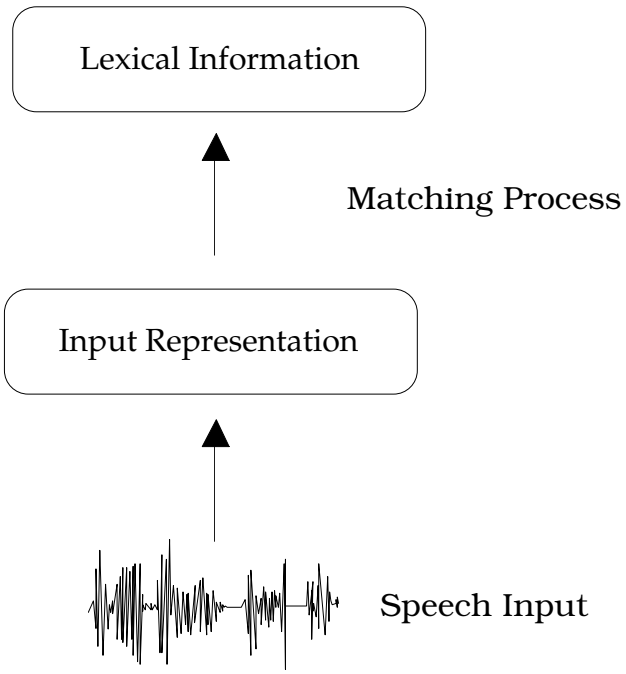

Figure 1.1. Auditory lexical access

I assume that speech is analysed into a detailed input representation, such as a phonetic featural representation, which is used as the basis for comparison with lexical forms. This comparison, the *matching process*, involves parallel assessment of multiple lexical candidates using a metaphor of activation (Morton, 1969; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986). Access to lexical information is granted to all candidates whose activation, whether transiently or ultimately, is sufficiently high (although in Morton's model, only the first candidate to reach the threshold activation elicits lexical knowledge). The activations of lexical candidates are a function of the

combined positive and negative perceptual evidence that the input representation provides for each word, and the expected outcome of the matching process is a set of lexical forms, each corresponding on a one-to-one basis to sections of the speech stream.

This description, loosely based on the Cohort model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987), is deliberately vague. I have attempted to make the minimum number of theoretical assumptions necessary to provide a basis for my research. For example, I am initially agnostic as to the mechanism and locus of segmentation in lexical access: it is sufficient to assume that such a process occurs, and thus that lexical information is stored and accessed word-by-word (or morpheme-by-morpheme)[2].

Nevertheless, there are assumptions in even this simple description that are not generally accepted. In the remainder of this section I will expand on this model and review the evidence accumulated on its properties.

1.4.2 The Input Representation

The input representation, or contact representation (Frauenfelder & Tyler, 1987), contains a pre-analysed form of speech, which is used to gain access to lexical knowledge through some kind of matching process. Many different kinds of units have been proposed for this representation, such as articulatory gestures (Liberman et al., 1967; Liberman & Mattingly, 1985), phonetic features (Stevens, 1986), spectral templates (Klatt, 1979, 1986), phonemes or segments (Pisoni & Luce, 1987) and syllables (Mehler et al., 1981). There are two issues involved in the discrimination between these options, one relating to the ease of detection of the representational units, the other relating to their value to the matching process.

The overriding problem with research into the pre-lexical units of speech is the lack of an invariant mapping between speech waveforms and linguistic units. Linguistic theory has led us to believe that speech is composed of discrete, relatively invariant units, reducing the process of pre-lexical analysis to a simple extraction of the discriminating information. Blumstein and colleagues (Stevens & Blumstein, 1981; Mack & Blumstein, 1983) have provided evidence for the invariance of a number of phonetic features in speech, but at present it seems likely that many features can only be identified in their phonological context. In particular, coarticulatory change, whereby the phonetic structure of a segment varies according to the identity of its neighbouring segments, makes the comprehensive use of invariant features by the perceptual system an unlikely possibility. This kind of problem led Klatt (1979, 1986) to propose a lexical access system based on lexical templates of diphones, reducing the scale of the problems that coarticulation causes.

The choice of input representation also affects the complexity of the matching process. A detailed representation swiftly cuts down the number of lexical candidates, whereas a more sparse representation allows a greater number of lexical candidates to remain active.

For the purposes of this research I shall assume an input representation analysed into perceptual features (Jakobson, Fant & Halle, 1963) since this appears to be the least controversial assumption. In fact, any kind of featural or segmental representation is sufficient for the purposes of both the simulations and experiments I report. However, in Chapter 5 I review the evidence, from phoneme monitoring and categorisation studies, that suggests that a simple phonemic representation is too coarsely grained to fully explain the experimental effects.

1.4.3 Lexical Representation

Lexical representations are assumed to consist of a form-based entry linked to the stored knowledge of a word. The representation of the phonological form of the word is used as the target representation for the auditory matching process and thus is the key to lexical access in speech perception. The research I report focuses on the *process* of lexical access rather than the information it retrieves; For this reason, I shall not discuss the structure of the semantic information in the mental lexicon.

---

[2]However, in Chapter 6 I shall argue that this research provides support for a segmentation process which combines pre-lexical and lexical cues to word boundaries.

To allow matching between speech input and lexical entries, the units of the lexical form representation must be compatible with the input units discussed above. But the *structure* of the lexical form representations is a quite separate issue. The standard assumption is that these representations contain a single fully-specified representation of each word, coded in the theorist's favourite units (e.g., features, phonemes). These representations can be compared to more structured representations, inspired by linguistic theories such as Autosegmental Phonology (Goldsmith, 1976). In Chapter 3 I review experimental and linguistic evidence relating to the issue of representational structure in speech perception.

1.4.4 The Matching Process

Many different mechanisms have been proposed for the process of mapping low level-units of speech onto lexical forms. Rather than listing the models themselves (see Norris, 1986; Klatt, 1989, for reviews) I shall examine a number of theoretical distinctions that these models have provided.

PARALLEL VS SERIAL SEARCH

A major segregating factor in models of lexical access is the degree to which parallel processing can occur during the matching process. The search model (Forster, 1976, 1981, 1989) uses serial searches of lexical lists as its basis for word recognition (although a small degree of parallelisation is allowed). The model assumes that normal auditory word recognition is the product of one of two serial searches: the first being a frequency ordered phonological list of word-forms, the second a semantic (or more correctly, associative) context-driven list. The order of the search makes it simple to explain the robust effects of word frequency on auditory word recognition (Rubenstein, Garfield & Millikan, 1970; Blosfeld & Bradley, 1981; Marslen-Wilson, 1990), and the inclusion of a similar context-based search in parallel explains the interaction between frequency and context effects (Becker, 1980).

Parallel search models (e.g., Morton, 1969; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986) allow matching between the input representation and any or all of the lexical entries at the same time, using the metaphor of activation to describe the state of the search for each candidate. Parallel search models do not accommodate frequency effects as directly as Forster's search model, since they must assume some kind of frequency bias, either in the resting levels of the activation functions (Morton, 1969) or in the response to facilitory or inhibitory information. However, connectionist learning algorithms offer a explanation for the development of these biases as a consequence of differential presentation proportions (Seidenberg & McClelland, 1989).

The major advantage of parallel models of word recognition is that they offer a simple explanation of the finding that, during the course of spoken word recognition, more than one word meaning is accessed (hence the distinction between access and recognition above). Zwitserlood (1989) used the cross-modal priming technique (Swinney et al., 1978; Onifer & Swinney, 1981) to examine the time-course of word recognition. Her experiments, carried out in Dutch, used prime words such as KAPITEIN (captain) which diverged phonologically from a competitor such as KAPITAAL (capital) towards the end of the word. These words were presented auditorily, in sentential contexts biasing subjects towards either the competitor or the prime word disambiguation, as well as in neutral contexts. The state of the word recognition process was assessed by presenting visual probe words related to either prime or competitor (e.g., *ship* for captain or *money* for capital) aligned to various points in the prime word. Zwitserlood found that early on in the presentation of the prime, the meanings of both words were accessed (in that the recognition of both probe words was facilitated). For the primes in neutral context the competitor word meaning was only inhibited when the disambiguating speech was presented, but for the biasing contexts, this inhibition could occur earlier (although it could not completely prohibit access to the contextually inappropriate meaning.

This transient access of multiple lexical entries is predicted by parallel models of word recognition (Marslen-Wilson, 1987; McClelland & Elman, 1986) but is incongruous in a serial search environment. These results could be accommodated in a serial model by conducting searches at regular intervals during the presentation of a word (e.g., on recognition of each successive segment) and accessing the meaning of all words failing to mismatch at each point. But all this achieves is a serial approximation to a parallel model. Word recognition, it seems, is best modelled by assuming that lexical candidates are assessed in parallel and that during recognition more than one lexical meaning is accessed.

AUTONOMOUS VS INTERACTIVE MODELS

The extent to which lexical access and its sub-components can be influenced by higher level processes has also been the subject of much debate. Many theorists (e.g., Forster, 1976; McQueen, Norris & Cutler, submitted) argue that information flow in word recognition is entirely bottom-up, with higher level processes being unable to influence the outcome of lower level processes. Similar ideas are expressed by Fodor (1983) when discussing the encapsulation of information in perceptual systems (although expressed in a looser sense). Other models vary as to the extent top-down processing may influence the perceptual system. The revised Cohort model (Marslen-Wilson, 1987) allows contextual constraints to operate at the higher levels of word recognition (i.e. in the selection process) whereas other theories (Marslen-Wilson & Welsh, 1978; Cole & Jakimik, 1980) allow interactive context effects to permeate the word recognition process. The TRACE model of McClelland & Elman represents an explicit exposition of the interactive extreme, with all levels of representation (phonetic features, phonemes and words) susceptible to the influence of information from both higher and lower levels.

The experimental evidence relating to this issue is confusing and inconclusive. There have been numerous demonstrations of the influence of top-down contextual information on both word recognition (e.g., Wright & Garrett, 1984; West & Stanovich, 1982; Zwitserlood, 1989) and segment recognition (Warren, 1970; Samuel, 1981; Foss, Harwood & Blank, 1980). However, these effects can often be explained as a result of decision bias (cf. Tanenhaus & Lucas, 1987), preserving the modularity of the lower level. For example, Warren (1970) examined the perceptual effects of replacing speech segments in sentential context with coughs or tones. Subjects were asked to determine the position of the non-speech sound, but found this task difficult and generally reported that the sentence was intact. It is tempting to conclude that this restoration of deleted segments is a top-down contextual effect, but it is equally plausible that in fact the word (or sentence) context biased the decision, which was made at the higher level, preserving the modularity of the processes.

This theoretically important distinction is extremely difficult to isolate experimentally. Samuel (1981a, 1981b) used signal detection methods to investigate the Warren (1970) phonemic restoration effect further. According to signal detection theory, perception involves the probabilistic discrimination of signals from noise. Comparison of response proportions in situations where the stimulus is present and where the stimulus is absent allows measures of *bias* and *discriminability* to be made.

Samuel presented subjects with noise-replaced and noise-added segments, embedded in words and nonwords in sentential context. Their task was to decide which of the stimuli were noise-replaced and which were noise-added. Samuel found that the effect of lexical status on subjects' responses was to reduce the discriminability of the target, whereas the sentential context of the words affected the bias measurement. He argued that the change in discriminability implied that the lexical effect on phoneme perception was a true perceptual (and thus top-down) effect, whereas the effect of context implied a change in decision bias.

While these results do indicate a difference in the effects of lexical and sentential context on phoneme recognition, it is not clear that the lexical effect is necessarily evidence for top-down processing. If the phonemic decision for real words is based at least in part on a lexical readout of phonological information, the reduction in discrimination may occur at this level, preserving the modularity of the system.[3] Massaro (1989) also points outs some methodological problems with this research.

Another attempt to isolate top-down context effects, while avoiding these problems, used the Ganong effect (Ganong, 1980) as its basis. This is a simple lexical effect on the processing of ambiguous phonemes. For example, when subjects are presented with a token which could be categorised as either a /d/ or a /t/ they are influenced by the lexical status of the carrier word. So if the ambiguous token is followed by *ash*, subjects are biased in their responses towards the /d/ because of the lexical status of *dash*. The opposite effect is found when the following context is *ask*, where *task* is the only real-word option. This effect could be explained as a top-down effect of lexical information on

---

[3]See Chapter 5 for a fuller discussion of the level or levels of representation of phonological knowledge.

phonemic processing, but it could also be explained in non-interactive terms if phonemic information is made available by lexical access.

Elman & McClelland (1988) used an ingenious modification of this task in an attempt to ensure the lexical read-out could not be used. In their experiment the word creating the lexical bias and the carrier were two different words, implying that any lexical effects found must be truly interactive. Their stimuli consisted of word-initial ambiguous phonemes, as above but where the carrier was such that both alternatives formed a real word (e.g., [d/g]*ates*). The lexical bias employed a coarticulatory compensation effect whereby subjects alter their perceptual boundaries between phonemic categories according to the identity of the preceding consonant (Mann & Repp, 1981). So if the ambiguous [d/g] above is preceded by a /S/, subjects are more likely to perceive it as a /d/, but a preceding /s/ pushes the boundary the other way. To make this effect a lexical one, the preceding biasing phoneme was also made phonetically ambiguous, with its identity decided by the preceding word, as in the simple Ganong effect. So the stimuli involved two ambiguous segments with a lexical effect on the first one causing a compensatory effect on the second one; for example, the stimulus *Engli*[s/S][d/g]*ates* should be perceived as *English dates*).

Elman and McClelland found that there was a lexical effect on the perception of the word-initial segments, which they attributed to a genuine top-down effect on phonemic processing. However, even here there are bottom-up alternatives. Shillcock, Lindsey, Levy & Chater (1992) showed that Elman & McClelland's findings could have been due to a confounding effect of the phonotactic regularities of the stimuli. Norris (1992) also showed that these effects could be captured in a bottom-up connectionist model of speech perception, provided the processes of phoneme and word recognition were able to interact.[4]

It is becoming clear that the debate about autonomous and interactive stages of processing is less significant when discussing distributed connectionist models. The popular model of human processing when these issues were first applied was the serial computer. Thus, processors were represented as isolated and the degree to which they interacted was a vital part of their description. In distributed models such as the Norris (1992) network, the concept of a level of processing is blurred, since knowledge is often represented over a number of layers of nodes, and these same nodes often encode many different types of information. In this case, it may be more helpful to determine the circumstances under which processes interact, rather than examining the flow of information itself (cf. Tabossi, 1993).

DYNAMICS OF THE MATCHING PROCESS

Irrespective of whether word recognition occurs as a result of a serial or parallel search, an explicit model of lexical access requires a mechanism by which the results of comparisons between input and lexical forms can be assessed. In serial models (e.g., Forster, 1976), this comparison process is relatively unimportant, since the recognition time is more dependent on the number of items searched than the comparison itself. However, in parallel activation models, the state of the comparison itself, as measured by the relative or absolute activations of the word candidates, dictates the duration of the process.

Using the activation metaphor, there are two components to this process. One is the algorithm used to compute the activations of the lexical candidates, the other is the criterion used to make a decision based on these activations. Some models of word recognition, notably TRACE (McClelland & Elman, 1986), use a competitive activation algorithm, allowing the activations of word candidates to affect each other. In TRACE, this has the effect of increasing the discriminability of the word candidates, since the stronger candidates (those with a higher activation) inhibit the less active candidates.

Other theories assume that each word is assessed independently of its competitors, with competition occurring as a product of the decision criterion used. The fuzzy logical model of perception (Oden & Massaro, 1978; Massaro, 1987) proposes bottom-up, independent processes of feature evaluation and integration, before a decision is made using the Luce choice rule (Luce, 1959). This rule states that the probability of a response of any particular word is proportional to its activation relative to the

---

[4]See Chapter 4 for a full discussion of both the Shillcock et al (1992) and Norris (1992) work.

summed activation of all word candidates. The Luce rule thus indirectly incorporates competitor effects in the word recognition process by requiring responses to be dependent upon the activations of both the candidate and its competitors.

The early Cohort model (Marslen-Wilson & Welsh, 1978) also employed competitor effects only at the decision stage. This model used a dichotomous activation system, whereby at any point in the processing of a word there was a set of candidates that matched the speech input so far (the cohort), and a second set of rejected, mismatching candidates. The candidates were assessed independently, using both sensory and contextual constraints, and recognition was defined as the point where the cohort was reduced to one candidate.[5]

Theories also differ according to the role bottom-up information plays in the matching process. In this respect, TRACE (McClelland & Elman, 1986) is unusual since it only allows facilitory effects between levels. For example, evidence of an incoming /p/ will activate words such as *play*, *pure* and *positive* but it will not directly inhibit words such as *sad*, which do not contain that phoneme. In Chapter 2, I discuss this issue at some length, using TRACE simulations to examine the consequences of this arrangement.

SEGMENTATION

A necessary component of spoken-word recognition is the process of segmentation. Segmentation of the speech stream allows the discrete phonological forms contained in the mental lexicon to be compared to the continuous information carried in speech waveforms. A number of theories have assumed that segmentation occurs as a natural by-product of word recognition. Marslen-Wilson & Welsh (1978) and Cole & Jakimik (1980) argued that the recognition of each word in an utterance allows the onset of the following word to be identified. This hypothesis was motivated by a number of experiments (Marslen-Wilson & Tyler, 1980; Tyler & Wessels, 1983) showing that words can often be recognised well before their offset. However, Grosjean (1985) and Bard, Shillcock & Altmann (1988) have shown that a significant proportion of words are recognised well after their acoustic offset.

The occurrence of post-offset recognition points does not necessarily contradict the above segmentation hypotheses, but it does indicate that a word-by-word segmentation strategy must have the ability to store and re-analyse unsegmented speech when this occurs. The word-by-word strategy is also unreliable when speech information is lost due to misperception or imperfect auditory conditions. Once the hearer fails to access one word properly, its offset cannot be identified and all following speech will remain unrecognised. In cases like this listeners may require other cues to allow segmentation to continue.

Sub-lexical cues to segmentation. Strategies that do not depend on word-by-word recognition must specify the points in speech at which lexical searches are carried out. Klatt (1979) proposed that lexical searches were carried out automatically every 5 ms, whereas others have used the presence of boundaries between linguistic units as cues to word onsets. McClelland & Elman (1986) suggested that each successive phoneme triggers a new lexical search. Other possible cues are syllable boundaries (Church, 1987), stressed or strong syllables (Grosjean & Gee, 1987; Cutler & Norris, 1988) and phonotactic boundaries (Harrington, Johnson & Cooper, 1987; Lamel & Zue, 1984).

These theories differ in the number of lexical hypotheses they create during the perception of an utterance and in the extent to which these hypotheses must compete. Generally speaking, the more comprehensive strategies (e.g., Klatt, 1979; McClelland & Elman, 1986) place a greater emphasis on competition between the matching lexical candidates, whereas the theories utilising fewer lexical searches are in greater danger of missing word boundaries and thus requiring secondary recovery strategies.

A large proportion of the studies relating to these issues have been statistical or linguistic. Harrington, Watson & Cooper (1989) examined the reliability of phonotactic cues in the detection of word boundaries. In any language only a fraction of the possible permutations of segments are actually realised within words. For example, many words in English contain the sequence /str/ but none

---

[5]In activation terms this corresponds to a value of 1 for the recognised word and 0 for all other words.

contain the sequence /tsr/. Between words these phonotactic constraints do not apply, so the sequence /tsr/ can occur, for example, in the phrase *The cats ran off.* Sequences like the latter are strong cues to the presence of word boundaries. Harrington et al. used a computer analysis of a large corpus of phonologically variant speech to estimate the proportion of word boundaries that could be detected in this way. They found that when the strategy was applied to a narrow (detailed) transcription of slow speech 41% of word boundaries were detectable, but when a broader transcription was used (Huttenlocher & Zue, 1984) only 2% of word boundaries were detectable.

Briscoe (1989) compared the efficiency of the word, segment, syllable and strong syllable strategies by estimating the number of lexical matches each strategy would create when parsing a typical sentence. He found that for a narrow transcription of speech all strategies were reasonably effective, producing a small number of lexical matches. But when the strategies were applied to a broad-class transcription, only the strong syllable strategy maintained a low lexical match count (since the transcription preserved the distinction between strong and weak vowels).

This suggests that a segmentation strategy based on prosodically strong syllables[6] is a robust strategy when faced with noisy or relatively unanalysed speech. There is also a certain amount of experimental evidence supporting this model. Cutler & Norris (1988) used a word detection task to examine the segmentation of nonsense words. Their experiments used monosyllabic word targets forming the onset of two types of bisyllabic nonwords. In the first set, both syllables were prosodically strong (e.g., *mint* in [mnteɪv]). In the second set, the second syllable was weak (e.g., *mint* in [mntəS]). The strong syllable segmentation hypothesis would predict that for the former targets, detection time should be greater since the strong syllable following the target should also trigger a lexical search, meaning that the final segment of the target (/t/ here) can only be recognised as such when the second lexical search has failed. In contrast, the weak syllable in the latter group should not provoke a lexical search and so should not cause mismatch. Cutler and Norris found that these predictions were confirmed: subjects responded to the strong-strong stimuli almost 100 ms more slowly than to the strong-weak stimuli.

Although this result provides support for the role of syllabic units in segmentation, it is not clear that a more simple segmentation strategy cannot explain these data. Cutler & Carter (1987) found that only 27% of content words in English have initially weak syllables. This suggests that the cohort (the set of competitors matching the initial segments of the speech) of a token with an initially weak syllable should be smaller than the cohort for the corresponding strong syllable. Therefore, a segmentation process that conducts a lexical search *whenever* a syllable onset is encountered would predict that a nonword with an initially weak syllable should be rejected as a word more quickly than its strong counterpart due to its smaller cohort size. Thus the Cutler and Norris results could be tapping a relative difference in inhibition for the second syllables in their test words, rather than the dichotomous effect they argue for.

Further evidence relevant to this issue comes from the study of misperceptions. Cutler & Butterfield (1992) collected a number of reported misperceptions in which the word boundaries were misplaced. They found that the majority of boundary insertion errors occurred before strong syllables, whereas the majority of boundary deletion errors occurred before weak syllables. They also found that laboratory induced misperceptions (in which sentences were presented to subjects at sound levels allowing only 50% identification) showed the same pattern of results.

Competition and Segmentation. Competition models of segmentation are generally regarded as an alternative to lower level segmentation strategies, yet there seems no need for this distinction. It is implausible that all word boundaries can be discerned on the basis of low-level judgements or strategies, so it is more relevant to ask how strong the component of competition is rather than whether it exists. The best known model of competition in speech segmentation is the TRACE model of McClelland & Elman (1986; see Chapter 2). In TRACE each phoneme encountered is treated as a possible word-onset and triggers a lexical search. As well as competition between word candidates in order to decide the closest matching candidates in each lexical search, the word candidates must compete with the candidates of other lexical searches, with the level of competition dependent on the amount of phonological overlap between the candidates. For example, given the input *bar tea*, with

---

[6]A prosodically strong syllable is one that contains an unreduced vowel.

no gap between the words, TRACE correctly identifies the two words, but only after a process of competition with other cross-boundary candidates such as *art* and *party*.

Frauenfelder & Peeters (1990, 1992) provided a comprehensive description of the segmentation properties of TRACE. They showed that the outcome of the segmentation process was dependent on a number of factors, including the word length of the candidates, the degree of overlap between them and the degree of match between input and the word candidates. These properties will, for example, cause TRACE to prefer to segment ambiguous speech as one long word rather than a number of short words (e.g., *porcupine* rather than *pork you pine*). Frauenfelder and Peeters point out that their simulations, along with all other simulations involving TRACE, make a number of simplifying assumptions. In particular, the size of the lexicons used in these simulations is rarely more than a few hundred words and it is quite plausible that the same processing principles applied to a realistic set of word candidates would produce quite different results.

Norris (submitted) has argued that a much simplified version of this mechanism can deal with competition of this type. His model, SHORTLIST, assumes that, as in TRACE, each successive phoneme initiates a lexical search. However, only the most active candidates from each search compete at the lexical level for their share of the speech stream. Norris argues that these simplifications allow the model to cope with a more realistically-sized lexicon.

Conclusions. There remains much research to be carried out in the field of human speech segmentation, but some general properties are beginning to emerge. The segmentation strategy with the most experimental support is the metrical segmentation strategy (Cutler & Carter, 1987; Cutler & Norris, 1988). However, this is not a complete theory since it does not specify how lexical candidates compete (Shillcock, 1990, estimates a false alarm rate of 16% using this strategy). Interactive competition networks employed by TRACE (McClelland & Elman, 1986) and SHORTLIST (Norris, submitted) give a plausible account of competition between overlapping word candidates.

Arguing from a purely functional point of view, I believe that there may be an element of all the theories described above in the segmentation of speech. Cross-linguistic studies (Cutler et al., 1986, 1992) indicate that the cues people use to segment speech depend on the properties of the language itself. This suggests that segmentation strategies are to a great extent learned, and develop to take advantage of the statistical properties of the language. It seems likely that such a learned strategy would use the best cues available at each point in time. This may involve combinations of cues, using phonotactic constraints where possible, and word offsets where pre-offset recognition points exist. In Chapter 6, I examine the uptake of cues of this type in a connectionist model of segmentation.

1.5 Thesis Structure

The above sections outline the current state of research into spoken word recognition. In the following chapters I shall return to these issues, providing a more detailed review of selected areas. In Chapter 2, I examine the matching process for isolated words, presenting two simulations of recent experiments that examine the dynamics and mechanics of lexical access. The remainder of this thesis is devoted to the role of phonological variation on the word recognition process, re-examining the issues of word recognition in sentential and utterance contexts.

The structure of this thesis does not adequately reflect the structure of the work it reports: my research has combined both experimental and theoretical computational work and these tasks have been developed in parallel, so it is difficult to depict the chain of thoughts involved in a composition of this nature. As an approximation to this parallel structure, chapters describing experimental work are interleaved with chapters on computational work. In Chapter 3, two cross-modal priming experiments are reported examining the effects of phonological change on lexical access. Chapter 4 introduces a connectionist learning model of speech perception which incorporates the kind of phonological inference implied in Chapter 3. In Chapter 5, I examine a prediction of this model, using phoneme monitoring to examine the perception of phonologically changed segments embedded in both words and nonwords, and in Chapter 6 I examine ways in which the performance of the model can be improved, in particular examining lexical effects in phonological processing and the issue of segmentation in lexical access. Finally, in Chapter 7, I discuss the implications of my research for the model of spoken word recognition.

Chapter 2 — Mismatch and Lack of Match

2.1 Introduction

The previous chapter outlined some of the issues involved in the study of spoken word recognition. In this chapter I shall investigate one of these issues in detail, namely the matching process between sensory input and lexical representations. As I have discussed, two of the leading models of lexical access, TRACE (McClelland & Elman, 1986) and Cohort (Marslen-Wilson, 1987), propose quite similar mechanisms for selection of the most appropriate lexical candidate, when presented with speech input. Both models use the metaphor of parallel activation of multiple candidates, with access to stored knowledge granted only to highly activated candidates. Both models also allow the influence of matching and mismatching information to alter the activations of each candidate. However, the details of the matching process[7] differ between the two accounts and I intend to focus on three aspects of the matching process for which these models make different predictions:

1) The directionality of the matching process

2) The mechanism by which mismatching information affects activations

3) The goodness-of-fit required between input and lexical representations in order to access stored information

I shall review the experimental research relevant to these questions and then present two simulated experiments; the first using the standard TRACE network and the second using a modified network in which inhibitory links are allowed from phoneme to word nodes. These simulations demonstrate that the competitive system TRACE uses to simulate the matching process is unable to accommodate experimental findings demonstrating the immediacy of mismatch effects.

2.1.1 The Cohort Model

In early versions of the cohort model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980), the temporal order of incoming information is crucial to the matching process. The initial stage of lexical access involves the construction of a word-initial cohort, consisting of all lexical items matching the first 100 to 150 ms of the speech input. At this point the lexical information relating to all members of this cohort becomes available. Subsequent input has the effect of eliminating mismatching candidates from the cohort until one candidate remains, at which point recognition is said to have occurred. If all candidates are eliminated from the cohort by mismatching information in the input, the nonword recognition point is said to be reached. Thus, the effects of sensory information can be both facilitory and inhibitory during the construction of the word-initial cohort, but beyond that point only inhibitory effects are allowed. Additionally, once a word-initial cohort is formed, the syntactic and semantic constraints imposed by the word's sentential context can also have the effect of eliminating candidates from the cohort of activated words.

The model was later revised (Marslen-Wilson, 1987) in response to a number of criticisms, mostly concerning the all-or-nothing nature of the matching process. Norris (1982) argued that the inhibitory effects of mismatching information were too strong in the original cohort model. He cited two cases where the failure of the cohort model to recognise words seems counter-intuitive. Firstly, words presented in unlikely sentential context will be eliminated from the cohort by the inhibitory effects of the context and will therefore be unrecognisable. Also, words with an initial mispronunciation such as *shigarette* for *cigarette* will not be recognised, since the mispronunciation

---

[7]The term *matching process* refers to the goodness-of-fit calculation between input and lexical representations. The metric on which this calculation is based — the vector dot product — is equivalent across the models examined in this thesis.

will cause an incorrect word-initial cohort to be built up. Although, as I shall discuss in this chapter, the status of the second case as a drawback of the model is questionable, there is much experimental evidence to back up the intuitive claim that words in unlikely contexts can still be recognised. For example, Marslen-Wilson, Brown & Tyler (1988), using an auditory word monitoring task, compared latencies to nouns in a number of different sentential contexts. They found that even nouns rendered unpredictable by both syntactic and semantic constraints (e.g., *The nurses walk to work each morning. They yawn the BEACH on their way to the hospital*, where BEACH is the target word) were recognised quickly (mean response time 320 ms), albeit significantly more slowly than the same targets in a normal sentential context.

The all-or-nothing nature of activations was also problematic in the modelling of frequency effects. The frequency effect (Howes & Solomon, 1951; Savin, 1963) is a widely studied phenomenon whereby frequently used words are recognised more quickly than less frequent words. The locus of frequency effects in word recognition has been the focus of much debate. Balota & Chumbley (1984) found no frequency effect in a semantic categorisation task, which is assumed to quite directly reflect lexical access. On the other hand Balota & Chumbley (1985) did find a frequency effect in a delayed naming task, when one might expect any initial frequency bias in lexical access to have diminished. These studies question the status of the frequency effect as a reflection of pre-lexical processing. However, it seems likely that frequency effects are at least partly the product of the normal process of lexical access. A series of studies reported by Marslen-Wilson (1990) used cross-modal associative priming to examine the effects of word frequency on lexical access. The experiments compared the priming effects of base words and their cohort competitors (e.g., *street* and *streak*) using the base word as a visual target. The relative frequencies of both base and competitor words affected response times to the target, although the competitor effects were found to be very transient, with the effects minimised by the time the prime word was complete. This transient effect is particularly difficult to accommodate in a model that represents all cohort members as equally active and all outcasts as equally inactive. The experiments seem to be picking up frequency effects early on in lexical access, while selection is still taking place, and so an account proposing a purely post-lexical locus is unlikely.

For these reasons, the cohort model was revised (Marslen-Wilson, 1987) in a number of important ways. The dichotomous nature of the activation of candidates was removed, so that membership of the cohort of matching candidates is dependent not on a complete match between auditory input and the lexical representation but on an overall goodness-of-fit measure. The sensitivity of this measure was increased by assessing the match in terms of phonetic features rather than the larger unit of phonemes. This means that mismatches such as *shigarette* for *cigarette*, which differ according to the phonetic feature system described in Jakobson, Fant & Halle (1952) on the value of just one feature, are less disruptive to the matching process than more fundamental deviations.

The role of context in the model was also altered, with the sentential context of a word playing no part in the goodness-of-fit calculation. This alters the status of the Cohort model, making the matching process completely bottom-up, although the context is still taken into account in the assessment of the *output* of the matching process. These alterations successfully meet the criticisms levelled at the cohort model whilst retaining the primary attributes of the original model, namely parallel activations of multiple lexical candidates with reduction of activation a consequence of incoming mismatching information. The directionality of the model and the inhibitory effect of mismatching information are explicitly stated in the new formulation, but what is lost to some extent is a precise indication of what makes a good match. The revised model concedes that candidates' activations can recover from the inhibitory effects of small mismatches, but does not state what size of mismatch is required to exclude a candidate from the cohort permanently.

2.1.2 The TRACE Model

One of the motivating forces behind the creation of TRACE (McClelland & Elman, 1986) was to provide a computational account of some of the processes the original cohort model introduced. It is therefore unsurprising that there are a great number of similarities between the two models. These similarities have increased since the modifications described above were made to Cohort. However, some differences remain and in this discussion of the TRACE model I shall concentrate on the differences between the two models with respect to issues of lexical access.

TRACE is based on the principles of interactive activation and competition. Each hypothesis the model makes as to the identity of a segment of speech is represented by the activation of a single node, and information is processed by interaction between these nodes, using facilitory and inhibitory links. In TRACE, there are three levels of nodes: the feature level, the phoneme level and the word level. Connections between nodes within a level are inhibitory whilst connections between relevant nodes at different levels are facilitory. The passage of time is represented by parallel duplication of these networks, so that each phoneme of input goes into its own interactive activation network, with word nodes taking their input from a number of these networks, depending on the size of the word.

This architecture captures the Cohort-like time course of processing using activation and competition between nodes. TRACE also manages to resolve the problem of segmentation of the speech input that Cohort faces by allowing multiple word candidates with different starting points to compete for each section of the speech input. This approach to segmentation means that for TRACE, word beginnings do not acquire a special status. Instead of a two-stage process, comprising activation of a word-initial cohort followed by inhibition as mismatching information is encountered, TRACE treats each incoming feature and phoneme in the same way.[8]

However, TRACE retains a strong directionality in the extent to which any phoneme can influence word-level activations. The competitive nature of processing in TRACE means that early information has a stronger effect on word-level activations than late information. To illustrate this point, consider the response of TRACE to the mispronounced tokens /mækɪntot/ and /bækɪntoS/ for /mækɪntoS/ (*Macintosh*). For the token with a word-final mispronunciation, by the time the mispronunciation is processed, the word node for *Macintosh* will already be the most active candidate due to the influence of the previous seven matching phonemes. Because the inhibitory power of a node is proportional to its activation, the distortion will have little effect on the activation of the *Macintosh* node, since all word candidates matching the final phoneme will have low activations. In contrast, the mispronounced word-initial phoneme is processed at a point where all word nodes have equal activations. Thus words beginning with /b/ become dominant and by the time the /k/ is processed, words such as *back* will be dominant. The remaining phonemes will activate the *Macintosh* node but because the *Macintosh* node is also inhibited by its more active neighbours, it should take longer to emerge as the best match to the input. In terms of empirical predictions, this suggests that the recognition of the token with initial mispronunciation should be delayed compared to the word-finally mispronounced token. This argument leads to the conclusion that TRACE does have directionality effects, but these are gradual and continuous and are a natural product of the architecture of the model rather than a discontinuous processing strategy.

The above discussion has touched on another difference between the matching processes of TRACE and Cohort. The effect of mismatching information in Cohort is simple: it directly inhibits active candidates. In the terminology of the original Cohort it removes candidates from the cohort of active words. In TRACE, the effect is less direct, since there are no inhibitory links between its layers. The only way mismatching phonemes can reduce a word node's activation is indirectly, by facilitating a competing word, which then increases its inhibition of the original word. McClelland & Elman (1986, p. 73) argue that prohibiting between-level inhibition protects the model from the criticism of being too unconstrained and they add that they have "no reason to feel that we could improve the performance of our model by allowing either between-level inhibitory interactions or within-level excitation".

The third aspect of the matching process in which differences can be found between TRACE and Cohort is the goodness-of-fit required for lexical access. Again, the original Cohort theory has a very specific formulation in this respect — a perfect match is needed for membership of the cohort, and thus lexical access is prohibited if the phonemic input and the lexical template of a word differ. The revision of Cohort allows for some deviation, with selection depending on the "relative goodness of fit to the sensory input" (Marslen-Wilson, 1987, p. 95). This brings the model into line with TRACE, which also allows small deviations, with word recognition depending on the relative activations of the various candidates using the Luce (1959) choice rule to transform activations into probabilities. Cohort does not define the relative activations of candidates needed to make a word recognition decision and neither model describes fully what the relative values should be to prohibit lexical access,

---

[8]See also Chapter 1 for a discussion of segmentation in TRACE.

leaving the fine-tuning of the models to empirical investigation. But assuming the models use equivalent activation functions in their assessment of candidates, we must return to the effect of mismatch in the models to find a difference between their predictions. As I have argued, the effect of mismatch in TRACE is strongly dependent on the activation of word-competitors and predicts a strong bias towards word-initial input in the goodness-of-fit comparison. In the revised Cohort, mismatching information is always inhibitory and although it is difficult to infer from the model how much deviation is needed to interrupt lexical access, it should be the same amount whatever the position of the deviation within the word.

This discussion has focused on three highly interrelated aspects of the matching process. The surprising result of this discussion is that despite the fact that Cohort proposes a highly directional two-stage process of word-recognition, the effect of deviant information on the matching process is in fact less directional than in TRACE. Cohort employs direct bottom-up inhibition from mismatching featural information, whereas TRACE depends on the facilitation of active competitors to provide inhibition of mismatching candidates — a dependence more likely to be satisfied when the mismatching information occurs early on in the word.

These differences in matching and mismatching effects have been exploited in some of the experiments described below, and are illustrated in the simulations I shall report later in this chapter.

2.2 Studies of Mismatch in Lexical Access

The effect of deviation of phonemic or featural input on speech processing has been studied using a wide variety of tasks. My purpose in this chapter is to examine the effect of match and mismatch on the lexical activations of the relevant lexical entries. For this reason I shall not discuss work using phoneme monitoring, which may only indirectly reflect lexical activations.[9] Here I shall review only the tasks that do attempt to investigate lexical activations, working in roughly chronological order.

2.2.1 Shadowing

The speech shadowing task (Cherry, 1953) has been used to demonstrate many properties of human on-line processing of information. Subjects are required to listen to speech presented in one or both ears and to repeat it as they hear it, with as short a time delay as possible. Marslen-Wilson (1973) reports that subjects can often accomplish this task adequately with delays as short as 200 ms. Marslen-Wilson & Welsh (1978) used shadowing to examine the performance of subjects when presented with sentences with deliberate mispronunciations, as in the *macintot*/*Macintosh* example above. They found that subjects, when repeating the words, would often restore them to their correct form, particularly when the mispronounced words occurred in predictable context, when the mispronunciation occurred late in the word, or when the mispronunciation was phonetically maximal.

This kind of evidence suggests that the lexical entry of a word can be adequately accessed despite the mismatching information contained in the mispronounced segment. One could therefore draw the conclusion that the goodness-of-fit calculation allows for variation in at least one segment without disrupting lexical access. However, there are a number of features of this paradigm that require caution when drawing such a conclusion. Firstly, there is the problem of the acoustic conditions in which experiments of this kind must take place. Because the subject is simultaneously listening and speaking, the words are heard in a noisy (although perhaps not unnaturally noisy) environment. This is likely to cause subjects to increase their tolerance for deviation, in an attempt to compensate for the noisy input. Secondly, the nature of the task, in which subjects are required to respond as quickly as possible, is likely to encourage subjects to work with incomplete information, especially in the case of word-final mispronunciations, where the subject will often be in the middle of producing the word by the time the mispronunciation occurs.

---

[9] See Chapter 5 for a full discussion of this issue.

2.2.2 Mispronunciation Detection

In contrast, the mispronunciation task (Cole, 1973; Cole, Jakimik & Cooper, 1978; Marslen-Wilson & Welsh, 1978) forces the subject to focus on mismatching segments. Subjects are again presented with speech in which a small proportion of segments are mispronounced but here, the task is to respond as soon as a mispronunciation is heard.

The proportion of mispronunciations missed by subjects in this task is generally smaller than for shadowing. For example, Marslen-Wilson & Welsh (1978) found that only 6% of three-feature deviations were missed by subjects in this task, compared to a restoration rate of 24% in the shadowing task. Similarly, Cole & Perfetti (1982) examined subjects' responses to (mostly) single-feature mispronunciations embedded in the context of a children's story. They found that a set of college students detected over 95% of mispronunciations and that detection rates were higher in predictable words. This high detection rate reflects the task difference between mispronunciation detection and shadowing. In mispronunciation monitoring, the subject is specifically directed to listen for mispronunciations and so the small proportion of misses here compared to the proportion of restorations in the shadowing task is hardly surprising.

It is tempting to conclude from the mispronunciation monitoring studies that the matching process does indeed require a perfect or near perfect match between the speech input and the lexical forms. The vast majority of the imperfect matches (the mispronunciations) are detected as such and therefore cannot be effectively activating the lexical entries. However, this argument hinges on the assumptions we make regarding the process underlying the recognition of a mispronunciation. One possibility is that a subject makes a mispronunciation response when a section of the speech stream fails to activate any lexical representation adequately. This would lead to the conclusion drawn above. An equally plausible alternative, however, is that the mispronounced words do successfully access the lexical information and that a post-access comparison between phonemic input and lexical form shows up the mismatch. This leads to the exact opposite of the above conclusion and so we must look to different tasks to resolve this conflict.

2.2.3 Gating

One way to find out whether lexical access has been successful is simply to ask subjects to identify what they heard. This is the basis of a study by Salasoo & Pisoni (1985) which had the primary aim of comparing the relative importance of word-initial and word-final information in lexical access. They used a variant of the gating paradigm (Grosjean, 1980) in which sections of either the onset or the offset of a spoken word were replaced by noise. The target words were presented either embedded in short sentences or in isolation, and the subjects were instructed to identify the words as they were presented. Subjects were presented with successive trials using the same sentence in which the size of the noise window decreased in 50 ms gates from the length of the target word to nothing. The study showed that for both backward (word-onset noise replaced) and forward (word-offset noise replaced) gated words, much less than complete sensory information was needed to correctly identify the target words.

The gating study is strong evidence that the human speech mechanism is able to access lexical information with incomplete sensory information. What is in doubt, however, is whether this ability is part of the normal first-pass processing of speech or whether it is more of a retrieval strategy activated when the primary system fails to parse speech. Two criticisms have been levelled against the gating task as a measure of on-line speech processes. Firstly, there is no time pressure on subjects to respond and so they are able to subject the stimuli to much more than immediate perceptual analysis. This problem is exacerbated by the way the stimuli are often presented in gating studies. In most experiments, subjects are presented with the same sentence or word repeatedly, with more and more of the stimulus presented at each trial. This gives subjects even longer to analyse the stimuli and allows them to use previous responses to influence later judgements. The result is that the use of the gating task in auditory word recognition seems to be the equivalent of using a crossword puzzle to investigate visual word recognition.

Salasoo and Pisoni themselves addressed the second criticism, comparing performance on the usual multiple presentation format to a condition where each subject was presented with a single presentation of each sentence. For the gated words in predictable contexts there was a non-significant

5 ms difference between the identification points for the two different formats. For the words in anomalous context there was a surprising *reduction* in the identification thresholds for the single presentation condition. If anything, the multiple presentation format used in gating may overestimate rather than underestimate the amount of sensory information needed for lexical access.

The effect of the lack of time constraints in the gating task was also addressed in a study by Tyler & Wessels (1985). Their study again used the single presentation format, but as well as the normal gating response, a condition was included in which subjects produced a timed naming response. They found that the mean naming latency was under 500 ms, comparable to many other naming experiments and that the isolation points extracted from the responses were of similar latency and equally affected by the constraints of the sentential context. This study is good evidence in support of the hypothesis that gating studies measure on-line processing of speech. Tyler and Wessels demonstrate that the minimal sensory information needed for lexical access in gating is not an artefact of the extra processing time allowed in normal gating studies. But they do not demonstrate that all gating studies can be treated as measures of on-line processing, particularly when the studies employ unusual or unnatural stimuli, when post-perceptual processing could play a large part in the responses of subjects. The backward-gated stimuli in the Salasoo & Pisoni (1985) study are an example, as the authors admit, of a gating task with unusual stimuli and so some caution must be exercised in interpreting the results as evidence of a tolerant matching process.

2.2.4 Priming

Recently, a number of priming studies have examined the matching process in auditory word recognition. The task involves the presentation of a prime word or sentence, which is quickly followed by a target or probe word, related in some way to the prime. The subject is then required to make a speeded response to the target word, normally in the form of a lexical decision. The rationale behind this task is that the auditory prime should facilitate the recognition of the target word, resulting in a quicker response, or *priming*, when compared to some unrelated control word. For the purposes of investigating lexical access, the amount of priming is assumed to be proportional to the goodness-of-fit between the token of the prime and its lexical representation. If the prime and target are presented in the same modality (either auditory or visual), the priming is described as intra-modal. Cross-modal priming refers to the case where the prime and target differ in modality.

A number of auditory intramodal priming studies have shown priming when the prime and target have surface similarities. For example, Slowiaczek, Nusbaum & Pisoni (1987) showed that prime-target pairs with a single segment in common (e.g., bald-bank) facilitate responses, compared to unrelated controls. However, this phonological priming may occur at a fairly low level, rather than reflecting lexical activations. Slowiaczek and Hamburger (1992) found no facilitory effects of this type when the prime word was presented visually, suggesting that the priming found in intramodal experiments occurred at a pre-lexical, modality-specific level of processing. Similarly, a series of studies by Marslen-Wilson, Tyler, Waksler & Older (1994), using auditory priming of visual targets, showed that even when the phonological overlap between prime and target is considerable (e.g., *tinsel-tin* or *gravy-grave*) recognition of the target word is not facilitated.

These results illustrate the value of cross-modal priming for the examination of lexical activations. The use of different modalities for prime and probe ensures that any effects found are true lexical effects rather than a reflection of lower-level perceptual processes. A further advantage is that the temporal position of the visual target with respect to the presentation of the prime can be manipulated, allowing the time course of lexical activation to be closely monitored (e.g., Zwitserlood, 1989). Like other speeded response tasks, the cross-modal priming task is assumed to be dependent upon the first-pass processing of speech rather than error-correction or higher level processes.

Marslen-Wilson & Zwitserlood (1989) used cross-modal priming to study the effect of word-initial mismatch on lexical access. The experiments, conducted using the Dutch language, compared the priming effects of rhyme primes to the original words (for example, comparing *money*-BEE to *honey*-

BEE).[10] They found that rhyme primes produced little or no facilitation of the lexical decision to the original word's associate, even when the rhyme made up a nonword (as in *noney*-BEE).[11]

A study by Marslen-Wilson, Moss & van Halen (in press) used both intramodal and cross-modal priming to examine the effect of word-initial mismatch. In the cross-modal experiment, they used monosyllabic primes in which the deviation was carefully controlled. They found that even a mismatch of one phonetic feature in the word-initial segment (e.g., *task*-JOB vs *dask*-JOB) suppressed the priming effect of the original word. Their experiment also used prime tokens in which the voice onset time of the word-initial segments was manipulated. This allowed them to create tokens such as *blank/plank* where voicing of the word-initial plosive was perceptually ambiguous between the voiced (/b/) and the voiceless (/p/) forms. They found that even these "half-feature" deviations removed the priming effect when the end-points of the voicing continuum were both real words (as in the *blank/plank* case). Only when the word-initial deviation formed a token half-way between the original word and a nonword (e.g., *dask/task*) was there significant priming for the changed words.

Their intramodal (auditory-auditory priming) study also found a mismatch effect for both single-feature and multiple-feature deviations. However, despite the mismatching effect of the deviations, there was a significant residual priming found of 39 ms, implying that the agreement between input and representation was still good enough to gain access to meaning. The differences between the results of the two experiments were explained with respect to the time course of processing in the two tasks. In the cross-modal study, the visual target was presented at the offset of the prime and the mean reaction time was 515 ms. In the intramodal study, there was a 250 ms inter-stimulus interval (ISI) and the reaction times were longer at 715 ms, reflecting the greater amount of time needed to recognise the auditory target words. This means that responses in the intramodal experiment are half a second slower than the responses in the cross-modal experiment. This may be long enough for the activation of the lexical entry to recover to some extent from the immediate effect of the mismatch.

Other studies have also found residual priming for changed tokens. Connine, Blasko & Titone (1993) also examined the effect of phonologically maximal and minimal phonetic changes using the cross-modal priming task. They found no priming for phonologically maximal word-initial changes, but a robust 22 ms priming for words deviating by a single feature (e.g., "nenny" for *many*). They also found a similar pattern when the change occurred in the middle of the word. These findings conflict with the cross-modal priming results of Marslen-Wilson, Moss & van Halen, but again, the critical factor may be the amount of time available for response. In the Connine et al. study, subjects' response times averaged roughly 650 ms, compared to 515 ms for the Marslen-Wilson et al. study. This suggests that the initial effects of mismatch in word recognition are robust, but that given time these effects are weakened, perhaps using some contextual recovery process.

Finally, a set of experiments by Marslen-Wilson & Gaskell (1992) used the same technique to study the effect of word-final deviations on lexical access. They found that, even for trisyllabic words, a mismatch of just one segment (e.g., *apricod* for *apricot*) effectively blocked lexical access[12].

2.2.5 Summary

The variation found in this review highlights the complexity of the process we are trying to understand. The crucial factor linking these results seems to be the amount of time available for analysis of the speech. Off-line techniques such as the Salasoo & Pisoni (1985) gating study suggest that words can be identified using noise-replaced word-initial or word-final segments. But, using on-line priming studies, we find that the alteration of a single segment at any position in the word

---

[10]In general, I shall represent auditory stimuli with italicised words and visual stimuli with capitalised words.

[11]Marslen-Wilson & Zwitserlood did find some residual facilitation for items with a particularly sparse phonological neighbourhood.

[12]A study by Seybold (1992) did find some residual priming for word-finally deviating tokens. However, it is difficult to compare the studies. Seybold used reaction times to the target presented in isolation in a separate experiment as his control condition whereas all the above experiments used unrelated prime words as controls, intermixed with the test words in each experimental session.

hinders access to meaning. The studies with the fastest reaction times (Marslen-Wilson, Moss & van Halen, in press; Marslen-Wilson & Gaskell, 1992), and hence the most immediate measures of the effects of mismatch, suggest that even a single-feature deviation is enough to halt lexical access.

There appear to be a number of processes at work here. It seems that an extremely good match between the speech stream and lexical representations is necessary to gain immediate access to the stored information on the words. If this initial matching fails, the system is able to recover — given time — either by lowering the standard required for match or simply by granting access to the most active candidate. Interestingly, there is no evidence of the position of the deviation in the word having any effect on the mismatch effect. Minimal changes produce mismatch whether they occur at the beginning, in the middle or at the end of a word. This contradicts the predictions of both Cohort and TRACE that word-initial information is more important in the matching process, but seems to be more compatible with the direct inhibitory processes of mismatch proposed in Cohort.

2.3 TRACE Simulations

Section 2.1 showed that the effects of mismatching input on word level activations in TRACE depends strongly on the position of the deviation within the carrier word. The review of experimental evidence, however, suggests that even at the end of multi-syllabic words, the effects of deviation from the canonical pronunciation of a word are immediate and strong (Marslen-Wilson & Gaskell, 1992). In this section I test TRACE's response to deviations of this nature, presenting two simulations of an experiment from Marslen-Wilson & Gaskell. The first employs the standard TRACE architecture, and fails to simulate the finding that phonetically changed words activate the lexical representation of the base word very poorly, whereas tokens with the final segment removed activate the base word representation very well. This failure is corrected in the second simulation, by the use of direct inhibitory links from phoneme to word levels.

2.3.1 Experimental Data

The experiment reported in Marslen-Wilson & Gaskell (1992; see also Marslen-Wilson, 1993), which does not form part of this thesis, examined the effect of word-final mismatch in bisyllabic and trisyllabic words using cross-modal priming. These are the conditions in which TRACE would predict the least mismatch effect for a single segment deviation, due to the dominance of the base word in terms of word-level activation. Subjects were presented auditorily with a mispronunciation of a base word (e.g., [sosɪn] for *sausage*) and at the offset of this prime word were presented visually with an associate of the base word (e.g., MEAT). The priming effect of the token was measured by comparison with an unrelated prime condition (e.g., *tulip*-MEAT). The experiment manipulated 2 variables for 4 different types of prime word. The within-item variables were:

1) Phonetic Change: Unchanged prime vs mismatching final segment vs control prime. E.g., [sosɪdz] (*sausage*) vs [sosɪn] (*sausin*) vs [tjulɪp] (*tulip*).
2) Fragmentation: Whole prime word vs missing final segment. E.g., [sosɪdz] (*sausage*) vs [sosɪ] (*sausi*).

The 4 different prime types manipulated the lexical status of the distortion, the uniqueness point of the word (the point in the word at which the word's cohort reduces to one) and the number of syllables in the word. The combinations used were:

a) Bi/E/NW Bisyllabic words, early uniqueness point, nonword mismatch (e.g., *sausage/sausin*).
b) Bi/L/NW Bisyllabic words, late uniqueness point, nonword mismatch (e.g., *bandage/bandin*).
c) Bi/L/RW Bisyllabic words, late uniqueness point, real word mismatch (e.g., *cabbage/cabin*).
d) Tri/E/NW Trisyllabic words, early uniqueness point, nonword mismatch (e.g., *apricot/apricod*).

The priming found for each word type is shown in Figure 2.1. Ignoring the differences between word-categories, the general finding was that the fragmented primes (the words with the final segment cut off) produced as much of a priming effect as the base words, but the mismatching tokens produced no

priming at all. There is suggestive evidence that there was residual priming for the mismatching tokens that form a real word but this difference was still not significant.

Thus, it seems that the absence of mismatching information (as measured by the fragmented primes) is perceptually tolerated, but that the presence of mismatching information (as measured by the changed tokens) swiftly and effectively inhibits the lexical representation of the base word.

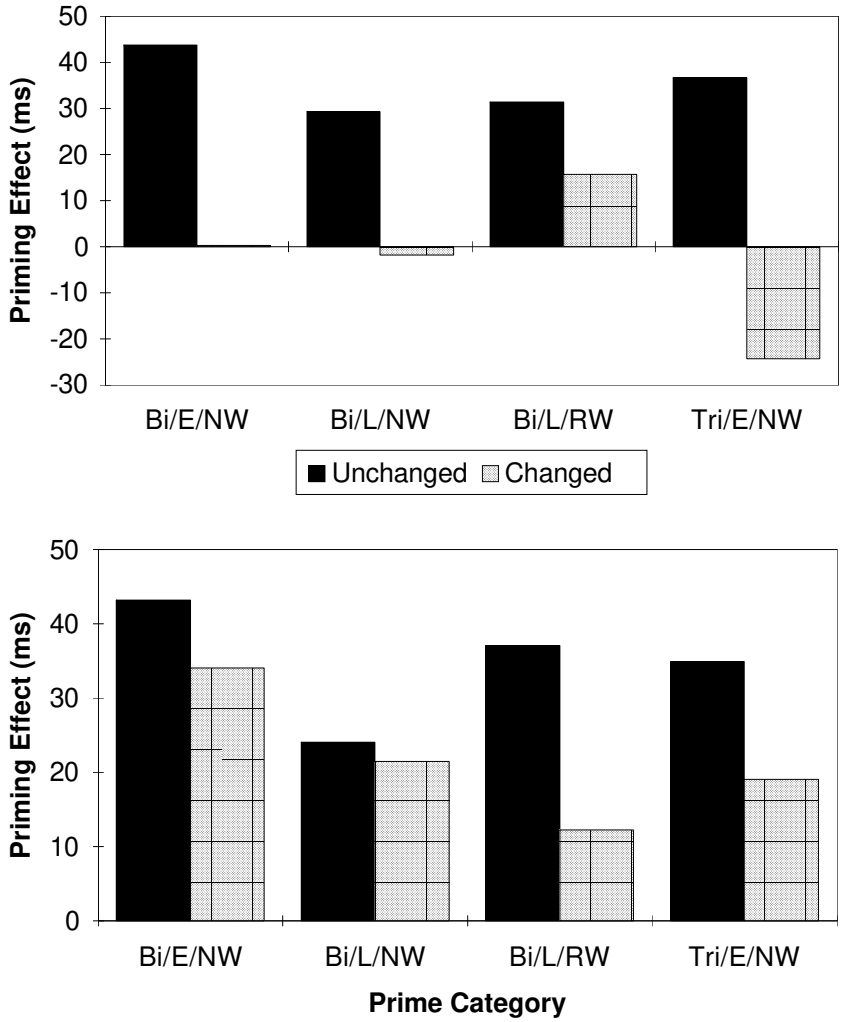

Figure 2.1. Experimental results for Marslen-Wilson & Gaskell. The upper graph shows the unfragmented conditions, the lower graph shows the results for the fragmented conditions. The categories represent a) to d) above and the values plotted are difference scores between test and control: the higher the value, the better the priming

2.3.2 TRACE Simulation 1

MATERIALS

A number of simplifying assumptions were necessary in order to simulate the experimental results using TRACE. The reduced set of phonemes defined in TRACE precluded the transcription of the original experimental materials for use in the simulations. Instead, a set of new words was created and added to TRACE's standard lexicon for use in the simulations. Bisyllabic words were represented by strings of 5 phonemes with CVCVC structure and trisyllabic words were given a CVCVCVC structure.

In the original experiment, early uniqueness point words diverged from their word-initial cohort at the third or fourth segment, whereas the late uniqueness point words diverged on the final segment. The competitor environment used in these simulations is vital to the outcome, but because of processing time limitations, the use of a realistically sized lexicon was impractical. Instead, the lexicon used in McClelland & Elman (1986) was used, with extra words added to create an appropriate cohort competitor environment for the words. The original lexicon consisted of 211 frequent uninflected words. To this, four competitors were added for each test word. For the early uniqueness point words, these diverged on the third phoneme. For the late uniqueness point words, two competitors diverged at the penultimate phoneme and two on the final phoneme. After the divergence point, the phonemes were randomly chosen within the CVC syllabic structure. This environment is likely to overestimate the level of competition among the test words' close competitors, giving TRACE a better chance to produce mismatch effects, and should to some extent compensate for the unrealistic overall size of the lexicon. Example stimuli with their close competitor environment are given in Table 2.1 (see Appendix D for a full list of stimuli).

Table 2.1. Example stimuli used in the simulation, with their close competitor environment. The words are presented using the ASCII phonetic transcription of TRACE. Bracketed stimuli are not included in the TRACE lexicon (i.e. they are nonwords).

Category	Bi/E/NW	Bi/L/NW	Bi/L/RW	Tri/E/NW
Unchanged	bat^S	puriS	giruS	r^sakiS
Changed	(bat^t)	(purit)	girut	(r^sakit)
Fragment	(bat^)	(puri)	(giru)	(r^saki)
Competitor environment	bagis	purip	girut	r^g^d
	bapub	purik	girul	r^pub
	badal	pur^d	girak	r^dal
	bakar	puras	gir^s	r^kar

In the Marslen-Wilson and Gaskell experiment, the mismatching phonemes were not controlled because earlier experiments on monosyllabic words had found that the featural size of the distortion had no effect on priming values. In TRACE, however, the identity of the phonemes used to create the deviations is critical. The simulation involved three item triplets for each category (i.e. original unchanged word, phonetically changed word and fragment). Between triplets, the featural correlation between the original and changed final segments was varied (using the feature values defined in Jakobson, Fant & Halle, 1952). One triplet employed high correlation word-final consonants for the original and mismatch (r/l; Pearson r = 0.80), another used medium correlation consonants (b/k; r = 0.46) and the third used low correlation consonants (s/t; r < 0.2).

TRACE does not model priming directly, but it seems plausible that priming reflects the activation level of the base word. The relationship assumed here was a linear one. That is, $A = kP + c$, where P is the experimentally observed priming effect, A is the word level activation of the base word in TRACE, and k and c are constants used in the linear transformation. The constants were chosen so that the means and standard deviations of the experimental and theoretical priming effects were equated.

The activation values were recorded after 90 processing cycles. This is rather a late testing point, since the last phonemes of the words were input at cycle 36 for the bisyllables and 48 for the trisyllables. This test point was chosen as it was a point when the activations had generally settled, and because it gave TRACE the best chance to produce a mismatch effect for the distorted phoneme strings.

DESIGN

The design of the simulation mimicked that of the original experiment in all respects except that instead of two conditions — fragmented original word and fragmented changed word — there was one fragmented condition. The two fragmented conditions were required in the original experiment to control for coarticulatory effects of the mismatching consonant on the preceding segments, but in TRACE this is not necessary. Therefore, there were 4 word categories as in a) to d) above which were presented either unchanged, with word-final change or fragmented (without the final phoneme). The activation of the base word (summed over adjacent time slots) was taken to be the dependent variable.

PROCEDURE

The test stimuli for each condition were presented as input to the TRACE II computer programme using the standard weighting parameters (see Table 2.2) and a lexicon consisting of 211 uninflected words with a frequency of 20 or more in the Kucera & Francis (1967) word count, plus the phoneme strings described above, designed to provide a more realistic close competitor environment. Each stimulus was presented to TRACE as a silence "phoneme" at cycle 6, followed by the other phonemes of the word, one by one, every six cycles. At each cycle, the activation of the target (i.e. original) word was recorded. Ninety cycles of processing in all were carried out for each test word and then the network was reset. The simulations were carried out using a VAXStation 3200.

RESULTS

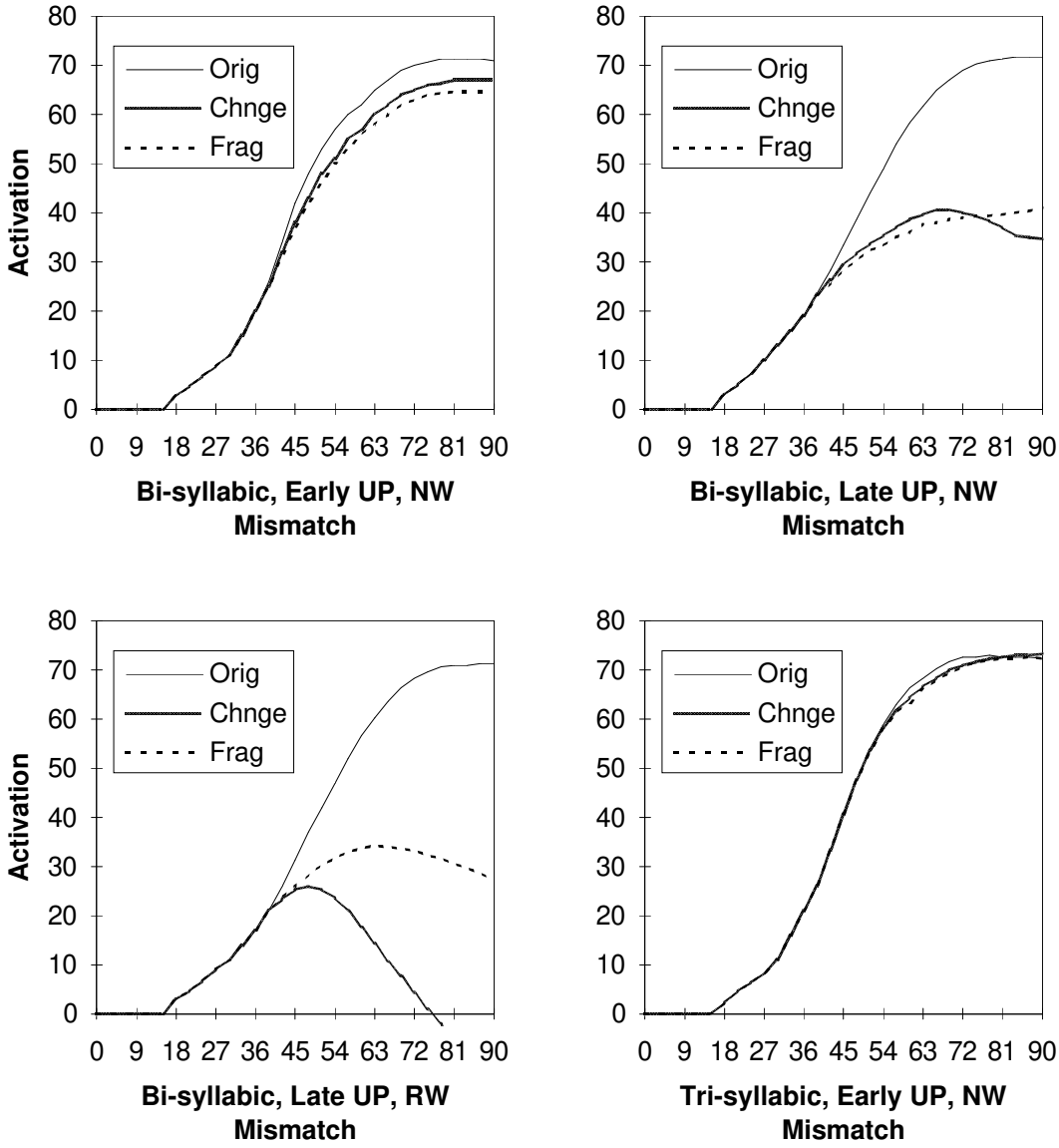

Figure 2.2. Activation curves for the TRACE simulation. The x axes represent the number of cycles of processing. Each graph shows the results for one prime category. The three curves correspond to the three types of input: Original (unchanged), Changed and Fragmented.

The mean activations for each prime type are displayed in Figure 2.2. The endpoints of the activation curves were taken to be the predictions of the priming effects and were tested for correlation with the experimental facilitation times. The fragmented values in the experiment (fragmented unchanged and fragmented changed) were collapsed to one value for comparison, controlling for residual coarticulatory information relating to the absent word-final consonant. The correlation between experimental and predicted results was not significant (Pearson $r = 0.14$, $p > 0.1$).

Comparison of the results for the four different categories in Figure 2.2 illustrates the problems TRACE faces when modelling these data. The experimental results dictate that in all cases the changed tokens should strongly reduce the activation of the underlying word. Yet in the early uniqueness point conditions (the top left and bottom right panels of Figure 2.2) this does not occur. This is because, in these cases, no lexical competitors remain active enough to inhibit the base word

effectively. Thus, the presence of mismatch in these cases is treated in much the same way as the absence of match (the fragmented primes).

For comparison, the activation end-points were subject to a linear transformation, as described above, to produce predictions for priming and are illustrated in Figure 2.3 (grey bars) with the corresponding experimental priming effects (black bars). The simulated priming effects for the original (unchanged) words agree with the experimental results. However, for the changed words the simulated priming values in the early separation categories (Bi/E/NW and Tri/E/NW) are much higher than the experimental findings, whereas for the fragmented words the simulated priming values in the late separation categories (Bi/L/NW and Bi/L/RW) are much lower than the experimental values.

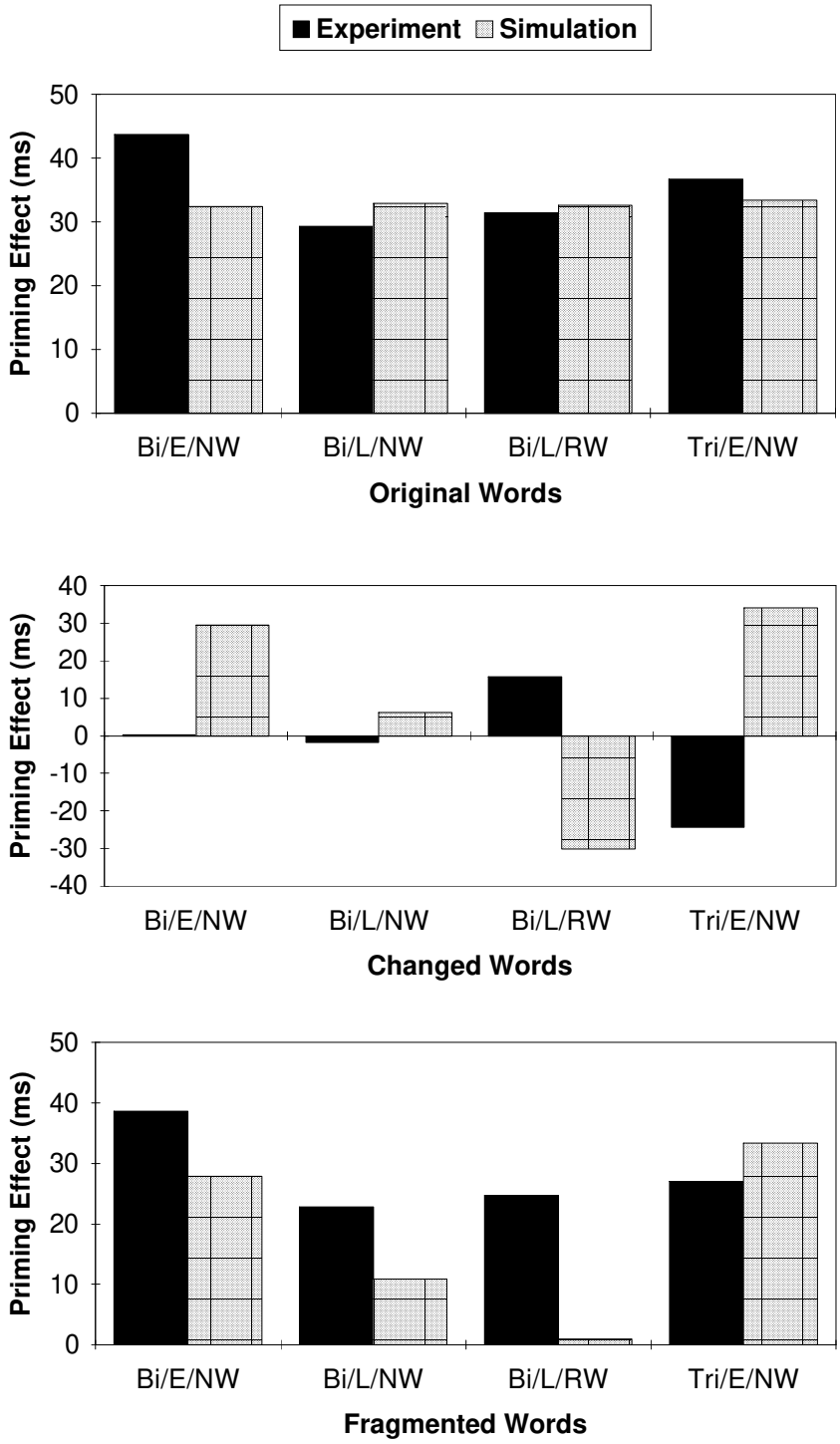

Figure 2.3. Comparison between experimental results and transformed TRACE predictions, shown separately for each of the three prime types. The upper panel refers to the responses to the unchanged words, the middle panel illustrates the responses to the phonologically changed words and the bottom panel shows responses to the fragmented words.

DISCUSSION

The TRACE simulation revealed no significant correlation with the experimental results. Here I examine the reasons for this failure, focusing on the distinction between mismatching information and missing information.

The graphs in Figure 2.2 show that almost all the strings presented to TRACE strongly activated the base word representations. The changed strings reduced activations but in most conditions by no more than the fragmented strings. In some cases, changed strings were *more* effective primes than fragments in the simulation, because of the activation of phonemes with similar features matching the original word. This contrasts strongly with the experimental results where fragmented tokens, particularly the fragmented base words, had a strong priming effect but the distorted tokens produced no facilitation.

In one condition, TRACE did manage to produce a strong mismatch effect for the changed string. This was when there was an equally active competitor up until the final phoneme, which was then facilitated by the final phoneme of the distorted string and was thus able to inhibit the base word laterally. Ironically, this is the condition in the experiment for which the mismatch effect of the changed tokens was weakest.

The two conditions in which the uniqueness point occurred early in the word provide the strongest evidence that the way TRACE deals with mismatch is not sufficient. In these conditions, the closest competitors are all inhibited early on in the processing of the word. By the time mismatch is encountered, the competitors are all minimally activated and the mechanism by which inhibition can occur is effectively blocked. The base word therefore continues its dominance throughout the remaining cycles.

These results are fairly robust to alteration of the parameters controlling TRACE's processing. The relative influence of the different connection types can be varied by changing the weighting factors in TRACE. The only factor relevant here is the *gamma* weighting, controlling within word-level inhibitions. Increasing this allows lateral inhibition to play a stronger part in the calculation of activations, but since in the early separation conditions the base word quickly becomes dominant at the word level, the effect of increasing the value of this parameter speeds the inhibition of the competitors which would otherwise be able to inhibit the base word. Conversely, reduction of the value of this parameter would allow competitors to remain active for longer but would reduce their ability to inhibit the base word.

Another possibility is that this lack of mismatch is simply due to the choice of function used to transform activation values into priming effects. It is conceivable that if we choose a non-linear transformation, the small activation differences could be amplified into large priming differences, thus modelling the results more closely. However, comparison of the fragmented and changed conditions reveals that this would not work. Any transformation which increased the mismatch effect of the changed condition would equally increase the mismatch effect of the fragmented condition. But the experimental results show that the fragmented words prime almost as strongly as the base words themselves. For a model of lexical access to accommodate these results it must be able to differentiate between lack of match (as in the word fragments) and mismatch (as in the changed tokens). In the cases examined here, this seems to require bottom-up inhibition.

2.3.3 TRACE Simulation 2

The second simulation was a direct replication of the first but with a modified version of the TRACE network, allowing bottom-up inhibition from phonemes to words. The algorithm at the heart of TRACE was essentially unchanged. However, instead of just facilitory links between phonemes and words, each phoneme was linked to each word, with facilitory links as before but with inhibitory links between each phoneme and all words not containing that phoneme. This alteration was intended to allow TRACE to accommodate the experimental results on the effects of mismatch.

Because of the architectural changes in this version of TRACE, a number of the weighting parameters had to be altered. The values used in the simulation are listed in Table 2.2, along with the values used in Simulation 1 (the default values). One important difference is that because of the large additional source of inhibition in the new model, the resting level for words, to which all word activations

gravitate over time, was raised from 0 to 0.5. This was raised to give word activations an initial boost to allow them to rise to similar levels to those found in Simulation 1, but it was not the best solution since it meant that the activation of inhibited words, if left long enough, would start to rise again. Ideally the initial activation should be separated from the resting level to stop this occurring, but to simplify matters this was not implemented.

Table 2.2 Parameters used in TRACE simulations

Parameter	Value	
	Simulation 1	Simulation 2
Feature - phoneme excitation	0.02	0.02
Phoneme - word excitation	0.05	0.05
Word - phoneme excitation	0.03	0.00
Phoneme - feature excitation	0.00	0.00
Feature-level inhibition	0.04	0.04
Phoneme-level inhibition	0.04	0.25
Word-level inhibition	0.03	0.00
Phoneme - word inhibition	N/A	0.25
Feature-level decay	0.01	0.01
Phoneme-level decay	0.03	0.03
Word-level decay	0.05	0.50
Resting level	0.00	0.50

Other changes were more minor. The between-word and word-to-phoneme connections were switched off because they were not needed. The within feature-level inhibition was increased and the decay rates for features and phonemes were increased slightly. These values were largely decided by trial and error, and little attempt was made to optimise the performance of the network. One interesting by-product of these changes was that the removal of facilitory links from words to phonemes meant that the modified TRACE used only bottom-up and within-level links, removing the top-down element that has been the object of much debate (see Elman & McClelland, 1988; Norris, 1992, 1993).

Details of Materials, Design and Procedure were the same as for Simulation 1.

RESULTS

The end-points of the activation curves were compared to the experimental priming values using the Pearson correlation coefficient. This time a strong correlation was found ($r = 0.81$) which was highly significant ($p < 0.01$). The data were again transformed into priming values for comparison and are illustrated in Figure 2.4.

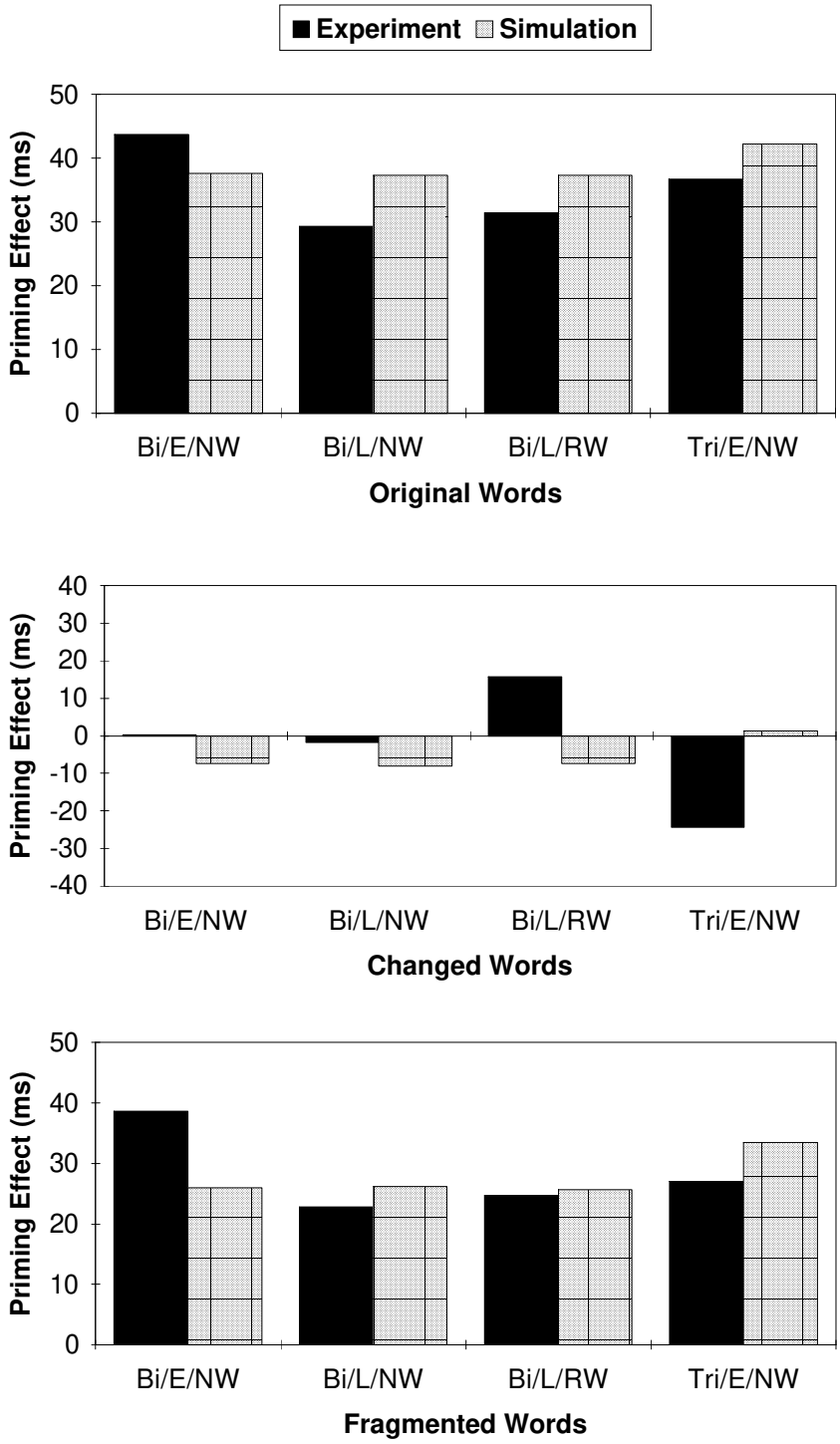

Figure 2.4. Comparison between experimental results and revised TRACE predictions, shown separately for each of the three prime types. The upper panel refers to the responses to the unchanged words, the middle panel illustrates the responses to the phonologically changed words and the bottom panel shows responses to the fragmented words.

DISCUSSION

The results of Simulation 2 agree with the basic findings of the experiment. As before, fragmented phoneme strings activate the lexical representation almost as well as undistorted strings. However, in

this simulation the mismatching phonemes reduce the base word activation strongly, as predicted by the experimental results. Only two of the 12 conditions showed a discrepancy between the experimental results and the TRACE simulation. These were both for the changed tokens, for the Bi/L/RW and Tri/E/NW categories. These deviations were not significant in the experimental analysis, and the discrepancies between the model's predictions and the experimental findings in these cases may be due to the greater variation found in human experiments in comparison to computer simulations.

The similarity between the activation patterns for the different categories of words reflects the removal of within word-level inhibitory links, meaning that competitor environment no longer has an effect on the matching process. This again agrees with the findings of the experimental study.

The success of the revised model in modelling the experimental results should be interpreted with some caution. I have shown that a TRACE-type network with bottom-up inhibitory nodes produces a plausible model of the matching process in lexical access. However, this is certainly not the only way these data can be modelled. TRACE has been used to model a wide variety of different psychological data, with an impressive degree of success. It is likely that the revisions proposed here limit the model's ability to deal convincingly with other aspects of lexical access. In particular, the removal of word-level inhibitory nodes is likely to remove the capacity of TRACE to segment the speech stream as word recognition takes place. The objective of these two simulations was not to come up with a rival interactive activation model of auditory lexical access. It was simply to point out a deficiency in the original model's matching process and to demonstrate how this deficiency may be resolved. The addition of inhibitory links between phonemes and words also increases the complexity of the model, and in subsequent chapters I shall argue that the architecture TRACE (even in this modified form) is inadequate for the modelling of new data relating to the perception of phonological variation.

2.4 Conclusions

The aim in this chapter was to review the evidence from previous research on the matching process in auditory lexical access. The most reliable evidence we have comes from a number of studies using cross-modal priming to tap activation levels. This evidence suggests that the matching process needs a good fit to activate lexical representations, with all studies finding no priming for tokens deviating by more than a few phonetic features in one segment. Even a single feature deviation inhibits lexical access, although there is some evidence that a mismatch of this size does not *prohibit* lexical access. The discriminating factor seems to be the time interval between distortion and response — the shorter the response latency, the stronger the mismatch effect.

Somewhat surprisingly, this mismatch effect is not dependent on the position of the deviation within a word. Studies using word-medial (Connine, Blasko & Titone, 1993) and word-final (Marslen-Wilson & Gaskell, 1992) deviations find no differences between the mismatch effect in these positions and the mismatch effect for word-initial deviations.

The results have been interpreted using the concepts of parallel activation of multiple hypotheses common to models such as TRACE and Cohort. Simulation of the effect of word-final mismatch show that the mechanism used in TRACE to discriminate between competing hypotheses, namely lateral inhibition between active candidates, is untenable. However, a variant of TRACE, which retains the general framework but has added inhibitory connections from phoneme to word level, is able to model the experimental data closely.

One of the appealing features of TRACE was its symmetry of information flow. The modifications suggested here break this symmetry unless we also allow direct inhibitory links from word to phoneme levels. However, a study by Frauenfelder, Segui & Dijkstra (1990), carried out in French, suggests that even the indirect top-down inhibitory effects predicted by the standard TRACE model are not present. They looked for evidence of these effects in a phoneme monitoring task, by comparing monitoring latencies for segments in 'inhibiting' nonwords such as /t/ in *vocabutaire* to control nonwords such as *socabutaire*. Although it has no direct inhibitory links, TRACE predicts that the /t/ detection time should be longer in the former case since there is inhibition from the /l/ phoneme node which is activated by the *vocabulaire* word node. They found no difference between the two cases and concluded that there must be little or no top-down inhibition between word and phoneme nodes.

The model that emerges from these revisions has a strictly bottom-up flow of information form sensory levels to lexical access. Both the intolerance to variation and the lack of top-down influence in lexical access suggest that the perception of words in humans cannot be deceived. Mispronunciations and other irregular deviations do not gain immediate access to higher level information. It seems likely that lexical access can occur with these tokens after some delay, but these tokens will still be perceived as mispronounced versions of the closest word.

The view of the lexical matching process as intolerant to deviation is not without its problems. The primary problem is that speech is variable. At the levels of representation psychologists are concerned with, there are numerous sources of variation in the form of phonetic and phonological changes. The remainder of this thesis is devoted to the effects of these regular changes, concentrating on the mechanisms by which phonological variation is dealt with in lexical access.

Chapter 3 — Mismatch and Phonological Variation

3.1 Introduction

The review of research in the previous chapter led to the conclusion that the matching process between auditory input and the lexicon allows little room for variation. Word meanings are only accessed effectively when there is a close fit between a token of speech and the canonical pronunciation of a word. However, speech is an extremely variable medium and so on face value, these results seem to imply that humans have difficulty understanding much of what is said to them. This seems a rather pessimistic conclusion, and from introspection is almost certainly wrong, so the challenge is to explain how our perceptual mechanisms overcome the problem of variation in speech.

Firstly, it is important to note that the intolerance of variation found in earlier experiments (e.g., Connine, Blasko & Titone, 1993; Marslen-Wilson, Moss & van Halen, in press; Marslen-Wilson & Gaskell, 1992) is at a specific level in the perceptual process. The deviations used in these experiments were at the phonetic and phonemic levels of analysis. It is with this kind of variation I shall be concerned here, rather than variation generally assumed to be at a lower level in the perceptual process, such as voice quality, pitch and speech rate.

Phonological change is rife in normal conversational speech. Consider this example from Hawkins (1984).

[vjəfawndzəkiz]

This is a transcription of the phrase "Have you found your keys?" as it may be spoken in normal conversation. The canonical transcription, however, is:

[hævjufawndjokiz]

Phonological processes such as segment deletion, assimilation and vowel reduction are applied to the canonical or underlying representation to produce a very different surface structure (the actual utterance). How can we understand speech like this when the lexical access system makes so little allowance for error?

The vital difference between this type of variation and the variation used in studies of mismatch is that the latter is random, unnatural variation, using mispronunciations or noise, whereas the former is regular, rule-bound variation. In this chapter I shall present two experiments examining the effect of phonological variation on lexical access. I shall argue that the input lexicon *is* intolerant to variation, but that this intolerance causes no problems in the perception of normal connected speech.

3.2 Phonological Theory

This section is not intended to describe the massive developments in phonological theory over the past thirty years. That would be decidedly beyond the scope of this thesis. Instead I shall summarise enough of the major theories to provide an adequate background to the experiments I have carried out.

Phonology is the study of the speech sounds and the way they interact. If we think of linguistic structure as a hierarchy, this puts phonology just above phonetics in the study of speech, and below morphology, syntax and semantics. The aim of phonology is to provide a description of the structure of speech which allows the various processes that occur within and between languages to be simply and coherently explained. Jakobson, Fant & Halle (1952) and Chomsky & Halle (1968) proposed that segments of speech are constructed by the combination of a number of binary features. These features were assumed to be universal throughout language and to represent either perceptual or articulatory differences. The identity of a segment would thus depend on its position in this phonetic feature space.

This type of formalism, presented by Chomsky & Halle in their book *The Sound Pattern of English* (SPE; 1968), allowed phonological changes to be described simply, in the form of phonological rules relating sets of phonetic features. For example, the nasalisation of vowels in English can be described using the rule:

$$\begin{bmatrix} +\text{vocalic} \\ -\text{consonantal} \end{bmatrix} \rightarrow [+\text{nasal}] / -[+\text{nasal}]$$

This means that a segment with the feature values [+vocalic] and [-consonantal] will be nasalised when followed by a nasal segment. The explanatory power of a feature set can thus be measured by the simplicity of its rules. The above rule defines the subset of segments it operates on by extracting all segments with the specified values. These form a *natural class* of segments and help to validate the descriptive system. If the rule had required exceptions or additions to form the feature sets used, the validity of the system would have to be questioned.

The formation of natural classes in descriptions of phonological processes is the most attractive element of classical generative phonology. However, as a theory, it is still relatively unconstrained. Natural classes can be chosen at random, but phonological rules only apply to some classes with certain values. The theory needs extra constraints in order to increase its explanatory power.

Chomsky and Halle used the notion of markedness to add explanatory power to generative phonology. This proposed that all binary features have a default (unmarked) value and a marked value, and that only the marked value is present in a lexical specification. They used this asymmetry to explain the tendency for certain feature values to occur together (see Kaye, 1989, for a fuller discussion).

More recently a number of theories have been developed that add more structure to the phonetic feature space used in generative phonology. Autosegmental phonology (Goldsmith, 1976) takes the SPE features of Chomsky and Halle and assigns them to a hierarchical skeleton. Phonological processes are, according to this framework, represented as deletions and re-associations within this skeleton. Autosegmental phonology is an attractive theory, since it abandons the almost certainly spurious notion that words consist of one-dimensional strings of phonemes or segments. Indeed, the phoneme as a unit of speech has no place in autosegmental theory. Segments of speech, being less rigidly defined structures of features, are normally preferred.[13] However, psychologists have been slow to follow up the predictions made by this theory, and only recently has there been experimental work carried out using non-linear phonology as its basis.

Closely linked to the development of more structured representations has been the development of less specified representations. Markedness is one way in which a lexical representation can be reduced, by only specifying the non-default values of a feature. Radical underspecification (Kiparsky, 1982; Archangeli, 1988) goes one step further. As well as eliminating unmarked features from the lexicon, marked features are also left unspecified if they can instead be predicted using context-sensitive redundancy rules. This results in an abstract, minimal and invariant set of features defining any lexical entry.

For example, in the representation of the place of articulation of a segment, one could assume that a coronal place is the default (unmarked) place (cf. Paradis & Prunet, 1991). Thus the lexical representation of segments whose place of articulation is coronal (as in /d/, /t/ and /n/) will not be specified for this feature. A default rule is then used to give unspecified segments their normal place of articulation:

$$[\ ] \quad \rightarrow \quad [+\text{coronal}]$$

But in addition, context sensitive assimilation rules can give the same segments a labial or velar place.

$$[\ ] \quad \rightarrow \quad [+\text{velar}] \ /\!- \ \#[+\text{velar}]$$

---

[13]Here I shall use the term segment to describe these units, except when discussing theories or models in which phonemes are explicitly stated.

[ ] → [+labial] /– #[+labial]

These rules state that the unspecified segment can gain a labial or velar gesture provided the following phonological context is appropriate.

Radical underspecification is most readily interpreted in the context of speech production, as a hypothesis about the mapping from lexical entry to surface form (Keating, 1988). For this theory to have value in human lexical access, we must assume that the input lexicon is specified in this way, and that just as phonological variation is linked to unspecified features in production, tolerance to variation must depend on whether the variant feature is specified in the lexicon. I shall develop this argument in the following sections, using specific examples of phonological variation.

3.3 Natural Variation in Speech

Phonological variation is often described in terms of a mapping between surface and underlying forms of speech. The surface form of a word is a representation of the way the word is produced in normal speech and is therefore variable, depending on its structure and environment. The underlying form is assumed to be the invariant core of the word, and as such is heavily dependent on phonological theory, as discussed above. For example, the word *cat* can be represented as /kæt/ underlyingly but may be produced as the surface variant [kæp][14] in certain situations.

3.3.1 Allophonic Variation

Types of phonological variation can be classified according to the level of description at which the changes occur. Allophonic variation occurs at the phonetic level and is non-distinctive, in that the product of the change is still perceived as the same segment. For example, the segment /p/ can be produced as one of two allophones depending on its context. In *pin*, the /p/ occurs as an aspirated [p^h], whereas the /p/ in *spin*, occurs in its unaspirated form, [p]. In English this change is non-contrastive, so both forms are perceived as the same segment; indeed most English speakers find it impossible to discriminate between the two forms even when the difference is pointed out. In other languages aspiration is contrastive, for example, distinguishing the Thai words [p^haa] (forest) and [paa] (to split).

This kind of variation is often thought of as noise, a complicating factor in the recognition of segment. Indeed, if segments are to be recognised in isolation, this must be true. But as Church (1987) points out, allophonic variation provides useful information as to the identity of surrounding segments. This is because allophonic variation conforms to rules rather than fluctuating freely, and because the product of the variation is unambiguous. Church showed that allophonic variation could be used by a chart parser in the segmentation of phonetic transcriptions into syllables and other higher level structures.

3.3.2 Phonemic Variation

Other types of variation alter the phonemic structure of words or phrases, and often result in surface ambiguity. Some forms of these changes are optional, depending on speech rate and/or care of the speaker. At other times, the changes are compulsory. The major forms of variation are outlined below.

1) Deletion: segments are often missing from words when produced in connected speech. For example, in the phrase *send the letter off*, the word-final /d/ can be deleted.

2) Assimilation: this involves the alteration of a segment in certain contexts to become more like its neighbouring segments. Place assimilation, for example, causes the /t/ in 'sweet girl' to be realised as a [k]. Assimilation is similar to types of allophonic variation discussed above, but the product here is a different segment.

---

[14] I use slash marks here to denote underlying forms and square brackets to represent surface variants.

3) Reduction: vowels in connected speech are often reduced to the schwa segment. For example, the *a* in *I ran into a shop* will often be reduced from /æ/ to [ə].

4) Epenthesis: segments will sometimes be inserted between or within words, probably to ease articulation from the speaker's point of view, or segmentation from the perceiver's point of view. For example, an /r/ will be inserted between the words *vodka* and *and* in connected speech.

Again, like allophonic variation, these types of variation can be useful in disambiguating the structure of a sentence. But unlike allophones, these variants do not have a one-to-one mapping onto the underlying representations. In terms of speech perception, this results in ambiguity, which must be resolved by using some or all of the potential phonological, morphemic, semantic, syntactic and even pragmatic constraints on the underlying hypotheses.

PLACE ASSIMILATION

The phonological change examined in detail in this thesis is the assimilation of place of articulation in English. This conforms to the phonological rule:

$$\begin{bmatrix}\text{coronal}\\ \text{plosive}\end{bmatrix} \rightarrow \begin{bmatrix}\alpha\ \text{locus}\\ \text{plosive}\end{bmatrix} / - \begin{bmatrix}\alpha\ \text{locus}\\ \text{plosive}\end{bmatrix}$$

$$(\alpha\ \text{locus} = \text{labial or velar})$$

This rule states that coronal segments, such as /d/, /t/ and /n/ can gain the place of the following segment when the following segment has either labial place (e.g., /b/, /p/ or /m/) or velar place (e.g., /g/, /k/ or /N/). In most cases these changes occur between words (e.g., [swikkɪd]; underlyingly *sweet kid*), with the place of articulation of the word-initial segment migrating across the word boundary. However, word-internal assimilation is also possible, normally occurring at morpheme boundaries in words such as *inconsistent* ([ɪNkənsɪstənt]) and *notebook* ([noωpbωk]).[15]

In fact, this depiction of assimilation is over-simplistic. Firstly, in the assimilation of stop consonants, the assimilated segment will not contain a burst, due to the influence of the following burst. A more realistic representation of the above assimilation is [swik°kɪd] (Ladefoged, 1982).

Secondly, place assimilation is a more graded process than is suggested by the above description. Articulatory analyses of place assimilation (Barry, 1985; Kerswill, 1985; Nolan, 1992) show that place assimilated segments can contain residual coronal information. For example, Nolan (1992) showed that for coronal to velar assimilations (e.g., t → k) the resultant segment contains varying degrees of both coronal and velar place. Indeed, even when there is no residual acoustic evidence of coronal place a hidden articulatory gesture may remain. (Browman & Goldstein, 1990). In fact, Browman & Goldstein (1990) argue that these results support an account of place assimilation in which the degree of assimilation of a segment depends on the temporal overlap between the two place gestures. By this account, complete assimilation is just an extreme case, where two gestures overlap to such an extent that one becomes acoustically hidden by the other.

Hayes (1992), however, argues that the facts of assimilation can also be accommodated by a model of assimilation incorporating both a phonetic coarticulatory element as well as a true phonological change. This standpoint is supported by the work of Holst and Nolan (in press), showing that full assimilations within a *single* articulator (e.g., the movement of the tongue tip in /s/ to /S/ assimilation) show no evidence of any blending or overlap as would be predicted by a gestural overlap account of assimilation.

Place assimilation is a form of neutralising variation, in that it neutralises the phonemic contrasts between segments, creating ambiguity as illustrated in the examples below, which all contain a potentially assimilated word-final [k]:

1) It was a black cat

---

[15]In some cases, the assimilation passes into the orthography of the word, as in words such as *important* and *impractical*.

2) She's such a sweek girl

3) He's such a weak boy

4) She's such a weak girl

5) They took a lake cruise

In all cases, the listener must resolve the ambiguity, either opting for an assimilated underlying /t/ or a regular underlying /k/. In sentences 1) and 2), the lexical status of the two hypotheses can be used to disambiguate. In 1), *blat* is not a word, so *black* must be the correct parse, whereas for 2), *sweek* is not a word, so the surface [k] must be part of an assimilated form of the word *sweet*. In sentences 3) to 5), both underlying alternatives are real words, so other information is needed to resolve the conflict. In 3), the following context of the [k] violates the above assimilation rule, suggesting that the surface [k] is in fact an underlying /k/. In 4), the syntactic and pragmatic constraints suggest that *weak* is the more likely hypothesis than *wheat*. Finally the [k] in sentence 5), is ambiguous at all levels, since *They took a late cruise* and *They took a lake cruise* are both plausible sentences.

The process of place assimilation produces featural changes in words. Apart from the problems of disambiguation discussed above, this is of particular relevance to the issues highlighted in the previous chapter. Place assimilation produces exactly the kind of alternation which was found in studies on isolated words to prohibit lexical access.

For example, place assimilation, when applied to the word *bad* can produce the variants *bab* and *bag* in the contexts of *He's a bad boy* or *She's a bad girl*. A matching process intolerant of deviation would reject both these tokens as candidates of *bad,* and so cause problems extracting the meaning of the sentences. Clearly this view of lexical access must be refined, and in this chapter I report experimental evidence suggesting that spoken word recognition involves a process of phonological inference in the evaluation of match and mismatch.

3.4 Psychological Studies of Phonological Variation

Despite the problems that neutralising phonological variation generates for lexical access, surprisingly little psychological research has investigated it. Many studies have examined the use of coarticulatory information in lexical access (e.g., Repp, 1978, 1983; Whalen, 1983; Warren & Marslen-Wilson, 1987) showing that, for example, cues to the identity of a consonant in the preceding vowel can have both facilitory and inhibitory effects in the on-line processing of speech. As described in Chapter 1, Elman & McClelland (1988) showed that the lexical status of a word can influence the coarticulatory influence of one segment on another. But the allophonic processes examined in these studies, whilst valuable to our overall picture of speech recognition in humans, do not cause the kinds of problems I have outlined above.

3.4.1 Place Assimilation

Recently, a number of studies have examined the effects of place assimilation on speech perception. An important first question to ask is whether assimilated segments really are acoustically the same as their labial or velar neighbours, and if they are not, can subjects perceive these differences? If they can, the problems caused by assimilation are reducible to the level of the allophonic and coarticulatory changes described above, since there is no need to use context to disambiguate assimilated segments. Nolan (1992) looked at this question for assimilation of place using an electropalatogram. This involved placing a metal artificial palate in subjects' mouths to record the positions of tongue contact on the palate. Since place assimilation involves the transition between segments differing only in their place of articulation, the electropalatogram provides a sensitive measure of the changes caused by place assimilation.

Analysis of electropalatographic maps of tongue contact in speech showed that assimilation was not a discrete change in terms of speech production, in that supposedly assimilated segments varied continuously in their point of articulation, from normal coronal contact to tokens with no coronal contact (as in true velar segments).

A word-identification task was used to examine whether fully assimilated surface velar segments were distinguishable from normal velar segments. Sentences such as *They did gardens for the rich* and *They dig gardens for the rich* were created in which the target word (*did/dig*) contained a word-final plosive with a surface velar gesture, but was either underlyingly coronal (*did*) or underlyingly velar (*dig*). These sentences were presented to naive subjects and trained phoneticians, with instructions to identify the target word. Subjects' responses to the fully assimilated underlyingly coronal targets were found to be no different to the underlyingly velar targets for either set of subjects. These findings suggest that in the normal processing of speech, assimilated segments are truly ambiguous. There is some evidence to suggest that subtle differences do remain between fully assimilated segments and underlyingly velar segments. Subjects performed significantly better than chance at a discrimination task when presented with both forms of each sentence one after the other. However, the fact that these differences were not picked up by subjects except when a direct comparison was allowed, suggests that the first-pass speech processor does not use these differences to distinguish between different lexical hypotheses.

Koster (1987) investigated place assimilation using a phoneme detection task. Subjects were presented with pairs of words in sentential context in which the segments around the word boundary either shared a labial or velar place of articulation (e.g., /k#g/) or consisted of a coronal segment followed by a labial or velar segment (e.g., /t#g/). The task was to identify both segments. He found that the former were harder to identify than the latter, due to ambiguity over the underlying form of the word-final segment, and that there was an effect of the lexical status and the semantic appropriateness of the carrier words. The effect of semantic appropriateness could only come into play after the second of the carrier words was heard, suggesting some post-lexical processing in the perception of assimilated words.

Koster also examined the time taken to monitor for coronal segments presented either unassimilated or assimilated to a labial or velar place. For example, asked to monitor for /n/, subjects would be presented with a sentence such as *Was the chain broken?* in which for the unassimilated condition, the /n/ in *chain* was realised as an [n] and for the assimilated condition was realised as an [m]. He also compared items which were lexically unambiguous (i.e. the surface non-coronal segment could not be treated as underlyingly non-coronal, as in *chain/chaim*) with items which did have a lexical alternative (e.g., *sun/sum*). An effect of both variables was found; unassimilated coronal segments were responded to more quickly than assimilated coronals, and segments in items which had no lexical alternative were recognised more quickly than those with an assimilated competitor.

Another experiment by Koster used a gating task to look at the effect of word-final assimilation on the recognition of the initial consonant of the following word. For example, subjects might hear a phrase such as *white gold* presented either in unassimilated form ([waɪtgoʊld]) or with a place assimilated word-final consonant ([waɪkgoʊld]). He found that the place of a consonant following an assimilated segment was recognised more quickly than following an unassimilated segment, but that this facilitory effect was countered by an inhibitory effect on the recognition of the voicing of the following consonant.[16]

These experiments give some clues as to the role of assimilation in speech perception. Firstly, there is ambiguity: surface velars and labials are harder to identify as underlyingly coronal than unassimilated coronals; and that, at least in an off-line task such as identification, word and sentence context are used as cues to the underlying identity of an ambiguous segment. Also, the results of the gating task suggest that in some circumstances subjects use the presence of assimilation to gain information about the following segment. This supports the idea that the matching process is sensitive to the regularities of phonological variation, and that the presence of an assimilated segment constrains its following context.

A recent study by Nix, Gaskell, & Marslen-Wilson (1993) also looked at the effects of various constraints on the perception of place assimilation. Their study used the gating task to examine the

---

[16]Koster attributed this inhibitory effect to the use of segments differing in voicing around the word boundary (as in [k#g]). Subjects used the assimilated segment to predict the place of the next segment but would normally choose a segment with the same voicing as the assimilated segment (in this example [k]).

time course of lexical, phonological and higher-level effects on the disambiguation of phonologically ambiguous segments. For example, sentences with a potentially ambiguous segment, such as the /k/ in *lake*, in the sentence *They thought the lake cruise was rather boring*, were presented to subjects using successively increasing gates. In the example above, there are no constraints biasing subjects towards either the underlying coronal form (*late*) or the underlying non-coronal form (*lake*). Consequently, the responses of subjects were divided, roughly half opting for the non-coronals, half for the coronals. Where there were biasing factors, the results showed these constraints to be swiftly utilised by the subjects in their responses.

3.4.2 Assimilation of Nasality

A cross-linguistic gating study by Lahiri & Marslen-Wilson (1991, 1992) looked at the perception of a different type of assimilation; assimilation of nasality in vowels. This change occurs word-internally across languages when a nasal consonant (e.g., /n/, /m/, /N/) is preceded by an underlyingly oral vowel. The nasality of the consonant spreads to the vowel so that, for example, /ban/ is produced as [ba^{N}n]. The languages used in this case were Bengali and English. In Bengali, vowel nasality is a distinctive feature, for example, contrasting [pa^{N}k] (*slime*) with [pak] (*cooking*). This means that the process of assimilation of nasality is neutralising in Bengali: in many cases it produces ambiguity in the surface forms of vowels. But in English this process is allophonic, since English does not have underlyingly nasal vowels.

Using the gating technique, Lahiri and Marslen-Wilson presented 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. The second type of item, denoted CVN, were triplets ending in a nasal consonant which contained an assimilated nasalised vowel. The Bengali subjects were also presented with CVC triplets containing an underlyingly nasal vowel followed by an oral consonant. As the triplets were presented, subjects were instructed to predict the identity of the complete word. For the nasal vowels, English subjects used the presence of vowel nasalisation to predict CVN words, whereas Bengalis initially interpreted the presence of nasalised vowels as underlyingly nasal vowels. For triplets with oral vowels, subjects almost never produced words with underlyingly nasal vowels by vowel offset, but responded with both CVC and CVN words. This suggests that subjects were often unable to use surface information in their judgements, instead operating on an abstract underlying representation.

The idea of abstraction away from the sources of variation is one way in which the problem of phonological variation could be solved. In the next section I shall develop this idea in more detail, along with a number of other possible solutions to the problem.

3.5 Models of Variation in Speech

A number of models of perception of variant speech have been proposed which quite neatly map out the relevant possibilities for theories of lexical access. These can be split into three categories; models treating variation as noise, those dealing with variation in the lexical representation and those compensating for variation using inferential processes. I shall briefly review these approaches here with emphasis on the predictions for cross-boundary variation such as assimilation. In this review I shall use the place assimilation process as a test case of phonological variation.

3.5.1 Phonological Variation as Noise

The simplest approach to variation is just to treat it as noise. This is the theoretically neutral approach adopted by most general models of word recognition. For example, the Cohort model (Marslen-Wilson, 1987) proposes that salient binary features are extracted from the speech wave and that these are directly mapped onto feature-based lexical representations of words. The only way for phonologically modified words to produce sufficient match in this type of model is either to tolerate small mismatches or allow top-down influences to interact with the sensory input.

This possibility appears to contradict the findings reported in Chapter 2, where cross-modal priming experiments found little or no tolerance for error. However, these studies all manipulated isolated

words. The kinds of deviations we are concerned with here occur in connected speech, and often do not occur unless the speech is relatively fast. It is quite conceivable that experiments on isolated words can only capture a subset of the properties of the speech recognition system and that where phonological variation is concerned, it is necessary to study larger units of speech. In other words, the same kind of experiment carried out within the context of a sentential structure may yield quite different results.

3.5.2 Representational Models

An alternative approach to the problem of variation in lexical access is to construct lexical representations that accommodate surface variation. This creates a distinction between phonological variation and noise; since only phonologically legal variants, constituting a subset of the possible variants of a word, will be tolerated by its lexical entry. The most obvious way to do this is to allow a number of alternatives for any lexical item, each one corresponding to a different phonetic variant of the base word (Harrington & Johnstone, 1987). For example, the lexical item *sweet* would need a representation of its canonical form [swit] plus two more variants [swip] and [swik] to ensure place assimilated forms are accessed without mismatch. Such a model, whilst allowing for the effects of phonological context on a word, does not *use* the context in the matching process and so it makes no difference to the matching process whether or not the context of the change makes it viable. For example [swik] when heard in a unviable context for assimilation (e.g., [swikbɒɪ], where the velar place of the [k] contrasts with the following labial segment) will be accepted as a token of *sweet*.

In contrast, the Lexical Access From Spectra model (LAFS; Klatt, 1989) allows for phonological variation by hypothesising a phonetic lexical network into which phonological rules are pre-compiled. Each variant of a word is represented by a path through this network. Even at the ends of words, each segment is linked to all possible following segments, allowing word boundary effects to be specific to certain contexts, by only linking variants of a word to the following contexts that allow such variation. Lexical access involves finding a path through this network that matches the sensory input. For example, a section of the LAFS network for the word *sweet* would include variants linked as in Figure 3.1. The [t] node at the end of sweet would be connected to all the possible word-initial segments whereas the [p] would only be linked to labial segments such as [b]. This type of model allows deviation to occur but only when its context makes it viable.

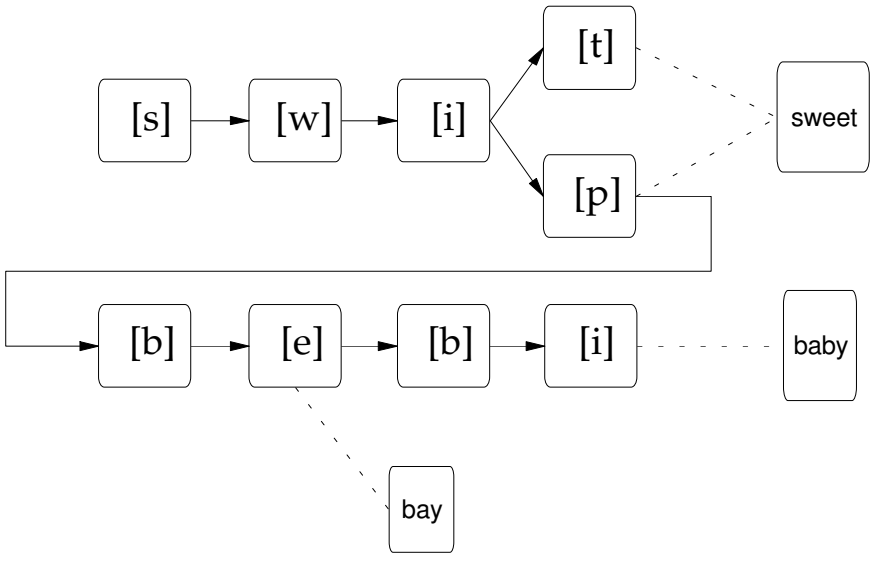

Figure 3.1. A LAFS Network for the perception of place assimilated words.

Instead of expanding lexical representations to allow for variability, it is possible that lexical representations of words contain only the features that do not vary from token to token. The Lexical Access From Features (LAFF) model (Stevens, 1986) uses this method to reduce the problem of within-word variability, by extracting binary invariant features from the speech signal, which are mapped onto the lexical representations of words. It is not clear how this model deals with between-

word variation, as is the case for assimilation, although Stevens states that features susceptible to spreading "should be indicated so that assimilation phenomena may be accounted for in a natural manner."

A similar approach has been advocated by Lahiri and Marslen-Wilson (1991), supported by the experimental evidence on the assimilation of nasals, described above. They proposed that form-based lexical representations conform to the theory of radical underspecification. The gating results they found, using vowel nasalisation as an example of phonological processing, were interpreted by assuming a simple matching process onto an underspecified lexicon containing only the marked, non-redundant features of a word. The possible outcomes of this matching process are:

1. A lexical *match,* in which the input matches a specified feature in the lexicon
2. A *mismatch,* in which the input is inconsistent with a specified feature in the lexicon
3. A *lack of mismatch*, in which the feature in the input is irrelevant since the lexical entry for that feature is unspecified, either through its default status or its redundancy

Figure 3.2 illustrates this hypothesis using the underspecification of place of articulation as an example (i.e. velar and labial segments are specified in the lexical representations of words, but coronal segments are not).

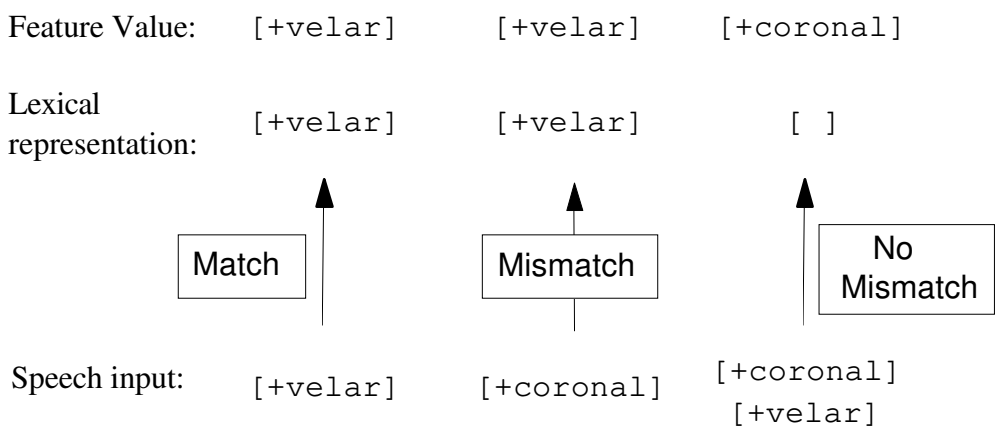

Figure 3.2. Matching an underspecified lexical representation.

The second and third possibilities again provide a distinction between random and phonologically regular variants of a word. The underspecified input lexicon allows for phonological variation, which alters only unspecified features and therefore causes no mismatch. However, deliberate mispronunciation generally alters specified features, causing mismatch with the lexical entry.

3.5.3 Inference Models

All the models described so far have dealt with variation at the lexical level, but another class of models deals with variation in the sensory input by proposing a more complex mapping process onto the lexicon. Using this approach we need only assume a single canonical representation of a word onto which variable speech is mapped. The best known example of this kind of model is the TRACE network of McClelland and Elman (1986), described in detail in Chapter 2. Regular variation in the input is dealt with in two ways. Firstly, active phoneme units can alter the feature to phoneme connections for subsequent input to compensate for coarticulatory effects (Elman & McClelland, 1986). Secondly, top-down activation from the word to phoneme level biases the activations of the phoneme nodes in favour of phonemes forming part of currently active words. Elman and McClelland (1988) showed that this structure was able to model the lexical and sub-lexical effects of compensation for coarticulation found in phoneme categorisation experiments.[17] Although

[17]See Chapter 1, for details of this work.

neutralising variation such as assimilation was not dealt with in TRACE, one would imagine that it would require hard-wired connections between phonemes and features, similar to those used for coarticulatory compensation. For example, to deal with a sequence such as [swikgɜl], where the [k] is an assimilated [t], TRACE requires biasing connections from [g] nodes to the *acute* and *diffuse* feature-phoneme connections (the features on which [k] and [t] differ in TRACE) in the previous time slice.

However, the type of variation we are interested in here has a number of properties which cause problems for the mechanism outlined above. Consider the behaviour of a TRACE model which has been set up to compensate for place assimilation as described above. There would again be a lexical effect on the activations of relevant phoneme candidates. For example, given the input [swik], there would be activation of the [t] node due to top down facilitation by the word 'sweet'. But the sub-lexical effects found are more difficult to produce in this case. The effects Elman and McClelland modelled were ones in which the perceptual boundaries between segments were shifted, rather than the more extreme transformations found in place assimilation. More importantly, place assimilation is a regressive effect; the influential segment is the one that follows the ambiguous segment. In the processing domain of TRACE, it is easy to influence a node which is not already strongly activated, as is the case for progressive effects, but very difficult to affect the activation of an already dominant phoneme node. So given the input [swip beɪbi], by the time the [b] input starts to become active, and thus able to influence the activations of the previous phoneme's nodes, the [p] node is already the dominant node and so fairly unassailable.

Nonetheless, the TRACE model would be able to compensate for assimilation to some extent and may predict a small effect of phonological context. More importantly, it is a representative of an approach to variation in which the effects occur pre-lexically. A more effective hypothesis is to assume a rule-based compensation strategy, whereby phonological rules are encoded in the access process. Pulman & Hepple (1993) have applied this approach to various phonological processes using a two-level parser. The important point about this approach is that it predicts a strong influence of phonological context on the mismatching effects of a deviation.

3.5.4 Summary

This review has produced a number of testable theoretical differences between the various accounts. The variation as noise account predicts no difference between random distortions and phonological changes, but all other accounts predict some asymmetry between the mismatch effects of phonological variation and random variation. The representational theories discussed above either add to lexical entries to cover all the possible phonological forms of a word or reduce the representational content to phonologically invariant features. The inferential approach depends to a greater extent on the phonological context of a variant, allowing phonological changes to be "undone" during the process of lexical access.

It is easy to view these issues of lexical representation and inferential processing as mutually exclusive. A strong phonological inference component in a theory of lexical access eliminates the need for anything other than a single fully-specified representation. Equally, a representational theory that encodes all possible variants of a word along with their validating contexts requires no more than a simple matching mechanism. However, there are reasonable intermediate positions, for example, employing a representational theory to deal with variation within a word and using phonological inference to compensate for between-word variation. In the rest of this chapter I report two experiments carried out to investigate the various predictions of these different approaches.

3.6 Experimental Considerations

This research is an attempt to answer a number of questions that arise from the issues discussed above. First, do minimal changes produce mismatch when presented embedded in a sentence? If the apparent conflict between phonological variation and intolerance of deviation is an artifact of single word experiments, then phonological changes in the surface form of a word might either go undetected or simply be treated as noise, with other factors, such as contextual constraint, compensating for any resulting ambiguities.

To assess this question, these experiments are a refinement of earlier studies on mismatch in lexical access (Marslen-Wilson & Zwitserlood, 1989; Marslen-Wilson & Gaskell, 1992; Marslen-Wilson, Moss, & van Halen, in press; Marslen-Wilson, 1993), using cross-modal priming to examine the effects on the matching process of single-feature word-final changes in sentential context. There are two aspects to the use of sentential context. Firstly, its presence may change subjects' tolerance of mismatch relative to the single-word experiments, perhaps by a shift in decision criteria. Secondly, there will be some reduction in the strength of the phonetic cues carried by the changed segments, especially the voiced and unvoiced stop consonants, where the final burst is not normally released in context. This might also contribute to a reduction of mismatch effects in context.

The second aim was to assess the sensitivity of the lexical access system to the phonological representation of words. The phonological changes used in these experiments could, in the appropriate context, occur naturally as a result of place assimilation. Thus, according to a pure representational account of variation, based on underspecification theory, the changed features should be unspecified in the lexicon and cause no mismatch with lexical entries.

Thirdly, I shall examine the extent to which phonological inference affects the perception of cross-boundary phonological phenomena. In particular, does the phonological viability of a feature change, as determined by its segmental right context, interact with the presence or absence of mismatch effects in lexical access? These experiments compared sentences with phonologically viable alternations to sentences in which the same changes occurred in a context making assimilation unviable. The unviable contexts were created by switching the place of the following segment from labial to velar or vice versa. For the base word *wicked*, for example, a viable change would be [wɪkɪbpræNk] (*wickib prank)*, where the places of articulation of the [b] and the [p] match. An unviable change would be [wɪkɪbgem] (*wickib game*), where the labial place of the [b] could not have spread from the following velar [g].

The experimental paradigm used to make these comparisons was cross-modal repetition priming. Subjects were presented auditorily with sentences containing a prime word in either phonologically changed or unchanged form, and at the offset of the prime the visual target was presented. The target was always the intact prime word, to which the subject made a lexical decision response. The single-word experiments all used cross-modal semantic and associative priming. Form-based repetition priming was chosen for use here for two reasons. The first was that cross-modal semantic priming in a sentence context is not a robust technique, and indeed preliminary versions of these experiments, using semantic priming, failed to elicit reliable priming effects. The second, and perhaps related reason, is that sentential context can interact in complex ways with responses to semantically or associatively related targets (e.g., Williams, 1988). These problems are avoided by the use of the more robust identity priming task.

To evaluate the effects of right context, as well as to examine the time course of possible effects, two experiments were conducted, as well as a initial pre-test of the stimuli. The first experiment presented the prime sentences with the speech following the offset of the prime word removed. This allowed examination of the effects of the deviation before their viability can be assessed. The second experiment presented the whole of the prime sentences. In each case the target was presented at the offset of the prime word.

3.7 Experimental Data

MATERIALS

The 48 sentences used in the pre-test (of which 42 were selected for the two main experiments; see Appendix A) were made up of, on average, 14 words and contained an embedded prime word. The sentences were manipulated with respect to three factors:

1. The absence or presence in the prime word of a word-final phonological change, which in the appropriate context could occur naturally as a result of place assimilation
2. The viability of the phonological context for assimilation

3. The prime — target relationship (either identical or unrelated control)

For each test item, six sentences were constructed. The sentences all had a common beginning and then diverged at the prime word according to the test condition. An example of each type of sentence is shown in Table 3.1.

Table 3.1. Sample prime words with phonological context. The sentential context in this case was "We have a house full of fussy eaters, Sandra will only eat ...".

		Phonological Change	
		Changed	Unchanged
Sentence Type	Viable	leam bacon	lean bacon
	Unviable	leam gammon	lean gammon
	Control	browm loaves	brown loaves

The prime words were all one or two syllables long and ended with a vowel followed by a coronal segment (/t/, /d/, /n/). For the phonologically changed primes, the place of the final consonant was changed from coronal to either labial (/p/, /b/, /m/) or velar (/k/, /g/) as would occur naturally in place assimilation. The result of this change was always a nonword. The viability of the following context of the prime was manipulated by following the prime word with a word whose initial segment was either labial or velar. For those sentences in which the place of articulation of the changed target and the following context matched (i.e. velar/velar or labial/labial), the context was phonologically viable. For the sentences in which the places did not match (i.e. velar/labial or labial/velar) the context was phonologically unviable for place assimilation. The following assimilation rules were used to create the viable changes:

$t \rightarrow k$ / -# g
$t \rightarrow p$ / -# (b,m)
$d \rightarrow b$ / -# (p,m)
$d \rightarrow g$ / -# k
$n \rightarrow m$ / -# (p,b)

These combinations avoided the use of the same segment for both the phonological change and the following context (as in [swikkɪs]; underlyingly *sweet kiss*). This was to make the offsets of the prime words easier to identify from study of the waveforms.

The unviable environments employed the same segments (/p/, /b/, /m/, /k/, /g/) but violated assimilation rules by pairing labial changes in word-final place with velars and vice versa. This made legal and illegal contexts equally phonetically similar to the unchanged segment (although not to the mismatch).

The sentences were between 10 and 20 words long with the prime words positioned towards the end of the sentences. The prime words were a mixture of nouns, verbs and adjectives and, although they were completely natural in the sentential context, were chosen to be difficult to predict given the preceding context. This was to reduce the possibility that any priming found in the main experiments might be caused by the relationship between the sentential context of the prime and the target. The sentences were also constructed to ensure that major prosodic boundaries did not split the prime words from their following context, since regressive assimilation does not usually cross such boundaries (Holst & Nolan, in press).

The control words were all words unrelated to the test primes. They were matched with the test primes on their number of syllables, word frequency and word-final consonant. They were also designed to be equally appropriate in the sentential context. The phonologically changed control prime was always followed by an unviable context for assimilation.

Because of the complex nature of the sentence construction, the sentences were recorded using a non-naive speaker. The sentences were recorded through a pre-amplifier onto DAT tape. They were then

filtered and digitised at 20 kHz using a CED Alpha mini-computer. The start and end points of each sentence and the offsets of the prime words were identified by analysis of the speech waveforms, using the CED SEDIT speech analysis tool.

3.7.1 Pre-test

The pre-test was necessary to ensure that the speech tokens used in the priming experiments had the correct surface place of articulation, and to check for confounding factors in the experimental design. My aim was to use tokens of speech which fell at either end of the assimilation continuum. Thus, the tokens employed in the unassimilated condition should have an unambiguous coronal surface form and the assimilated conditions should use word-final tokens which were unambiguously labial or velar. To check whether the intended surface forms were actually present a forced-choice phonetic decision task was used, in which the two alternatives were the unassimilated and assimilated surface forms of the prime words. To ensure subjects' decisions were not influenced by the following contexts of the changes, the stimulus sentences were presented with all speech following the offset of the prime words spliced out.

There were two possible confounding factors that might have arisen in the production of the stimulus sentences. Firstly, the greater articulatory ease of the viable context sentences (where the changed segments and their right context have the same place of articulation) may alter the quality of the word-final segments in these conditions. Secondly, there is a possibility that speaker bias could produce a difference in stimulus quality between conditions. The pre-test examined the effects of these factors on the properties of the word-final segment of the prime.

The perceptual quality of the following context of the prime word is also important in these experiments, particularly the place of articulation of the consonant directly following the word-final changes, as this determines their viability. However, since CV place cues are generally richer and more reliable than VC place cues (Repp, 1978; Ohala, 1990) I assumed that the context segments were unambiguous and did not pre-test them.

DESIGN

The pre-test manipulated two independent variables: Phonological Change and Sentence Type (see Table 3.1). Subjects were given a forced-choice test between the unchanged and changed version of the prime word and asked to rate the confidence of their responses. Thus, the dependent variables were the response of the subjects (changed or unchanged) and the confidence ratings (1 - 9).

In all there were 6 versions of each of 48 test items. These 288 sentences were split into 4 test versions, each including 72 test items. This meant that the control items doubled up with test items within a version, but as the control items had different prime words this avoided any repetition effects. No version of the experiment contained more than one test condition for each item.

SUBJECTS

Forty-eight subjects from the Birkbeck Speech and Language subject pool were tested. All were native British English speakers with ages ranging from 18 to 45.

PROCEDURE

The subjects were tested in groups of 2 to 4, sitting in booths in a quiet room. They were given answer sheets on which, for each sentence, two versions of the final word (the prime word) were printed, one corresponding to the changed version of the word and the other being the unchanged version. There was also a confidence scale for each sentence consisting of the numbers from 1 to 9. The sentences were played from DAT tape through BeyerDynamic 770 headphones to the subjects. Each item consisted of a warning tone followed by 5-15 words of preceding context and then the ambiguous word. The subjects then had 3 seconds to indicate the word on the answer sheet that best matched the word they heard, and to circle a number according to how confident they were about their choice. They were instructed to vary their ratings from 1 for a complete guess to 9 for a certain response. The subjects were given 10 practice sentences and then a break. The 72 test items were then presented. Each session lasted about 20 minutes.

ANALYSIS CONVENTIONS

Throughout this thesis, the analyses of experimental data were carried out after removal of outlying data points. The procedure for removing outliers consisted of first excluding the data of subjects who either responded too slowly or produced too many errors. These data played no further part in the analyses. Secondly, individual data points above a certain cut-off value were removed from response time analyses but were included in error analyses as normal responses.

Apart from Experiment 3 which required special treatment due to the nature of the task, there was little variation in cut-off limits between experiments. The subject cut-off points were decided by examination of the frequency distribution of subjects responses; on average, 16% of subjects were eliminated at this stage. The individual data point cut-offs were also chosen after analysis of the frequency distributions and, since the frequency distributions of Experiments 1, 2 and 4 were similar, a fixed cut-off point of 1200 ms was chosen, eliminating less than 1% of the data.

The data were analysed in terms of the mean error proportions and the mid-mean response times. The mid-mean statistic is the mean of the central 50% of the data. This reduces the influence of outlying data points whilst retaining sensitivity to the distribution of the data. Values quoted in tables or illustrated in graphs are the mean of the item and subject mid-means.

Analyses of variance were carried out on both item and subject mid-mean data. Effects were assumed to be significant if their p-value was below 0.05 in both analyses. Effects were labelled marginal if either p-value was between 0.05 and 0.10. Where possible, the experimental version was included as a factor in the analyses, reducing the estimate of random variation. However, effects or interactions involving this factor were assumed to be irrelevant and are reported, but not discussed.

Post hoc comparisons of condition means were carried out using both Newman-Keuls and Tukey HSD analyses of both item and subject data. Where effects were significant using both analyses, only the Tukey probabilities are reported (either 0.05 or 0.01); otherwise the Newman-Keuls probability is quoted.

RESULTS AND DISCUSSION

Four subjects did not complete the test due to faulty equipment and were rejected from the analysis. Of the remaining 44 subjects there were 12 subjects each for versions 1 and 2 and 10 each for versions 3 and 4. Each subject produced two response measures for each item, a forced-choice response to the two possible interpretations of the prime word and a confidence rating. For the purposes of this analysis, a correct response was defined as a response that agreed with the intended pronunciation of the final word (whether as a word or as a nonword). All subjects produced responses that were at least 80% correct.

Table 3.2. The pre-test scores for the 42 remaining test items. The column headings refer to the sentence type (Viable, Unviable or Control).

		Changed			Unchanged		
		V	U	C	V	U	C
% Correct	Mean	87.9	89.6	90.6	90.8	92.5	95.5
	S.D.	13.3	15.2	16.7	12.3	12.6	9.3
Confidence	Mean	7.76	8.09	8.09	7.93	8.04	7.80
	S.D	0.77	0.75	0.79	0.76	0.74	0.82

For each item in each condition a mean percentage correct and mean confidence rating across subjects was calculated. After examining these values and the digitised speech 6 items were excluded. The rejected items were ones in which low response rates or confidence ratings coincided with poor quality stimuli. The means and standard deviations for the cleaned data are displayed in Table 3.2.

Two 2-way analyses of variance (ANOVAs), over items and subjects, were carried out for both measures on the means of the 42 remaining items, with the factors of Phonological Change (2 levels)

and Sentence Type (3 levels). There were no significant effects of either factor, indicating that there were no systematic differences between the clarity of word-final consonants in the test materials.

A further analysis examined the variation in the results according to the manner of articulation of the word-final segment of the prime (nasal, voiced stop or unvoiced stop). All three groups produced low error scores and high confidence ratings although there were small differences between the groups. The unvoiced stops were the most clearly recognised in the analysis (5% error, 8.3 confidence), with 8% error and 7.8 confidence for the voiced stops, and 12% error and 7.9 confidence for the nasal segments.

The results of the pre-test show, first, that the vast majority of word-final consonants, whether changed or unchanged, were produced as intended, and perceived as such. The mean proportion of correct responses was 91% and the mean confidence rating was 7.95. These results suggest that the surface places of articulation of the critical segments were reasonably unambiguous, meaning that in the appropriate contexts they would be treated as either unassimilated or fully assimilated. Second, there were no significant differences across conditions, ruling out any interpretation of the subsequent results in terms of confounds in the production of the stimuli.

3.7.2 Experiment 1

Experiment 1 presented the stimulus sentences up to the offset of the prime word, making the phonological right contexts of the prime words unavailable to the subjects. In Experiment 2, the subjects will hear all of the prime sentence. This allows evaluation of the perception of mismatch under different phonological conditions.

DESIGN AND MATERIALS

The test items used were the 42 items that performed satisfactorily in the pre-test (see Appendix A). Since the viability of the phonological context could play no part in this experiment, the design was reduced to three conditions per item, representing three levels of the variable Prime Type. Two conditions contained related primes: one with a word-final phonological change (i.e., where *leam* is followed by the visual target LEAN) and one in which the related prime was presented unchanged. These were compared to a control condition in which the prime word was unrelated to the target (i.e., where *brown* is followed by LEAN). For 50% of the control sentences a phonologically changed prime (e.g., *browm*) was used, and in 50% the primes were intact.

The sentences were split into three test versions, with one sentence from each test item in each version. To reduce possible strategic effects, 150 filler sentences were interspersed with the test sentences. Of these, 100 were accompanied by nonword visual targets, with two-thirds following an intact prime and one third following a phonologically changed prime. In 14 of each of these subtypes the primes and nonword targets were related in form, preventing subjects from associating similarity between prime and target with the real-word status of the target. A further 50 fillers (also with one-third changed and two-thirds intact primes) with unrelated real word targets were added, to reduce the proportion of sentences with a strong prime-target relationship. The fillers were matched with the test sentences for sentence length and prime frequency.

As well as the test items and fillers there were 25 practice sentences and 10 dummy sentences, played after breaks in the test sequence to allow for settling in. This brought the total number of sentences in each test version to 227, with the proportion of related items being 25% of the pairs with real word targets.

To encourage the subjects to pay attention to the auditory stimuli, a secondary recognition task was used at the end of the experimental sessions. This employed written forms of ten filler sentences from the experiment and ten sentences not present in the experiment. The task of the subjects was to identify the sentences they had heard in the experiment.

SUBJECTS

Thirty-six paid subjects from the Birkbeck College Speech and Language subject pool were tested. The subjects were in the age range 18 to 45 and there was a roughly equal proportion of males to females. None of the subjects had taken part in the pre-test.

PROCEDURE

Each subject was tested on one of the three test versions, in groups of 1 to 4. They were warned that they would be given a recognition test on the auditory stimuli after the main experiment, but that they should not try to memorise the sentences. The lexical decision experiment was then carried out in three blocks: first the 25 practice sentences, followed by two experimental blocks of 101 sentences. There was a short break between each block. Each sentence was preceded by a warning tone and a short interval. At the offset of the prime word, the monitor in front of the subject displayed the target word for 200 ms and the subject was required to press the "Yes" button if the target was a word or the "No" button if it was not. The reaction time was measured from the offset of the prime (which corresponds to the onset of the target). The button box was set up so that the subjects always responded "Yes" with their dominant hand. Once all the subjects had responded, or the 3 second time-out was reached, there was a short interval and the procedure was repeated. The test session lasted about 45 minutes.

At the end of the lexical decision experiment, the subjects were given a recognition sheet, containing twenty sentences, ten of which were filler sentences in the experiment. The subjects were instructed to circle the numbers of any sentences that seemed familiar to them. There was no time limit in this part of the experiment but most subjects completed the task in two or three minutes.

RESULTS AND DISCUSSION

Of the 36 subjects tested, five were dropped due to high error rates (over 20%) or high mean response times (over 750 ms). A further 16 individual response time scores over 1200 ms were also excluded. As the control condition was made up of both phonologically unchanged and changed items a preliminary analysis was carried out to test whether the phonological change in the control words was significant. The changed items (e.g., *browm*) provoked longer reaction times (654 ms compared to 644 ms for the unchanged items) but this difference was not significant in a one-way ANOVA ($F_2[1,40] < 1$). The results were also analysed with the manner of articulation of the prime word-final segment as a factor, but no significant effects involving this factor were found. Because of this, the remaining analyses are presented collapsed across this variable.

The results for the three conditions are summarised Table 3.3 and Figure 3.3. Two-way subject and item ANOVAs were performed on the data, with the variables Prime Type (unchanged-related, changed-related or unrelated control) and Version, the test version. The effect of Prime Type was highly significant ($F_1[2,56] = 48.3$, $p < 0.01$; $F_2[2,78] = 27.8$, $p < 0.01$) with Tukey HSD comparisons showing both related conditions to be significantly different to the control condition ($p < 0.01$).[18] The difference between phonologically changed and unchanged related conditions was not significant.

A two-way ANOVA on the error proportions revealed a significant effect of Prime Type ($F_1[2,56] = 8.8$, $p < 0.01$; $F_2[2,78] = 6.2$, $p < 0.01$). This showed that the control items provoked more errors as well as longer responses.

Table 3.3. Results of Experiment 1. The response times quoted are the average of the item and subject mid-mean values.

	Unchanged	Changed	Control
RT.	578	584	648
S.D.	63.8	71.6	87.1
% Error	5.3	4.9	11.3

[18]The interaction between Version and Prime Type was also significant ($F_1[4,56] = 11.1$, $p < 0.01$; $F_2[4,78] = 3.78$, $p < 0.01$).

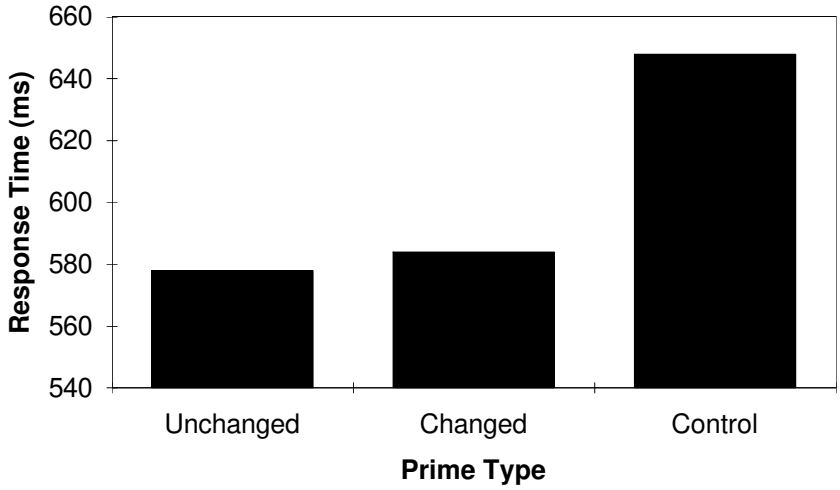

Figure 3.3. Mean response times for the prime types in Experiment 1

Experiment 1 showed a strong priming effect for both changed and unchanged prime words with little evidence of a mismatch effect for the changed words. This result contradicted the findings of the experiments on single words, but by itself is not enough to isolate the mechanisms involved in this effect. It could be that the small deviations used here were treated as noise in the matching process, but were insufficient to disrupt access to lexical information. But these findings also conform with a representational account of variation such as underspecification, since the phonological changes used are assumed to be underspecified in the lexicon. These issues will be resolved in the second experiment, where the effects of phonological viability on lexical access can be assessed.

3.7.3 Experiment 2

In Experiment 2, the full prime sentences were presented, enabling the following context of the phonological changes to affect the matching process.

DESIGN

The design of Experiment 2 was as of Experiment 1 except that since following context was expected to be influential in Experiment 2, the full six-condition design was used as shown in Table 3.1. The independent variables were Phonological Change (unchanged vs changed) and Sentence Type (related-viable vs related-unviable vs control). There were 6 test versions, each including one condition from every test item.

The prime words were presented embedded in full sentential context, such as "The house was full of fussy eaters. Sandra would only eat *leam bacon*". The target word (in this case LEAN) was presented at the offset of the prime. The breakdown of the fillers used was the same as for Experiment 1.

SUBJECTS

Forty-six members of the Birkbeck College Speech and Language subject pool were tested in this experiment. Of these, three had taken part in the pre-test but as this had been six months earlier, it was assumed this would not affect their performance. The subjects were paid for their participation.

PROCEDURE

The procedure was the same as for Experiment 1 except that all sentences were presented in full. The target word was presented at the offset of the prime, which was now mid-sentence.

RESULTS AND DISCUSSION

Of the 46 subjects, five were excluded due to high error rates (over 20%) or high mean response times for the test items (over 800 ms). In addition, individual item response times over 1200 ms were excluded from analysis. The item and subject mid-means were calculated along with the error rates and are summarised in Table 3.4.

Table 3.4. Means, standard deviations and error rates for Experiment 2. The values quoted are the mean of the item and subject mid-mean values. The column headings refer to the sentence type (Viable, Unviable or Control).

	Changed			Unchanged		
	V	U	C	V	U	C
Mean	624	655	679	625	615	651
S.D.	89.0	83.1	90.4	85.4	70.2	98.9
% Error	6.3	6.5	7.8	4.9	4.5	9.0

The reaction times in Experiment 2 are slightly longer than those found in Experiment 1, with the overall mean rising from 607 ms in Experiment 1 to 642 ms in Experiment 2. This may reflect an increase in the processing load when performing the lexical decision, due to the continuation of the auditory stimulus sentences in Experiment 2.

As in Experiment 1, a preliminary analysis was carried out using the manner of the word-final segment of the prime as a factor but no significant effects of this factor were found. The results were therefore collapsed across this variable.

Three-way ANOVAs were performed on the data using the independent variables Sentence Type, Phonological Change and Version (the number of the test version in which the data was collected). There were significant main effects of both Phonological Change ($F_1[1,35] = 6.93$, $p < 0.05$; $F_2[1,36] = 10.89$, $p < 0.01$) and Sentence Type ($F_1[2,70] = 8.83$, $p < 0.01$; $F_2[2,72] = 17.65$, $p < 0.01$).[19] These show firstly that across conditions, phonologically changed words are responded to more slowly than unchanged words (651 ms as opposed to 628 ms) and that the control items (661 ms) were responded to more slowly than both the viable context conditions (623 ms) and the unviable context conditions (636 ms).

What is most relevant here is whether the context of a prime affects this mismatch effect. The ANOVAs showed the interaction of Phonological Change and Sentence Type to be only marginally significant ($F_1[2,70] = 2.64$, $p = 0.08$; $F_2[2,72] = 2.451$, $p = 0.09$). However, this analysis is not the most direct way of addressing the question. The factor Sentence Type is an amalgamation of two variables: prime-target relatedness (i.e. test versus control prime) and phonological context (viable versus unviable). In an ANOVA on just the four related-prime conditions, using the factors Phonological Context and Phonological Change, the interaction was significant ($F_1[1,35] = 4.70$, $p < 0.05$; $F_2[1,36] = 5.165$, $p < 0.05$). A post hoc comparison using the Tukey HSD statistic showed a significant mismatch effect for unviable related items (see Figure 3.4).

An ANOVA on the error proportions in each category revealed no significant main effects or interactions.

---

[19]There was also a significant interaction between Version and Sentence Type ($F_1[10,70] = 7.67$, $p < 0.01$; $F_2[10,72] = 2.18$, $p < 0.05$).

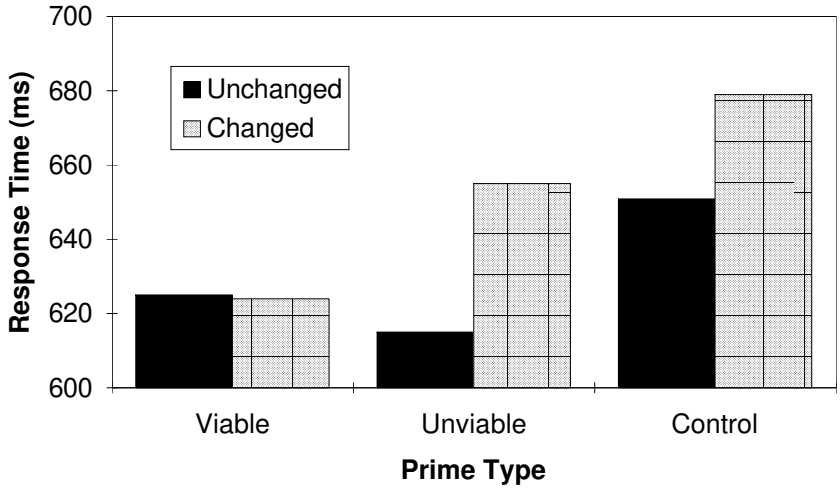

Figure 3.4. Mean response times for Experiment 2.

The results of Experiment 2 highlight the important role of phonological processes in lexical access. Changed primes, when presented in viable context for assimilation, produced a strong cross-modal priming effect and showed no mismatch effect in comparison to unchanged tokens. The same primes, presented in unviable context produced a reduction in priming of 40 ms.

In order to interpret these results as a product of the perceptual mechanism's sensitivity to phonological changes, we must first rule out a more simple explanation of the findings. The phonologically changed segments and their following contexts varied along three phonetic dimensions: place of articulation, manner of articulation and voicing. Of these, differences in place of articulation were fully controlled, with viable changes and their following contexts sharing the same place of articulation and unviable changes mismatching the place of the following contexts. The differences in voicing and manner were more variable, although the viable changes all differed from their following contexts on at least one of these dimensions (i.e. the changed segment was never phonemically identical to its following context).

It is possible, therefore, that the difference in priming between the viable and unviable changes could be explained in terms of a phonetic masking effect of the context segments on the changed segments (Kallman & Massaro, 1979). The viable contexts may be, overall, more similar to the changed segments than the unviable contexts and so mask the mismatching effects of the phonological changes more effectively. Similarly, these effects could be the consequence of differential phonetic contrast effects between the viable and unviable conditions (Repp, 1978, 1983).[20]

To examine these possibilities, the stimulus items were categorised according to the phonetic similarity between the changed segments and their viable and unviable contexts (in terms of the features place, voicing and manner). The masking and contrast interpretations would predict that when the viable contexts were phonetically more similar to the changed segments than the unviable contexts, the mismatching effect of the changed segments should be smaller, and therefore the difference between response times to the viable and unviable changed conditions should be greater. Twenty-two items had phonologically changed segments which were phonetically more similar to the viable following contexts than the unviable following contexts (e.g., the /m/ and /b/ in the viable assimilation *leam bacon* differ only in manner whereas the /m/ and /g/ in the unviable assimilation

---

[20]A phonetic contrast explanation is weakened by the findings of Repp (1983) that for VCCV sequences involving stop consonants, contrastive effects are due mainly to the information given by the duration of the silent closure interval between consonants. If this closure is of a duration normally associated with two different stop consonants (such as d-b) rather with geminate stops (such as b-b) subjects will use this information to make a response consistent with the perception of two different stops. In these experiments, however, no conditions contained identical phonologically changed and context segments and so the closure duration does not help in the evaluation of these segments.

leam gammon differ in both place and manner. In the remaining 20 items the phonologically changed segment was either equally phonetically similar to both viable and unviable context segments (19 items) or was more similar to the unviable context (1 item).

An item ANOVA on these groups revealed no effect of phonetic similarity on the viability effects: the former group showed a 29 ms viability effect and the latter a 28 ms effect ($F_2[1,40] < 1$). A correlational analysis of these data showed a small and insignificant correlation in the opposite direction to the predictions of the masking or contrast hypotheses (Pearson's $r = -0.03$, $p > 0.10$). It seems, therefore, that the effects found here are best described in terms of a context-sensitive inference process rather than a more low-level phonetic masking.

These results support the claim that the lexical access process is intolerant of small deviations. The changes used in the unviable context conditions were all single feature deviations, but they still produced a 40 ms mismatch effect.

Given that these minimal deviations have such a strong effect on response times, it is relevant to ask whether there is any residual priming for the phonologically changed words in unviable context. In this experiment we have two types of control words: changed and unchanged. Most studies quoted in Chapter 2 (e.g., Marslen-Wilson, Moss & van Halen, in press; Marslen-Wilson & Gaskell, 1992) found no residual priming for phonologically changed words in cross-modal priming tasks, but these were in comparison to unchanged control words. If we make the same comparison here, there is again no apparent priming (test-control difference = 4 ms). This apparent lack of priming may be due to the influence of two competing factors, one being the actual priming, due to the relatedness of prime and target, and the other being inhibition or slowing simply due to encountering a nonword. For this reason the phonologically changed control word is the fairest baseline for assessing priming effects for changed words. When the reaction times for the changed related words in unviable contexts are compared to the changed control words there is a suggestion of priming for the changed words in unviable context, with a marginally significant 24 ms effect (Newman-Keuls $p > 0.10$ by subjects; $p < 0.05$ by items).

3.7.4 General Discussion

The two experiments reported here provide empirical evidence relating to a number of interconnected issues in the area of auditory lexical access. Firstly, the mechanics and dynamics of lexical access and secondly, the effect of phonological processes on lexical access. Here I shall discuss both issues in the light of the new evidence these results provide.

GOODNESS OF FIT IN AUDITORY LEXICAL ACCESS

One motive for these experiments was the finding that, for isolated words, minimal distortion of the tokens of speech disrupted lexical access. Considering the unviable context conditions of the two experiments, the only difference between the sentences in these conditions was in the place of articulation of the word-final consonant of the prime word. In Experiment 1, with no following context this made no difference, but in Experiment 2 it was enough to reduce the priming found by 40 ms. The results of Experiment 2 are therefore additional evidence for a lexical matching process that needs an extremely good fit to activate lexical representations.

The presence of a mismatch effect confirms that the repetition priming technique is sensitive to small changes in the speech signal and is therefore a valid tool for examining the process of lexical match. Thus, the absence of this effect in Experiment 1, where the sentences were cut at the offset of the prime word, itself requires explanation, given the results of experiments on isolated words where mismatch effects were found for all distorted tokens. One possibility is that the additional processing load created by the preceding speech forces subjects to process speech with increased tolerance for deviation. But this cannot be a general tolerance for deviation, given the results of Experiment 2, which suggest instead an account based on phonological processes and representations. I will return to this question in the next section.

Turning to the dynamics of the mismatch effects, comparison of the two experiments implies that phonological context is brought into play very rapidly. Experiment 1 shows that, at the offset of the changed word, there is strong activation of the base word by a phonological variant, but Experiment 2 shows that an unviable following context quickly inhibits this activation. This kind of result is

difficult to model using the winner-takes-all processing environment of models like TRACE (McClelland & Elman, 1986), since again, we find fast inhibitory effects on the dominant word candidate, this time due to the influence of the following context. As I showed in Chapter 2, this type of behaviour, if interpreted in TRACE-like terms, require direct inhibition from featural information, so that mismatching information can have strong effects in a short time.

PHONOLOGICAL PROCESSES AND LEXICAL ACCESS

The most striking finding of this research has been the strong effect of the phonological context of deviations found in Experiment 2. This shows that the matching process is sensitive to phonological interactions. If we are to accommodate this result into the standard paradigm of lexical access, the definition of a mismatch must be revised.

When sentences were presented with a phonologically changed prime in a context in which the change was phonologically viable, there was no mismatch effect – the target was facilitated as strongly as for the unchanged primes. However, the same change, in circumstances where it could not occur naturally, strongly disrupted the priming effect. A matching process that analyses segments or features without reference to their neighbours cannot cope with results like this. Neither the changed segments nor the context segments by themselves create mismatch, only when the two elements combine does mismatch occur.

I can now offer an explanation of the absence of a mismatch effect found for distorted words in Experiment 1, which contrasts with the earlier results using isolated words. At the point the sentences were cut off, the word-final change had been presented but the following context was unknown. The preceding sentential context means that the subjects had engaged the processes normally involved in the interpretation of connected speech. Within the framework of this type of analysis, the underlying identity of the word-final deviation in the prime word is ambiguous. If the deviation is in fact underlyingly coronal it should not mismatch the lexical entry for the word, but if it is underlyingly non-coronal it should cause mismatch. This ambiguity can only be resolved once the following context of the changed segment is known. Thus, a matching process that is intolerant of mismatch but at the same time acts conservatively, only rejecting candidates when there is unambiguous mismatching information, would predict precisely the results found in our experiments: no mismatch in Experiment 1, where the phonological viability of the change is unknown, but a strong mismatch effect in Experiment 2 when a minimal but unambiguous mismatch is perceived.[21]

MODELLING VARIATION IN LEXICAL ACCESS

Earlier in this chapter, I presented three major classes of models of variation in lexical access: variation as noise accounts, lexical representation accounts and pre-lexical processing accounts. I shall now discuss the plausibility of these accounts with reference to the new experimental data.

It is clear from the results of Experiment 2 that treating regular variation as noise is untenable in a model of auditory lexical access. The same deviations can either cause mismatch or no mismatch, depending on the viability of the context in the following word. Phonological processes must therefore be involved in the lexical access process, perhaps mirroring to some extent the processes that occur in speech production.

Two classes of theory remain: representational theories and inferential theories. The representational approach assumes phonological variation is accommodated either by adding to the lexical representation (e.g., Harrington & Johnstone, 1987; Klatt, 1989) or abstracting to invariant features (e.g., Stevens, 1986; Lahiri & Marslen-Wilson, 1992).

Of these, the Klatt network is the only model that can accommodate these results since it is the only one that specifies the context that allows phonological change. None of the other models predict the effects found in Experiment 2 of phonological context on word recognition. But the Klatt network lacks plausibility on a number of accounts. For phonological change across word-boundaries to be

---

[21] It is also likely that the reduction in the strength of phonetic information on the place of the changed segments, compared to the isolated word case (see Section 3.6), also reduces their mismatching effect.

accommodated, connections between every word and its viable context must be made. This requires some kind of inferencing mechanism, which discovers new rules and applies them to the words in the network. This inferencing mechanism is left unspecified and, as Klatt himself conceded, reduces the credibility of the model.

The lexical abstraction theories, such as underspecification, are theoretically more attractive, since they result in a more compact lexicon rather than multiplying the lexical representation of each word with more than one phonological form. But how can these theories deal with cross-boundary effects such as the ones found here?

One retreat from a pure representational theory would be to retain the abstract lexical specification but to add a more sophisticated matching strategy. For example, the effects found in these experiments could be explained if we assumed that surface features being mapped onto underlyingly unspecified forms are not simply discarded, as assumed up to now. Instead, these features are retained in some "input buffer" to be matched with the following segment. But to some extent, the value of underspecification to psychological theory is lost if we take this additional step. Underspecification as a theory of lexical *representation* makes strong claims as to the effects of deviation on the matching process. But the addition of a more complex matching process obscures this representation and weakens the theory.

Turning to the purely inferential models, I have pointed out earlier that the kinds of interactions involved here are unsuitable for modelling using TRACE. But the results are easily accommodated by assuming some kind of rule-based system that disambiguates surface input before, or in the process of, lexical access. Pulman & Hepple (1993) implemented this approach using a two-level phonological parser, mapping surface input directly onto lexical representations using a set of phonological rules.

3.8 Conclusions

This research was motivated by the intriguing finding that small deviations from the canonical tokens of isolated words are enough to disrupt lexical access. This is despite the fact that the human speech recognition system must cope with a wide variety of sources of variation without a significant loss of comprehension. The experiments have examined this result drawing on a particularly regular natural form of variation, phonological variation. The results of the two experiments agree with the supposition that the matching process involved in lexical access requires a good fit for the lexical representation of a word to be accessed. However, I have shown that this matching process must be sensitive to the phonological processes present in the speech stream, allowing deviations as long as they conform to the phonological rules of assimilation. This result is also evidence for the psychological reality of phonological interactions in speech perception.

These results point to a model of auditory lexical access that incorporates context-dependent phonological inference to assess the validity of a phonological change in lexical access. The structure of the lexical representation onto which this inference process maps remains unclear. It is possible that there is a single fully-specified lexical representation of each word we are familiar with and that phonological inference compensates for all regular variation we encounter. However, these results are also consistent with a hybrid approach in which some forms of variation are accommodated by the structure of lexical representations and others by inference processes. It is possible that within a lexical unit, phonological variation is accommodated by a process of lexical abstraction, but across word or morpheme boundaries phonological inference processes act to compensate for change. This approach thus combines the simplicity and cognitive economy of the representational account with the flexibility of the inferential account.

However, these results can only be accommodated by a model of spoken word recognition that involves a process of phonological inference. The only psychological model that explicitly incorporates such a process is TRACE, which is unable to produce the required effects. This is because it lacks the flexibility to alter lexical and phonemic hypotheses swiftly, as new information is encountered. In the next chapter I present an alternative model of phonological processing, using distributed connectionist learning algorithms to overcome the limitations TRACE encounters.

Chapter 4 — A Connectionist Model of Phonological Inference

4.1 Introduction

The experiments reported in Chapter 3 indicate that auditory lexical access involves a component of context-dependent phonological processing. The mechanism underlying this process is still largely unknown, but in this Chapter, I present an explicit connectionist model of one possibility. This is the hypothesis that phonological inference is carried out pre-lexically, as part of the matching process between speech input and lexical forms.

Connectionist models have had a powerful impact on the modelling of psychological processes in the last ten years. In the first section of this chapter I briefly review the issues involved in the application of connectionism to psychological theory. I evaluate the main architectures and learning algorithms employed by connectionists and describe the properties that these networks hold. I then review the way these different architectures have been used in the modelling of speech perception, where sensory information is spread over the temporal domain. In the final section I describe and evaluate a simple recurrent network model of speech perception, in which the network learns to compensate for phonological change by identifying the contextual situations in which phonological rules are applied. The evaluation takes the form of three simulated experiments which examine the success of the model at learning and exploiting the various cues to the presence of place assimilated segments of speech.

4.2 Connectionist Modelling

Psychological theory has, in the recent past, looked to the computer as a metaphor for the mind, and in particular to the model of the computer usually attributed to von Neumann. This supposes that information is operated on sequentially by a single processor. Connectionism, instead, takes what we know about the brain as its basic foundation, resulting in a massively parallel, distributed information processing environment. Connectionist networks, in common with other models (e.g., Selfridge's Pandemonium system, 1959), use the activation of units, either singly or across a group of units, to represent hypotheses. These units, or *nodes*, interact via weighted connections which are often modifiable, allowing learning to occur. The reasons for the use of this substrate and its implications are examined in the following sections.

4.2.1 The Roots of Connectionism

Early roots of connectionism are found in work by McCulloch & Pitts (1943) and Hebb (1949). Both studies draw heavily on the link between neural functioning and psychological processes. McCulloch & Pitts carried out the first formal analysis of simple networks of this kind, showing how simple idealised neurons can be added together to perform logical functions such as AND, OR and NOT. From this we can infer that any task that can be broken down into logical units such as these can, in theory, be carried out using a neural network. This finding was important since it showed how complex thoughts could be broken down into simple elements with a neural metaphor. The smallest unit of psychological processing was termed the *psychon* and corresponded to the firing of a single neuron.

Hebb's work modelled the mechanism of synaptic change in the brain, showing that repeated co-ocurrence of activation between cells leads to a permanent modification of the synaptic link between them. This has the effect of ensuring that the cells will continue to become active simultaneously. An idealised version of synaptic modification has been described using the Hebb learning rule, stating that the weight change for a connection between two units, *i* and *j* is given by:

$$\delta w_{ij} = \varepsilon . a_i . a_j$$

where ε is the fixed learning rate and a_i and a_j are the activations of the relevant units.

This rule is powerful enough to allow learning of associations between patterns in many non-trivial cases. However, it will fail to learn correctly if the input patterns it receives are correlated. The delta rule (Widrow & Hoff, 1960) overcomes this problem by comparing the output of a network to training values in order to direct the modification of weights as shown:

$$\delta w_{ij} = \varepsilon.(t_j - a_j).a_i$$

where t_j is the idealised or training activation of the *j*th unit.

This algorithm loses the strong neural analogy that Hebb's rule enjoyed since it necessitates comparisons with training patterns, but it is a much more powerful rule. In simple terms, it adjusts each weight a small amount each iteration so as to reduce the mean square error between the actual output of the network and the desired output.

Rosenblatt (1962) carried out a mathematical analysis of simple networks using this type of rule. The networks, which he termed *perceptrons*, consisted of a single layer of input units connected to a single layer of output units. He showed that if there is a network state (i.e. set of connection weights) that will satisfy the constraints of a problem, the perceptron will converge on this state. In other words, if the simple perceptron can solve a problem, it will.

Unfortunately, the range of mappings a simple perceptron can learn is still limited. Minsky & Papert (1969) in their book *Perceptrons*, analysed the kinds of problems simple perceptrons could solve. They showed that if a set of patterns is not *linearly independent*, or the input-output mappings are not *linearly separable*[22] the perceptron will be unable to converge on the correct state. The most famous example of a simple problem the perceptron cannot solve is the exclusive-or (XOR) problem. This involves learning the mappings shown in Table 4.1.

Table 4.1. The mappings required to model the exclusive-or function. Two inputs must map onto one output unit.

Input 1	Input 2	Output
0	0	0
0	1	1
1	0	1
1	1	0

Here, the two patterns that require a 0 to be output are the most dissimilar. So what is needed in order to solve the problem with a network is a certain amount of recoding. For example, McClelland & Rumelhart (1988) showed that a network with one hidden (intervening) unit can solve the problem. But the addition of hidden units violates the definition of the simple perceptron, and thus falls outside the scope of the Minsky and Papert analysis. Minsky and Papert also studied more complex multi-layered networks, but found that, unlike the delta rule for simple perceptrons, there existed no algorithm that could be proved to converge on the solution.

Largely as a consequence of this finding, research in this area dwindled in the seventies. The resurgence of connectionist networks has been due mainly to the introduction of new learning algorithms that attempt to resolve the problems posed by Minsky and Papert.

---

[22] A linearly independent set of patterns is one for which no element can be made from a combination of other elements. The most common input set satisfying this constraint is a localist one. If this constraint is not met, the input-output mappings must have a linear relationship with each other for the delta rule to work (for example, if $i_1 + i_2 = i_3$ then o_3 must be $o_1 + o_2$, where i_x is an input pattern and o_x the corresponding output pattern).

4.2.2 Learning Algorithms

HOPFIELD NETWORKS

There have been a number of responses to Minsky and Papert's critique. One has been just to use the simple perceptron as a model of psychological processes despite its limitations (e.g., Rumelhart & McClelland, 1986). Another popular response has been to avoid the problem of learning, by hard-wiring the connections in an interactive network. Hopfield (1982) used the physical analogy of the energy of a spin glass to study a system of this type. A Hopfield network consists of a set of binary units, interconnected with symmetrically weighted links as illustrated in Figure 4.1. Input to the network is in the form of a set of initial activations to the nodes. Nodes are then updated in a random order by assessing their summed input from the rest of the network and assigning an OFF value (0) if the net input is negative, or an ON value (1) if the net input is positive.

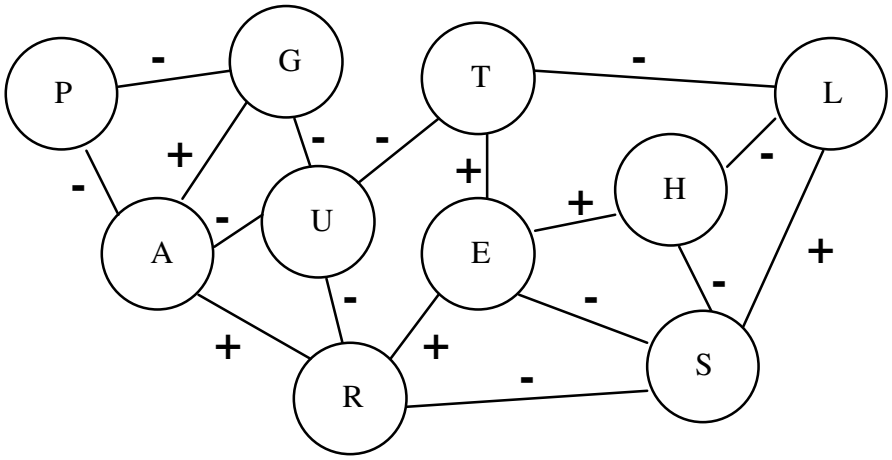

Figure 4.1. A simple Hopfield network.

Hopfield defined the energy of the system, E, by the equation shown below

$$E = -\frac{1}{2}\sum_{i}\sum_{j\neq i} w_{ij}.a_i.a_j$$

where *a* is the activation level of a node and *w* is the weight of a link. The use of this equation implies that nodes with the same value connected by an inhibitory link add to the energy of the system, as do nodes with opposite values connected via a facilitory link. If the weights of the links are thought of as representing logical constraints we can see that only constraints which are not satisfied add to the energy, so Hopfield's E is a kind of negative goodness-of-fit value.

Hopfield also showed that all changes of state caused by the asynchronous update method will have the effect of decreasing the energy of the system, and that the system will converge on a stable state in a finite number of updates. This is a point where the energy of the system is at a local minimum and thus the goodness-of-fit is at a local maximum.

This behaviour makes Hopfield networks valuable for modelling constraint satisfaction and pattern completion problems. The time taken to settle into a stable state is used as an analogue of recognition time or facilitation strength, depending on the process being modelled.

Hopfield networks are particularly valuable if the network can develop a set of learned weights rather than using pre-set values. Hopfield (1982) proposed a variant of Hebbian learning which reduces the weights of connections between conflicting units and increases the connection strength between units of the same value. This allows the network to build up attractors corresponding to the patterns used in training. However, a number of problems emerge when using this learning system. Firstly, the network has a tendency to develop spurious attractors (Hopfield, 1984), consisting of blends of two or more input patterns. Also, the number of patterns a network can learn is extremely limited compared to other types of network. Campbell, Sherrington & Wong (1989) estimate that a network of N units is capable of storing roughly 0.15N patterns. Attempts to exceed this ratio result in catastrophic loss

of all memory rather than just a degraded performance on a few patterns. Adaptations of the learning algorithm can improve this ratio (see Amit, 1989) but the storage capacity remains poor compared to other distributed systems.

A serious problem for the modelling of temporally dependent processes such as speech perception using these systems is the tendency of a Hopfield network to quickly settle into a stable state. This creates difficulties when information is presented to the network over a substantial period of time. Seybold (1992) examined a number of systems for the representation of real speech tokens in a Hopfield network model of spoken-word recognition, but found that the network would settle into a stable state early on in the processing of a word. This behaviour is especially problematic when compared to human data (Marslen-Wilson & Gaskell, 1992) showing that even at the end of a word, small deviations in surface form have strong effects on the mental representations activated.

BACKPROPAGATION

The problem of learning in connectionist networks can be envisaged in terms of the descent of an error curve. Consider a network with *n* variable connections. The error curve of the network is the (*n+1*)-dimensional map of the mean-square error values for all weights. So if we have two weighted connections the error curve will be a three-dimensional planar surface, two dimensions representing the values of the weights and the third, the corresponding error value. The object of learning is to find the minimum on this surface. This is a fairly simple operation for the standard perceptron since the gradient increases monotonically. This means that the shape of the error curve is roughly an elongated u shape in all dimensions, so the learning algorithm just has to choose the direction on the slope that is the steepest, and head that way and it will reach the bottom. The problem comes when a more general error curve is considered, corresponding to a multi-level network.

Rumelhart, Hinton & Williams (1986) present a modification of the standard delta rule known as backpropagation, to allow learning in multi-level networks. The standard architecture for a backpropagation network is much the same as a simple perceptron except for a set of hidden units that intervene between input and output. A typical network is illustrated in Figure 4.2.

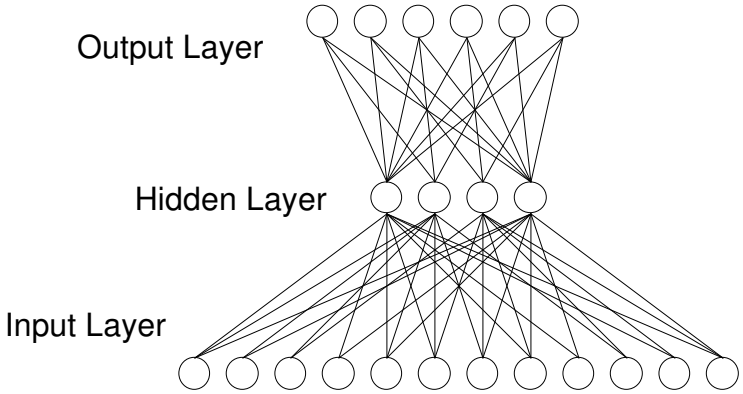

Figure 4.2. A simple feedforward network

The generalised delta rule allows modification of weights between both input and hidden units, and hidden units and output units (see Rumelhart et al., 1986 or Bechtel & Abrahamsen, 1991 for a fuller description). For the rule to work, a number of assumptions are needed. The main proviso is that information flow (i.e. update of activations) is strictly one way, from input units to hidden units to output units. This means that there are no backward connections and no connections within a level as there are in interactive networks (an exception in the case of recurrent networks is examined below). The activation function for the nodes must be continuous and non-linear and is normally chosen to be the logistic activation function.[23] Learning occurs in cycles of activation-update and error

---

[23]The output, o according to this function is given by $o = \frac{1}{1+e^{-x}}$ where x is the summed weighted input to the unit.

propagation. First, the input pattern is presented and a forward sweep alters the activations of all nodes, then a backward sweep alters the weights to reduce the output error compared to a training pattern.

Backpropagation is a powerful tool for learning in connectionist networks and is probably the most popular algorithm currently used in psychological models. However, it does not guarantee to find the best solution given a set of input-output patterns. The problem is that the algorithm used is still a gradient-descent procedure. This means that the algorithm will not necessarily find the lowest point on the error curve; it just goes downhill until it can go no further, at which point it could be trapped in a local minimum rather than the global minimum (the best solution). Figure 4.3 illustrates this point for a simple error curve. Fortunately for more complex networks with many weights, local minima are rare. This is because for a local minimum to occur, it must be a minimum in every dimension of the error space. Normally there will be at least one dimension through which the algorithm can "escape" to a lower error value (see Elman, 1993, for a more detailed discussion of this problem).

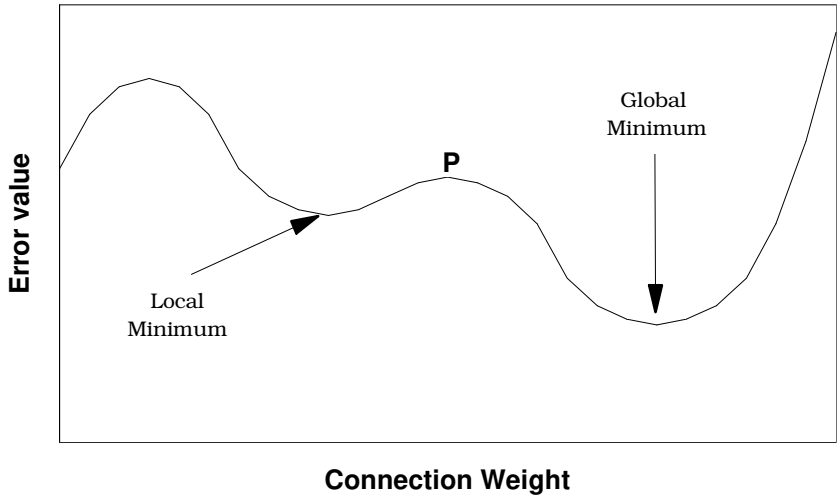

Figure 4.3. A simple error space for a backpropagation network. A gradient descent algorithm starting to the right of point P would find the global minimum. To the left of point P, the algorithm would get stuck in a local minimum.

A limitation of the standard backpropagation network is that it does not represent the temporal order of the patterns it encounters, so there is no way for one pattern to have any effect on the network's response to the next pattern. These kinds of interactions are particularly important in tasks involving speech, because of the temporal nature of the information. Sejnowski & Rosenberg (1987) resolved this problem in their NETTalk model by proposing a temporal window which maps onto the input nodes. Their research demonstrated the ability of a backpropagation network to learn a complex mapping — the mapping from text to speech — in this way. Their network received as input a localist representation of letters and was trained to produce the correct phonetic form corresponding to the input (the training patterns were actually a transcription of the speech of a child). A network with no knowledge of the context of a letter would perform poorly on this task since there is no reliable mapping from single letters to single sounds in English. Because of this problem, the input for each letter consisted of a localist representation of the letter itself, plus its six nearest neighbours. For example, the h in *neighbours* would be presented as [EIGHBOU]. This gave the network enough contextual information to perform reasonably accurately in their task. The trained network was 95% correct on the text used in training and 78% correct on novel text.

There are problems with this approach, since the choice of the size of input window is somewhat arbitrary, and the equal status given to all letter positions in the input window has no psychological grounding. An alternative solution is to build a certain amount of memory into the network itself, rather than the input. Jordan (1986) and Elman (1990) present a variant of the standard feedforward network, the simple recurrent network, that allows some contextual memory.

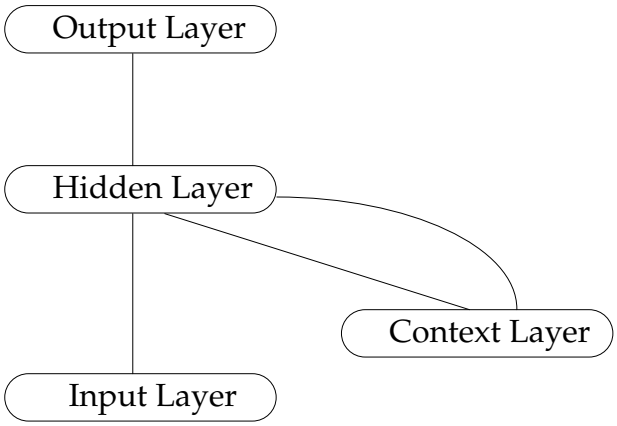

Figure 4.4. A simple recurrent network (Elman, 1990). The curved line indicates that the activations of the hidden units are directly copied onto the context layer.

The Elman network, a variant of Jordan's original architecture, is presented in Figure 4.4. The network is a standard feedforward network, trained using backpropagation, with an added set of units known as the *context* or *copyback* units. These units act as normal input units, but they contain a direct copy of the hidden units at the previous cycle. Thus, the hidden units are always presented with the current input plus a copy of their previous state. This allows context-dependent processing to occur: the output of a trained recurrent network depends on the network's internal representation of the previous input, as well as the pattern currently presented. The advantage of this representation over the static window approach used in NETTalk is that, at least theoretically, there are no limits to the distance over which context effects can occur.

Elman used this architecture to model a number of linguistic processes, finding that recurrent networks are able to discover structure in sequences of letters and words. For example, when trained to predict the next word in simple two or three word utterances, the network developed a representation that discriminated words by their syntactic class as well as semantic category. Servan-Schreiber, Cleeremans & McClelland (1988) examined the learning capacity of the Elman network and found that long-distance dependencies (in terms of temporal separation) could be learned, but that the strength of the context effects and the memory capacity was highly task-dependent.

Pearlmutter (1990) discusses a variant of this system, backpropagation through time. The recurrent links in this system are unravelled, turning the network into a standard feedforward system with multiple copies of the network representing time. This increases the power of the network to capture long-distance dependencies but also increases the complexity of the learning algorithm.

BOLTZMANN MACHINES

All the learning algorithms discussed so far are gradient descent algorithms and thus cannot escape from local error minima. Boltzmann machines offer a solution to this problem, using an algorithm known as simulated annealing (Kirkpatrick, Gelatt & Vecci, 1983; Hinton & Sejnowski, 1986). The algorithm is based on an extension of the Hopfield network described above. The goodness of fit can again be measured using an energy function, but here units have a probabilistic activation function, depending both on their summed input and the "temperature" (T) of the system.

The addition of a temperature factor allows an analogy to annealing crystals. When the temperature of the system is high, the activation function of a unit has a large random element, but at low temperatures this function approximates to a binary, deterministic step function, as in Hopfield networks. The probabilistic element in the activation function allows the system to escape from local minima provided the learning procedure is carried out correctly. Simulated annealing involves initial learning with a very high T value and gradually decreasing the temperature until the network behaves deterministically. Provided the temperature reduction is carried out slowly, allowing equilibrium to be reached at each point, the network should settle into the best solution to the problem.

Hinton & Sejnowski found their network could solve some quite complex problem (see Hinton & Sejnowski, 1986; Ackley, Hinton & Sejnowski, 1985), but that the learning involved was a very time consuming process.

4.2.3 Properties of Connectionist Networks

NEURAL PLAUSIBILITY

The initial reason for studying neural networks was to exploit their similarities with the functioning of the brain. Both Hebb (1949) and McCulloch & Pitts (1965) based their work on their knowledge of the brain. At the simplest level, nodes represent neurons, activation levels represent firing rates, and connections represent synaptic links. The massive parallelisation of information processing is another way in which connectionist models mirror processing in the brain. Feldman & Ballard (1982) argued that since neurons have a relatively slow firing rate, of the order of a few milliseconds,[24] a serial psychological process could have around 100 steps at the most: too few for all but the simplest of operations. This implies that at some level of representation, information is processed in parallel.

The process of learning in connectionist networks also has a strong neural analogy. In Hebb's view, the process of association between nodes mapped directly onto the strengthening of synaptic connections between the relevant neurons. As noted earlier, more complex learning algorithms are more difficult to map onto known neural processes, but the way connectionist networks learn seems to be closer to the constantly changing neural substrate than other learning procedures.

But are all these properties important to psychological modelling? It is quite simple to transform a serial process into a parallel one, or a localist representation into a distributed one. One criticism of connectionist theory is that it is addressing issues at the wrong level, and that psychological theory is not advanced by this rather simplistic analogy between mental models and the substrate of the brain. The validity of this criticism rests on the question of whether operating under these assumptions makes psychological processes more understandable. In other words, do connectionist models have any inherent properties that reflect what is known about psychological processes? This is the question I address in the next section.

PSYCHOLOGICAL PLAUSIBILITY

One problem with any discussion of the properties of connectionist models is that there are so many different types of model, all with distinctive properties. I have already stated that one very general property of connectionist models is that they represent information in a distributed fashion, yet there are many localist models in the field of connectionism, of which TRACE (McClelland & Elman, 1986) is a notable member. This kind of variation causes problems, since it means that the connectionist approach becomes too powerful, being able to account for almost any data and thus almost untestable (Massaro, 1988). Nevertheless, in this section, I outline the general features of most connectionist networks that have found use in psychological models.

Constraint Satisfaction. One of the main traits of connectionist models considered valuable in the modelling of psychological processes is the ability to work with *soft constraints*. This is the ability of a network to handle a large number of possibly conflicting constraints, coming up with the best overall solution to the problem. This kind of model has been very useful in visual perception (McClelland & Rumelhart, 1981), speech perception (McClelland & Elman, 1986) and memory research (McClelland, 1981) since it is able to replicate subjects' ability to handle noisy or incomplete information.

A good example of a soft constraint model is the "Jets and Sharks" model described in McClelland & Rumelhart (1988). This is a localist interactive activation model, hard-wired to model aspects of conceptual memory. The actual information used was a set of properties of the members of two imaginary gangs, the Jets and the Sharks. Six groups of nodes represent the name, gang, age, education, marital status and occupation of each gang member. A set of central nodes represents the gang members themselves, and is connected via facilitory and inhibitory links to the relevant traits.

---

[24]Modern computers, by comparison, can carry out millions of instructions per second.

Activation of a central node causes activation and inhibition of the property nodes, allowing the properties of that particular member to be retrieved. More interestingly, the information can also be retrieved by activation of the property nodes themselves, an example of pattern completion. For example, activation of the name node "George" retrieves the properties of the Jets member George (a divorced burglar).

As well as retrieval from incomplete information, the model is able to deal with noisy information in an informative manner. For example, there is no gang member with the attributes (Sharks, 40's, College, Married, Bookie) but given this input, the network will isolate Don as the gang member most closely fitting this description.

The Jets and Sharks model also demonstrates a number of other features common to many connectionist models. The memories encoded in the network are content-addressable, meaning that they can be accessed via the content of the memory itself, rather than via some arbitrary index. Memories stored on computer are normally accessed by their address (their serial position in the computer memory). Access via some feature contained in this memory normally requires cross-referencing or some similar intermediate process. This aspect of computer architecture has featured in a number of models of psychological processes (e.g., Forster's search model, 1976, 1979), but has acquired little empirical support.

Degraded Performance. As well as viable performance given degraded stimuli, connectionist models (and in particular distributed models) provide a useful analogue of the degradation of psychological *mechanisms*. Most neuropsychological deficits surface as graded loss.[25] For example, in the storage of lexical information, patients will have a general reduction in performance over a large number of stimuli rather than complete loss of a few items and normal performance on the rest. This kind of loss falls naturally out of a system representing information in a distributed fashion. If neurological loss is represented by random damage to a proportion of all nodes or connections, the result will be an overall reduction in performance since any one node or connection plays a small part in the representation of many items.

Similarly, psychological loss at more than one level and through interactions between levels are easily modelled in connectionist networks. Few patients are found with a pure and specific deficit to one psychological mechanism, leaving all others functioning normally. For example, Coltheart (1980) describes some of the errors made by deep dyslexics when reading aloud as visual (e.g., *crag* → "crab"), semantic (e.g., *negative* → "minus"), visual *and* semantic (e.g., *amount* → "account"), visual *then* semantic (e.g., *pivot* → "airplane" [via pilot]), function word substitution (e.g., *or* → "with") and derivational errors (e.g., *grown* → "growing"). Interactions such as these have been modelled using a variant of feedforward network which use recurrent clean-up units to assist in the formation of attractors[26] (Hinton & Sejnowski, 1986; Hinton & Shallice, 1991; Plaut & Shallice, 1993). Hinton and Shallice used this architecture to map from orthography to meaning, looking at just half the transformation needed in reading aloud. Errors such as semantic errors were easy to model since random damage to semantic nodes or connections resulted in small random movements in the mapping onto semantic space. However, the presence of basins of attraction, created by the recurrent links to clean-up units, produced visual and mixed visual and semantic errors as well. To understand how these occurred, we need to look at the map of semantic space in the network shown in Figure 4.5.

---

[25]Specific deficits are claimed, for example, in the naming of animate objects (Hillis & Caramazza, 1990; Sartori, Miozzo & Job, 1993).

[26]In terms of the state space of a connectionist network, an attractor is a point in this space towards which similar patterns are drawn. The area around the attractor is known as its *basin of attraction*.

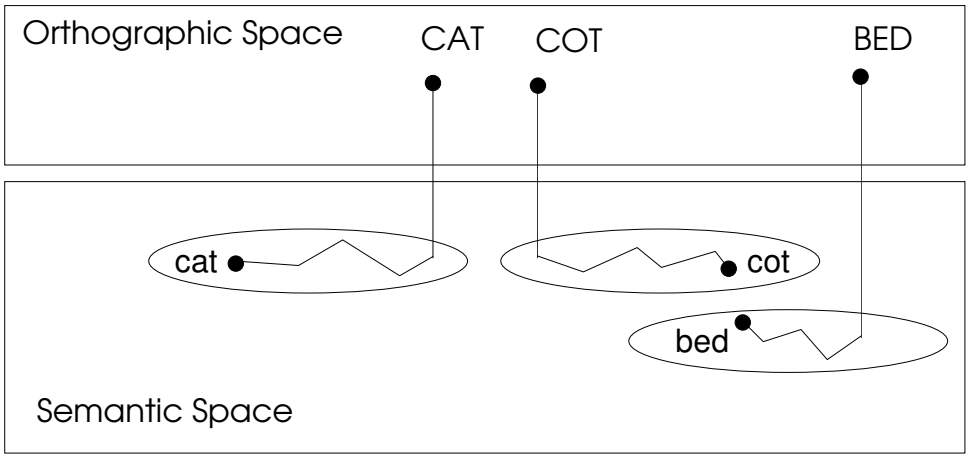

Fig 4.5. Mapping orthography onto semantics (adapted from Plaut & Shallice, 1991). The oval shapes represent basins of attraction formed by the clean-up units. This allows orthographically similar patterns to map onto semantically dissimilar patterns and vice versa.

The purpose of the clean-up units in this network is to create basins of attraction, allowing visually very similar items, such as COT and CAT, to map onto very different areas in semantic space. The oval lines show the edges of the attractor basins — anything mapping onto one of these areas will be drawn towards the centre of the basin. The effect of damage to the semantic units is to blur the edges of the basins, so that input falling near the edge of another word's basin (being a basin for a visually similar word) may be drawn into that basin rather than the correct one. The input is then drawn to the centre of the wrong basin, resulting in an output semantically distant from the correct output. Interaction of this type of error with the standard semantic error results in the mixed visual and semantic errors found. Hinton and Shallice found that a number of different analogues of lesioning such as removal of connections or adding random noise to the weights on various sets of units all produced these dyslexic errors with a probability much greater than chance.

Rules versus Connections. Perhaps the most contentious property of connectionist systems is that they are purported to offer an alternative to the classical symbolic description of a process. Cognitive processes have generally been modelled using abstract manipulations of symbols, related by rules. Mappings that do not conform to these rules are accommodated either by secondary rules or by lists of exceptions. A good example of this kind of description, and one that has been used as a battleground by connectionists and their detractors, is the relationship between an English verb and its past tense.

The majority of English verbs have a very regular relationship with their past tense: the past tense is formed by adding /t/, /d/ or /ɪd/ to the stem, depending on the final segment of the verb. Exceptions to this rule either have no phonological relationship between present and past tense or form a sub-class with separate phonological rules. Rumelhart & McClelland (1986) trained a simple two-level connectionist network to learn the mapping between regular and irregular verbs and their past tense. Their aim was partly to show that a single mechanism could be used to model these relationships and partly to model the developmental literature on past tense learning, which involves a u-shaped learning curve, due to over-regularisation as rules are acquired (see Brown, 1973).

Despite its lack of hidden units, Rumelhart and McClelland's network achieved these objectives with a fair degree of success. The network exhibited over-regularisation during learning; and when fully trained, achieved a reasonable performance in the mapping of both regular and irregular verbs to their past tense. Their model thus represents an alternative to the standard explanation of past tense learning, employing a single mechanism to exploit the regularities between all verbs and their past tense.

Pinker & Prince (1988) attacked this model in a general criticism of connectionist accounts of linguistic processes. The thrust of their argument is that reduction of linguistic processes to rules is the only relevant linguistic description. They criticised the Rumelhart and McClelland model on a

large number of counts, arguing that it is unable to cope with many features of the mapping it is trying to model. However, their analysis takes the model much more literally than its authors originally intended. For example, they use the fact that the trained network produces no past tense at all for a small number of verbs as evidence against the model as a plausible model of linguistic competence. They also point out that the original model is over-simplistic in a number of respects: for example, due to the lack of hidden units, the model does not provide enough computational power to learn all the necessary mappings. The training set was also over-simplistic, and the features used to encode the verbs obscured some of the phonological information on which generalisations could be made. But these are very specific problems, relating to one model (see Plunkett & Marchman, 1991, 1993, for an improved methodology) rather than inherent failings of the connectionist approach.

Pinker and Prince conceded that a more powerful system could solve some of the problems that they cite, but argued that such a system would be merely a connectionist implementation of the rule-based system they propose. This seems to be the most important question that Pinker and Prince raise in their critique. If a single connectionist system is able to model a dual or multi-process mapping, does this make it a single process model? Certainly in terms of mechanism, the same hidden units are employed in both regular and irregular mappings (Seidenberg, 1993). But what functional differences do the connectionist models provide? A connectionist model of a function is able to describe the process using principles of statistical regularity; a rule-based model suggests a more context-free processing environment. Both forms of description are valid and distinct, and each could obscure regularities that the other would emphasise. The vital point is that connectionism does offer a different account of mental processes, which applied to the appropriate task should advance our knowledge of the mechanisms of the mind.

Connectionist explanations appear to be most relevant when exceptional mappings form subgroups, which can be described in terms of secondary rules. In the English inflectional morphological system there are groups of exception words with similar phonological forms which share a common mapping onto their past tense forms. A example of such a group is the set of semi-weak verbs (e.g., sleep-slept, keep-kept, weep-wept). A rule-based description of these regularities is exactly that; it describes the subregularities, but is unable to offer any explanation for these clusterings. A connectionist model provides a description of these regularities in terms of phonological similarity, but it also has the potential to make predictions about the circumstances in which such clusterings should occur, in terms of type and token frequencies of the verbs; as well as offering an explanation of the way these clusterings change over time (Hare & Elman, 1992, 1993) and how they may affect responses when errors are made (Seidenberg, 1993).

The types of processes that are best described in connectionist terms are still a subject of much debate. The argument over the past tense of English verbs rages on (see for example Daugherty et al., 1993; Marcus et al., 1993; Prasada & Pinker, 1993). A similar debate exists over the pronunciation of written words (e.g., Coltheart, 1985; Seidenberg & McClelland, 1989; Plaut & McClelland, 1993; Coltheart, Curtis & Atkins, in press). However, even at this early stage, it seems that connectionist models provide a valuable alternative to rule-based accounts in the description of human performance in these areas.

4.3 Connectionist Models of Speech Perception

In this section I shall review a number of recent models of speech perception based on connectionist networks. TRACE has already been dealt with in previous chapters so I shall not return to it here. The emphasis in this review is on the qualities the chosen architectures embrace, with special consideration of the representation of time and the effect of variation on the models' performance.

4.3.1 Waibel's Phoneme Recognition Model

The lower levels of speech recognition, mapping from spectral information to phoneme identity have been explored by Waibel and colleagues (Waibel & Hampshire, 1989; Waibel, 1989) using a variant of a feed-forward backpropagation network. This task can be viewed as a necessary precursor to lexical access, since most models of word recognition in speech assume some kind of phonemic or featural input representation. The model was presented with an input representation consisting of

frequency spectra of real speech; thus the only pre-processing assumed in this model is a Fourier-type analysis (a frequency transformation of the wave), as is believed to occur early on in the perception of sounds.

The input nodes of Waibel's model were arranged in an array of 16 x 15 units, with the axes of the array representing time and signal frequency. These nodes fed sequentially into two sets of hidden units, each set condensing the information from the input level, with the second set of hidden units connected to a localist output layer of phonemes. The connections between levels in this model were not universal — again the hidden units were organised into arrays, with one column in the first set of hidden units receiving input from three 10 ms time-slices of the input units. The next column would receive input from three time-slices, but shifted along by 10 ms; so the information stored in the first set of hidden units contained time-condensed spectral information with a large overlap between columns. The second set of hidden units compressed the representation of the first set in a similar way, but in this layer, there were only three rows of units, each row connected to a phoneme units in the output layer (in the initial simulations the network was trained to recognise three phonemes: /p/, /t/ and /g/).

The object of this rather complex system of connections was to reduce the dependence of the network's internal representation on low-level temporal factors, which obscure the information required to identify speech sounds reliably. The learning algorithm used — a variant of the back-propagation algorithm — was also chosen to facilitate time-independence. Because of the massive overlapping of connections between layers, the weight changes for one time-slice could be used to constrain weight-changes on others. This reduced learning time and allowed further generalisation over time (see Waibel, Sawai & Shikano, 1989).

The network was initially trained on a small number of tokens of the phonemes /p/, /t/ and /g/. As the output error reduced, the number of tokens used in the input was increased, until the network achieved a good performance (98.5% correct) on 780 tokens. Waibel et al. found that the network was able to generalise well to novel tokens produced by the same speaker, but not to different speakers. They also found that the network performed better on recognition than a hidden Markov model (see, for example, Giachin, Rosenberg & Lee, 1991) trained on the same data.

The work of Waibel et al. emphasises the difficulties involved when dealing with real speech. Their task, a simple one by comparison with the task of human listeners, was to train a connectionist network to recognise tokens of three phonemes produced by the same speaker. Despite the reduction in complexity of the task, the learning time was extremely high (three days of super-computer time). Nevertheless, the performance of the trained network was impressive and the network out-performed a popular and effective stochastic method of learning, the hidden Markov model. Waibel & Hampshire (1989) report a similar network trained to recognise six phonemes but at the expense of a six-fold increase in training time.

This kind of network, however, is not useful in the modelling of the variation found in Chapter 3, since it cannot even generalise well across speakers and was not designed to take into account the interactions between speech segments within or between words. The other models I shall discuss ignore the lower level problems of phoneme or feature recognition, assuming an input representation already analysed along these dimensions.

4.3.2 Recurrent Network Models of Speech Perception

I have already mentioned some promising characteristics of the simple recurrent network (Elman, 1990) for modelling aspects of language perception. Norris (1990, 1992) has developed a similar architecture but applied it more directly to the problems of speech perception. Norris's initial model (see Figure 4.6) examined the mapping between the phonemic/featural level and the lexicon in spoken-word recognition. The representations he used were highly simplified — input consisted of a featural representation of the orthography[27] of the training set, and the output level contained a

---

[27]Although the network was designed to model auditory processes, Norris used an orthographic representation of words, assuming that for the properties he was examining, the form of the representation would be irrelevant.

localist representation of words. As in the Elman network, a set of hidden units, connected to context units, intervened between input and output. The network was trained on a set of fifty words, chosen to allow minimal comparisons between word-initially and word-finally deviating pairs. These words were presented phoneme-by-phoneme to the network, with no gaps between the words. The output was trained to activate the correct word node, with the target pattern present throughout the word. Testing was then carried out on the ability of the network to recognise words.

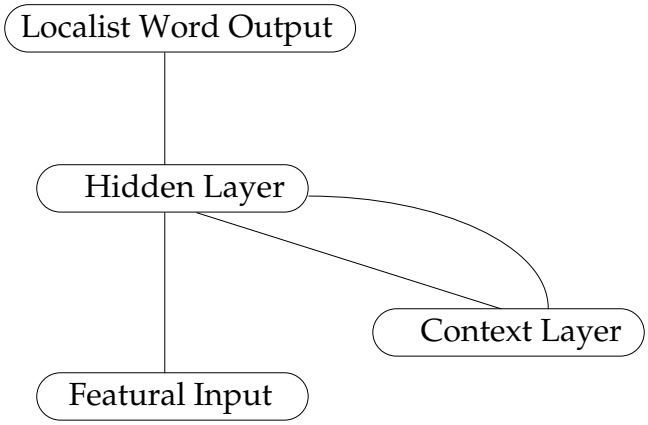

Figure 4.6. The word recognition network of Norris (1990).

The performance of the network fulfils expectations based on the Cohort model (Marslen-Wilson, 1987). In the early stages of the presentation of a word, all the matching candidates become active to some extent; but as soon as the recognition point of the words is reached, the activations of the mismatching candidates drop sharply and the activation of the remaining matching candidate rises to nearly 1 (the maximum activation). Norris also found that the network was intolerant of small deviations. For example, given the input *horonet* (where *coronet* is a member of the training set), the base word's activation never rises above 0.1. However the input *goronet*, where the featural representations of *g* and *c* are similar, activates the *coronet* node almost as well as the base word. While these results may indicate that the network is still not quite as selective as humans appear to be (see Chapter 2), this behaviour is still a great improvement on the performance of TRACE.

The main disadvantage of this model compared to TRACE, as Norris conceded, is that it cannot segment speech in a plausible manner; the problem being that the output of the word units is still in a fairly rudimentary form. Each presentation of a phoneme creates a set of output activations, representing the lexical hypotheses at that point in time. These hypotheses constantly change, and there is no way of knowing from the output of the network whether a reduction in activation is caused by a mismatching segment of input or because the input has gone on to the next word. This problem is compounded by the proliferation of embedded words in speech. Using Norris's example, /kætəlog/ (*catalogue*) contains the embedded words *a*, *at*, *cat*, *cattle* and *log,* and in connected speech may contain more words embedded across the word boundaries. Many of these words will be highly active at some point, and without any information about word length, the network cannot hope to produce a reasonable segmentation of its input.

To solve this problem, Norris (1991) proposed a hybrid architecture, taking a TRACE-like interactive activation network and attaching it to the output of his original recurrent network. The interactive activation network functions in the same way as the word level of TRACE, but with a vastly reduced set of word nodes. Only the most strongly activated words from the output of the recurrent network were used as input to the segmentation network[28]. This implies that the construction of the interactive activation network must be carried out "on the fly", when the intermediate activations of the recurrent network are known.

---

[28]Norris experimented with the number of active candidates used in the segmentation network, finding little effect on the resulting performance. In most cases, using the three most active candidates at each point produced a reasonable segmentation of the speech.

The main advantage of the modified network is that it can deal plausibly with the segmentation of speech, even when the vocabulary of the model is of a realistic size (the lexicons used in most TRACE simulations generally consist of a few hundred words). However, although Norris correctly identified a number of problems with recurrent network models of speech recognition, the solution he proposes seems to share many of the drawbacks of TRACE that led to the recurrent network model being proposed. The real-time construction of interactive networks greatly reduces the simplicity and elegance of Norris's original model and removes the attractive quality of learnability that the simple recurrent network held. In Section 6.6, I examine an alternative solution to this segmentation problem.

Norris (1992) also examined the interactions between phonemes and words using a recurrent network. The network he used was similar to his original model but mapped to a phonemic output as well as the localist word output (see Figure 4.7). This allowed the network to capture interactions between a word and its constituent phonemes during learning. Norris showed that a network trained with this information produced the variant of the Ganong effect found by Elman & McClelland (1988, see Chapter 1), often cited as evidence for top-down processing in spoken word recognition. Norris, however, showed that the essentially bottom-up model in Figure 4.7 could also capture these effects. Coarticulation was simulated by modifying the feature values of the relevant phonemes and a network trained to recognise simplified forms of the stimulus words produced the desired lexical effects.

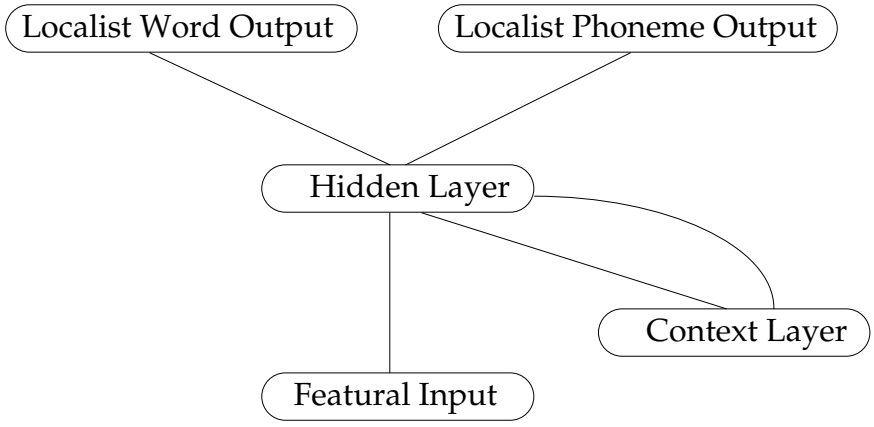

Figure 4.7. Combined phoneme and word recognition (Norris, 1992).

The lexical effects in this network are not simple top-down effects, since during testing, the information flow is strictly bottom-up. But two points are worth noting; firstly, the network could not learn the dependencies needed without backpropagation of error, which *is* top-down information flow, so in one sense the network cannot produce the effects without top-down processing. But if we imagine a non-learning model "hard-wired by God" with exactly the same weights as the trained model, we can only classify this as a strictly bottom-up processor. But Norris draws the distinction between processing flow (top-down vs bottom-up) and the interaction of knowledge sources (interactive vs modular). In the past these concepts have been confused, but here we have a bottom-up processor that shows interactions between the processes of phoneme and word recognition. Indeed, the diagram in Figure 4.7 is somewhat misleading, since the words and phonemes are to a great extent represented in the activations of the hidden units and the connection weights from input and context phonemes to these units. By this interpretation, words and phonemes are largely represented at the same level and on the same units, so that it is unsurprising that they are processed in an interactive manner.

Shillcock and colleagues (Shillcock, Levy & Chater, 1991; Shillcock, Lindsey, Levy & Chater, 1992) have addressed very similar questions with a model based more on Elman's (1990) original network. Their approach to the perception of speech is to take as their null hypothesis the premise that psycholinguistic findings are the product of low-level structural and statistical regularities. Their research is therefore aimed at discovering what can and cannot be modelled in this way, with recourse to higher level structures and processes, such as lexical access, only when a lower level explanation

fails. With this idea in mind, they developed a simple recurrent network to map from featural input to phonemes. The network they devised is illustrated in Figure 4.8.

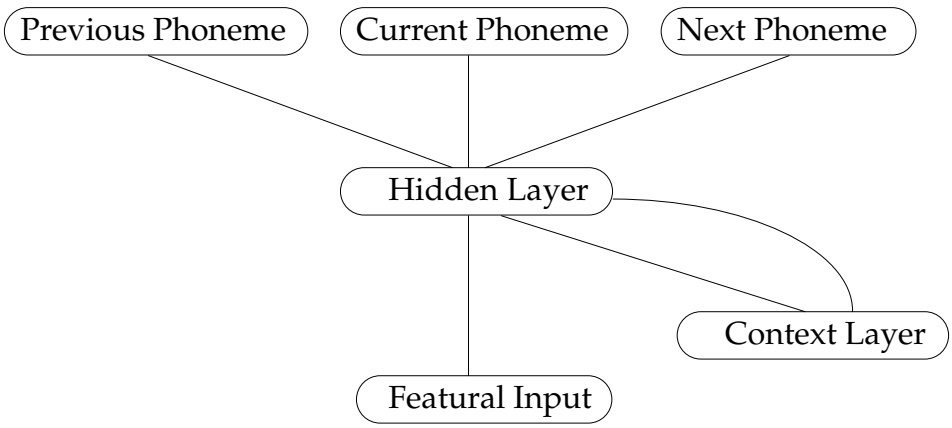

Fig 4.8. The speech recognition network (Shillcock et al., 1991, 1992).

The input to the network was originally represented in terms of standard phonetic features, presented one segment at a time (Jakobson, Fant & Halle, 1952). The task of the network was to identify the phonemes presented as input using three output "windows". The current phoneme window simply identifies the phoneme presented at the input. The previous phoneme window is trained to identify the phoneme presented one time-step back, forcing the network to retain information as it is presented. The next phoneme window is trained to predict the next phoneme in the sequence. Obviously, this task is the most difficult — the only way the network can improve performance on this task is by learning the regularities in the speech stream and applying them to the recent input.

Because the regularities of the speech stream are vital to the performance of the model, Shillcock et al. used a transcribed corpus of natural speech as their training data. The speech was automatically translated from an orthographic corpus of speech, mostly the LUND corpus (Svartvik & Quirk, 1980). However, their training set was quite a small section of this corpus: less than 10,000 phonemes.

Shillcock et al. used the trained network to study the Ganong effect. They found that, if a word was presented to the network with word-final phonemic change (e.g., *gop* for *got*), there was little restoration. But if the word-final token was ambiguous, equally similar to both alternatives, the network chose the option completing a real word.

This restoration effect for ambiguous segments, if not necessarily interpreted as a top-down effect, has previously been explained at least as a lexical effect. But here, the effect is found in a model with no explicit word level: the model is demonstrating an effect of familiar sequences of segments over unfamiliar ones.[29] However, the words Shillcock et al. used were all high frequency short words, which maximised the chances of the effect being found.

Using a revised form of the same model, Shillcock et al. (1992) examined the Elman & McClelland (1988) effect as described above. The network they used was essentially the same as before but input and output representations were in terms of Government Phonology elements (Kaye, Lowenstamm & Vergnaud, 1985, 1990) rather than the more traditional features. The authors contend that the new representation improves the performance of the network, in terms of modelling psycholinguistic data, although they do not explain how this representation confers such an advantage.

The network was trained using the same corpus of spoken discourse, using the back-propagation through time algorithm (Rumelhart, Hinton & Williams, 1986; Pearlmutter, 1990). The trained network was then presented with a transcribed version of the six biasing words used in the Elman and McClelland experiment. Again they found that the network tended to favour the real-word

---

[29]I shall refer to this familiarity effect as a *pseudo-lexical* effect.

completion of the ambiguous tokens. So for the test words *Christma?*, *copiou?* and *ridiculou?* (where the question mark represents a token ambiguous between /s/ and /S/), the response was closer to the output /s/ on the ambiguous word than for *fooli?*, *Engli?* and *Spani?*[30].

Shillcock et al. explained this difference not as a lexical bias as Elman & McClelland had argued but as an effect of the low-level statistics of speech. This explanation could be empirically tested by replicating the Elman and McClelland study using more carefully controlled biasing words, ensuring that the immediately preceding speech is matched across conditions. For example, the carrier words *hospi?* and *waspi?* are matched on the four segments preceding the ambiguity, but should still induce different lexical biases (towards *hospice* and *waspish* respectively).

4.3.3 Recurrent Network Models of Phonology

Although not directly addressing the questions of speech perception, Gasser & Lee (1989) have reported an interesting model of phonological processes in speech. Their model uses an Elman (1990) recurrent network in an attempt to model the acquisition of phonological processes. They address two phonological processes in this way: vowel harmony and pluralisation. Of particular interest in this discussion is the model of pluralisation in English since it employs a network architecture similar to the Shillcock et al. model above (see Figure 4.9). For regular plurals, the process involves affixing either /s/, /z/ or /ɪz/ depending on the final segment of the singular form.

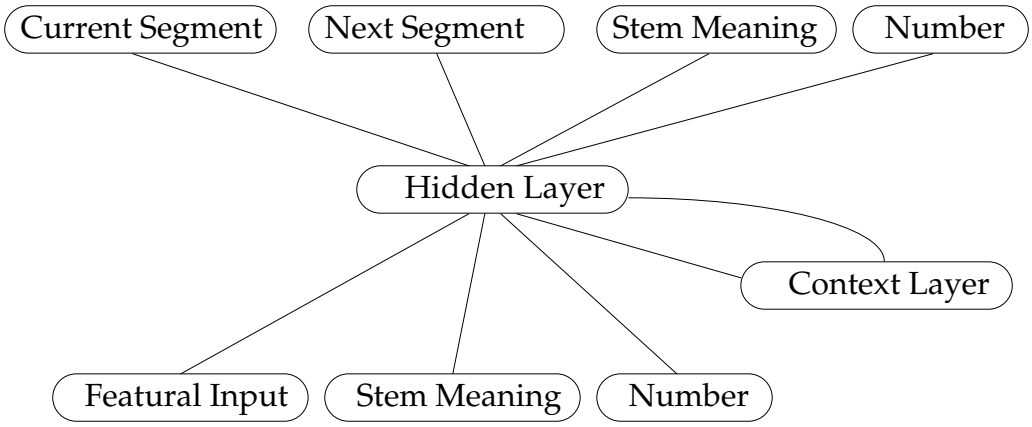

Fig 4.9. Gasser & Lee's (1989) phonological processor.

The training involved presentation of single words segment-by-segment, along with an arbitrary meaning and a value corresponding to the plurality of the word. Singular words were represented with a number of 0 and plurals with a 1. Initially the network was trained on an auto-association and prediction task, with all information available at input. Then the network was trained on a regime where for one word in 5, the plurality input was treated as unknown and the network was trained to identify the number from segmental information. Testing then consisted of presentation of segmental and meaning input with the plurality again unknown. Gasser and Lee found that plural forms that had been present in the training set were correctly classified as plurals and that novel plurals (i.e. those that had only been presented as singular in the training) were correctly classified 7 times out of 8.

This work shows the value of recurrent networks for modelling phonological processes. The network learned to identify the plurality of words in a task similar to the perceptual process of word recognition. However, a number of features of the model and the training corpus make the task much easier than the corresponding task in human speech processing. The training words were presented separated by word-boundary patterns, avoiding the problem of segmenting the input. The amount of information available, both in training and testing was also unrealistic; especially the use of meaning

---

[30]There was also a general bias towards the /s/ response which was attributed to the predominance of this phoneme in the training corpus.

information as input in the testing procedure. Furthermore, as the authors point out, as a model of plurality classification, the model is incomplete since it only addresses regular pluralised words.

4.4 Simulating Phonological Inference

The previous section has uncovered a number of properties of recurrent networks that are important in the modelling of phonological processes in speech perception. The representation of time in this domain is particularly important and simple recurrent networks seem to offer a neat and reasonably effective method of processing a continuous stream of information. The architecture allows on-line generation of hypotheses; and "windowing" the output (e.g., Shillcock, Levy & Chater, 1991; Gasser & Lee, 1989) allows these hypotheses to be modified as new information is encountered. This ability to use close preceding and following context to influence decisions is particularly important for the task I aim to model here: perception of place assimilation. Also, the findings of Norris (1990), at the lexical level, and Shillcock et al. (1991), at the phonemic level, that the processing of recurrent networks is heavily dependent on bottom-up influences suggests that the simple recurrent network architecture could closely model experimental findings that lexical access is intolerant of slight deviation.

In this section I present a model of pre-lexical phonological inference using a network architecture very similar to that of Shillcock et al. to examine the process of compensation for place assimilation. The primary aim of the simulations I report here is to identify the kinds of information a simple connectionist network can pick up and exploit when processing phonologically ambiguous segments.

CUES TO ASSIMILATION — A LOW LEVEL ANALYSIS

The task of the network I intend to examine is the mapping from a raw phonetic transcription of speech to an invariant phonological representation, on which lexical access can be performed. The details of lexical access itself are largely unimportant; the only relevant features are that lexical access is fairly intolerant to variation and that information flow in lexical access is strictly bottom-up: both features argued for in Chapter 2. This view requires that the processor I propose must produce a representation containing the underlying structure of the speech, with the results of phonological processes disentangled.

The provision that information flow is bottom-up excludes one major source of information relevant to this disambiguation, lexical information. Nevertheless, a number of cues can be gleaned from the surface structure of a section of speech. Consider the phrase [əswikg3l] (underlyingly, *a sweet girl*). The task of the network when presented with this phrase would be to output all the segments without change apart from the /k/, which it should recognise as an assimilated /t/ and compensate for accordingly. So the first cue to assimilation is the presence of the /k/ itself. Only a subset of segments are potentially assimilated (i.e. plosives and nasals with a non-coronal surface place), so any segment not a member of this set can immediately be ruled out as a place assimilated segment. The second source of information is in the following context of the ambiguity. For assimilation to occur, there must be a following segment from which the place of articulation migrates. So the /g/ above is another cue to the assimilation of the preceding segment.

A third source of information is contained in the preceding context of the /k/. Assimilations generally occur across morpheme or word boundaries. Thus an ambiguous segment can only be assimilated if it is part of a real word and forms either a morpheme-final or word-final segment. So in the example above, *sweet* is a real word but *sweek* is not, therefore the surface /k/ must be an assimilated word-final /t/.

However, at a pre-lexical stage this information is not available, so a cue of this kind could only be used if the word derivation of the ambiguity was a more familiar sequence of segments than the nonword completion. This is the *pseudo-lexical* effect found in the Shillcock et al. studies, where an ambiguous token half-way between *gop* and *got* was perceived by the network as *got*. It is unclear whether this kind of cue is generally reliable as a source of disambiguation since the two demonstrations of the effect either used small and frequent tokens (Shillcock, Levy & Chater, 1991), maximising the chances of an effect being found, or used a small number of tokens as a test (Shillcock, Lindsey, Levy & Chater, 1992).

4.4.1 A Model of Pre-lexical Compensation

OBJECTIVES

The work reported in the rest of this chapter is an attempt to develop a computationally explicit connectionist model of pre-lexical processing. The main precursor to this model is the TRACE model (McClelland & Elman, 1986), which allowed contextual interactions between temporally adjacent phoneme nodes in its processing of speech. As I have argued in Chapter 3, this system is unable to cope with phonological variation. The model I propose aims to provide a plausible account of the development of phonological rules in speech perception and assesses the viability of a low-level approach to compensation for phonological change. The tests of the model I report here are simplified simulations of experimental work, chosen partly to assess the viability of the model when applied to psycholinguistic data, and partly to examine the relative strength and importance of the various sources of information in phonological processing.

The approach I shall take here is to impose on the network the minimum level of external structure necessary to explain the experimental findings about the perception of phonologically variant speech. For this reason, the model I describe here does not incorporate word or morpheme units in the training of the network. Instead, I examine the extent to which phonological inference in speech perception can be explained simply by exploiting the surface properties of speech. The input to the network is assumed to be a product of a low-level featural analysis of speech and the output is the canonical form of the speech. These two levels correspond to the representation used in standard linguistic rules, normally applied to speech production: the network input is the variable surface representation and the output is the underlying form. Although the input and output of the network is specified and discussed in terms of atomic segments, these are merely convenient shorthand for the feature bundles that make up these units. Indeed, since the phonological process being modelled is a feature-changing process, it is important that the featural information is specified so that the network can make the appropriate generalisations.

Although the model contains no lexical entries, lexical information is used in training the network. Training involves presentation of the underlying form of the 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.[31] 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 utilised. But the experimental evidence from Chapter 3 suggests that lexical access will only be successful if the phonological inference mechanism is already at work. This is a "bootstrapping" problem, common to many learning processes in perception. A solution could 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 involves 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 allow phonological compensation processes to be learned.

NETWORK ARCHITECTURE

The network used in these simulations is a slightly modified version of the Shillcock, Levy & Chater (1991) network illustrated in Figure 4.8. The main modification is 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 I have used the more traditional Jakobsonian features in both input and output representations. This gives the representation enough resolution to capture the phonological processes

---

[31] An alternative approach, more compatible with the model described in Section 6.6, allows underlying representations to develop as a consequence of the mapping between the surface form of speech and representations of the *meanings* of the words. This avoids the use of underlying form information in the training phase.

in speech, and in terms of these simulations seems to have no disadvantage compared to any other phonetic system.

The network uses 11 input nodes, corresponding to the 11 phonetic features used. These are connected to 100 hidden units, which have recurrent connections to 100 context units, allowing context dependent processing to occur. The hidden units are fully connected to 33 output units, organised 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 used here 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 recognised. 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.

The network was trained using the standard backpropagation learning algorithm rather than backpropagation though time, to reduce training time and to keep the architecture of the network simple, and more psychologically plausible, in terms of on-line processing.

TRAINING PROCEDURE

The network was trained to learn the mapping between a phonetic surface form of speech and its underlying representation. Rather than attempting to study phonological variation in general, I chose to model the process explored experimentally in Chapter 3 — place assimilation. In order to allow dependencies between place assimilated segments and their neighbours to be learned, a realistic corpus of speech is necessary, the larger the better. Unfortunately, speech corpora are normally transcribed orthographically, and phonetic transcriptions where they do exist usually do not carry enough detail to include phonological variation.

Because of these limitations 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 & Quirk (1980). The corpus was translated from orthographic to phonetic form using a translation program and errors were corrected by hand. The corpus was unfortunately rather small, being only 3719 words (roughly 12000 segments) but was expected to be large enough for the coarser statistics of speech to be learned.

Place assimilation was artificially introduced in the corpus by randomly selecting 50% of the segments in viable context for assimilation and altering their place to become the same as the following context (using the phonological rules described in Chapter 3). This proportion is close 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 do not occur in a viable context for assimilation) and the proportion of altered segments overall was minimal (less than 0.5%). To simplify the mapping, 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 as input to the network 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, 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. Because of the number of nodes, and thus connections, in the network this was a computationally expensive procedure, taking roughly 20 hours of CPU time on a SiliconGraphics Iris Indigo. 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). The performance of the prediction output window is assessed in Chapter 6.

4.4.2 Simulation 1 — Phoneme Monitoring

In order to test the performance of the trained network with human experimental data, a simulated part-replication of a phoneme monitoring experiment by Koster (1987, Experiment 6) was carried out. This was expected to be a fairly gentle test of the network's ability to model real data.

Experimental Data. The experiment in question (see also Section 3.4.1) used a phoneme monitoring task to examine the effects of place assimilation on subjects' perceptions of word-final coronal segments. Subjects were presented with sets of sentences and asked to press a button when they heard a word containing the segment /n/. The test sentences all contained an assimilable word as the second word, some of which had a lexical competitor in the assimilated form (e.g., *The line/lime broke because it was too tight*). The second set of sentences had no such competitor (e.g., *Which green/greem book do you mean?*). Each sentence was presented in both assimilated and unassimilated forms.

The results of the experiment showed that unassimilated segments were detected significantly quicker than assimilated segments (a mean advantage of 149 ms for native English speakers). There was also a significant advantage of the no-competitor sentences over the competitor sentences. However, it was expected that this advantage would be restricted to the assimilated conditions where the competitor was actually produced. In fact, the advantage found was a general advantage for both assimilated and unassimilated forms and so could be due to some uncontrolled difference between the sets of sentences used in the two conditions.

The aim of the simulation was therefore to replicate the main finding of the experiment, that word-final coronals are recognised more quickly when presented in their canonical form as opposed to assimilated form. Partly due to the inconclusive results for the competitor variable and partly due to the scarcity of *line/lime* type pairs in the speech corpus used, no attempt was made to examine the effect of competitors in the simulation.

DESIGN

The simulation involved 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 or with velar or labial assimilation. The following word always provided a viable phonological context for assimilation. 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, Place (unassimilated coronal vs assimilated labial vs assimilated velar) and Context (unknown vs known).

The concept of a monitoring response time has no direct analogue in the model, since the output of the network is in terms of activations of features and no decision criterion has been specified. It is plausible to assume that the time needed to detect a word-final segment is proportional to the goodness of fit between the output of the network in one of the two relevant windows and the ideal activation pattern of that segment. In the feature set used here, the difference between coronal and labial place is represented by the value of the *grave* feature: a value of 1 representing labial place and 0 representing coronal. The difference between coronal and velar place, however, is represented by the value of two features: *grave* again and *diffuse*. To control for these representational differences, the measure I chose as the dependent variable was the mean deviation from coronal (the underlying place) on the distinguishing features. For labials this was therefore just the deviation for the *grave* feature and for labials it was the average of the deviations found on the *grave* and *diffuse* features.

This deviation was assumed to be negatively correlated to the human response time to monitor for the coronal segment.

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 (see Appendix D). The test words were at least three segments long and had 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. The five context words were, in the representation system used, all three segments long (some were one consonant followed by a diphthong, some consonant-vowel consonant) 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 segment if necessary (e.g., *quite go* vs *quike go* vs *quipe by*).

PROCEDURE

The stimuli were presented in one long list to the trained network. Because of the danger 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.

RESULTS

In general, the performance of the network was accurate. The response of the current and previous windows to the non-test segments were almost all correct. For the test segments, the deviations from coronal response for the six conditions are summarised in Table 4.2.

Table 4.2. Results of Simulation 1. The figures quoted are the mean deviation from coronal response (standard deviations in parentheses).

	Without Phonological Context		With Phonological Context	
Coronal (unassimilated)	0.006	(0.015)	0.023	(0.096)
Labial (assimilated)	0.949	(0.093)	0.511	(0.370)
Velar (assimilated)	0.855	(0.217)	0.451	(0.290)

A two-way item[32] ANOVA was carried out on the data using the variables Place and Context. The effects of both variables and their interaction were highly significant (Place: $F(2,34) = 84.7$, $p < 0.001$; Context: $F(1,17) = 52.4$, $p < 0.001$; Interaction: $F(2,34) = 20.2$, $p < 0.001$). The results are illustrated in Figure 4.10.

[32]For obvious reasons, no subject analyses were possible for the simulations.

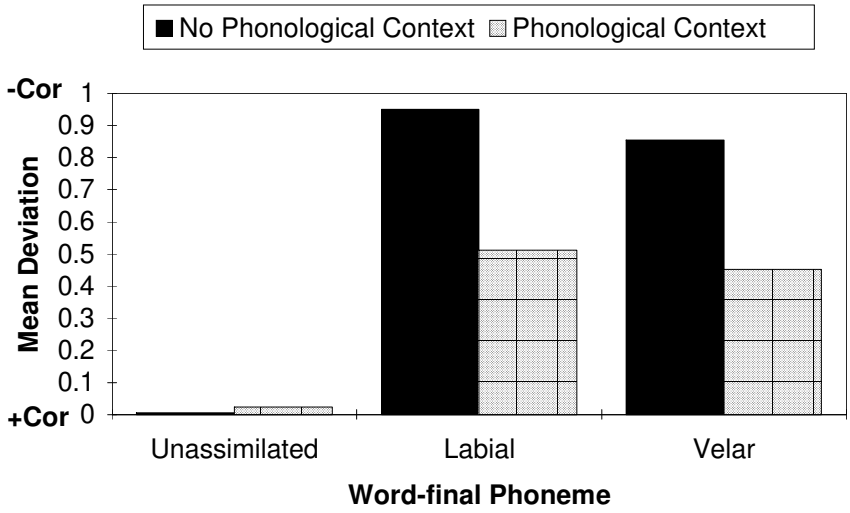

Figure 4.10. Results of Simulation 1. The values are the deviation score for each condition. The higher the value, the less coronal the response.

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 initial weights (generated pseudo-randomly with different starting seeds). The replications showed almost exactly the same error curve during learning and were tested after 100 sweeps through the training corpus. Each replication was analysed 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.

DISCUSSION

As Figure 4.10 shows, the error scores for unassimilated (i.e. coronal) conditions are roughly zero: the network treats coronal segments as completely unambiguous, and so in phoneme monitoring terms, recognises them without delay. The assimilated segments are treated as much more ambiguous: before the phonological context of the segments are known, the network tends to treat assimilated segments at face value, giving a non-coronal response. Once phonological context is known, the deviation scores drop sharply, to 0.51 for the surface labial segments and 0.45 for the surface velars. The network is obviously using phonological context as a cue to the presence of an assimilated word, but equally obviously the network is not using this cue to make an unambiguous underlying coronal response.

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 ***amb****iguous* as well as across word boundaries. Unless lexical information is exploited, the network can at best produce only an ambiguous response. The high standard deviation values for the assimilated conditions with phonological context (0.37 and 0.29) suggest that the output of the network fluctuates strongly between highly coronal and highly non-coronal responses. The distribution of responses bears this out to some extent: 61% of the responses in these conditions were above 0.75 or below 0.25 and only 39% were between these values. This suggests that the network uses the preceding context of the ambiguity to make a best guess as to the underlying place of articulation and selects this as its response.

The variability found in this simulation gives some support to the idea that the network has learned to build up an underspecified representation of the place feature (Archangeli, 1988; Lahiri & Marslen-Wilson, 1991). One empirical implication of an underspecified lexical representation of place is that a segment of speech containing the default feature, [+coronal], can only be mapped onto an unspecified representation, since it mismatches any representation specified for place. In other words, surface coronals can only be perceived as coronals. The simulation shows that this is what happens in

the case of the network: when presented with a [+coronal] segment the network gives a strongly coronal response with minimal variation. For specified features, such as [+velar] and [+labial], to some extent free variation is allowed. They match lexical representations specified for the feature, but they also do not mismatch unspecified representations. Again, the network mirrors this, allowing variation between strongly coronal responses and strongly non-coronal responses. The critical issue that this simulation does not address is the question of how much this variation depends on the viability of the following context. In other words, is the variation found here truly free variation or does it rely on a following context that licenses the change? The next simulation I report addresses this question.

Turning to the ability of the network to model the experimental phoneme monitoring results, we find that both in terms of the amount of information needed to identify a coronal segment and the final activation of that segment on the output nodes of the network, the model would predict a strong response time advantage of unassimilated coronals over the assimilated conditions. The coronal segments were recognised by the network unambiguously before the following context was presented, but the assimilated non-coronals needed the following context for any compensation to occur. Even when the following context was known, the assimilated segments had a much higher deviation score than the unassimilated coronals. This predicts that subjects should be slower to monitor for assimilated segments than unassimilated segments, as was found, but also predicts that there should be a high proportion of non-responses, where the output of the network is closest to a non-coronal. In Koster's (1987) experiment, there was a small increase in errors in the assimilated conditions (14% compared to 10% for the unassimilated conditions) but this was not significant and is nothing like the change in error the model would predict. Taking the network's response to be the segment most closely matching the output of the network once the phonological context is known, the network predicts 47% error in the assimilated conditions compared to 0% error in the unassimilated conditions. Compared with Koster's results, the network tends to under-compensate for assimilation.

However, other studies have found evidence that assimilated segments *are* treated as ambiguous. The gating study of Nix, Gaskell & Marslen-Wilson (1993) found that potentially assimilated tokens similar to the ones used in Koster's study (e.g. *They thought the late/lake cruise was rather boring*) were treated by subjects as ambiguous, with the ratio of coronal to non-coronal responses being roughly 1:1. In this light, the results of the Koster study, particularly in the conditions where there was a lexical competitor (e.g., *line/lime*), are surprising. In Chapter 5, I report a phoneme monitoring study, based on predictions of the model described here, which supports the findings of this simulation, showing that subjects do in fact treat place assimilated segments as underlyingly ambiguous.

In conclusion, the network mirrors human behaviour in the strong asymmetry of its processing, with coronal segments treated as unambiguous and non-coronals evoking more variable responses. Before their following context is known, labial and velar segments are treated on face value, producing a non-coronal response. But once the following context is known to validate place assimilation, these segments are treated as underlyingly ambiguous, sometimes producing a coronal response and sometimes a non-coronal response. These results conform with the finding of Koster that subjects are quicker to respond to unassimilated coronal segments than assimilated ones.

4.4.3 Simulation 2 — The Effect of Phonological Context

Simulation 1 indicates that the presentation of a viable context for assimilation is used as a cue to the presence of an assimilated segment. This effect was found by comparing the responses of the network before and after the following context was presented, so it does not tell us how much this effect is due to the viability of the following context and how much is due simply to the extra time and information another segment of speech allows. Simulation 2 allows a more direct examination of the viability of the following context by comparing the responses of the network to an assimilated segment when both viable and unviable following contexts are presented. This simulation also allows comparison with the cross-modal priming experiments reported in Chapter 3.

DESIGN & MATERIALS

The simulation used the same carrier words as Simulation 1 (see Appendix D). However, the combinations of carrier word and following context were altered so that each item was presented in

six ways. The manipulations involved the place of articulation of both the word-final segment of the carrier word (the target) and its following context. Thus within each item, the variables Following Context (labial or velar) and Assimilation (unassimilated, viable assimilation and unviable assimilation) were manipulated. As in Experiment 2 of Chapter 3, the unviable contexts for assimilation were constructed by matching labial final segments with velar context words and vice versa. An example of each combination is given in Table 4.3. 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.

Table 4.3. Example stimuli for Simulation 2. The target segment is the word-final consonant of the first word.

	Following Context	
Assimilation	Labial	Velar
Unchanged	quite pay	quite go
Viable	quipe pay	quike go
Unviable	quike pay	quike go

RESULTS

The results of Simulation 2 are summarised in Table 4.4 and illustrated in Figure 4.11.

Table 4.4. Mean coronality scores and standard deviations (in parentheses) for Simulation 2.

	Following Context	
Assimilation	Labial	Velar
Unchanged	0.023 (0.096)	0.007 (0.021)
Viable	0.519 (0.365)	0.451 (0.290)
Unviable	0.649 (0.325)	0.577 (0.349)

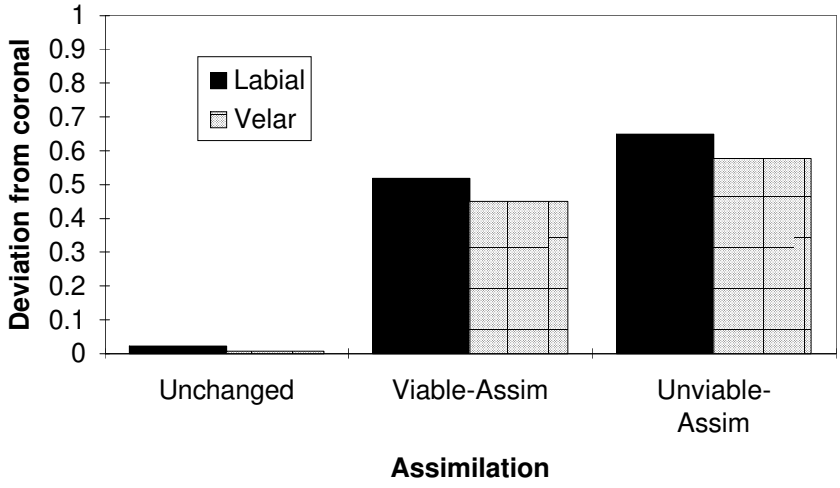

Figure 4.11. Results of Simulation 2.

The data were subject to a two-way item ANOVA, using the variables Context (labial or velar) and Assimilation (unchanged, viable assimilated or unviable assimilated). There was a highly significant effect of Assimilation $F(2,34) = 78.1$, $p < 0.001$ but all other F-ratios were less than 1. A post hoc

Tukey HSD test on the variable Assimilation showed that all levels differed significantly at the 5% level. Most critically, the responses to the viable conditions were significantly lower than the responses to the unviable conditions.

DISCUSSION

Simulation 2 shows 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 assimilated segments, as compared to an unviable context. This implies that the network has learned to identify the phonetic context that allows assimilation to take place in connected speech. Indeed, the unviable contexts were a stiff test of the network's knowledge, since they differ from the viable context only on the value of the *grave* feature.

The similarity between the viable and unviable contexts may explain why the unviable context still provoked a certain amount of compensation for assimilation, reducing the coronality of the responses compared to the no-context conditions of Simulation 1. This is another example of the graded response 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. If I had chosen a vowel, for example, to represent the unviable phonological context it is likely that the amount of compensation found would be much reduced, since the phonetic representations of most vowels are very dissimilar to non-coronal stop consonants.

This graded behaviour may seem undesirable when comparing the simulation results to the experimental results in Chapter 3. Experiment 2 used the same combinations of non-coronal segments to construct the unviable sentence conditions used in the cross-modal priming experiment. These sentences produced a strong mismatch effect when compared to both unchanged sentences and viably assimilated sentences. So there is little evidence of a graded response in the experimental results so far. However, the experimental task makes a direct comparison with these data rather difficult. Cross-modal priming is a measure of lexical access and so the graded effects this model predicts pre-lexically may be obscured by the more dichotomous nature of lexical access. This is a possibility that is explored further in the phoneme monitoring experiments reported in Chapter 5.

4.4.4 Simulation 3 — Lexical Effects in Pre-lexical Processing

The third potential source of evidence for the presence of assimilation comes from the preceding context of a segment. This is a *pseudo-lexical* effect caused by the preference of the network for familiar sequences of segments over unfamiliar ones when encountering an ambiguous segment. When confronted with two options, as is the case for an assimilated segment, the network will prefer the option completing a real word (a familiar sequence) over a nonword (an unfamiliar sequence). The two demonstrations of this effect in a recurrent network of this type were by Shillcock et al. (1991, 1992). The first used short frequent words such as *got* and *yes*, maximising the chances of observing the desired effects. The second demonstration was intended as a replication of Elman & McClelland (1988) and so used their stimuli. These words were longer and less frequent than the words used in the original study, (e.g., *Christmas*, *copious*), but only six words were used.

It is still questionable, therefore, whether dependencies such as these provide reliable and effective cues to the resolution of ambiguities in speech. This simulation is a more rigorous test of the abilities of a simple recurrent network to resolve surface ambiguities using preceding context. The ambiguity used in this simulation is caused by the neutralising effect of assimilation, rather than the phonetic ambiguity used in the Shillcock et al. experiments, but this is not expected to have any influence on the effects of lexical status.

DESIGN & MATERIALS

The test items used in this simulation consisted of pairs of words, of which the first word was the ambiguous test word and the second provided viable phonological context for assimilation (see Table 4.5). The 88 test words (Appendix D lists the stimulus words) were drawn from the training set and

were divided equally into two groups. The -COR group had a word-final labial or velar segment and so would provide a lexical bias towards the non-coronal surface form of the word-final segment (e.g., *break*). The +COR group had word-final coronal segments but were presented in assimilated form, providing a lexical bias towards the coronal underlying form (e.g., *sweek*). 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*). Half the ambiguous segments had a labial place and half velar. The frequencies of each test word in the training corpus were calculated. The +COR group had a mean frequency of 3.04 with s.d. 3.57 and the -COR group had a mean frequency of 2.84 with s.d. 3.43. This means that, as the corpus was presented to the network 100 times during training, the words used in the simulation had been presented to the network roughly 300 times. The conditions were matched for word length, with the mean number of segments in the biasing word being 4.4.

The network's responses to the ambiguous segments were measured as the following context was presented. Details of procedure were the same as Simulation 1.

Table 4.5. Example stimuli for Simulation 3. The italicised phrases are the underlying words assuming a place assimilated coronal.

Place	Lexical Bias	
	+Cor bias	-Cor bias
Labial	* nime pay (*nine pay*)	name pay (* *nane pay*)
Velar	* dake go (*date go*)	luck go (* *lut go*)

RESULTS AND DISCUSSION

The response of the network to the phonologically ambiguous word-final targets were recorded at the point of presentation of the following segment, allowing the phonological viability of the context to influence the responses. The mean deviation from coronal scores are presented in Table 4.6.

Table 4.6. Deviation from coronal scores. 0 = coronal response, 1 = non-coronal response.

	Lexical Bias	
Place	+Cor bias	-Cor bias
Labial	0.398	0.573
Velar	0.616	0.67

The data were analysed using a two-way item ANOVA with the factors Bias and Place of articulation. There was a significant effect of the place of articulation of the ambiguity ($F[1,84] = 4.59$, $p < 0.05$), with labial segments being perceived by the network as less coronal than velars. There was no effect of the lexical bias of the carrier word ($F[1,84] = 2.44$, $p > 0.10$) and the interaction between the variables was not significant ($F[1,84] < 1$).

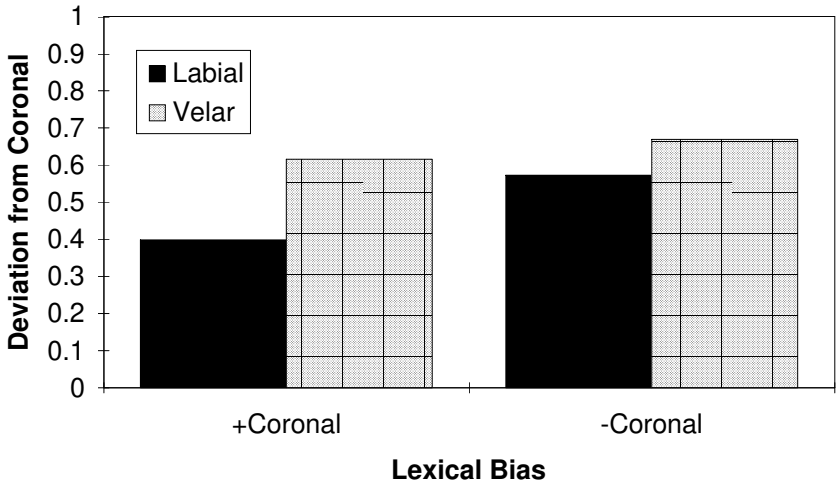

Figure 4.12. Simulation 3 results.

There is a trend (see Figure 4.12) for the +coronal biased conditions to be treated by the network as underlyingly more coronal than the -coronal conditions. However, this effect is mainly restricted to the surface labial conditions, where the compensation process involves alteration of just one phonetic feature (the value of the *grave* feature), and is not statistically significant.

This result contradicts the findings of Shillcock et al. (1991, 1992) who were able to explain lexical effects using only low-level statistics. One important difference between this simulation and the Shillcock et al. studies is that their bias effects were only found when the test segments were phonetically ambiguous. Here the test segments are phonetically unambiguous but, as Simulation 2 shows, are treated as ambiguous given viable phonological context. Therefore, as in the Shillcock et al. studies, the network must rely on context to resolve the ambiguity.

The lack of any biasing effect may reflect the more stringent test posed by the stimulus materials in this simulation. The biasing words here were not selected to be particularly short or frequent, as was the case for the Shillcock et al. (1991) study, and the size of the test set and the range of words used makes this simulation a more general test of the network's abilities than the Shillcock et al. (1992) study which used only 6 test words.

Three explanations of these findings seem plausible. One is that the size of the corpus used in training the network is too small to allow reliable dependencies to be built up. A second possibility is that there is insufficient information in the surface structure of speech to allow lexical effects to be found, regardless of the size of corpus selected. A third possibility is that the architecture or memory capacity of the simple recurrent network used in these simulations is not powerful enough to exploit the statistical patterns embedded in speech. At present it is impossible to choose between these alternatives, but in Chapter 6 I report some further simulations exploring these and related issues.

4.5 General Discussion

The three simulations reported here assess the uptake of various cues to the presence of assimilation by a recurrent network model of phonological processing. The results of the simulations show that the place of articulation of a segment strongly affects the network's response: coronal segments are invariably mapped onto underlying coronal representations, but labial and velar segments are treated as underlyingly either unchanged, ambiguous or coronal. The viability of the phonological context of a non-coronal segment also affects the operation of the network, with changes conforming to assimilation rules producing more compensation than those violating those rules. But these effects, while significant, are much smaller and appear more graded than the results of Simulation 1. Finally, Simulation 3 failed to find a clear pseudo-lexical effect of preceding context on the perception of phonologically ambiguous segments. This finding may be due to the inadequacies of the training corpus or recurrent network employed in these simulations, or may be an indication that the

relationships between word-final ambiguous segments and their preceding context are too variable to be usefully employed as cues to disambiguation.

Where does this leave the model of phonological disambiguation proposed in this chapter? The behaviour of the network as it stands suggests that a pre-lexical phonological inference model can use phonological context to identify likely points in speech in which assimilation and other phonological processes occur, and represent these in the output of the process as ambiguous segments. The cues to assimilation learned by the network allow the model to minimise the number of phonemic ambiguities, with remaining ambiguities resolved by a process of selection based on lexical information such as lexical status, and semantic and syntactic biases. Thus, the network represents an explicit model of pre-lexical phonological inference in spoken-word recognition.

However, the application of this model to the cross-modal priming studies of Chapter 3 causes some problems. The phonological viability effect found here agrees in direction with the effect found in Experiment 2, but it seems unlikely that the small effect the network displays could translate to the very strong viability effect found in the experiment. Similarly, the results of Experiment 1, where no mismatch effect was found when the following context of the phonologically ambiguous segment was spliced out, are difficult to explain. The probe point in Experiment 1 is analogous to monitoring the output of the network to an assimilated segment before the following context is presented. At this point, the network almost invariably opts for the surface form of a non-coronal segment. The coronality scores in Simulation 1 before following context was presented were 0.95 and 0.86 for labial and velar segments respectively. These segments are only interpreted as ambiguous by the network once the following context is known to validate assimilation. Thus the model would predict a mismatch effect for Experiment 1, which would only be reduced in the viable context condition of Experiment 2.

Nevertheless the predictions of the model described here seem worth pursuing. The inconclusive nature of the above predictions are partly due to the problem of mapping these predictions onto experimental findings using lexical access as their basis. The effects may become clearer if the process of phonological inference is separated as far as possible from lexical access. Arguably, phoneme monitoring allows this separation to be made, by forcing subjects, at least in part, to respond to sub-lexical or non-lexical units. As such, phoneme monitoring may allow a more direct mapping between the process being modelled here and experimental results, and so should be a more reliable test of the model's predictions.

So what predictions can be gleaned from the behaviour of the network in these simulations? One predominant property of this model is the graded quality of the effects found here, rather than the dichotomous nature of rules and representations common to standard phonological theory. For example, Simulation 2 implies that the validity of a phonological context for assimilation appears to depend, to some extent, on the number of features by which the segment deviates from the ideal viable context. Similarly the interaction of cues to the presence of phonological change seem to be additive, rather than all-or-nothing as linguistic theory would predict. These properties are typical of the processing environment a distributed connectionist model provides, and produce testable predictions of human performance.

Most critically, this model predicts that phonological inference will occur irrespective of whether lexical access is successful. Even if this model had exhibited a pseudo-lexical effect, it would still show phonological inference effects in nonwords, especially given a viable phonological context. The next chapter explores this prediction, using phoneme monitoring to examine subjects' perceptions of phonologically changed speech.

Chapter 5 — The Locus of Phonological Effects

5.1 Introduction

Chapter 4 presented a model of speech perception in which compensation for assimilatory phonological change is not an all-or-nothing process. Partial compensation occurs when some, but not all, the conditions necessary for assimilation are satisfied. This approach predicts that phonological compensation should occur when lexical access fails, in the perception of nonsense words. This prediction is tested in the two experiments reported here.

A second aim of these experiments is to look at what people actually *hear* when they are presented with phonologically changed speech. Experiments 1 and 2 showed that phonological inference is a vital component of word recognition, but there is more to speech perception than merely extraction of the underlying meaning of an utterance. Spoken words produce an auditory *percept*, which involves the form of words as well as their meaning. Given that phonological inference occurs during word recognition, it is relevant and important to find out how much this inference process affects peoples awareness of the words they hear. Do listeners have access to the surface form of phonologically variant speech, or do they rely on a more abstract underlying representation when making judgements about what they hear? These questions are particularly pertinent in the perception of nonwords, where the auditory percept involves little or no aspect of meaning.

To investigate these issues, the two experiments I report here employ the task of monitoring for segments in phonologically variant speech: Experiment 3 involves monitoring for place assimilated coronal segments, and in Experiment 4, subjects monitor for the following context of these changes. The experiments show that, to some extent, subjects rely on an abstract phonological form of speech when making monitoring judgements. As in Experiment 2, there are strong effects of phonological viability, this time in both words *and* nonwords, supporting the prediction of the network model that phonological inference occurs even when lexical access fails. However, the phoneme monitoring experiments also find evidence suggesting that phonological inference involves interaction between lexical and phonological constraints.

Both the representational units and the mechanisms underlying phoneme monitoring are the subject of much controversy. The experiments in this chapter rely on the assumption that decisions based on information about phonological form are not completely dependent on the outcome of the word recognition process. The following section, therefore, reviews the psychological literature relating to this issue.

5.2 Phoneme Monitoring

The phoneme monitoring task has been used for many years to measure aspects of speech comprehension. Subjects are presented with sentences or lists of words and asked to press a button when they hear a particular speech sound; for example, monitoring for the /k/ in *She ran into the kitchen*. The task was developed by Foss and colleagues, initially as a tool to assess processing load at different points in the comprehension of a sentence. For example, Foss & Lynch (1969) showed that the monitoring time for a phonemic target increased as a function of the structural complexity of the preceding sentential context. Other studies have used the same tool to examine the effects of word frequency (Foss, 1969), lexical ambiguity (Foss, 1970) and verb complexity (Hakes, 1971). For the purposes of these studies, the actual mechanisms by which a subject made a response to a phoneme were assumed to be largely unimportant. Only when focus was turned to questions of lexical access did the mechanisms involved become important. The traditional view of the status of the phoneme is that words are made up of strings of phonemes, which are in turn composed of phonetic features. Thus, it is natural to assume that lexical access mirrors this structure: feature analysis is followed by phoneme recognition, which allows matching with lexical entries. According to this view, phoneme monitoring reflects the state of activations of these pre-lexical units.

A number of alternatives to the pre-lexical status of the phoneme are possible. One is that phonemic or phonological information only becomes available once lexical access is complete (e.g., Klatt, 1989). By this account, lexical access is mediated by lower-level structures such as phonetic features or spectral information, and phoneme monitoring is only possible once the phonological information about the recognised word is retrieved and post-lexically compared to the phonemic target. Other researchers have proposed that phonemic information is available from both pre-lexical analysis and lexical knowledge (e.g., Cutler & Norris, 1979; Foss, Harwood & Blank, 1980). These theories predict that a phoneme monitoring decision utilises information from either or both of these routes, depending on their relative strength and speed in each case.

These models address two separable issues. The first is the level in the system at which phonetic and phonemic information becomes available (pre-lexical, post-lexical or both). The second issue relates to the form of this information, which is often assumed to be phonemic, but which could quite plausibly be in terms of finer grained units (e.g., features) or higher level units (e.g., syllables).

ROUTES TO PHONOLOGICAL INFORMATION

Many studies have attempted to answer the first question by looking at the effects of lexical information on phoneme monitoring times. Strong lexical effects indicate that the task depends on a post-lexical read-out mechanism whereas no lexical effects suggest a pre-lexical status. Morton & Long (1976) provided evidence for the post-lexical account when they demonstrated that subjects were influenced by the predictability of the carrier word (i.e. the word carrying the target phoneme) in its preceding context. For example, subjects were quicker to monitor for the /b/ in sentence (1) below than in sentence (2).

(1) He sat reading a book until it was time to go home for his tea.

(2) He sat reading a bill until it was time to go home for his tea.

This finding seems to imply that a monitoring response is at least partially dependent on the recognition of the carrier word, which in turn depends on the predictability of the word in its preceding context. However, Foss & Gernsbacher (1983) showed that this result was in fact due to a confounding effect of vowel length. They replicated the Morton and Long study, but presented the target words (e.g., *book* and *bill* above) without the biasing context. The results showed that the supposed biasing effect was still present, which Foss and Gernsbacher attributed to differences in vowel length between the two conditions.

The effects of word frequency are also unclear. In Foss's original study (1969), he found a significant effect of the frequency of the word preceding the carrier word. Foss, Harwood & Blank (1980) replicated this finding but found that there was no frequency effect of the carrier word itself. Segui & Frauenfelder (1986) also found no effect of the carrier word frequency on monitoring time.

These findings, among others, led Foss et al. to propose the dual code theory of speech perception. This theory predicts that two types of phonological information become available during lexical access. One is a pre-lexical representation, derived from the speech waveform and loosely corresponding to the phonetic level of description. This representation is assumed to preserve non-distinctive information, such as aspiration in English. The second becomes available after lexical access and represents the phonemic form of the word. The extent to which these two codes are used in phoneme monitoring depends on the time available for lexical access. Foss et al. predicted that the manipulation of factors such as task difficulty should alter the relative strength of each route, but they found little empirical support for this hypothesis.

Cutler & Norris (1979) proposed a similar model of retrieval of phonological information, the race model. This again hypothesises a pre-lexical and a post-lexical mechanism for phoneme identification, with a "race" between the two processes to identify the segments of speech. Their model is not specific about the types of information available from the two processes, but outlines the conditions under which each of the two routes will be dominant. Their rules state that in normal circumstances, the pre-lexical route will dominate, but if word identification is particularly fast, either due to contextual facilitation or by using a short carrier word, the lexical mechanism will be dominant.

A crucial factor, both in the debate over mechanisms for phoneme monitoring and for the experiments I report here, is the effect of lexical *status* on response times. For nonwords, lexical information is unavailable, and so according to a dual-route account, the phoneme monitoring responses should reflect only the output of the pre-lexical mechanism. Foss et al. (1980) examined the effect of the lexical status of both the carrier word and the preceding word on response times. Their reasoning was that the presence of a nonword immediately before the word-initial target should slow responses, since subjects were asked specifically to monitor for word-initial targets, and the position of the target within the carrier word is made ambiguous by the previous nonword. But in the condition where the target itself is part of a nonword, a pre-lexical mechanism can respond before the full nonword is perceived, suggesting that the inhibitory effect of the nonword should be smaller. This is the pattern Foss et al. found: when the lexical status of the carrier word was manipulated there was no effect on the response times, but when the lexical status of the preceding word was manipulated the nonword condition produced responses 100 ms slower than the real-word condition.

The lack of an effect of lexical status of the carrier word is very difficult to explain on the basis of a lexical code alone. But again there is evidence that contradicts these findings. An earlier study by Rubin, Turvey & Van Gelder (1976) found a significant inhibitory effect on monitoring response times when the target formed the initial phoneme of a monosyllabic nonword. Foss et al. explained this as an effect of increased task difficulty — the Rubin et al. experiment required subjects to monitor for two targets at the same time (/s/ and /b/) whereas the Foss et al. experiment used only one target. Indeed, a replication of the Rubin et al. study with only one phoneme as a target (Rubin, 1975) found no effect of lexical status. Also, the length of the words used in the Rubin et al. experiment may be an important factor. The carriers were all monosyllables, which according to the Cutler and Norris race model should facilitate the lexical route to phoneme identification. Segui, Frauenfelder & Mehler (1981) examined the word/nonword effect using bisyllables and, like Foss et al., found no effect. Also, a series of studies by Cutler, Mehler, Norris & Segui (1987) examined the influence of word length and other factors on the lexical status effect. They found that even for monosyllables, there was sometimes no inhibition for the nonword condition. The discriminating factor they found was the monotony or homogeneity of the test and filler set. They argued that when a stimulus set is particularly monotonous (for example, when all test and filler words are monosyllables) and the task has no element of comprehension, subjects tend to ignore the lexical output, concentrating solely on the pre-lexical analysis of the speech in order to make their judgements.

In some circumstances, the use of nonsense words can actually facilitate recognition of a target phoneme. Marslen-Wilson (1984) reports a set of experiments in which subjects were required to monitor for phonemes at various positions in words and nonwords. For the words, there was a strong effect of the position of the target within the carrier word: word-initial phonemes were monitored most slowly, and response times decreased as the position of the target was shifted towards the end of the word (as measured from the onset of the target phoneme). This was explained in terms of variation in the amount of lexical information available at each particular point. For the nonwords (assessed in a separate experiment), the response times remained constant over various target positions: subjects were actually quicker to monitor for initial phonemes in nonwords than for the corresponding phonemes in the real word experiment, but there was no decrease in response times for the later targets. Again, this is evidence that subjects are able to adopt a monitoring strategy to fit the situation. In the word experiment, the lexical route to phonological information predominated; meaning that, even though word-initial targets would be more quickly identified using a non-lexical route (as was shown by the nonword experiment), this did not occur.

What conclusions can be gleaned from this rather confusing set of results? Firstly, the studies showing little or no effect of lexical status on response times (e.g., Foss et al., 1980; Cutler et al., 1987) indicate that there must be a mechanism by which information about phonological form can be retrieved when lexical access fails. So it seems untenable to suggest that all the results reviewed here can be explained on the basis of a purely lexical analysis. The opposite conclusion, that all the results can be explained pre-lexically, relies on an explanation of seemingly lexical effects as artifacts of regularities in the speech stream. A similar argument was employed by Shillcock, Levy & Chater (1991) in their simulations of the Ganong (1980) effect. This is an engaging possibility, but at the moment is highly speculative. A more likely explanation is that phonological form decisions can be made using both lexical and non-lexical mechanisms, as in the dual-route or race model. The extent

to which these routes contribute to any particular response depends on numerous factors, such word-length, lexical status, word frequency and even the monotony of the test stimuli.

FORMS OF PHONOLOGICAL INFORMATION

It has been traditionally assumed that the phoneme is the dominant unit of phonological form in speech perception, but there is increasing evidence that lexical access is based on more fine-grained units such as phonetic features. Phoneme monitoring studies (Newman & Dell, 1978; Dell & Newman, 1980) have shown that false alarms (positive responses to non-target segments) and response rates are directly related to the featural similarity between the target phoneme and the phoneme eliciting the false alarm. For example, Dell & Newman (1980) compared responses to sentences such as (3) and (4) below where the target phoneme is the word-initial /b/ of *beach,* which is preceded by a word with either a similar initial phoneme (*private*) or a dissimilar one (*secret*).

(3) The surfers drove out to a *private beach* to try out the waves.

(4) The surfers drove out to a *secret beach* to try out the waves.

They found responses in the similar condition were roughly 60 ms slower than the baseline condition, and that the preceding word-initial phoneme (/p/ here) provoked a 4.2% false alarm rate, compared to none for the dissimilar condition. Not only is this further evidence that sub-lexical information is used in phoneme monitoring (since the similarity that causes these effects is between the distracter and target *phonemes* rather than the words themselves), it also suggests that featural or feature-like information is still present in the system when phoneme monitoring is carried out.

This research conforms with a set of gating studies carried out by Warren & Marslen-Wilson (1987, 1988). These examined the uptake of phonetic cues in lexical access by presenting subjects with gradually increasing gated sections of monosyllabic words. The responses of the subjects provided no support for a view of word recognition in which information is segmented in some way prior to lexical access (such as into phonemes or syllables). Instead they proposed that speech information is mapped onto lexical representations continuously, so that word candidates can be selected as efficiently as possible.

A valuable source of information on the identity of pre-lexical units comes from research into subcategorical mismatch (Streeter & Nigro, 1979; Whalen, 1982, 1984). Subcategorical mismatches are the effects on the perceptual system of conflicting phonetic cues to phoneme or word identity. Streeter & Nigro (1979) used splicing techniques to create tokens of speech with conflicting transitional cues to the identity of critical consonants. For example, the initial consonant and vowel of the word *faded* was spliced onto the final syllable of the word *fable*. Thus, the initial vowel contained cues to the place of articulation of the following consonant which conflicted with the information from the transition out of the consonant. In an auditory lexical decision task, these subtle mismatches were found to increase response times for words, but not for nonwords. Streeter and Nigro concluded that the lexicon is searched using a detailed representation of speech rather than some pre-categorised form (but see Whalen, 1991).

Marslen-Wilson & Warren (submitted) used a modification of this technique to address the same issues. Their study also employed cross-spliced tokens, this time examining all combinations of words and nonwords in the two components of each token. The critical tokens were the ones made up from two nonwords, for example, combining the initial consonants and vowel of *smod* with the final consonant of *smob*. If features are pre-lexically integrated into phonemes these tokens should cause mismatch due to the conflicting nature of the cues to the identity of the final consonant. But Marslen-Wilson and Warren found, both in a lexical decision task and a forced-choice phonetic categorisation task, that these tokens caused no disruption of the perceptual processes, compared to an unambiguous nonword baseline. They argued that these results can only be accommodated by a model of lexical access that maps directly from featural information to the lexicon.

SUMMARY

The above research indicates that subjects' judgements about phonological form appear to be based on a detailed, possibly featural, representation of the speech they hear. It also appears that this featural representation is mapped directly onto the lexicon during word recognition, with no mediating phonemic structures. This result seems at odds with the finding that, in certain circumstances,

phoneme monitoring can reflect non-lexical routes to phonological information. How can phoneme monitoring tap non-lexical or sub-lexical processing of speech, when the phoneme seems to have no basis as a unit of speech perception?

One possibility is that phoneme monitoring does not probe activation of phonemic units, but that it does reflect the overall state of featural activations prior to lexical access. The phoneme, according to this view, is merely a localist approximation to the detailed information actually used in the process of lexical access. A possible mechanism underlying this kind of position is presented in the connectionist model of Norris (1992; see Chapter 4) in which phonemes and words are represented at the same level, and phoneme recognition can occur independently of word recognition, but will normally be strongly influenced by the word recognition process.

The account I have argued for here, where detailed featural information is mapped directly onto the mental lexicon also seems to conflict with the notion that phonological inference can be pre-lexical. It is usually assumed that the more detailed a representation is, the more variable it is. This account was used to explain the presence of two processes by which phoneme monitoring can be carried out in the dual-route model of Foss, Harwood & Blank (1980). They proposed that the pre-lexical code was a detailed phonetic one, preserving non-vital information such as aspiration in English. The lexical code, in contrast, was assumed to eliminate redundancies, containing only the phonemic representation of words. By this account, effects of phonological inference should only be found at the lexical level of representation. However, this aspect of their model was largely based on a desire to maintain cognitive economy by avoiding duplication of information rather than any empirical data.

5.3 Experimental Considerations

The previous section indicates that phoneme monitoring can reflect both lexical and non-lexical phonological activations, depending on a number of factors. It also seems that the representation of speech used in lexical access is feature-based and detailed, with the concept of the phoneme only existing in terms of a localist approximation to a distributed featural representation. In this section I intend to use phoneme monitoring to assess, among other issues, the degree of abstractness of the phonological code. Following Koster (1987; see Chapter 4), I intend to use as my experimental task phoneme monitoring for underlying coronal segments. The experiments involve presentation of sentences such as (7) below, which contain a word-final consonant embedded in a carrier word (*clean* here).

(7) The city got two awards for its *clean* parks.

In Experiment 3, the target is either presented in canonical form ([n]) or with a place change such as could occur as a result of place assimilation (i.e. [N] or [m]). Experiment 4 uses similar sentences to examine the effects of assimilation on recognition of the phonological context of assimilation (/p/ here).

One aim of this work was to replicate the main finding of Experiments 1 and 2 that the perceptual system is sensitive to the viability of deviations in their phonological context. If a viability effect is found here, using a different experimental task (phoneme monitoring rather than cross-modal priming) it will support my assertion of the significance of phonological inference in speech perception. For this reason the experimental contrast was made between sentences with viable assimilation (e.g., [klimpaks], *cleam parks*) and those with phonologically unviable ones (e.g., [klimgesthaʊzɪz], *cleam guesthouses*).

Phoneme monitoring also allows a more direct mapping between the computational model proposed in Chapter 4 and the experimental data, since both provide measures assumed to correlate with featural activations. A further aim of these experiments was therefore to test a prediction of the network model, that viability effects should be found at pre-lexical levels as well as in the lexicon. The experiments therefore contrasted sentences such as (7) with corresponding nonword carriers such as (8).

(8) The city got two awards for its *threan* parks.

These sentences were designed to be similar to the real-word carriers but with sufficient mismatching information to prohibit access to the stored information regarding the base word. This was done by

replacing the word-initial consonant cluster of the base word with a featurally dissimilar cluster. In the example here, the /kl/ is replaced by /θr/. Studies of the effects of word-initial deviation on lexical access, using cross-modal and intra-modal priming (e.g., Marslen-Wilson, Moss & van Halen, submitted; Connine, Blasko & Titone, 1993; Marslen-Wilson & Zwitserlood, 1989), agree that deviations of more than a couple of features are enough to block initial perceptual access to the base word. This means that to monitor for segments of the nonwords subjects cannot rely on post-lexical constructs. Any effects of phonological viability found in the nonword conditions here would be support for the network model of phonological inference before lexical access.

Given the contrasts described above a number of predictions can be made regarding the outcome of Experiment 3. A lexical theory of phonological inference is one in which phonological inference is dependent upon successful access to lexical information. This theory predicts an effect of phonological viability on the responses for the real word carriers but not for the nonword carriers. In addition, I would expect a general advantage for the real word conditions over the nonword conditions due to the availability of two types of information (pre-lexical and lexical) in the former case.

The predictions of the network model are made explicit in Simulation 2 of Chapter 4. The network predicts an effect of phonological viability on the responses for both nonwords and words. Whether this effect interacts with the lexical status is less clear. Certainly the model as it stands predicts no difference between the responses for nonwords and words, but the network is an incomplete model of speech perception since it does not cover lexical access itself. Both lexical and pre-lexical theories would predict that, when monitoring for a coronal segment, unassimilated coronal segments will be more easily monitored than assimilated ones.

The predictions above lead to a difficulty in the interpretation of Experiment 3. Both theories predict that in some conditions of the experiment the correct response to a target word is no response at all. This implies that in some conditions only a few response times will be recorded; these being false positive responses to non-coronal segments. For this reason, the actual response times are of limited value. The most appropriate tests of the differing predictions will be based on analysis of the response proportions.

5.4 Experiment 3

TEST MATERIALS

Forty-eight sets of test sentences were constructed (see Appendix B for a full listing). Each set consisted of 8 sentences, across which three binary variables were alternated. The target segment was presented either with unassimilated coronal place (–Phonological Change) or with a non-coronal place (+Phonological Change). In addition, the carrier word and the context word were manipulated. The context word (the word immediately following the target) was varied so that the assimilation was either phonologically viable or unviable and the carrier word (the word containing the target) was presented either as a real word or as a nonword. In all conditions, subjects were required to monitor for the coronal form of the target (see Table 5.1).

Table 5.1. Example critical stimuli for Experiment 3. The preceding context in all cases is 'The city got two awards for its...' and the target segment is /n/.

Phonological Change	Lexical Status	Viability	Example
+	+	+	cleam parks
+	+	-	cleam guesthouses
+	-	+	thream parks
+	-	-	thream guesthouses
-	+	+	clean parks
-	+	-	clean guesthouses
-	-	+	threan parks
-	-	-	threan guesthouses

The test sentences were between 5 and 14 words long with a mean length of 8.9 words. The preceding context of the critical words was the same for all conditions, and was designed to make all conditions equally plausible, but so that neither the carrier word nor the context word could be predicted from its preceding sentential context. This is an important feature of the design, since predictability of the carrier word from its preceding context might well allow partial recovery of the base word in the nonword carrier conditions. The assimilation rules employed to create the viable and unviable contexts were the same as for Experiment 2, avoiding situations where the target and its following context were the same segment. The target occurred only once in each sentence.

The real word carriers were all monosyllables ending with a coronal segment (16 each of /d/, /t/ and /n/). The words were chosen so that the changed forms were also nonwords (as in *clean/cleam* rather than *sane/same*). As in Experiments 1 and 2, the carrier words were a mixture of nouns, verbs and adjectives. The nonword carriers were created by altering the initial consonant or consonants so that both the +Phonological Change and the -Phonological Change conditions were nonwords (as in *threan/thream*). The deviations used to create the nonwords were designed to maximise the featural difference between the nonwords and the real words they were based on. Twenty-three of the 48 items had deviations involving more than one segment (e.g., /fl/ → /p/, /sm/ → /br/). The remaining items had single segment deviations. For these items the featural correlations between the original segment and the nonword deviation were calculated using the Jakobson, Fant & Halle (1952) feature set (mean $r = 0.34$).

The sentences were recorded using a non-naive speaker. They were filtered at 10 kHz and digitised on an Apple Macintosh computer using Sound Designer software. The start and end-points of each sentence were identified and the onset of each target segment was marked by placing a tone on an inaudible channel. The onsets of the context segments were also identified for use in the pre-test and Experiment 4. The sentences were then played out onto DAT tape in experimental order using the Experimenter software package.

5.4.1 Pre-test

The pre-test was necessary to check the surface place of articulation of the word-final targets in each condition. As in Experiments 1 and 2, my aim was to use unassimilated and fully assimilated tokens as stimuli, and speaker bias or ease of articulation could again be confounding factors in this design. The pre-test presented subjects with the stimulus sentences up to the offset of the carrier word and a forced-choice task was used to identify the word-final segments.

DESIGN

The design of the pre-test was much the same as the main experiment. Eight versions of each test sentence were used, one for each combination of the binary variables Phonological Change, Lexical

Status and Viability (see Table 5.1). The subjects were given a forced-choice test between the changed and unchanged versions of the carrier word, and asked to rate the confidence of their responses; so there were two dependent variables: the response of the subjects (either the changed or unchanged version) and the confidence rating (1 - 9).

The 384 test sentences were split into 4 test versions, with each version containing one real word condition and one nonword condition from each test item. The conditions that were paired within one version were rotated between items so that subjects could not use their response for one sentence to predict the response to the other sentence within that item (e.g., subjects hearing *cleam* could not predict that the corresponding nonword was *thream*). The items were recorded in a pseudo-random order and this order was maintained throughout testing.

SUBJECTS

Twenty-eight subjects from the Birkbeck Speech and Language subject pool were tested. Subjects were allocated to the 4 test versions in the order they arrived. Seven subjects per version were tested.

PROCEDURE

The subjects were tested in groups of 2 to 4 on one of the four versions of the experiment. The subjects were given answer sheets on which, for each sentence, two versions of the final word (the carrier word) were printed, one corresponding to the changed version of the word and one being the unchanged version. There was also a confidence scale for each sentence consisting of the numbers from 1 to 9. The sentences were played from DAT tape through headphones to the subjects. Each item consisted of a warning tone followed by 5-15 words of left context and the ambiguous word. The subjects then had 3 seconds to circle the word on the answer sheet that most closely represented the word they heard and to circle a number corresponding to their confidence in their decision. They were instructed to vary their ratings from 1 for a complete guess to 9 for a certain response. The subjects were given 10 practice sentences and then a break. The 72 test items were then presented. Each session lasted about 20 minutes.

RESULTS

There were data from 7 subjects in each test version with no subjects rejected from the analysis. Eight test items were rejected from further analysis due to high error rates or poor stimulus quality. These items were also excluded from the main experiment. The results (correct response proportions and confidence ratings) for the remaining 42 test items are summarised in Table 5.2.

Table 5.2. Mean response rates and confidence scores for the pre-test. Example critical words (i.e. carrier and context words) are given for each condition. The target segment in this example is /n/.

Change	Lexical Status	Viability	Example	% Correct	Confidence
+	+	+	cleam parks	91	7.9
+	+	-	cleam guesthouses	90	7.8
+	-	+	thream parks	94	7.5
+	-	-	thream guesthouses	98	8.1
-	+	+	clean parks	92	7.6
-	+	-	clean guesthouses	94	7.6
-	-	+	threan parks	90	7.5
-	-	-	threan guesthouses	93	7.9

These data were subject to four-way item and subject ANOVAs using the factors Phonological Change, Viability, Lexical Status and Version. The error analysis revealed a significant interaction between Lexical Status and Phonological Change ($F_1[1,24] = 9.41$, $p < 0.01$; $F_2[1,35] = 8.89$, $p < 0.01$). In the confidence ratings analysis there was a significant effect of Viability ($F_1[1,24] = 20.96$, $p < 0.01$; $F_2[1,35] = 4.90$; $p < 0.05$) and a significant interaction between Phonological Change and Viability ($F_1[1,24] = 8.77$, $p < 0.01$; $F_2[1,35] = 9.18$, $p < 0.01$). There was a marginal effect of Lexical Status ($F_1[1,24] = 6.25$, $p < 0.05$; $F_2[1,35] = 3.09$, $p = 0.087$). All other effects were not significant.

DISCUSSION

On the whole, the pre-test showed that the place of articulation of the target segments were perceived as intended — subjects correctly identified the target words 93% of the time, suggesting that the surface places of articulation were unambiguous. However, there were a number of significant effects in the analyses suggesting that there were systematic differences between the clarity of the targets in some of the conditions. The strongest deviation was for the phonologically unviable changed nonwords, which were identified more often (98% correct) and with a greater degree of confidence (8.1) than the other conditions. This may be because these tokens are the least like real words and so were articulated with more care. It is possible that these systematic differences are enough to affect monitoring response times, but because the differences were small it was decided to continue with these stimuli in the main experiment, using the pre-test scores for each item as covariates in the analyses of the main experiment.

5.4.2 Main Experiment

DESIGN AND MATERIALS

The independent variables; Phonological Change, Lexical Status and Viability; were as in the pre-test (see Table 5.1). Here, the task was to monitor for the underlying coronal target; so in each trial, subjects were first presented with the target visually and then the sentence was presented auditorily. The dependent variables were the proportion of correct responses for each condition and the mean response times. As explained in Section 5.3, it was anticipated that because of the wide variation in response proportions the reaction times would be of secondary importance.

The test sentences were the 42 items remaining from the pre-test (see Appendix B). The full sentences were used in this experiment and are as described above. In addition a number of other sentences were used. Twenty-four sentences were presented at the start of the experiment to give subjects practice at the task: of these, 12 contained a target segment (/d/, /t/ or /n/) and 12 did not. Ten filler sentences, constructed in the same manner as the practice sentences, were used after the break between the practice and the main session to prepare subjects for the test sentences. Interspersed with the test sentences were 60 filler sentences, which were used to disguise any regularities in the test materials. The breakdown of these sentences was as follows.

1. Six sentences contained targets followed by segments other than labials or velars. All targets in the test sentences were followed by labial or velar segments so it was envisaged that without these fillers, subjects might start to ignore all segments not followed by a labial or velar segment (although there were many labial and velar segments in the experiment which were not preceded by a target).

2. Six sentences contained a place-changed version of the target embedded in a real word and followed by viable context for assimilation (e.g., *speak quickly*, where *speat* is a nonword). These ensured subjects could not just respond when they heard two consecutive labial or velar segments.

3. Twelve sentences contained a target segment and a nonword elsewhere. These sentences discouraged subjects from learning to correlate the occurrence of a nonword and the occurrence of a target.

4. Thirty-six sentences, with no target present, were used to reduce the number of "Yes" responses. Of these, roughly a quarter contained a nonword. If subjects were

to respond only when presented with unassimilated coronals, then roughly 50% of sentences would provoke a response.

In all cases there was an equal proportion of each of the three target segments, and the fillers were matched with the test sentences in terms of sentence length.

Eight versions of the sentences were recorded onto DAT tape for use in the experiment. These ensured that for each test item only one of the 8 conditions was used in any one experimental version. The test items and fillers were pseudo-randomly ordered, with the test order maintained across experimental versions.

SUBJECTS

Eighty-nine subjects from the Birkbeck Speech and Language subject pool were tested, none of whom had taken part in the pre-test. They were paid £4 for their participation.

PROCEDURE

Each subject was allocated to one of the 8 test versions in the order they arrived and tested in groups of 1 or 2. The subjects were warned that some of the sentences may contain nonwords but that this should not influence their decisions.

The phoneme monitoring experiment was then carried out in two blocks with a break after the practice sentences. Each sentence was preceded by a warning tone, followed by the target for that sentence, presented as a capital letter on a computer screen for 1.5 seconds. The sentence was then presented to the subjects auditorily through a pair of headphones and the subject was instructed to press a button on the button box in front of them as soon as they heard the target, or to do nothing if no target was presented. The reaction time was measured from the onset of the target. At the end of each sentence there was a three second interval and the procedure was repeated. The experiment was controlled using DMASTR experimental software on PC microcomputers. The experiment lasted approximately 15 minutes.

RESULTS AND DISCUSSION

Subjects generally found the monitoring task difficult, and monitoring times and error rates were, compared to other phoneme monitoring studies, very high. For the purposes of data exclusion, subjects' response times and error rates were calculated for the unchanged conditions, where the target was present in surface form and the correct action was unarguably to make a response.

Subjects' data were excluded from the analyses if their mean response time to the unchanged conditions was over 1400 ms or their miss rate for these conditions was over 50%. These figures, especially the miss rate cut-off, are much higher than the values I have used in other experiments, but I shall argue that the nature of the task in this experiment warrants the change.

These cut-off limits excluded 16 subjects' data from analysis, with 73 subjects remaining. One test item was excluded from analysis at this point because it contained two tokens of the target segment. The number of subjects remaining ranged from 7 to 10 per version. There was no further exclusion of data due to high response times other than the 3 second time-out of the experimental software. Item and subject mid-means were calculated for both sets of data. These data, in the form of the proportions of test items provoking a response, are summarised in Table 5.3.

Table 5.3. Mean response rates and reaction times for Experiment 3. Example critical words (i.e. carrier and context words) are given for each condition.

Change	Lexical Status	Viability	Example	% Response	RT
+	+	+	cleam parks	59.2	947
+	+	-	cleam guesthouses	34.0	997
+	-	+	thream parks	33.0	1162
+	-	-	thream guesthouses	20.2	1118
-	+	+	clean parks	77.3	725
-	+	-	clean guesthouses	75.6	712
-	-	+	threan parks	87.5	865
-	-	-	threan guesthouses	82.6	852

Four-way item and subject ANOVAs were carried out on the data using the independent variables Phonological Change, Lexical Status and Viability, as well as the experimental version.

There was a significant main effect of Phonological Change ($F_1[1,65] = 515.5$, $p < 0.01$; $F_2[1,31] = 201.1$, $p < 0.01$), with unchanged (i.e. surface coronal) conditions (80.4%) provoking far more responses than changed (i.e. surface non-coronal) conditions (36.0%). The effect of Lexical Status ($F_1[1,65] = 16.7$, $p < 0.01$; $F_2[1,31] = 8.6$, $p < 0.01$) showed that there were more responses to real words (62.3%) than to nonwords (54.1%) across conditions. There was also a main effect of Viability ($F_1[1,65] = 57.0$, $p < 0.01$; $F_2[1,31] = 32.5$, $p < 0.01$), which was mainly restricted to the phonologically changed conditions.

The interaction between Phonological Change and Lexical Status ($F_1[1,65] = 173.5$, $p < 0.01$; $F_2[1,31] = 59.2$, $p < 0.01$) showed that the effects of phonological change were strongest for the nonword stimuli. The effects of phonological change were also stronger for the changes in unviable contexts, as reflected in an interaction between Phonological Change and Viability ($F_1[1,65] = 32.5$, $p < 0.01$; $F_2[1,31] = 18.4$, $p < 0.01$). Finally, the interaction between Lexical Status and Viability ($F_1[1,65] = 4.21$, $p < 0.05$; $F_2[1,31] = 4.76$, $p < 0.05$) showed that viability effects were stronger for the real words than for the nonwords. These results are illustrated in Figure 5.1.

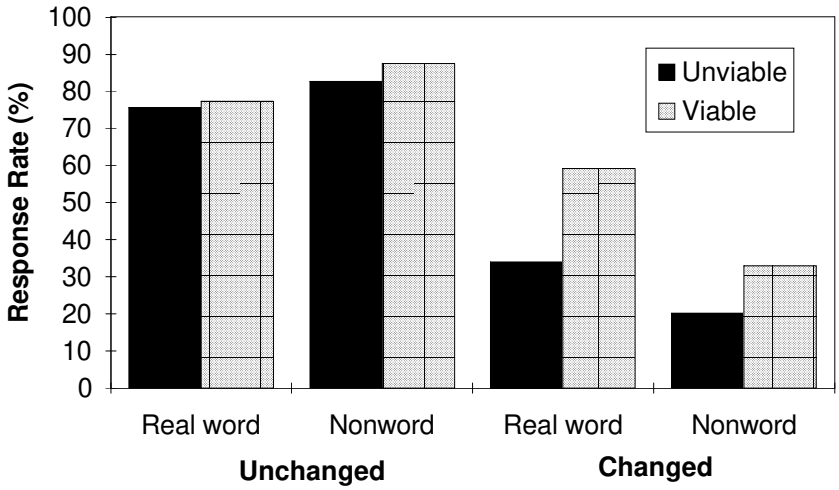

Figure 5.1. Effect of phonological viability on the response proportion in Experiment 3.

Post hoc Tukey HSD tests were carried out on the effects of Viability for the phonologically changed conditions. For the real-word carriers, the 25% effect of viability was significant at the 1% level in both item and subject analyses. For the nonword carriers, the 13% effect was significant at the 1% level for the subject analysis but not in the item analysis (although still significant at $p < 0.01$ in a Newman-Keuls item analysis).

Because of the variations found in the pre-test a number of further analyses were carried out on the data. An analysis of covariance was carried out on the item means, using as covariates the pre-test error and confidence scores. The analysis found that all the effects listed above were still highly significant, apart from the interaction between Lexical Status and Viability, which was insignificant ($F_2[1,302] = 1.74$, $p = 0.19$). For the most crucial comparison, the effect of viability on phonologically changed nonwords, a Pearson correlation test was carried out to check whether the viability effects for each item correlated with the pre-test viability effects for either confidence ratings or error rates. No significant correlations were found. The effects, therefore, seem to be dependent on the contrasts intended, rather than on confounding effects of target clarity.

The response time data were difficult to interpret, since the response rate varied between conditions from 20% to 88%. The data were also heavily skewed (skewness = 1.40) and so item and subject ANOVAs were carried out on the response times following a logarithmic transformation. The means for each condition (after an inverse transformation) are presented in Table 5.3. The analyses revealed a significant main effect of Lexical Status ($F_1[1,23] = 18.9$, $p < 0.01$; $F_2[1,9] = 12.4$, $p < 0.01$), showing that across conditions, subjects responded to targets embedded in words more quickly than to targets embedded in nonwords. A main effect of Phonological Change ($F_1[1,23] = 23.8$, $p < 0.01$; $F_2[1,9] = 44.1$, $p < 0.01$) showed that subjects also responded more quickly when the target was presented without a surface change of place. No other effects approached significance.[33]

General performance. The response times and error rates found in this experiment were much higher than those normally found in phoneme monitoring tasks. Obviously, factors such as the change of place of articulation of the target and the use of nonword carriers greatly increase the difficulty of the task, since in four of the conditions the target is not actually present in its surface form. But even in the conditions where the targets were unchanged and embedded in real words, the mean response times for the cleaned data were over 700 ms with a miss rate of over 20%. A number of possible explanations for this performance are discussed below.

The position of the target in its carrier word was unusual, since most phoneme monitoring tasks use word-initial targets. However, Koster (1987) used word-final targets and found response times were roughly 300 ms for unassimilated targets and 400 ms for assimilated targets (compared to 725 ms and 947 ms in the equivalent conditions here). Why should subjects be able to respond to the targets in Koster's experiment so much more quickly than here? One possibility is that experimental procedure used in Koster's experiment allowed the targets to be predicted more effectively. In an attempt to control for frequency differences between conditions, Koster presented subjects with a list of all the critical carrier words and nonwords (with the assimilated alternatives) and subjects were asked to read aloud all the words, but not the nonword alternatives, before the experiment was carried out. This episodic facilitation could have the effect of greatly speeding response times, possibly converting the phoneme monitoring task to something more akin to a word-monitoring task.

Another reason for this seemingly poor performance may be the way the stimulus materials were created. My goal was to use sentences in which the occurrence of between-word phonological changes would appear natural. For this reason, the sentences were quite long (up to 14 words, mean length 9) and were spoken at a normal conversational rate, with no attempt to introduce unnatural clarity. Additionally, the position of the carrier word in the sentence was not kept constant, reducing the predictability of the targets.

The use of three target segments rather than one is also an unusual feature of this experiment. I chose to employ more than one target to assess the generality of any phonological effects found. However, the drawback of the testing procedure chosen is that subjects must constantly switch between three different sets of response criteria. Blocking the test sentences according to target type may have

---

[33]The reduced values for the degrees of freedom of the error terms here (23 and 9, compared to 65 and 31) are due to the deletion of missing data where no responses were made.

produced swifter responses, although this procedure may also encourage subjects to pay less attention to the lexical route to phonological information (Cutler et al., 1987).

The amount of information available in the speech signal is also critical. In this experiment, three coronal targets were used: /t/, /d/ and /n/. Of these, only the /n/ is fully articulated in connected speech. Stop consonants are normally unreleased when followed by another stop consonant (as is the case here) and so the identity of the segment can only be found out from analysis of the preceding vowel transition. Particularly in the case of the unvoiced stop /t/, this means that subjects must base their phonetic decisions on a very small amount of information. The forced choice pre-test shows that there is sufficient information in the target segments to identify their place of articulation without the information in the release, but it is possible that this information was not always strong enough to provoke a response in the phoneme monitoring task.

Because of the variation of informational content between target types, a further item analysis was carried out on the response proportion data, with an extra between-item variable added: the manner of articulation of the target (either nasal, voiced stop or unvoiced stop). There were no main effects or interactions involving target type, although there was a marginal main effect in the response time analysis, ($F_2[2,14] = 3.16$, $p = 0.074$), suggesting that subjects were slower to respond to voiced stops (915 ms) than to nasals (877 ms) or unvoiced stops (883 ms). It was concluded that the informational content of the target segment did not strongly affect response times. The lack of interaction between manner and the other independent variables showed that the effects discussed here are not restricted to one particular segment type.

Perception of Phonological Form. The responses to the phonologically changed conditions show that to some extent, subjects rely on an abstract underlying representation when making judgements about phonological form. Both the lexical status of the carrier word and the phonological viability of the following context affect the proportion of subjects responding to labial or velar segments as if they were coronals. In the real word, viable condition, the proportion of items provoking responses is pushed up to 59%. Here, subjects are more likely to make a response based on the underlying, phonologically invariant, form of the speech than its surface form. The effect of phonological viability supports the findings of Chapter 3 that phonological inference is a vital component of speech perception. That this effect is found in a task that, by its nature, directs subjects to attend to the more low-level attributes of speech makes it all the more impressive.

The effects of lexical status, found in both response time and response proportion analyses, suggest that subjects' percepts are strongly influenced by lexical information about phonological form. Collapsing across viability conditions, subjects were roughly 170 ms faster and produced 20% more responses to the conditions in which the carrier word was a real word. It is clear that the process of access to the lexical entry for a word has a strong effect on the perception of the form of that word. Indeed, although the three-way interaction in the main analyses of response proportions was not significant, an analysis of just the phonologically changed conditions showed a significant interaction between Lexical Status and Phonological Viability ($F_1[1,65] = 4.61$, $p < 0.05$; $F_2[1,31] = 5.31$, $p < 0.05$). This interaction suggests that the process of phonological inference is to some extent dependent on successful access to stored information: the effect of phonological inference is stronger (25%) when lexical access is successful than when it fails (13%).

Theoretical Issues. The influential role of lexical status on the results of Experiment 3 seems difficult to incorporate in a purely pre-lexical model of phonological inference. Any effect of this kind relies on a pseudo-lexical effect of sequence familiarity, as described in Chapter 4, whereas Simulation 3 showed that these differences had no effect on the response of the model to assimilated segments. Leaving this problem aside for the moment, there are a number of findings that provide support for a connectionist model of phonological inference. Subjects exploit the cues to the underlying identity of assimilated phonemes in a graded manner. There is no all-or-nothing interaction between the cues of phonological context and lexical status on the responses of subjects, as might be expected from a rule-based lexical theory. Unviable assimilations in real words elicit 14% more responses than the corresponding nonword condition. Similarly, the viable assimilations in nonwords elicit 13% more responses than the corresponding unviable condition. Neither of these differences would be predicted by a model of phonological inference in which an underlying real word and viable context were *necessary* conditions for compensation. The former effect can, of course, be simply accommodated as a lexical bias effect, similar to the Ganong (1980) effect on phonetically

ambiguous segments. However, it is the finding of an inference effect in nonwords that provides the strongest support for the network model of Chapter 4. The initial deviations used for the nonwords prevent access to lexical information about phonological form, suggesting that the viability effect found in nonwords is due to a lower level compensation process, as predicted by the network model.

Interestingly, these results also map closely onto the predictions of Simulation 1. Assimilated segments in viable phonological context were treated by the network as ambiguous, provoking responses which were, on average, half way between coronal and non-coronal. This behaviour was interpreted as a drawback of the network model when compared to Koster (1987), who found the same segments were unambiguously perceived as coronal. But here, subjects' responses are much closer to the theoretical predictions: subjects produced coronal responses to 59% of the phonologically changed segments in the viable, real word condition; compared to the network prediction of 53%.

However, there is one alternative explanation of the compensatory effects in nonwords that requires consideration. The research on the mismatching effects of phonetic changes shows that single-feature deviations are often enough to nullify any priming effect of the base word. However, the general trend that can be gleaned from this research is that the more time the subject has to make a judgement, the more the lexical access process is able to recover from the initial mismatching effect of the change. In this experiment, the mean response time was 890 ms, much longer than most phoneme monitoring and priming experiments. Even though the phonemic changes used to create the nonword carriers were substantial, the phonological viability effect found for the nonwords may reflect a partial recovery of lexical information, which biases subjects towards the coronal disambiguation of the target words. The reduced viability effect for the nonwords would then be interpretable as an effect of the degree to which the carrier word matches the lexical form: the real words match well, allowing a strong viability effect, whereas the nonwords match less well, reducing the lexical viability effect.

This interpretation would predict that the viability effect in the nonword conditions should increase as a function of the time taken to respond to each particular item. I tested this by calculating the Pearson coefficients for the correlation between the viability effect for the nonword phonologically changed conditions (i.e. the difference between the viable and unviable response proportions for the nonword changed conditions) and the response times to the nonword unchanged conditions (since there was insufficient data to use the changed response times). This analysis showed a small correlation in the right direction, which did not reach significance (0.21 for the subject analysis, 0.18 for the item analysis; $p > 0.1$).

Since the deviations used in the word-initial clusters to create the nonword carriers varied between 1 and 2 segments, the recovery explanation would also predict more recovery of lexical information for the lesser deviations (i.e. the single segment deviations). To check this possibility, a between-item factor Word-initial Change (either 1 or 2 segments) was added to the analysis of the error proportions, but no significant effects involving this factor were found.

My conclusion is therefore that the viability effects found for the nonword carriers reflect phonological inference at a non-lexical level. But the length of the response times here remains a worrying aspect of the results. The next experiment addresses similar issues using the phoneme monitoring task under conditions designed to elicit shorter response times.

5.5 Experiment 4

Experiment 3 examined subjects' perception of assimilated segments, using phoneme monitoring, finding further evidence of phonological inference in speech perception. Here, I use the same task to examine the way these assimilations affect perception of their phonological context as speech is heard. The vast majority of phonological changes that occur in normal speech are viable ones. Therefore it is conceivable that a speech processor that is maximally efficient in the uptake of phonological information will learn to use the presence of phonological change to predict the context that validates the change. Consider the perception of the utterance [klimpAks] (*cleam parks*). By the time the assimilated [m] is perceived, the only lexical derivation for the surface speech is *clean* (since *cleam* is not a word). Unless the word has been mispronounced, the [m] must be an assimilated segment and so must be followed by a labial segment. The labial place of the following /p/ can, therefore, be predicted by the time the [m] is heard. Of course, the *identity* of the following consonant cannot be

predicted, since it could just as easily be another labial segment, such as /m/ or /b/, but knowledge of the place of articulation of the following segment could be enough to facilitate its recognition, particularly if the lexical representation of speech has a distributed, featural form.

However, it is also possible that the presence of a phonological change adds to the processing cost of extracting the lexical information about a word. This extra processing load may delay the recognition of the following segment (Foss, Harwood & Blank, 1980), cancelling any facilitory effect. In addition, the presence of a phonetically similar segment closely preceding the target (like the assimilated [b] in the example) may also inhibit the recognition of the following target, as was found by Dell & Newman (1980) for word-initial targets with a similar distracter phoneme in the previous word. With all these conflicting factors it is difficult to predict what effect assimilation would have on the recognition of the following segment. Koster (1987, Experiment 7) addressed this question empirically by asking subjects to monitor for word-initial targets such as the /g/ in (9).

(9) That's a very sweet girl.

The preceding word was presented either in canonical form or with a word-final assimilation (i.e. [swikgɜl]). Koster found that there was no difference between the response times of the two conditions. This was also true when the assimilation occurred at the end of a longer word with an early recognition point (e.g., *accurate*).

These conflicting effects may be avoided if we examine the situation where the presence of assimilation provides an incorrect cue to the place of articulation of the following segment. For example, in the utterance [klimgesthaʊzɪz] (*cleam guesthouses*) the presence of the assimilated [m] is a cue to the labial place of the following segment. Here, the following segment in fact has a velar place and so we would expect an inhibitory effect of the unviable phonological change on the recognition of the following segment. The results of Experiment 2 indicate that the unviable changes will mismatch the lexical entry of the underlying word preceding the target, which may also delay recognition of the target due to increased processing load (Foss, Harwood & Blank, 1980); and as the phonologically changed segment and its following context are no longer minimally distinct we would expect no inhibitory effect of the phonological similarity of the preceding segment (Dell & Newman, 1980). There are, therefore, two possible mechanisms by which the presence of a place assimilated segment may inhibit the recognition of the following segment when this segment renders the assimilation unviable.

Experiment 4, therefore, compared the effects of unassimilated, viably assimilated and unviably changed segments on the recognition of the following word-initial segment. Again, the contrast was made between real word carriers and nonword carriers.

5.5.1 Design and Materials

This experiment used stimulus sentences such as (10) and (11) with the task of monitoring for the word-initial following context of the phonological change (/p/ here).

(10) The film shows a *toad* pouncing on a fly.

(11) The film shows a *groad* pouncing on a fly.

I shall refer to the italicised word here as the carrier word to maintain continuity with Experiment 3, even though it no longer contains the target segment (although it does still contain the phonologically changed segment). In Experiment 4, all the independent variables manipulated this carrier word, which was presented either with no assimilation (e.g., [toʊdpaʊnsɪN]; *toad pouncing*), with assimilation made viable by the target segment (e.g., [toʊbpaʊnsɪN]) or with assimilation made unviable by the target segment (e.g., [toʊgpaʊnsɪN]). Both phonologically changed tokens had surface forms which were nonwords. As in Experiment 3, the initial consonant cluster of the carrier word was presented either unchanged or with word-initial changes, such that in all three conditions the token was a nonword (e.g., [groʊd] / [groʊb] / [groʊg]). Thus, the independent variables in Experiment 4 were Phonological Change (none, viable and unviable) and Lexical Status (real word or nonword).

The stimulus materials were an adaptation of the stimulus set used in Experiment 3 (see Appendix C). Since the original materials were designed to be used in this experiment as well, the target segment

for this experiment was not present anywhere else in the sentence. The rules used to create the viable assimilations ensured that the segment preceding the target (i.e. the phonologically variant segment) always had the opposite voicing to the target. This avoided sequences such as [swikkɪd] (underlyingly, *sweet kid*) where the assimilated segment is identical to the target. In Experiment 3, the word-final phonological change for any particular item was either labial or velar, with the following context varied to make the viable and unviable conditions. Here, however, *both* word-final phonological changes are required to allow the following segment, the target, to be held constant. For this reason, two extra sentences per item were recorded using the same recording conditions as Experiment 3.

The test items were divided into three groups according to the word-final manner of articulation of the carrier word (nasal, unvoiced stop and voiced stop). Within each group half the items had a labial target segment (either /p/ or /b/) and half had a velar target (either /k/ or /g/). All other aspects of design and stimulus preparation were as in Experiment 3. Examples of experimental stimuli are given in Table 5.4.

Table 5.4. Sample critical words (carrier word and target word). The sentential context in this case is "The film shows a ..... on a fly". The target is the word-initial /p/.

		Lexical Status of Carrier	
		Real Word	Nonword
Phonological Change	Unchanged	toad pouncing	groad pouncing
	Viable	toab pouncing	groab pouncing
	Unviable	toag pouncing	groag pouncing

5.5.2 Pre-test

As in Experiment 3, a pre-test was carried out on the test sentences to check the clarity of the word-final consonant of the carrier word (which in Experiment 4 precedes the target segment), using a forced-choice test.

Of the 6 test sentences per item, four were sentences used in Experiment 3, and two were new recordings, both of which contained phonologically changed final segments which rendered the phonological context of the change either viable or unviable. Only the new recordings were rated in this pre-test. The test sentences were split into two test versions, consisting of 48 test sentences plus 48 filler sentences, which were used to equate the number of changed and unchanged final segments in the pre-test. Fourteen subjects were tested, 7 per version, using the same procedure as the earlier pre-test.

A number of items were found to contain the target segment elsewhere in the sentence and were rejected from the main experiment (see Appendix C). After also excluding items that gained high error rates in one or more conditions of the pre-test, 36 test items remained. The cleaned data were analysed, together with the relevant ratings from the pre-test of Experiment 3, using two-way item and subject ANOVAs. The independent variables were the Lexical Status of the carrier word (two levels) and the word-final Phonological Change (unchanged, viable changed and unviable changed). No significant or marginal main effects or interactions were found, for either the error rates or the confidence ratings. The pre-test results are summarised in Table 5.5.

Table 5.5. Error rates (%) and confidence values (in parentheses) for the stimuli used in Experiment 4.

		Lexical Status of Carrier	
		Real Word	Nonword
Phonological Change	Unchanged	7.5 (7.6)	9.5 (7.8)
	Viable	11.1 (7.5)	6.3 (7.4)
	Unviable	10.3 (7.7)	8.7 (7.9)

The pre-test confirmed that there were no confounding factors of the clarity of the segment preceding the target. The word-initial targets, which were fully released in all cases, were assumed to be unambiguous and were not pre-tested (Repp, 1978; Ohala, 1984, 1990).

5.5.3 Main Experiment

SUBJECTS

Fifty-eight subjects from the Birkbeck Speech and Language subject pool were tested. None of the subjects had previously taken part in Experiment 3 or the pre-test.

PROCEDURE

The procedure followed that of Experiment 3, with the exception that there were only six test versions, and subjects were warned that the 4 targets that they would encounter were at the beginnings of words. Response times were measured from the onset of the target segment.

RESULTS

After examination of the response distributions, subject cut-off points were set at 850 ms mean response time and 15% miss rate. Ten subjects exceeded one or both of these limits and were excluded from further analysis. Individual response times above 1200 ms were also excluded. This left between 7 and 10 subjects in each of the six versions of the experiment. Item and subject mid-means were calculated for both the response time data and the error rates. These data are summarised in Table 5.6.

Table 5.6. Response times and percentage error (in parentheses) for Experiment 4.

		Lexical Status of Carrier	
		Real Word	Nonword
Phonological Change	Unchanged	487 (3.3)	566 (7.7)
	Viable	503 (4.1)	567 (3.7)
	Unviable	555 (7.4)	593 (5.6)

The response times here are much faster than in Experiment 3, with an overall mean response time of 546 ms and only 5% error. This superior performance reflects the difference between the two phoneme monitoring tasks: the word-initial targets are all fully released and contain more information on which judgements can be based (Repp, 1978; Ohala, 1984, 1990).

The data were subject to three-way item and subject ANOVAs using the variables Lexical Status and Phonological Change as well as the experimental version. The response time analysis showed highly significant effects of Lexical Status ($F_1[1,42] = 47.3$, $p < 0.01$; $F_2[1,29] = 16.0$, $p < 0.01$) and Phonological Change ($F_1[2,84] = 16.0$, $p < 0.01$; $F_2[2,58] = 8.8$, $p < 0.01$). These showed that the presence of a nonword immediately preceding the target segment delayed recognition, and that the surface place and viability of the segment preceding the target also affected response times. Within

the Phonological Change variable, post hoc analyses showed that unviable changes evoked significantly longer response times than both viable changes and unchanged conditions (Tukey HSD, $p < 0.01$). The viability effects for both real word and nonword conditions were tested separately, with the effect holding for the real word carriers (51 ms difference, Tukey HSD $p < 0.05$), but not for the nonword carriers (25 ms difference, Tukey HSD $p > 0.1$; see Figure 5.2).

An item analysis of the response times with the place of articulation of the target as a factor showed that, across conditions, subjects responded more quickly to labial segments (502 ms) than to velar segments (597 ms; $F_2[1,33] = 20.0$, $p < 0.01$). However, there were no significant interactions involving this factor.

A further analysis was carried out on the item response time data with the number of segment changes used to create the nonword carriers as a factor. As in Experiment 3, there were no main effects or interactions involving this factor.

The error analyses revealed no main effects and a marginal interaction between Lexical Status and Phonological Change ($F_1[2,84] = 3.76$, $p < 0.05$; $F_2[2,58] = 2.60$, $p = 0.082$). The error rates in the unviable conditions for both words and nonwords were higher than the corresponding viable conditions, showing that the response time effects for these changes were not due to a trade-off between accuracy and speed of response.

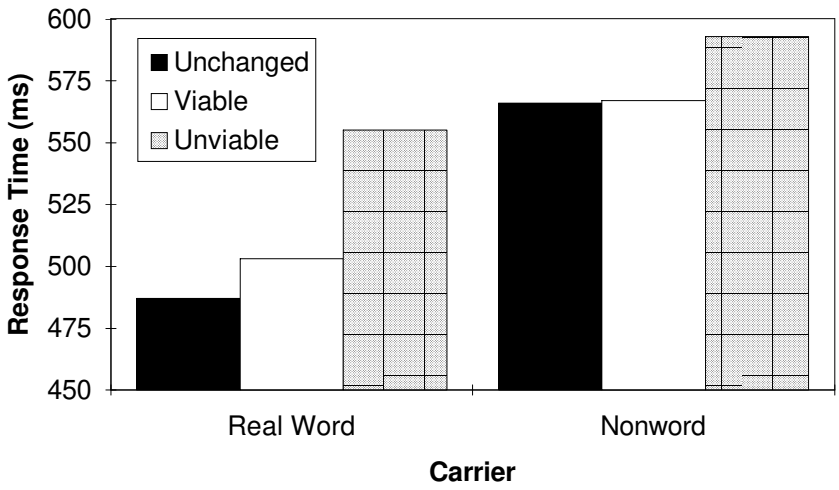

Figure 5.2. Effects of phonological change on response times in Experiment 4.

DISCUSSION

As before, Experiment 4 shows a robust effect of phonological viability on subjects' responses. Here though, the effect shows up in the response times rather than the response proportions, and is also confined to just the unviable conditions. As in Koster (1987), the viable phonological changes neither inhibit nor facilitate responses compared to the unchanged control, which suggests that either the speech processor does not use the presence of assimilation as a cue to the place of the following context, or that the facilitory effect of the cue is cancelled out by conflicting effects of phonetic similarity (Dell & Newman, 1980) or the inhibitory effect of the phonological change in the previous word (Foss, Harwood & Blank, 1980).

The unviable contexts show an increase in response times compared to both the viable contexts and the unchanged control conditions. This shows up in the highly significant effect of Phonological Change and the post hoc comparisons between the levels of this variable. The lack of an interaction between Phonological Change and Lexical Status suggests that this effect holds for both the real word and the nonword carriers. However, the separate comparison within the nonword conditions shows that the 25 ms viability effect is not significant on its own.

Lexical Phonological Inference. In order to explain the effect of unviable context in the real word condition, two factors require consideration. One is the effect of a nonword preceding the target, the other is the incorrect cue to the place of articulation of the target created by the assimilation. A lexical theory of phonological inference would predict that for the real word carriers, both factors would inhibit the recognition of the target. Phoneme monitoring studies (e.g., Foss, Harwood & Blank, 1980) have shown that the presence of a nonword immediately preceding a word-initial target slows recognition of the target. Here, the nonword can only be recognised as such once the target itself is known, so this effect may well be reduced. Nonetheless, the responses to the real word unviable condition are only an insignificant 11 ms faster than the nonword unchanged condition; indicating, as in Experiment 2, the immediacy of the mismatching effects of unviable phonological changes. It is also likely that the place change of the word-final segment is used as a cue to the place of articulation of the target. Since this cue is misleading, it interferes with the recognition of the target, assuming some kind of competition-based recognition system.

The pattern for the nonword carriers is less clear. The 25 ms increase in response times for the unviable condition is suggestive of an inhibitory effect, but is not statistically reliable. Indeed, given a lexical theory of phonological inference we would expect no inhibitory effect in this case. Both the phonologically changed and unchanged baseline conditions are nonwords and so there should be no difference between their mismatching effects; and since the nonwords do not map onto any lexical items there can be no phonological inference mechanism that uses the cue to the place of articulation of the target.

Non-lexical Phonological Inference. The implications of these results for the connectionist model of phonological inference are also unclear. To make any interesting predictions in this case the network needs to be able to recognise that a segment is assimilated *before* the following context is known, which the network in Chapter 4 failed to do. However if we imagine, for the moment, a more powerful network that is able to show pseudo-lexical effects in its processing of assimilated segments, a different set of predictions could be made. The effect of a known sequence could then be a strong enough cue to assimilation to induce the network to compensate. The network would then be in a position to use the assimilation as a cue to the place of articulation of the following segment, predicting an inhibition of responses for the real word unviable condition in this experiment. Importantly, this framework would also predict a (smaller) inhibitory effect for the unviable changes in the nonword condition, due to the graded behaviour of the network. In the nonword case, the preceding sequence is only slightly less familiar than the real word preceding context, since the deviation occurred some way back in the speech stream. This would suggest that the pattern of responses to the nonword conditions should be similar to the real word conditions, with only a slight reduction in viability effects.

Needless to say, the above explanation is extremely hypothetical, since it relies on a feature of the model that the simulations in Chapter 4 failed to show. Furthermore, the difference in behaviour predicted by the network model is only exhibited in the results as an insignificant trend. In Chapter 6, I return to this point when examining in greater detail the learning of lexical and pseudo-lexical cues in a connectionist model of phonological inference.

5.6 General Discussion

The experiments reported in this chapter address a number of issues concerned with both the representations and processes involved in phonological inference. However, I shall begin this discussion by examining how these results fit in with the debate introduced in this chapter about the mechanisms behind phoneme monitoring.

ONE ROUTE OR TWO

Both experiments showed a significant effect of the lexical status of the carrier word. In Experiment 3, this was the word containing the target, and subjects were both faster to respond and, overall, more likely to respond to real words (although in the phonologically unchanged conditions they were actually more likely to respond when the carrier was a nonword). In Experiment 4, subjects were again faster to respond when the carrier word (the word preceding the target) was a real word. This implies that in Experiment 3, subjects were able to make use of representations of lexical form to

facilitate their judgements, and in Experiment 4, their responses were disrupted when a nonword closely preceded the target. These findings imply that phonological judgements were influenced, at least in part, by a lexical route to phonological form information.

Since this lexical route is exploited in responses to words in Experiment 3, it is important to consider the possibility that *all* responses in this experiment are based solely on lexical information. By this account, the responses to the nonwords are based on the partial activation of lexical information rather than a non-lexical representation of phonological form. However, this possibility seems unlikely given the cross-modal priming results discussed earlier. Consider a subject in Experiment 3, responding to the /n/ underlying the /m/ of *thream* (derived from clean). In order to accomplish this task using only lexical information, the subject has to be able to access the base word from which the nonword originated (or any other word containing a word-final /n/). But in this case the token of speech only matches the base word on one of the four segments (the vowel) and in fact more closely matches words such as *cream*, which would bias judgements towards a non-response. Although this is a rather extreme example, there were no significant effects of the size of deviation used to differentiate words and nonwords in either experiment. It seems more likely that for the nonword carriers, the effect of lexical information on the responses was minimal[34], supporting a two-route model of retrieval of phonological information (Foss, Harwood & Blank, 1980; Cutler & Norris, 1979).

ABSTRACTNESS AND PHONOLOGICAL PERCEPTION

The phoneme monitoring task specifically directs subjects' attention to the phonological form of speech, as opposed to priming studies, which evaluate access to semantic information. Despite this, in Experiment 3, a large proportion of surface labial or velar segments were responded to as coronals. For the viable assimilated segments in real word carriers, this constituted nearly 60% of the responses.

Subjects seem to make use of a mixture of surface and underlying codes in their judgements about phonological form. However, these judgements do not conform in a standard way to lexical and pre-lexical judgements, since subjects are able to make responses to real words based on surface form, as well as using underlying representations of nonwords. The abstractness of subjects' responses depends on a combination of lexical and phonological factors.

These results illustrate the prevalence of phonological abstractness in the perceptual system: even when subjects are asked to focus on the substructure of words they show effects of phonological inference. It would be interesting to see whether this inference still occurs using an even more analytic task, such as a forced-choice test (with, unlike the pre-test, following context presented). If subjects cannot actually *hear* labial or coronal segments when they occur as a result of a phonological process we have good evidence that phonological inference occupies a central role in the perception of speech.

THE LOCUS OF PHONOLOGICAL EFFECTS

These results support the main finding of Chapter 3, that the phonological viability of deviations in the speech stream has a strong effect on their perceptual analysis. In Experiment 3, this showed up as a variation in the probability of spotting an underlying coronal when presented with a surface labial or velar segment. In Experiment 4, the viability of changes affected the recognition of the following segment. Because these changes occur across word-boundaries, in the former case regressively, in the latter, progressively, they are difficult to explain in terms of the lexical specification of the words, instead indicating the presence of a process of phonological inference in lexical access.

The experiments in this chapter also attempted to ascertain the locus of this inferential process by examining the effects of phonological viability in nonwords. Experiment 3 found such an effect, in that subjects, when presented with a phonologically changed coronal segment embedded in a nonword, were more likely to respond as if they had heard the unchanged coronal if the phonological

---

[34]I am assuming here that the mental lexicon is a unitary store, and that access to the canonical form information about a word occurs in the same manner as access to other forms of information, such as meaning and orthographic form. However, in Chapter 6, I discuss the possibility that these access routes can be separated in certain circumstances.

context of the change conformed to place assimilation rules. However, the length of the response times in Experiment 3 leave this finding open to interpretation as a post-perceptual recovery effect. Experiment 4, on the other hand, found more rapid response times, but only a suggestive effect of viability using nonword carriers.

Taking the results of Experiment 3 at face value, it seems that phonological inference can occur in nonwords, supporting as pre-lexical or non-lexical locus of inference. However, Experiment 3 also found that the degree to which compensation occurs is affected by the lexical status of the carrier word. This interaction implies that the inference process cannot be entirely independent of lexical constraints.

MODELLING ISSUES

The network model of Chapter 4 exhibited a number of properties which have found support in these experiments. The network learned and utilised cues to the presence of phonologically changed segments in a graded manner. Similarly, humans use lexical and phonological cues, not in an all-or-nothing manner, but as probabilistic weighting factors in the determination of phonological form. The network also does not rely on access to stored information on lexical form, and is therefore able to apply its learned statistical dependencies to words and nonwords alike. This prediction received some support in both experiments, although the evidence for non-lexical phonological inference is not yet compelling, due to the length of the responses in Experiment 3 and the lack of a statistically significant viability effect in the nonword conditions of Experiment 4.

However, the strong lexical effects in both experiments cause problems for the network model as it stands. Some of these effects could be explained by assuming a modular lexicon, based on the output of the pre-lexical network, but the evidence of interaction between lexical status and phonological inference in Experiment 3 constitutes a further obstacle to this approach. The discussion of autonomy and interaction in Chapter 1 shows that it is difficult to deduce the mechanism involved from an observation of interaction between two processes. Nevertheless, for a model to realistically represent the phonological inference process in speech perception it must be able to demonstrate interactive behaviour in its processing.

Two possible mechanisms for this interaction are discussed in Chapter 6. One is that the lexical effects found here can be explained in terms of the pseudo-lexical biases which were expected, but not found, in the network model of Chapter 4. The alternative is a retreat from the purely pre-lexical status of phonological inference hypothesised here, proposing that access to various aspects of lexical information are governed by different constraints. Specifically, aspects of word meaning are accessed using a matching process intolerant of deviation, whereas phonological information is gained more freely, using a more analogical process. This dissociation between the properties of form and meaning access systems is a radical departure, but is one that is hinted at by Marslen-Wilson and Warren (submitted) to explain their findings about subcategorical mismatch in words and nonwords.

Chapter 6 — In Search of Lexical Effects

6.1 Introduction

At this point it may be useful to review the relationship between the experimental data I have presented and the predictions of the network model of pre-lexical processing. The dominant finding from both priming and phoneme monitoring studies is that phonological inference plays a strong role in the compensation for between-word variation in lexical access. This is consistent with the predictions of the connectionist model described in Chapter 4. However, a number of experimental findings show that lexical context plays a part in the evaluation of phonologically ambiguous segments.

In Experiment 1, the prime words were presented with following context removed. In this situation both the phonologically changed and unchanged primes facilitated the visual targets, suggesting that an unviable phonological change will only inhibit lexical access of the base word once the following context of the change is perceived. The network failed to exhibit this behaviour since it would only compensate for assimilation once the following context was known. Experiments 3 and 4, using phoneme monitoring, both showed lexical effects on subjects' responses to phonologically ambiguous segments.

Findings from both experimental tasks show that preceding lexical context is used as a cue to the presence of phonological change. However, the network model shows little or no effect of preceding context on the evaluation of phonological ambiguity (Simulation 3). This lack of a pseudo-lexical effect raises questions about the plausibility of the network as a model of phonological inference.

One solution to this problem would be to deny responsibility for lexical effects in a model of pre-lexical processing. When outlining the network model, I assumed that the output of the network forms the input representation of a lexical matching process. More or less any model of word recognition would thus produce the lexical effect in Experiment 1, and provided phonological information was available at the lexical level, it would also explain the lexical effects for the real words in Experiments 3 and 4. But the phoneme monitoring experiments also revealed similar but weaker effects when the ambiguous segments were embedded in word-like nonwords, which I have argued cannot be explained in terms of a standard lexical bias.

How then can these effects be explained? In this chapter I review a number of simulations designed to explore further the acquisition and exploitation of lexical or pseudo-lexical effects in connectionist models of speech perception. Simulations 4 and 5 examine more closely the model developed in Chapter 4; looking at the strength of left context effects in a simple recurrent network and the nature of the representations built up from the training corpus. The remaining simulations involve modifications of the original network and training corpus, in an attempt to encourage the network to show the desired lexical effects. These modifications touch on the role of segmentation of the speech stream in word recognition, and I shall argue that the low-level statistical approach applied here to phonological inference is also well suited to the task of speech segmentation.

6.2 Simulation 4 — Memory Span

What are the differences between the phonological viability cues contained in the following context of an ambiguity (which the network learned to use) and the lexical cues contained in the preceding context of an ambiguity (which the network failed to detect)? One important distinction is in the temporal location and spread of these cues. The phonological viability of an ambiguous segment is

defined by the identity of the segment immediately following the ambiguity. However, lexical cues to the identity of a segment are potentially spread over many preceding segments.[35]

Although research has shown that long-distance dependencies can be learned in simple recurrent networks (Cleeremans, Servan-Schreiber & McClelland, 1989), the ease with which these dependencies are acquired depends strongly on the task involved and the structure of the network. In this case, it is plausible that either the learning algorithm or the size of the hidden unit layer used in these simulations was inadequate, leaving the network with an insufficient memory span for the task of encoding word-like sequences.

In this section, I report a simple test of the memory span of the network model described in Chapter 4. The test involves the systematic manipulation of a number of six-segment words from the training corpus. The effects of these manipulations were measured by examining the network's predictions for the final segments of the words.

MATERIALS AND DESIGN

Eight six-segment words were chosen at random from the training corpus as test stimuli for this experiment.[36] Each word was presented to the trained network in 6 different forms. The first form was the undistorted word, which served as a baseline for the remaining sentences. On each successive presentation, one of the first five segments of the word was replaced by a noise segment (all feature values were set to 0, making the stimulus dissimilar to all segments in the training corpus). These five stimuli varied with respect to the distance between the noise segment and the final segment. Each presentation was preceded by the filler word *then*. A set of example stimuli is illustrated in Table 6.1.

Table 6.1. Example stimuli ('then anything') for the memory span simulation. The asterisks denote 'noise' segments.

Distance	Example Stimulus
Baseline	/ðen enɪθɪN/
1	/ðen enɪθ∗N/
2	/ðen enɪ∗ɪN/
3	/ðen en∗θɪN/
4	/ðen e∗ɪθɪN/
5	/ðen ∗nɪθɪN/

The dependent variable was the network's prediction for the final segment of the word (in the example here, the output of the prediction nodes were recorded on presentation of the /ɪ/ before the /N/). For the conditions with a noise segment, the effect of the noise was measured by calculating the root mean-square (RMS) deviation from the prediction in the baseline condition. This gives a measure of the influence of each segment on the prediction for the word-final segment; and the comparison of the 5 conditions containing a noise segment shows the variation in influence as the distance between the noise and the final segment increases.

The stimulus lists were presented to the trained network twice, with the results from the first presentation discarded to eliminate any confounding effects of the initial state of the network. For details of the network training, see Section 4.4.1.

---

[35] The mean length of the words in the training corpus was 4.7 segments and the longest word, *pseudo-prepositions*, contained 17 segments (counting diphthongs as two segments).

[36] The test words were: *anything, always, pounds, country, lantern, minutes, Monday* and *suppose*.

RESULTS AND DISCUSSION

The RMS deviations from the baseline values were calculated for each of the 5 noise conditions. The mean deviations for each condition are summarised in Figure 6.1.

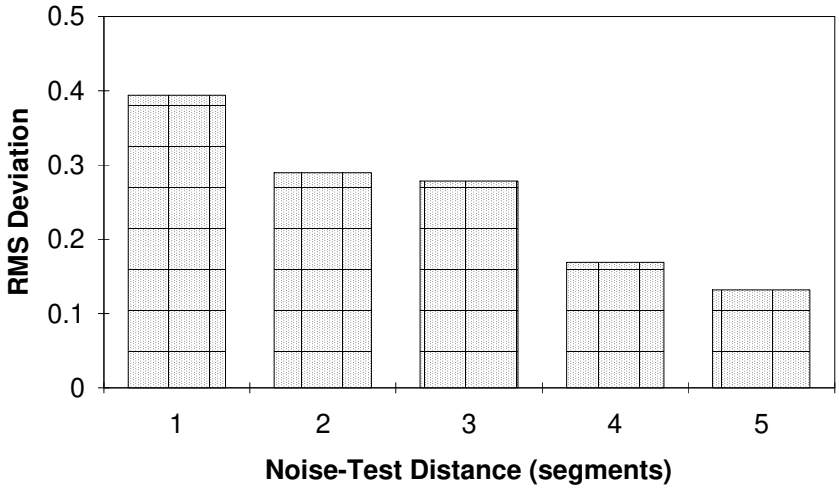

Figure 6.1. Results of Simulation 4.

A one-way ANOVA was carried out on the data, using the variable Distance (the distance in segments between the noise segment and the word-final test segment) with 5 levels. This revealed a highly significant main effect ($F[4,28] = 8.03$; $p < 0.01$).

A further analysis was carried out to check whether the deviations found in the individual conditions were significantly greater than zero. The differences for conditions 1 to 3 were significant ($p < 0.05$) and the differences for conditions 4 and 5 were marginal ($p = 0.079$ for condition 4, $p = 0.066$ for condition 5).

These analyses show that the network's predictions in this task depend on the identities of the previous three segments, and suggest that earlier segments are also likely to affect the network's response. In terms of utilising pseudo-lexical cues in the speech stream, this implies that the network has a sufficient memory span to allow lexical status to influence the processing of four-segment words, as well as possibly five or six-segment words. Since much of the training corpus (55% by type, 81% by token) consists of words of up to 4 segments, it seems likely that the lack of lexical effects in the previous simulations is not due to a simple lack of memory capacity.

However, this conclusion is qualified by the effect of distance found between conditions. The segment with the greatest influence on the network's predictions is the segment currently being processed, and the influence of preceding segments decreases in a roughly linear manner. Although the network's memory span may be reasonably large, the effect of segments some distance back in the speech stream is severely restricted. This recency effect can be understood by considering the way temporal information is represented in a simple recurrent net. At any point in the processing of the network, the hidden units receive information from two sets of units. One set represents the current input and the other represents the state of the hidden units at the previous time-step (i.e. a representation of all previous inputs). Assuming that all input is informative (as is true in this case), the hidden units can only encode the current input at the expense of losing some of the information about the previous input. Thus each time the information about a segment passes through this loop it is degraded, reducing the influence the information can have on the output of the network.

Although this simulation reveals a general trend of decay of influence over time, the influence of preceding context on any particular judgement is very much task dependent. In particular, the effect of word-boundaries on the formation of dependencies has not been tested here. Such structural relationships have been studied by Reilly (1993), who examined the representations built up by a recurrent network learning simple sentences word-by-word. Reilly's aim was to simulate findings of Jarvella (1971) and Caplan (1972) that short-term memory for words during the on-line processing of

sentences is strongly dependent on whether a clause boundary separates the material to be recalled and the probe. This kind of result has traditionally been explained by assuming that the perception of a clause boundary triggers a transfer of the information about the preceding clause from a highly accessible short-term memory to a more permanent (and less accessible) store (Fodor, Bever & Garrett, 1974; Garnham, 1985).

Reilly showed that these effects could be simulated using a network trained to encode simple sentences (see also Elman, 1991). The task relies on the use of preceding cues to sentence structure to optimise performance: here the most salient cues to the identity of a word are contained within the immediate clause. The network learned to retain information about the current clause until the end of that clause, and then cast off this information to allow a new representation of the next clause to be built up. The short-term memory effect was thus explained as a product of the connectionist learning approach, and reflected the structure of the language rather than the perceptual system.

The stimuli used for the current test of memory capacity were all single words, chosen so that word-boundary effects could not confound the results. However, it is also relevant to ask whether there are any word-like representations being built up during training of this network. The lack of any lexical effect in Simulation 3 suggests not, but this is a rather indirect test, since only word boundaries were examined, and the effects may have been masked by the presence of phonologically changed segments. In the next simulation I present a more comprehensive examination of the representations the network has built up.

6.3 Simulation 5 — The Representation of Words

Simulation 4 showed that the network has a sufficient memory capacity to encode the majority of words presented as training input. Here, I examine the structure of the representation the network has built up from the training set. Backpropagation networks are often investigated using cluster analysis or principal component analysis of the hidden units (e.g., Sejnowski & Rosenberg, 1987). This allows insight into the representations formed by the network in learning the its task. For simple recurrent networks this analysis is more complex, since the state of the hidden units depends on the previous inputs to the network as well as the current input. Also, the representation of a word can only be assessed by combining the states of hidden units as each segment of the word is presented. Because of these complicating factors, the study I report here uses a less direct, but much simpler method.

Examination of prediction error allows the internal state of a simple recurrent network to be analysed. In this case, the variation in prediction error across word position allows measurement of the extent to which word representations have been built up by the network. Consider, for example, the performance of a Cohort-like model of word recognition in this prediction task (analogous to gating) at various points during the processing of a sentence. Before the beginning of a word, the model can only make a prediction based on the lexical statistics of word onsets. Thus, the prediction error for word-initial segments would be fairly high.[37] As words are heard, a cohort of matching candidates is built up. The predictions can then be based on the statistics of the cohort, increasing the probability of a correct prediction. Towards the end of the word (if it is a long word) the recognition point is reached, and the remaining segments are completely predictable, implying no error until the start of the next word is reached (but see Grosjean, 1985; Bard, Shillcock & Altmann, 1989).

The predictions based on a model using co-occurrence statistics are more complex to analyse. This type of model is likely to base its responses on the phonotactic regularities in speech. Phonotactic constraints describe the permissible sequences of segments within a language, and since these constraints are much weaker across word-boundaries, a learning model trained to pick up on low-level statistical regularities will be most strongly influenced by within-word consistencies rather than between-word consistencies. When these regularities are applied to a random selection of words, the performance of the model on any particular segment depends critically on its distance from a word boundary. The prediction of word-initial segments will produce the greatest error since the learned

---

[37]Of course coarticulatory information (Fowler, 1984; Warren & Marslen-Wilson, 1987, 1988) and neutralising changes (this research) provide cues to the identity of word-onsets, but to allow comparison with the network simulation I am ignoring these factors.

regularities do not apply across word boundaries. For the second segment of a word bigram statistics[38] will apply, but not longer-distance regularities, implying that error will be somewhat reduced for these, compared to word-initial segments. Further into the word, longer-distance regularities will become useful suggesting that prediction error will continue to decrease.

The precise shape of the error function depends on the length of dependencies the model develops. A model using only bigram statistics will show a maximum error for word-initial segments, but then all other segments should produce roughly equal prediction error. A model exploiting longer-distance regularities would show a reduction in error stretching further into the word, with the onset of a plateau coinciding with the maximum length of statistical dependencies encoded by the network.

Simulation 5 is an attempt to examine these issues, using the network's predictions when presented with a list of words randomly chosen from the training corpus.

DESIGN AND MATERIALS

One hundred and ninety-one words were randomly selected from the training corpus for use in this test (see Appendix D). These words were presented segment by segment as input to the trained network. On each presentation, the network's prediction for the next segment was recorded. As before, the list of test words was presented twice with the results for the initial presentation discarded. There was a total of 907 segments in the test set.

RESULTS AND DISCUSSION

The RMS error for each prediction was calculated by comparison to the correct featural value. Each error value was then coded according to the position in the word of the predicted segment. A prediction for a word-initial segment was coded 1, the following prediction was coded 2 and so on. The mean error values for each word position using this coding scheme are summarised in Figure 6.2.

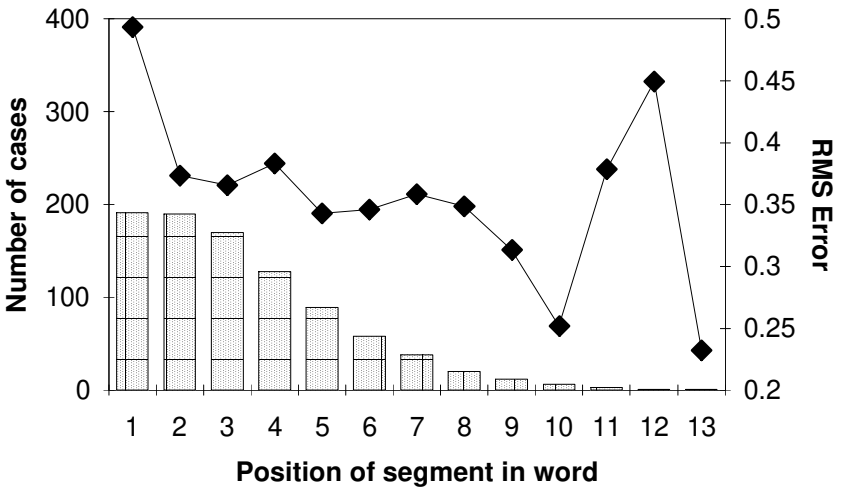

Figure 6.2. Results of Simulation 5. The line graph shows the mean prediction error for each word position; the bars represent the number of cases taken into account for each value.

Because of the variation in length of the test words, the error values become less reliable as the word position gets higher; hence the massive variation towards the right hand side of the graph (only three words in the test set contained more than 10 test segments). Disregarding this variation, a general

[38]These are the statistics of speech based on sequences of two segments.

downward trend can be discerned from the error values and a significant negative correlation between position in word and RMS error was found ($r = -0.26$, $p < 0.01$).

A second analysis examined a subset of these data. Much of the negative correlation appears to be due to the high error rate for the prediction of word-initial segments, which I have argued is predicted by both co-occurrence and word-based models. A correlational analysis carried out on the data excluding the word-initial predictions showed a significant, but weaker, negative trend ($r = -0.085$, $p < 0.05$).

The most conspicuous result of this study is the sharp rise in prediction error as a new word is encountered (Segment position 1 in Figure 6.2). This implies that the representations formed by the network allow phonotactic constraints to be detected and exploited in the prediction task. The smaller negative trend found over the remainder of the word suggests that longer distance dependencies can also be used but that these are either less effective cues or that the network is able to make less use of them.

A cohort-type model would predict that towards the ends of words prediction error would be reduced to near zero, but because of the variation of word length in the test set it is not clear from Figure 6.2 whether this occurs (since word-final segments are mixed in with segments in other positions). The data were therefore recoded in terms of their distance from the *ends* of the words and the mean values are illustrated in Figure 6.3. The correlation between segment position and RMS error was again significant in this analysis ($r = -0.095$, $p < 0.01$).

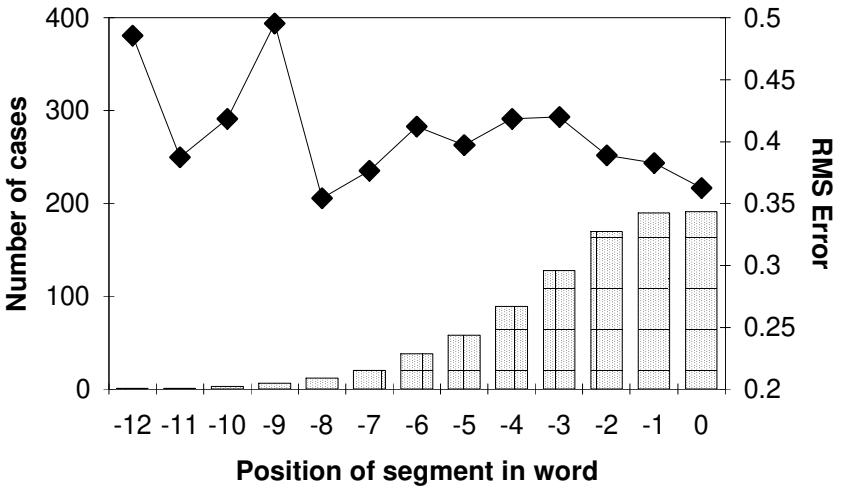

Figure 6.3. Mean error values as a function of distance from the end of the word. The zero segment is the end of the word.

Despite this small correlation (accounting for roughly 1% of the variation), these results indicate that the network does not develop word-like representations in order to carry out its task, since even word-final segments are still relatively unpredictable (mean RMS error = 0.36). The lack of word-like representations in the network is hardly surprising, given the scarcity of information presented to the network in training. Learning to understand speech involves the interaction of meaning with auditory percepts, which for this task, aids segmentation of the speech stream and allows unique lexical representations to become associated with sequences of speech sound. The simulations described in this chapter show that simple recurrent network have adequate raw memory capacity to encode word-like representations, but the fact that they do not implies that the information available during training is insufficient.

6.4 Speech Segmentation

In this section I describe a modification of the network presented in Chapter 3, in which the network is trained, as an additional task, to segment its input into isolated words. The use of segmentation information during training represents a minimal addition of lexical information, since completely correct word-boundary information can only become available once the word recognition process is complete. This allows a re-examination of the issue of lexical cues to the resolution of phonological ambiguity. It also provides an opportunity to evaluate the availability and effectiveness of low level cues in the process of speech segmentation.

6.4.1 Cues to Assimilation — A Re-analysis

In Section 4.3.3, I examined a number of cues to the disambiguation of phonological changes such as place assimilation. Three factors were identified: the surface form of the ambiguous segment, the place of articulation of the following context, and the lexical status of the preceding context. For example, the cues to the presence of place assimilation in the phrase [əswikgɜl] (*a sweet girl*) comprise the potentially assimilated surface velar segment [k], the validating following velar /g/ and the preceding context /swi/ which with an underlying /t/ forms a real word. In terms of statistical regularities, /swit/ is a familiar sequence whereas /swik/ is not.

However, this analysis assumes that the preceding context used to assess familiarity is of the right length. In the example above, it assumes that the three segments /swi/ are used in the assessment of the familiarity of the underlying readings, /swit/ and /swik/. However, the results of Simulations 4 and 5 suggest that these regularities are exploited in the network model, but along with both longer and shorter distance regularities which may produce conflicting cues. In the *sweet girl* example, /k/ is a valid completion of the embedded words /wik/ (*week, weak*) and /ik/ (*eke*), whereas /t/ is a valid completion of the words /wit/ (*wheat*) and /it/ (*eat*). Thus the pseudo-lexical cues to disambiguation are confused by the lack of segmentation information.

If instead, the network was able to learn regularities from a segmented transcript of speech, the problems of conflicting cues would be significantly reduced. The behaviour of the network would still be influenced by the coarser co-occurrence statistics of the training corpus, but given enough training time the network may be able to learn that the only reliable cues to the correct underlying form are those involving word-like representations.

Unfortunately, as the review in Section 1.4.4 shows, there are no such segmentation cues obvious from analysis of the physical attributes of speech. However, there is evidence to suggest that the statistical properties of speech can provide reliable cues to the presence of word boundaries. For example, syllable boundaries (Church, 1987), strong syllables (Cutler & Norris, 1988), and phonotactic constraints (Harrington, Johnson & Cooper, 1987) have all been suggested as cues to word boundaries. The approach I have advocated for the resolution of phonological ambiguity in speech perception should also be appropriate for modelling the acquisition of cues to speech segmentation. In particular, it provides a framework within which any or all of these cues can be used in parallel to identify word boundaries as speech is perceived. This approach contrasts with other models addressing the issue of word segmentation, which often advance one particular type of cue as the basis of speech segmentation and ignore all others.

A similar standpoint has been taken by Cairns, Shillcock, Chater & Levy (submitted). They showed that the errors of their recurrent network model of speech perception (see Section 4.3.2) in the prediction of the following segment were strongly related to the word-position of the segment (as in Simulation 5 here). Cairns et al. used this information to make predictions about the presence of word boundaries in the speech stream. In other words, they used the incidence of a poor prediction as a cue to the presence of a word boundary. This strategy allowed word boundaries to be identified at much better than chance levels. The vital difference between their strategy and the model I propose is that the Cairns et al. strategy does not use any segmentation information in training. Here I assume that some semantic information is available during the process of learning, which allows

segmentation information to be used to direct performance.[39] This means that I cannot claim that the processes of segmentation and word recognition develop independently, but it also means that performance of the trained network in segmentation should be superior using this strategy.

The interaction between cues to word boundaries and cues to phonological ambiguity resolution is also of relevance here. Church (1987) and Kaye (1993) have argued that a primary role of phonological change is to aid perceptual processes such as speech segmentation. Place assimilation is a good example of such a process since it occurs only at morpheme boundaries, which in the vast majority of cases coincide with word boundaries.[40] It seems plausible, then, that as well as segmentation cues being used to evaluate possible assimilation sites, the presence of assimilation will be used as a cue in the segmentation of speech into word or morpheme units (see Church, 1987).

6.4.2 A Connectionist Model of Speech Segmentation

OBJECTIVES

The model I propose here is a refinement of the earlier network model in which the task of word-boundary identification is added. The initial aim of this modification is to assess the ability of a simple recurrent network to both learn and exploit the low-level segmentation cues in speech. I also aim to examine the effect of segmentation on the network's response to phonologically ambiguous segments. In particular I wish to examine whether the use of segmentation information in training allows the network to form more word-like representations and to assess the interaction between cues to assimilation and cues to word boundaries.

NETWORK ARCHITECTURE & TRAINING

For ease of comparison, minimal changes were made to the architecture and training of the network. Three new output nodes were added to the network described in Section 4.4.1, one in each output window (see Figure 6.4). Each window now contains 11 nodes representing the featural description of a segment, plus one node representing the hypothesis that the segment is word-initial. No extra hidden units or context units were added.

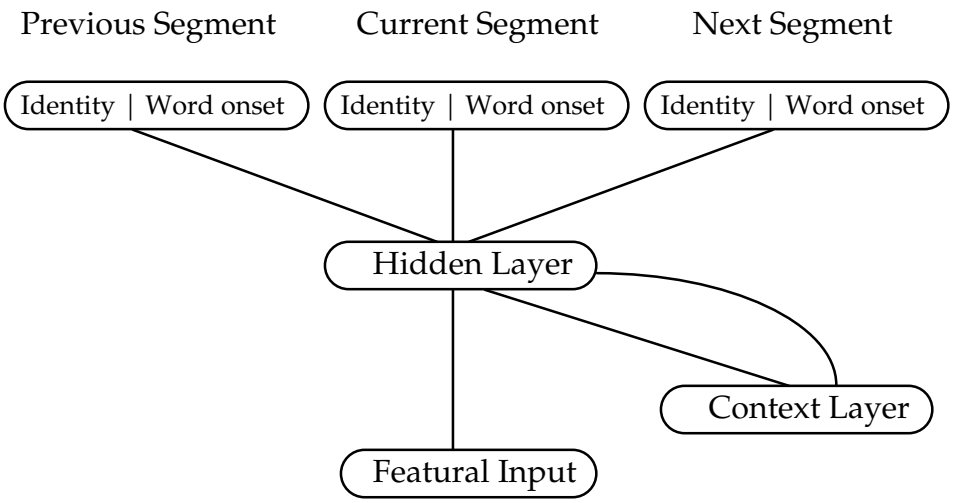

Figure 6.4. The segmentation network.

---

[39]In the terminology of Cairns et al., mine is a weakly bottom-up model compared to their strongly bottom-up model.

[40]Only four potential within-word assimilation sites were found in the network training corpus, all involving nasal segments. These were: *Edinburgh* (ten instances), *concatenation, concatenatives* and *Cannongate*.

The training regime was as described in Section 4.4.1, except that segmentation information was available for use in the adjustment of weights. At all points other than word-onset the segmentation nodes were trained to produce zero activation. For word-initial segments, these node were trained to produce a value of 1. For example, when presented with the input [i] from [swikg3l], all segmentation nodes are trained to produce a zero. On presentation of the [k], the segmentation node of the next segment window is trained to produce 1, since the next segment is word-initial. On presentation of the word-initial [g], the current segment node should become active; and on presentation of the [3], the third output window recognises the preceding word boundary. Thus, the segmentation nodes allow three different points at which hypotheses about word boundaries can be probed.

The corpus of speech used to train the segmentation network was the same as for the original network, and again 50% of potential assimilations were presented in fully assimilated surface form. The network was again trained on 100 sweeps through the speech corpus.

6.4.3 Simulation 6 — The Segmentation Network

Here, I examine the performance of the revised network on the recognition of word boundaries in a novel sequence of words. To allow examination of the internal representation developed by the segmentation network, as well as comparison with the performance of the original network, the words in this test are the ones previously used in Simulation 5 to examine word representations. These words are not representative of spontaneous speech, since they lack syntactic structure and contain far fewer function words than normal utterances. In some ways this makes the task of the network harder, since the test set violates the statistical regularities of the speech corpus on which the network was trained. In other ways segmentation is easier for this set of words, because of its relative absence of function words. These are generally short words composed of a single weak syllable, which makes their boundaries difficult to identify (Cutler & Butterfield, 1992). It was hypothesised that the acquisition of the ability to identify word boundaries would improve performance in the resolution of surface phonological ambiguity, by encouraging word-like representations to be developed.

MATERIALS AND DESIGN

Segmentation. As a test of the network's ability to segment speech, the network was tested on a set of words drawn randomly from the training corpus (see Appendix D). The test set comprised the 191 words used in Simulation 5 and is described in Section 6.3. These words were familiar to the network, since they were part of the training set but were presented in a novel order.

The test words were presented segment by segment as input to the trained network. The list was presented twice and the output of the network was recorded for each segment of the second presentation. These data contained information about both the word-position and identity of the test segments at three measuring points, corresponding to the three output windows. The segment prediction values were encoded according to their word position for analysis of the internal representations formed by the network as in Simulation 5. The segmentation values were categorised into hits, misses, false positives and correct rejections for signal detection analysis as described below.

Phonological Variation. In addition to the test of the network's segmentation properties, a replication of Simulation 3 is reported here, examining whether the uptake of segmentation cues has affected the network's use of preceding lexical context in its resolution of phonological ambiguity. All experimental details were exactly the same as for the original simulation reported in Section 4.4.3.

RESULTS

Segmentation. The segmentation data consist of three sets of figures; one for each output window. Each value corresponds to a single segment and represents the network's assessment of the probability of that segment being word-initial. If a decision criterion is chosen, these data can be transformed into discrete predictions and compared with the true boundary positions contained in the set of test words. This transformation allows the data to be categorised into hits (word boundaries correctly predicted), misses (word boundaries missed), correct rejections (CRs; non-boundary segments identified as such) and false positives (FPs; non-boundary segments incorrectly identified as word-initial) and analysed using signal detection theory measures.

The value of signal detection analysis (Grier, 1971; Aaronson & Watts, 1987) in this task is that it provides a quantitative measure of the network's ability to identify word boundaries. This is the A' statistic of discriminability: a high A' value reflects a good level of discrimination between boundary segments and non-boundary segments. The analysis also provides a second statistic, B", which is a measure of the bias of the decision criterion used. Two criteria are illustrated in Table 6.2. On the left hand side are the results when the critical value is arbitrarily set to 0.5, on the right hand side the critical value was adjusted to minimise the bias measurement, B". Figure 6.5 illustrates the receiver-operating characteristic (ROC) for the segmentation values at the latest monitoring point.

Table 6.2. Signal detection analysis of the segmentation responses. The three columns of data on each side represent the three positions of the probe point in relation to the segment in question (before the target, on presentation of the target and after the target).

	Monitoring Point						
CV = 0.5	Bef.	On	After	**B'' ≈ 0**	Bef.	On	After
Crit.Value	0.5	0.5	0.5	Crit. Value	0.4	0.16	0.13
Hits	133	132	138	Hits	145	157	161
FPs	137	61	53	FPs	164	122	110
Misses	58	59	53	Misses	46	34	30
CRs	579	655	663	CRs	552	594	606
A'	0.84	0.88	0.90	A'	0.85	0.89	0.91
B"	0.15	0.47	0.49	B"	0.02	0.02	0.01

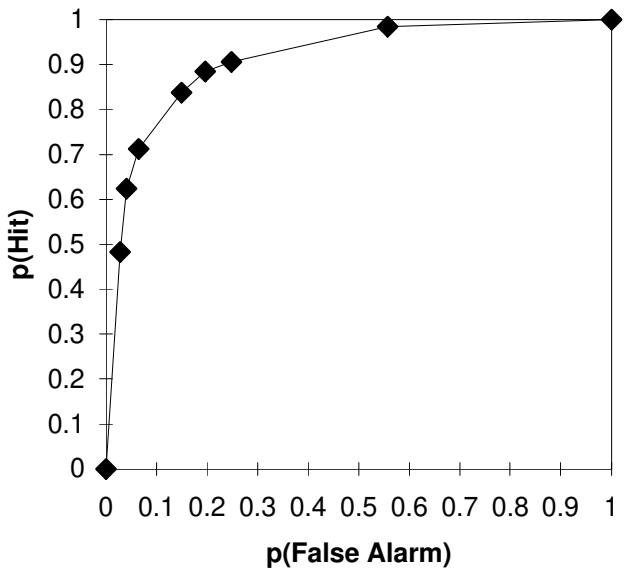

Figure 6.5. ROC for the segmentation responses when probed after target presentation (i.e. using the previous segment window).

The results of the analysis show that the network is able to predict word-boundaries in a novel sequence of words with an impressive degree of accuracy. The network's most accurate responses occur in the previous segment output window. Here (using the zero bias analysis), the network misses

only 16% of the word boundaries and correctly rejects 85% of the non-boundary segments. But even when the network is forced to respond *before* the word-boundary it only fails to predict 24% of the boundaries with 77% correct rejection.

Word Representation. As in Simulation 5, the data were subject to an analysis of the RMS prediction error for each segment, with the word-position as a factor. The results are illustrated in Figure 6.6.

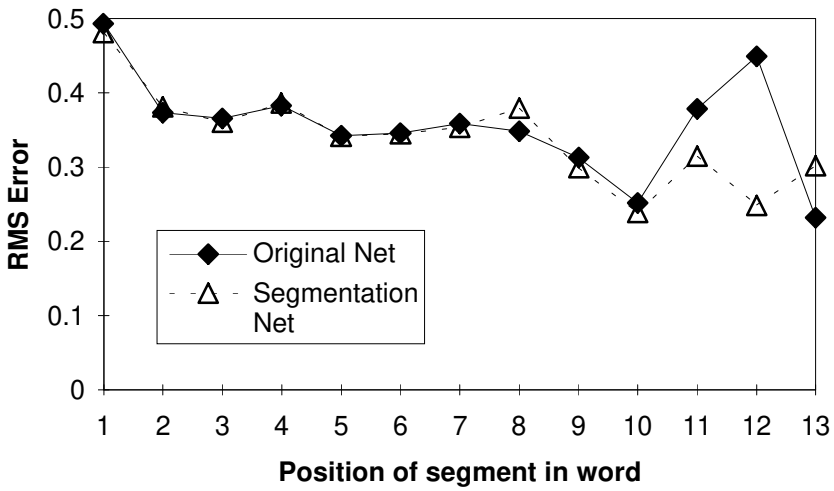

Figure 6.6. Comparison of Segment prediction error for the two networks.

As before, there was a small negative correlation between prediction error and word position ($r = -0.26$; $p < 0.01$). However, the close similarity between the shapes of the error curves for the two networks reveals that the segmentation information in training has not led to any change in the representations formed by the network. A paired sample t-test found no significant difference between the error values of the two networks ($t = 0.43$; $p > 0.1$).

Lexical Effects on the Resolution of Phonological Ambiguity. A replication of Simulation 3 using the segmentation network was carried out with all experimental details identical to those described in Section 4.4.4. The data consisted of a set of scores representing the amount of deviation from a coronal response for each ambiguous non-coronal segment. These were subject to a two-way ANOVA with the factors Bias (the lexical bias on the ambiguous segment) and Place (the place of articulation of the ambiguous segment). As in Simulation 3, there was a significant effect of Place ($F[1,84] = 11.76$, $p < 0.01$)[41] and no effect of Bias ($F[1,84] = 0.072$, $p > 0.10$) nor any interaction ($F[1,84] = 0.29$, $p > 0.10$). The mean deviation scores for each condition are illustrated in Table 6.3.

[41]This effect may reflect the representational difference between coronal-labial and coronal-velar contrasts in the Jakobson, Fant & Halle (1952) feature set (see Chapter 4).

Table 6.3. Deviation from coronal scores in the lexical effect test. 0 = coronal response, 1 = non-coronal response.

	Lexical Bias	
Place	+Cor bias	-Cor bias
Labial	0.48	0.42
Velar	0.67	0.69

DISCUSSION

Segmentation. The above analyses reveal a rather surprising picture. The network, when presented with a novel sequence of familiar words identifies word boundaries extremely accurately. This is true even when the network is forced to make its response before the boundaries themselves. However, the analysis of the prediction error in the network shows that the network is not using word-like structures to make these decisions: the predictions of the segmentation network are almost identical to the predictions of the original network, suggesting that the network's responses are largely guided by co-occurrence statistics.

It could be argued that the tasks of word-boundary detection and segment prediction are separate and therefore may operate on independent internal representations: for the segmentation task, using word representations, and for the prediction task, using co-occurrence statistics. However, an examination of the way networks learn renders this explanation unlikely. In general, it is difficult to train distributed backpropagation networks to perform two tasks without interaction unless they are truly independent. Indeed, the degree of interaction found in these networks is normally regarded as one of their attractive qualities (Norris, 1992; Seidenberg & McClelland, 1989; Rumelhart & McClelland, 1986). In this case, the two tasks require very similar types of information and so it is implausible that word structure would be discovered and utilised for one task but not for the other.

It follows that the information the network exploits to identify word boundaries must be non-lexical. One rich source of information on the localisation of word boundaries comes from the analysis of phonotactically legal and illegal segment sequences. Harrington, Watson & Cooper (1989) used a statistical analysis of phonologically variant speech to examine the value of these cues in the identification of word boundaries. Their analysis used a large speech corpus (12000 words) to identify all possible permutations of three segments across word boundaries. They found that if a narrow transcription of speech was used only 18% of these permutations were also found word-internally. For example the sequence /mgl/ occurs across word boundaries (as in *dim glow*) but not within words. Therefore the co-occurrence of these three segments in a stream of speech allows a word boundary to be identified between the /m/ and the /g/. Harrington et al. found that an analysis of a novel transcription of speech using these co-occurrence rules allowed 41% of word-boundaries to be correctly identified with only 9% false positives[42].

The strategy employed by the network in the current study can be seen as a more probabilistic version of the method used by Harrington et al. Although the many differences in experimental method make comparison of results difficult, it seems that the network is able to identify a greater proportion of word boundaries than the rule based model. I can see two reasons for this improved performance. One is that the network is not restricted to a co-occurrence window of three segments, allowing it to utilise longer and shorter distance regularities where applicable. Secondly, the co-occurrence statistics can be applied probabilistically, allowing more effective use of the information available during training.

For example, there are many sequences of segments, such as /kbæ/ in *kickback*, which occur within words, but only very infrequently. These will be designated as legal word-internal string in the Harrington et al. model, despite the fact that they occur much more often between words. But in the network model, each occurrence of such a string has only a limited effect. Positive tokens (those

---

[42]These are the figures for the most successful analysis, using a narrow-class transcription of citation form speech.

which occur across word boundaries) will alter the weights in the network to make the segmentation nodes' activations slightly stronger, and negative tokens will alter the weights in the opposite direction. But the final response of the trained network depends on the proportion of positive to negative tokens. This makes the network fallible, in that it will make incorrect predictions where a specific instance during testing goes against the trend, but it allows the network to make a best guess at any point in the processing of speech.

The accuracy of the network in its segmentation of the test set is in fact much closer to a probabilistic version of the Harrington et al. analysis (Cairns et al., submitted). Using only bigram statistics, their analysis identified 75% of word boundaries, at a hits to false alarms ratio of 4.7:1, whereas a similar analysis using trigram statistics considerably improved performance (93% detection at 9:1). The performance of the network model falls close to the bigram analysis, with 73% detection at 2.6:1. The network, however, is at a disadvantage in this comparison due to the limited size of the training corpus (3700 words) from which generalisations must be made, compared to the 460,000 words used in the Cairns et al. analysis.

The performance of the segmentation network when forced to respond before the word boundary (i.e. on presentation of the word-final segment) also merits examination. At this point the network cannot use cross-boundary phonotactic cues in its predictions since only one side of the word boundary has been presented. Instead, it must base its predictions on generalisations about word-length (i.e. the distance from the last identified boundary) and word-final sequences of segments. An informal analysis of the errors of the segmentation network when responding in this early position backs up this explanation. The network often missed word boundaries following words ending with vowels (e.g., *true*), and also erroneously placed word boundaries after consonant clusters such as /nt/ in *painted*.

The performance of the network reflects the richness of information contained in the statistics of the speech stream and shows how partial and probabilistic cues to the presence of word boundaries can be learned and exploited by a connectionist network. The network's segmentation of its input is by no means perfect, but the architecture chosen prohibits the use of certain types of information. In particular, the cues presented by prosodically strong syllables in the segmentation of speech (Cutler & Butterfield, 1992; Cutler & Norris, 1989) cannot be fully utilised by the network. The structure of the output windows of the network allows the first two segments of a word to be used in the assessment of a word boundary. Thus, if the first vowel of the word is one of these segments (as in *apart* or *tap*) the statistical properties of that vowel can be used in the evaluation of the presence of a word boundary. But in many cases (e.g., *trap* or *strap*) the first vowel occurs further into the word and so the prosodic strength of the syllable remains unknown to the network. This problem could be solved by increasing the temporal span of the output windows.

Internal Representation. Despite the fact that the network has learnt to segment speech into words, this is not due to any change in the structure of the internal representation built up by the network. The network shows exactly the same error behaviour in the prediction of segments as the original network not trained on word-boundary detection. The results of the replication of Simulation 3 also show that segmentation information during training is also insufficient to improve the application of co-occurrence statistics to the resolution of phonological ambiguity. These findings are both contrary to expectations.

It is possible that the network was, in fact, too good at picking up low-level segmentation cues to allow word representations to be developed. Harris (1991) and Elman (1993) have shown that the learning ability of backpropagation networks is not constant over time. The sigmoidal function used in the calculation of both activation values and weight changes has the effect of giving network a strong initial ability to learn. As the network learns and activations tend towards extreme values this ability decreases.

For this reason the representations formed early on in training are vital to the end performance of the network. In this case, the simplest cues to the presence of word-boundaries are those that require the least mediating structure. The network will therefore initially learn co-occurrence cues to segmentation. The ability of the network to develop more complex structures, such as word representations, depends on the success of the initial representations in carrying out the task. Here the segmentation output is reasonably accurate, meaning that output values are extreme (i.e., near 1 or 0) and the output error is low. Both these facts imply that even though the error function is not at a

global minimum (since presumably word representations would provide useful additional segmentation cues), any subsequent learning will be extremely slow.

Elman (1993), confronted with a similar learning problem when examining the formation by a network of long-distance dependencies in syntactically complex sentences, found two solutions. One was to simplify the initial patterns presented to the network and gradually increase the complexity of the training instances. His second solution was to initially reduce the memory capacity of the network so that only simple short-term dependencies could be learned and then gradually increase the memory span. He found that both these methods forced the network to learn the correct intermediate representations necessary for the resolution of the more complex dependencies.

Of these solutions, the latter does not appear to be applicable to the current problem, since the two types of representation involved (co-occurrence information and word forms) do not differ in the amount of memory capacity required. However, it is plausible that the initial training data could be adjusted to encourage the formation of word representations early on in learning. This would involve initially training the network using a small number of words and gradually increasing the size of the training corpus, ensuring that at any point during training, the internal representational structure most beneficial in terms of error reduction is word-based.

The drawback of this approach is that it is not clear that the experience of language learners supports such a training procedure. Although the speech encountered by young children is, in many ways, quite different to the experience of adult listeners, the vocabulary they encounter is still likely to be large (certainly larger than the 1000 words used in the current simulations). Starting the network on an even smaller vocabulary would therefore give the network an unrealistic advantage over human language learners.

It may be that more direct cues are needed in order to develop word-like internal representations. Cues to the meanings of words are often available via the integration of perceptual experiences (Nelson, 1974; Gentner, 1982). This is especially the case for nouns which form the bulk of the words children first learn (Gentner, 1982; Nelson, Hampson & Shaw, 1993). These cues may be augmented by the behaviour of adults in their interactions with children (Bruner, 1983). An obvious example of such behaviour is the Naming Game (Ninio & Bruner, 1978) in which adults direct a child's attention to an object and ask the child what the object is. The child's response is followed by either positive or negative feedback.

Although it is difficult to determine the exact role these cues have in the formation of word concepts, there is no doubt that they are influential. In simplistic terms, they correspond to semantic teaching input to a recurrent network of the type I have been using. This type of input forces the network to develop internal word representations, as this is the only way the correct semantic output can be determined. Unfortunately, the implementation of such a model, even in quite small scale simulations, is currently impractical due to the amount of computer time required to train large backpropagation networks. This problem has led Norris (1991) to abandon network simulations in favour of dictionary searches. In my final simulated experiment, I examine another way in which a network can be encouraged to form word representations, by the manipulation of the frequency of words in the training corpus.

6.5 Simulation 7 — Word Frequency and Lexical Effects

The above simulations show that the use of the simple co-occurrence information about speech segments is not sufficient to produce pseudo-lexical biasing effects. Furthermore, the use of segmentation information in training does not force word-representations to be learned by a simple recurrent network. It seems that more reliable information is needed in order to model the human behaviour exhibited in Chapters 3 and 5 using a connectionist learning approach.

In this section I examine the performance of the network on the resolution of phonological ambiguity in high frequency words. These are the words in which the information gained from the analysis of statistical regularities is strongest and so they are the words most likely to display lexical or pseudo-lexical effects. To some extent, the increased frequency of presentation may compensate for the inadequacies of the training regime, and in particular the size of the training corpus from which the statistical regularities must be derived. All the simulations presented here have used a training corpus

of just over 3000 words by token, and 1000 by type (albeit repeated many times). This amount of speech might represent an hour or two's gentle conversation, compared to the years of speech a child is exposed to before learning to understand even simple grammatical constructions. It is plausible that using such an unrepresentative sample of speech only the coarsest lexical statistics can be used. If so, this should show up as a lexical effect of the most frequent words.

The test I report is a replication of Simulation 3 in which the network is retrained on an adapted version of the speech corpus in which a number of test words are given an artificially high token frequency. The test words are a subset of the words used in the original simulation, of which half present a lexical bias towards a coronal resolution of a surface ambiguity and half bias in the opposite direction.

In expectation of an effect of the lexical bias of the preceding context, a further variable is considered in this simulation. One of the motives for the simulations in this chapter was to explain the findings of the phoneme monitoring studies of Chapter 5. These experiments found compensatory effects on subjects' perceptions of phonologically changed segments for both real words and nonwords (although the effects were weaker for the nonwords). To examine the network's predictions in this case the test items were presented twice, once with normal preceding context and once with the biasing word changed so that the context remained word-like but was not a word.

MATERIALS AND DESIGN

Because of the lack of high frequency words containing word-final plosives or nasals in the original training corpus, a new corpus was created in which the token frequencies of selected words were manipulated. The test items for this simulation were selected randomly from the two sets of test items used in Simulation 3 (see Appendix D). Twelve of these were items which contained a phonologically ambiguous surface 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 velar place. The training corpus for the network was altered so that the twenty-four biasing words in these test items had token frequencies of 30 (out of a total of just over 3000 words). This was done by, for each biasing word (e.g., *David*, *job*), 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 presented to the recurrent network described in Section 4.4 and trained using the procedure described in Section 4.4.1. The test items were then presented to the network as in Section 4.4.4.

The test words were presented in two forms. In the real word condition, the items were presented unchanged. In the nonword condition, the word-initial consonant clusters were changed so that the resultant sequence was a nonword. For example, the stimulus [deɪvɪbpeɪ] became [seɪvɪbpeɪ]. The procedure used for the segment substitution was similar to that used in Experiments 3 and 4. Single, word-initial consonants were replaced by single consonants and consonant pairs were replaced by consonant pairs. One test item contained a biasing word with an initial vowel which was replaced with another vowel. In all cases, the segments were chosen to maximise the featural difference between the real-word and the nonword segments. However, the deviations used in the simulation were on average more minimal than in Experiments 3 and 4, since only 2 out of the 24 nonword stimuli employed changes of two segments (compared to 23 out of 48 experimental stimuli).

In Simulation 3, the network's response to the ambiguous segment was only evaluated on presentation of the following segment. This was because Simulations 1 and 2 had shown that the network produced little or no compensation for phonological change before the following context was known to be validating. However, in this simulation a lexical effect is predicted, which may cause compensation before this point. Because of this expectation, 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.

RESULTS

The results of the simulation are in the form of a deviation from a coronal response, found by taking the mean deviation from the coronal values of the phonological features (Jakobson, Fant & Halle, 1952) that discriminate place of articulation (*grave* for the surface labial segments, *diffuse* and *grave* for the surface velars). A low value of this measure indicates that the network is treating the input as being underlyingly coronal. The mean deviation values for each condition are presented in Table 6.4.

Table 6.4. Deviation from coronal response scores in Simulation 7. Responses for the nonword conditions are presented in parentheses.

		Monitoring Point		
Lexical bias	Surface place	Before ambiguity	At ambiguity	After ambiguity
+Coronal	Labial	0.04 (0.12)	0.53 (0.62)	0.43 (0.60)
+Coronal	Velar	0.09 (0.11)	0.59 (0.74)	0.41 (0.35)
-Coronal	Labial	0.79 (0.52)	0.93 (0.86)	0.82 (0.82)
-Coronal	Velar	0.71 (0.63)	0.95 (0.96)	0.80 (0.80)

A four-way ANOVA was carried out on the data, using the variables Bias (the lexical bias), Place (the surface place of articulation of the ambiguity), Lexical Status (of the biasing context) and Monitoring Point (before, at or after the ambiguous segment).[43] 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. The interaction between Lexical Status and Bias ($F[1,20] = 6.18$, $p < 0.05$) showed that the lexical biasing effects were also stronger for the word carriers than for the nonword carriers. No other effects approached significance.

To check whether there was a significant biasing effect for the nonword carriers, the data at the three monitoring points for both real words and nonwords were analysed separately in two-way ANOVAs. At each monitoring point the effect of the Bias variable was significant for the real word conditions (Before: $F[1,20] = 75.22$, $p < 0.01$; At: $F[1,20] = 9.97$, $p < 0.01$; After: $F[1,20] = 8.78$, $p < 0.01$) For the nonword conditions, the effect of bias was significant for the monitoring points Before ($F[1,20] = 25.18$, $p < 0.01$) and After ($F[1,20] = 5.02$, $p < 0.05$), but only marginally significant for the monitoring position At ($F[1,20] = 3.60$, $p = 0.072$). Again, no effects involving Place were significant. These biasing effects are illustrated in Figure 6.7.

[43]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.

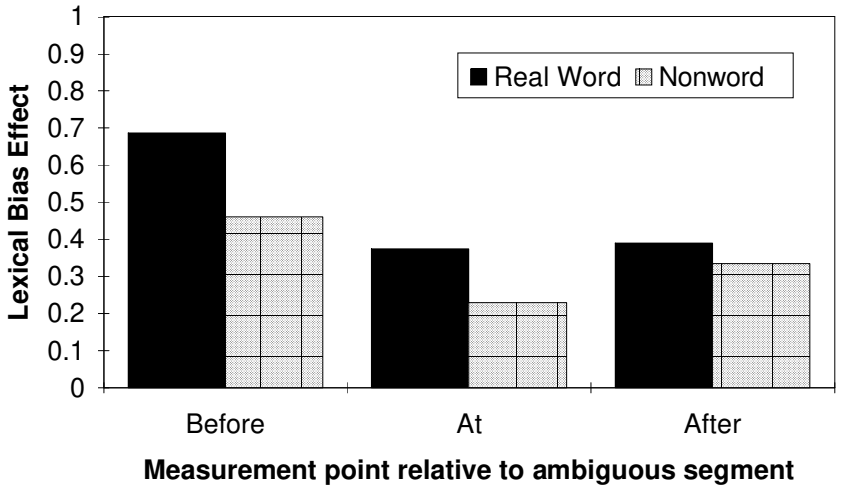

Figure 6.7. Difference scores ([-coronal] bias – [+coronal] bias) for the lexical bias simulation.

DISCUSSION

The results of Simulation 7, examining the effects of lexical context on the disambiguation of phonological change in high frequency words, form a clear and interesting pattern. Looking first at the responses in the real word conditions, 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 behaviour 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 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.

The nonword conditions, where the initial segments of the biasing words were changed, show a similar pattern. The effects of lexical bias are still prominent for these conditions, reaching significance for two of the three probe positions and showing a marginally significant effect at the third. But the interaction between Bias and Lexical Status found in the main analysis indicates that these biasing effects are smaller than in the real word conditions. Despite the fact that the preceding contexts of the target segments in these conditions are nonwords, they show reduced biasing effects of the real words they resemble.

Comparison with Experimental Results. These results provide a basis for the explanation of the inconsistent results described in the introduction. Experiment 1 revealed that tokens such as /swik/ for *sweet* did not mismatch the lexical representation of the base word when the following context was spliced out. The original network did not predict this result since surface non-coronal segments were treated at face value by the network until the following context showed assimilation of an underlying coronal to be viable. Here, the preceding biasing context of the ambiguous segment provides sufficient information to provoke some phonological compensation *before* the following context is known. The mean deviation for the coronal-bias conditions on presentation of the target segment is 0.56. While this value does not represent a clear coronal response, the target is at least treated as

ambiguous. The equivalent figure for the original network in Simulation 1 is 0.90 representing an unambiguous non-coronal (i.e. surface) response.[44]

Experiments 3 and 4 showed that the lexical status of the carrier word affected subjects' responses to phonologically ambiguous segments. In Experiment 3, subjects were more likely to respond to a surface velar or labial segment as coronal if the coronal segment completed a real word. For example, subjects monitoring for the segment /d/ were more likely to respond to the word-final labial in [bleibpiəs] (underlyingly *blade pierce*) than in [skeibpiəs]. This effect shows up in the network's responses to the coronal bias conditions in this simulation. At all monitoring points the network's responses to the surface non-coronal segments show less deviation (i.e. are more coronal) for the real word conditions than for the nonword conditions (see Figure 6.8). However, this effect does not reach significance level in an ANOVA on just the +coronal bias conditions ($F[1,10] = 2.12$, $p > 0.10$), perhaps because of the small number of observations in this subset of the data.

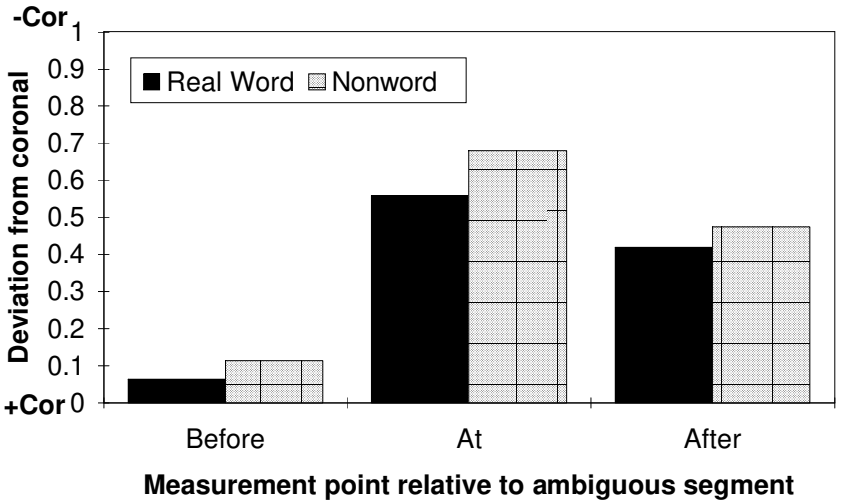

Figure 6.8. The network's response to the +coronal bias conditions collapsed across surface place of articulation.

In Experiment 4, the word-initial phonological context segments were used as targets in a phoneme monitoring task. Subjects responses were delayed when the target segment was preceded by an unviably phonologically-changed segment. For example, responses to the target segment /p/ were faster for phrases like [toωbpaωnsɪN] (underlyingly *toad pouncing*) than [toωgpaωnsɪN] where the change from /d/ to /g/ is rendered unviable by the following /p/ according to assimilation rules. There was some evidence that this effect held even when the initial segment or segments of the carrier were altered to make it a nonword. I have argued that for this result to be accommodated in a connectionist model of this type a number of conditions must be met. Firstly, the network must be able to show pseudo-lexical effects in its responses to ambiguous segments. Secondly, these effects must also be evident when the biasing context is a word-like nonword (e.g., *groad* for *toad*). These two properties allow surface velar segments to be perceived by the network to be assimilated underlying coronal segments. The third condition is that the model is able to use the presence of phonological change of this type to predict the following context that makes this change phonologically viable.

Of these properties, the first two have been demonstrated in the above simulation: the network shows pseudo-lexical effects on the processing of surface labial and velar segments *before* the following contexts of these segments are known. These effects are also displayed when the network is presented with words whose initial segments have been changed but are still word-like: the network shows a biasing effect of the familiar token the nonword resembles. The third property has not been examined

---

[44]Simulation 1 used different stimulus words, but since there were no lexical effects in the original network, this is assumed to be unimportant.

in the above analyses but can be assessed using the same data set. If the network is able to use preceding pseudo-lexical cues to predict the place of articulation of the following context of a phonological change, we would predict a difference in the responses of the network between the +coronal bias -coronal biasing conditions. Specifically, if we examine the network's prediction for the following segment on presentation of the word-final ambiguous segments we would expect a difference between the two conditions. For the -coronal bias conditions, the occurrence of a word-final labial or velar segment is uninformative and so we would expect no particular bias in the prediction of the place of articulation of the following segment. But for the +coronal bias conditions, the occurrence of a word-final surface labial or velar can be used as a cue to the place of articulation of the following segment. Therefore, the network's predictions for the place of the following segment should show a greater deviation from a coronal response in the +coronal conditions than in the -coronal conditions.

Analysis of the actual responses showed a small shift in the anticipated direction. In the coronal bias conditions, the network's predictions produced a deviation from a coronal response of 0.44 compared to 0.39 for the non-coronal bias conditions. However, this difference was not significant in an analysis of variance ($F[1,20] < 1$).

In summary, the effects displayed in this simulation allow a number of previously incompatible experimental findings to be explained. The most important difference between the behaviour of the network here and the previous results is that cues in the preceding context of a surface velar or labial segment are strong enough to force the network to compensate for assimilation before the following context of the segment is known. This makes the network's responses more compatible with the results of Experiment 1, where tokens containing phonological changes did not cause mismatch when presented without following context. It also offers an explanation of the lexical effect found in Experiment 3, in which subjects' responses to ambiguous segments were affected by the lexical status of the carrier word. The network shows graded effects of lexical cues to the identity of segments, with the size of effects depending on the degree to which the preceding context matches known strings. This property goes some way towards an explanation of the results of Experiment 4, where phonological viability affected response times to monitor for the following context of an assimilation. However, the network is still unable to use these cues to make firm predictions about the place of articulation of the context segment.

6.6 General Discussion

The simulations described in this chapter represent an investigation of the types of information needed for a recurrent network to show lexical or pseudo-lexical effects on the processing of phonologically ambiguous segments. Two possible types of internal representation have been discussed here; one developed to exploit inter-segment regularities in the speech stream and one encoding representations of word units. Simulation 5 has shown that the network advanced in Chapter 4 constructs hidden unit representations that allow it to make predictions based on co-occurrence statistics; largely using the regularities contained in two-segment and three-segment sequences (although some longer distance effects were observed in Simulation 4, which examined the memory span of the network). Simulation 3 in Chapter 4 showed that these regularities are not enough to produce reliable biasing effects on the disambiguation of phonologically changed segments.

Simulation 6 examined the performance of a similar model of speech processing which is trained to predict word boundaries in addition to the tasks required of the original model. This network learns to use the phonotactic regularities in words to accurately predict the location of word boundaries in novel word-sequences. However, despite the success of the network in this additional task, there was no significant difference in the way it represented these regularities. It seems that connectionist networks do not need to employ word-like representations in order to carry out the task of segmenting speech into words, due to the richness of phonotactic information in speech.

The final simulation showed that despite the fact that lexical effects on phonological disambiguation cannot be adequately explained in terms of the co-occurrence statistics of speech, the network did show pseudo-lexical effects when tested on a set of high frequency words. The performance of the network in this case more closely mirrored the performance of humans in the cross-modal priming and phoneme monitoring studies reported here.

These results force me to revise the theoretical standpoint I have taken with regard to the perception of phonologically variant speech. The low-level statistical approach to the resolution of phonological ambiguity, while fully able to cope with asymmetry in phonological information as well as context-dependent phonological inference, is unable to fully explain the pattern of results regarding lexical bias. It is clear that to maintain this kind of approach, word-like representations must be learnt and encoded in the hidden unit representations of the network. I see two possible procedures by which this encoding could be attained. One is that the response of the network to the high frequency words in Simulation 7 should be thought of as an approximation to the network's response to all words, given a more realistic, varied and extensive training corpus of the kind a human language learner would be exposed to. This interpretation assumes that word or morpheme structures can be learned from exposure to just the co-occurrence statistics of speech. Elman (1991, 1993) has shown that some grammatical structures can be learned from exposure to sequences of word units, but there is doubt about whether the simplifying methods employed to encourage the network to learn the more complex constraints (Elman, 1993) are applicable to the problems involved in the acquisition of word structures.

Alternatively, one could assume that aspects of meaning are used to direct the acquisition of word forms. This forces the network to learn distinct word forms, since it must be able to map similar sequences of segments onto very different areas of semantic space. For example, *sleep* and *leap* are very similar in terms of their phonological co-occurrence statistics but quite different in terms of their meaning. This alternative runs into problems of bootstrapping during learning, since word meanings can only be efficiently combined with their phonological forms once the speech stream has been segmented, whereas speech can only be fully segmented once the meanings of words are known. It is possible that segmentation can only develop once a sufficient number of words have been learnt in isolation (Suomi, 1993). However, it is more likely that these problems are solved gradually. For example, Simulation 5 and the work of Cairns, Shillcock, Chater & Levy (submitted) have shown that the prediction statistics of this type of network provide a reasonable cue to the presence of word boundaries. These and other low-level cues can provide a partial segmentation of the speech stream which in some cases will be matched onto the meanings of words. This improves the capabilities of the network in the extraction of semantic information, which in turn provides further cues to the identity of word boundaries.

Both the possibilities described above lead to similar states, in which word-structures are used as an additional source of information about the identity of phonological ambiguities. Indeed, the question of how these representations are derived is in some ways of secondary importance in this thesis. The primary objective of the modelling work was to construct an explicit connectionist model of the role of variation in lexical access. The model I describe here does not fundamentally differ from the one described in Chapter 3, but it was originally advanced as a *pre-lexical* model of phonological processing. The use of word-like structures in the processing of ambiguous segments means that the model becomes difficult to categorise in this way. Figure 6.9 provides a better illustration of my view of the interaction between the various forms of information in spoken word recognition.

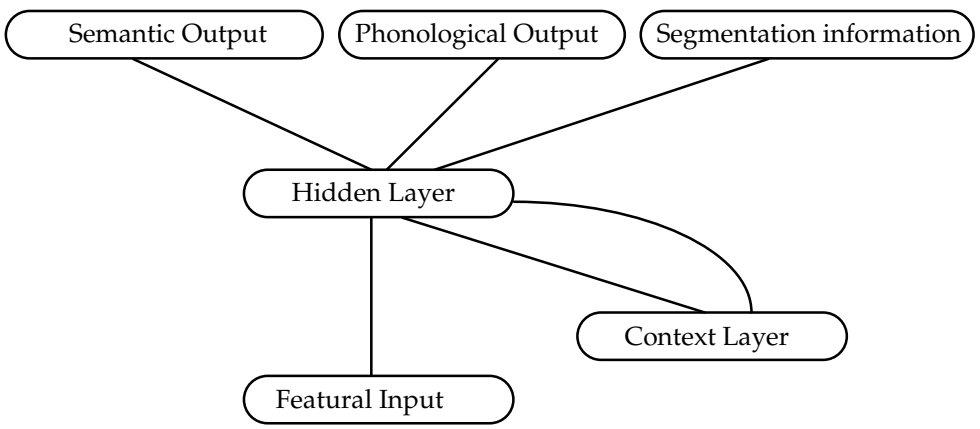

Figure 6.9. A connectionist model of spoken word recognition.

This model is like the segmentation network illustrated in Figure 6.4, with the addition of a third task, the retrieval of semantic information. In this model, the process of word recognition involves the mapping from speech input to all three types of information at the output level. Thus, the retrieval of phonological information (i.e. the underlying form of the words) will normally interact strongly with the retrieval of semantic knowledge, causing the lexical effects found in the phoneme monitoring experiments. But when access to lexical information is blocked by the use of mismatching segments, phonological output is still possible, allowing weaker interaction with the partial lexical representations formed in the hidden unit layer. By this account, phonological inference can be thought of as lexical, since it is influenced by the retrieval of word-identity information, but it may also be thought of as non-lexical, since it can occur independently of the recognition of words. But in both cases, the retrieval of the underlying phonological form of speech makes the best use of whatever cues are available.

Similarly, the segmentation of speech into words operates in a maximally efficient fashion. As well as phonotactic and other low-level regularities, this model allows interaction between lexical information and word-boundary recognition. In other words, learned knowledge about the length of particular words can be used, where relevant, to guide the network's recognition of word boundaries.

The interaction between the tasks of segmentation and word recognition may also enhance the performance of the network on the task of word recognition. Norris (1991) argued that recurrent networks do not satisfactorily model human spoken word recognition since the word-activations that they produce are not adequately segmented (see Section 4.3.2). A network, trained to map from temporal speech input to just word nodes, activates all the known words embedded in the speech, regardless of word boundaries. For example, presented with the sequence /kætəlog/ (*catalogue*), the network would activate, at some point, the word nodes for *a*, *at*, *cat*, *cattle* and *log*. Norris resolved this inadequacy by proposing that the output of the recognition network was fed into an interactive network that used competition between lexical candidates to segment the speech. However, the addition of segmentation nodes, as described above, allows the model to simulate lexical competition while retaining the simplicity of the recurrent network architecture. The network has the opportunity to learn that high segmentation values correlate with a shift in semantic output from the meaning of one word to another.

A more enriched segmentation representation, such as would develop if the network were trained to output the position in the current word of *each* segment it encounters, would increase the level of lexical competition even further. Consider the *catalogue* example again, when presented as input to such a network. For each segment, the network must predict its position in the current word. So the sequence of responses {1,2,3,4,5,6,7} would suggest that the network has recognised a single seven-segment word, *catalogue,* whereas the sequence {1,2,3,1,1,2,3} would be more compatible with the output *cat a log*. Concentrating on the 5th segment in the sequence (the /l/), the network could activate a node corresponding to word-position 5, which is compatible with the word outputs *cattle* and *catalogue*. Alternatively, the network could opt for word-position 3, as in *telepathy* or position 2, as in *along*. The response of the network is also likely to be influenced by its knowledge of co-occurrence statistics and phonotactic constraints. But whatever the response of the network, this information should influence the semantic output of the network with, for example, a response of 5 facilitating the output of the *cattle* and *catalogue* nodes and inhibiting other semantic outcomes. The addition of the segmentation task at the same level as the word recognition task thus allows the network to display effects of lexical competition in its responses to both tasks.

This analysis remains speculative, but I believe it is an approach worthy of further investigation. It forms part of a general framework for the modelling of psycholinguistic processes in which perceptual tasks are carried out in parallel, with interaction between the information types occurring as a product of perceptual experience. The recurrent backpropagation network offers a simple environment for modelling both the development and the adult performance of such processes.

Chapter 7 — Concluding Remarks

In conclusion to this thesis, I shall summarise the main findings of my research, with reference to theoretical models of variation in speech perception. I shall argue that the behaviour described here is best explained in terms of a context-dependent inference process, which forms the basis for speech perception. The connectionist model I have developed provides an explicit and plausible framework for the modelling of this process.

7.1 Models of Variation in Speech Perception

7.1.1 Variation as Noise

The simplest hypothesis considered was that all variation, phonologically viable or not, is treated as noise by the word-recognition process. In Chapter 2, I reviewed the experimental evidence on the tolerance of deviation in speech perception, finding that (at least in terms of the initial perceptual evaluation) very small deviations from the canonical pronunciation of a word have strong effects on the word recognition process. The two simulations reported in Chapter 2 showed that the competitive environment used by TRACE (McClelland & Elman, 1986) to model the matching process is unable to capture the dynamics of this behaviour. In particular, word-final deviations such as *apricod* for *apricot* (Marslen-Wilson & Gaskell, 1992) swiftly reduce the activation of the base word. TRACE is unable to exhibit this behaviour since it relies on lateral inhibition from other lexical candidates to eliminate a word candidate from the matching process. The second of the simulations showed that the response to mismatching information is better modelled as direct inhibition from featural or phonemic nodes to word nodes.

The remainder of the thesis was devoted to the effects of natural phonological changes on speech perception. Chapter 3 supports the findings of Chapter 2 — that minimal unnatural deviations have strong mismatching effects on the activations of lexical candidates — using a different experimental task: cross-modal repetition priming with sentential context. However, Experiment 2 showed that these same minimal deviations, when occurring as natural phonological changes, have no inhibitory effect on lexical activations. Thus, the variation as noise approach is untenable, since it makes no distinction between viable and unviable phonological changes. The fact that humans do not treat variation as noise is reassuring, but not surprising. There have been experimental demonstrations in all kinds of situations that the perceptual system is extremely sensitive to tiny manipulations of the speech signal (e.g., Whalen, 1984; Martin & Bunnell, 1982; Repp, 1983; Warren & Marslen-Wilson, 1987). The moral these studies provide is that when variation in the speech signal is informative — as most, if not all, natural variation is — the perceptual system will utilise this information to allow maximally efficient processing of speech.

7.1.2 Lexical Representation of Phonological Change

In Chapter 3, I made the distinction between two major classes of theory: those that deal with phonological change lexically, and those that rely on inferential processing to handle surface ambiguities. The lexical theories were further divided into three types. Firstly, there are theories in which lawful phonological changes are tolerated by adding to the lexical representation of a word (Harrington & Johnstone, 1987). This allows us to distinguish between random variation and natural phonological change, at the expense of a rather bulky lexical specification. A more attractive possibility is that phonological change is dealt with by lexical abstraction (Stevens, 1986; Lahiri & Marslen-Wilson, 1991). This allow the same distinction to be made, but offers the prospect of a more systematic explanation of why some changes should be tolerated, but not others, by linking abstraction to phonological theories of lexical representation. A third type of lexical theory (Klatt, 1979, 1989) assumes that phonological changes are dealt with by addition to the lexical form representation, and that these changes are linked to their validating contexts in a network. This view of lexical access challenges the standard assumption that phonological form representations are stored in discrete word or morpheme-like units.

This third theory is the only one that is able to accommodate the experimental findings of Chapter 3. Since they do not represent the context of a phonological change the two word-based lexical theories cannot explain the strong effects of phonological context found in Experiment 3. Phonological changes with viable cross-boundary context had no mismatching effects on the base word activations, but the same changes in unviable context mismatched strongly. Both word-based theories would predict that the phonological changes should be tolerated whatever the following context.

The Klatt network did accommodate the pattern of results found in Experiment 3, but had more trouble with the phoneme monitoring results of Chapter 5. These experiments showed that phonological viability effects can occur in the perception of nonwords. This result seems impossible to model in representational models incorporating the traditional view of a unitary lexicon, since, as Chapter 2 showed, the deviations in the nonwords strongly disrupt access to lexical knowledge.

It seems, then, that the accommodation of phonological change on a purely lexical basis is also untenable. However, this does not mean that the representational units of phonological form are irrelevant to this discussion. An important difference between representational and noise approaches, that this thesis did not address, is that a representational approach allows some within-word changes, but not others, whereas the noise approach doesn't care what the changes are, just how strong they are. Thus, a representational theory would predict a difference between the mismatching effects of [swik] (*sweek*) for *sweet* and [tSoɷt] (*chote*) for *choke*. As Experiment 1 showed, before its following context is known, [swik] is a viable token of *sweet*, since it could be the product of place assimilation. But irrespective of the following context, [tSoɷt] can never be a viable alternation of *choke*, since non-coronal segments do not assimilate to coronals. A representational approach would therefore predict a difference in the mismatching effects of these changes, whereas a noise approach would predict none.

Nevertheless, the experiments reported here create serious problems for a purely representational model accommodating phonological change. It seems that irrespective of the units of lexical representation, there is a need for a process of phonological inference in speech perception, which exploits phonological context in order to discern the underlying form of surface ambiguities.

7.1.3 Models of Phonological Inference

Context-dependent evaluation of segments of speech is not a new idea. The effects of coarticulation in speech perception have been widely studied, and a mechanism of compensation for these changes is made explicit in the TRACE model (Elman & McClelland, 1986). However, the changes I have studied here, while they can be seen as extremes of the process of coarticulation (e.g., Browman & Goldstein, 1991), create special problems in the perception of speech. TRACE works by using fixed links between the features underlying coarticulated phonemes to make them perceptually less similar. But in the case of place assimilated speech this mechanism fails, since the result of place assimilation, such as the [kk] derived from /tk/ in *sweet kid*, is underlyingly ambiguous: the underlying form could equally be /kk/ as in *weak kid*. Compensation in terms of a fixed link is therefore not an option: a more flexible approach is necessary to cope with the underlying ambiguity that neutralising phonological changes cause. Because of the lack of a viable model of phonological inference in the psychological literature, I have proposed and evaluated a connectionist model of phonological inference.

PRE-LEXICAL PHONOLOGICAL INFERENCE

The model described in Chapter 4 is a pre-lexical model of phonological inference. Based on a recurrent connectionist model of speech perception (Shillcock, Levy & Chater, 1991), the network was trained to identify the surface contextual cues that mark place assimilated segments, mapping from surface to underlying forms of speech. Simulations 1 to 3 showed that the network was sensitive to the phonological context of place assimilations, and was able to alter underlying hypotheses swiftly as new information was encountered. However, the network showed no effects of pseudo-lexical bias in its responses to assimilation.

The network, therefore, accommodated the strong effects of phonological viability found in the results of all four experiments. The model also predicted the graded effects of cues to assimilation found in the phoneme monitoring experiments, as well as the viability effects for nonwords suggested by

Experiments 3 and 4. However, the model did not exhibit the lexical effects found in these experiments, nor the interaction between lexical status and phonological inference.

Chapter 6 examined the reasons for the lack of pseudo-lexical effects in this model. The structures the network learned did not conform to word-like representations: instead, the network relied on fairly short-distance dependencies, similar to bigram and trigram statistics, in its processing. It is likely that the performance of the model would be improved by training on a more realistic corpus, both in terms of size and content. However, it is unclear, even using such a corpus, whether pseudo-lexical regularities are strong enough to model the effects found in Chapter 5, and in particular, the interaction between lexical status and phonological inference found in Experiment 3, with a similar trend in Experiment 4.

A number of possible improvements to this model were evaluated in Chapter 6. Firstly, the effects of word-segmentation information on the performance of the model were examined. The network was trained on the same corpus but was taught to segment the speech by activating a set of segmentation nodes at word boundaries. This extension of the network was driven by the expectation that the extra information would improve the intra-word regularities the network learned. In fact, despite the accurate performance of the network on the segmentation task, the representations built up the network were much the same. The segmentation performance, however, is an excellent illustration of the value of the overall approach taken here to the problems of speech perception. The network, was able to learn and apply, in a probabilistic manner, numerous phonotactic and other cues to word boundary position present in the surface structure of speech.

INTERACTIVE PHONOLOGICAL INFERENCE

The final model examined, although not in great detail, is a model of phonological inference which is no longer strictly pre-lexical, but which still shows phonological inference in nonwords. The model is an extension of the pre-lexical model above, which incorporates a third task alongside segmentation and the derivation of underlying phonological form. The final task is the mapping onto the semantic representation of the words contained in the stream of speech. This allows stronger and more interactive lexical effects on the inference process, while still allowing phonological mapping to occur when access to meaning fails. In this model, the phonological output is not the input representation to lexicon, but the lexical form representation itself: the phonological percept of a word. However, the model requires no secondary or additional mechanism for the perception of nonwords — the same representational space is used as the basis of the percept of a nonword.

This distribution of information allows the dichotomy between the results of the priming and the phoneme monitoring experiments to be accommodated. The priming studies are assumed to reflect the activation of semantic information — what is traditionally thought of as the lexical entry.[45] These studies show that although the matching process is tolerant of phonologically lawful variation, there is still a strong intolerance of unviable deviation. The phoneme monitoring studies examine the underlying phonological percept of speech, which is formed whatever the lexical status of the token encountered. Access to phonological information is simpler because of the obvious part-whole relationship between the access code and the underlying form: there is no useful partial meaning that can be extracted from the [swi] of *sweet*, but the partial phonological form is much more useful and is formed under all circumstances. Again, however, strong effects of phonological viability were found in these experiments, even in the perception of nonwords.

The ability to apply phonological constraints to nonwords in speech perception, may seem quite a surprising characteristic of the perceptual system. Yet much the same capacity has been taken for granted in the production of speech. The classic example is the study by Berko (1958), which found that children as young as 4 were able to apply morphological and phonological rules to novel verbs, for example, producing the past tense of [wʌg] (wug) as [wʌgd]. The phonological inference found here, is likely to be of similar value to the listener: most of the words we learn are not produced

---

[45]The visual target was in fact the orthography of the word. However, I assume that this information is part of the general lexical entry in terms of speech perception, since it seems unlikely that humans would develop a special sub-lexical mapping from the phonology of a word to its orthography (although the opposite mapping is more plausible).

clearly and in isolation; so the ability to use phonological inference when encountering novel words greatly simplifies the task of discerning their underlying phonological form.

7.2 Linguistic Issues

The application of a connectionist learning strategy to the problems of perception of phonologically variant speech has uncovered a number of issues relevant to the role of perception in phonological theory. Firstly, the way the network deals with surface phonological change is better described in terms of the satisfaction of multiple constraints than the application of discrete phonological rules. The priming experiments found no evidence of such graded behaviour (which, as discussed above, may be due to the different requirements of the phonological and semantic access systems). However, the phoneme monitoring results, which were designed precisely for that purpose, did reveal graded effects. The extent to which subjects relied on surface and underlying representations of the speech they heard depended, in a probabilistic manner, on the lexical and phonological cues available. Changes which violated phonological rules (either in terms of the viability of the phonological context or the lexical status of the underlying word) still exhibited some degree of compensation in subjects responses, just not as much as those that conformed to the rules. This behaviour suggests that a description of speech perception in terms of the application of standard phonological rules is inaccurate in detail, although it still provides a good generalisation of the overall pattern. Phonological theories that do not rely on such dichotomous change, such as optimality theory (Prince & Smolensky, in press) are able to accommodate these findings more directly.

The second implication of this approach relates to the role of the perceptual system in the 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. This contradicts theories (e.g., Kohler, 1990; Ohala, 1990) in which phonological change is seen as perceptually tolerated simplification. A connectionist network basically learns what it is trained to learn: there may be differences in the ease of learning different mappings, 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. So this approach does not permit any insight into the asymmetries found in phonology. In contrast, a perceptually driven theory of phonological change asserts that natural phonological changes occur as a result of articulatory simplification only when the perceptual system allows them.

However, much of the support for this kind of explanation comes from studies of the perceptual tolerance for change in the adult perceptual system. For example, Hura, Lindblom & Diehl (1992) showed that, in the context of a following word-initial stop, fricatives are more distinctive than nasals and unreleased stops. This was used as an explanation for the finding that in production, fricatives are less likely to undergo assimilation than nasals and stops.

But this type of finding is also evident in the behaviour of the network described here. The distinction between fricatives and nasals or stops was not examined here, but Chapter 4 showed that the trained network tolerated much more variation when presented with labial and velar segments than with coronals. This was consistent with a description of the representation of place by the network as underspecified, but was merely a consequence of the composition of the training set, which in turn was designed to represent the statistical properties of natural speech. Thus, it seems that variations in tolerance in the adult system do not necessarily imply that the phonological rules are shaped by perception: the perceptual tolerance may be shaped by the speech production system. This suggests that to show that the perceptual system truly plays a driving role in the shaping of phonological change, it may be more fruitful to study the properties of the perceptual system as it develops, by examining the distinctiveness of these kinds of changes to young children.

7.3 Future Directions

This research has demonstrated the importance of the role of phonological variation in speech perception. It is not sufficient to view the sensitivity of the perceptual system to phonological change as something that can be simply tacked onto our model of speech perception. I have shown here that

the way we deal with variation has implications for both the structure of lexical representations as well as the process by which these representations are retrieved.

However, I cannot claim that the model I have argued for in this particular instance is applicable to all variation humans encounter, or even to all phonological variation. Place assimilation, whilst an important and challenging example of phonological change, represents only a small fraction of the problems that natural phonological changes cause. For example, compensation for word-internal change may depend more on the structure of lexical representation than the inference process I have described here. The processes of elision and deletion also cause particular problems for speech perception since they violate the segment-to-segment mapping between surface and underlying forms of speech. These changes are problematic for the connectionist model I have proposed since its architecture relies on just such a mapping. A solution to this problem may be to find a network architecture or algorithm that is still more flexible in its representation of time.

The phonological changes examined here can be thought of as near the middle of a spectrum of regular changes in speech perception. Allophonic changes, at one end, cause the fewest problems of ambiguity and can be dealt with at a fairly low level. At the other end, morphological changes such as inflection can cause quite radical variation in the phonological forms of words. Compensation for changes such as these is likely to be a lexical process. It seems fitting then, that the changes studied here should require an interaction of lexical and non-lexical processes in order to derive the underlying form of the speech.

The connectionist approach I have taken here has, on the whole, proved to be a useful and realistic basis for theoretical modelling. The model has been able to make effective use of partial cues, both in the modelling of phonological inference and speech segmentation. The approach is flexible enough to allow swift changes in the hypotheses the network makes, whilst still maintaining an overall intolerance to deviation. But here, again, there remains much work to be done. Most prominent is the lack of an explicit implementation of the final model I have proposed. Because of computational limitations I have had to simulate the effects of lexical interaction by training using highly frequent sequences of segments. However, these limitations are likely to diminish quickly, as computing speeds increase, making proper implementation of this model possible.

The interaction between lexical and phonological constraints during learning is another area worthy of investigation. The training regime I have used here, in which the underlying form of the speech is available for comparison, is unrealistic. This comparison can only occur if the lexical information about the words is available, but these experiments have shown that lexical access will often be successful only if phonological inference is used to disambiguate the surface changes. I have hypothesised that the general intolerance to deviation of the matching process is a property of the learned system and that learning involves a gradual tightening of constraints on the access to semantic knowledge. However, this hypothesis requires both empirical and computational testing if it is to gain support.[46]

7.4 Summary

The pattern of behaviour found in this research demonstrates the complexity of the word recognition process for speech. This process is generally intolerant to small deviations in the speech signal, yet accommodates natural variation, seemingly without cost. People are also able to make phonological decisions based on a surface representation of speech, even when lexical access succeeds, but can employ a more abstract underlying representation, even when lexical access fails. At the heart of this behaviour is a process of phonological inference, which utilises contextual and lexical cues to deduce the underlying form of the speech we hear. I have described two connectionist models incorporating a process of phonological inference: one purely pre-lexically, the other in a more interactive manner.

---

[46]A further factor worthy of investigation, but one which has not been tightly controlled in the simulations and experiments reported here, is the relationship between the Cohort properties of a word and phonological inference. For example, it is plausible that the process of phonological inference is dependent on the distance between a phonological change and the uniqueness point of the carrier word. Similar effects are possible in the network models examined here.

The influence of lexical factors on the results suggests that phonological inference employs interaction between lexical and phonological information, as reflected in the second model.

Appendix A — Materials for Experiments 1 and 2

The 48 items included in the pretest for Experiments 1 & 2 are listed below. Items 12, 23, 24, 32, 38 and 46 were not included in the main experiments. The format of the stimulus material list is: Word-final segment: (UNCHANGED / CHANGED), preceding context, PRIME, (viable following context / unviable following context), control PRIME and following context.

1. D/B I don't see how we can miss it. The field has a BROAD (path across it / gate at one end), WOOD right next to it.

2. D/B The conditions were changing. After a few minutes the CLOUD (melted away / grew larger), FLOOD started to subside.

3. D/B I wouldn't bet my life on Chelsea winning, they CONCEDE (penalties all the time / goals all the time), DIVIDE their midfield too much.

4. D/B The attraction of the game was the price. After all, the CROWD (paid just two pounds to enter / got in free of charge), FOOD was included in the price.

5. D/B Finally the starter arrived. It turned out to be a HORRID (mixture of celery and tomato / cauliflower salad), SALAD with French dressing.

6. D/B There's little point trying to cover up. I think I will PLEAD (manslaughter / guilty), GUIDE the press towards the story.

7. D/B I would say you got what you deserved. That was a WICKED (prank / game), VALID sentence.

8. D/G I was soon almost overcome with nostalgia. The BALLAD (kept bringing back the memories / brought tears to my eyes), CHILDHOOD visions all came back.

9. D/G Don't take any chances on the summit. I think you BLEED (copiously at high altitude / more at high altitude), SLIDE easily up there.

10. D/G It all looked suspicious. There was a note hidden under the BREAD (counter / board), SHED door.

11. D/G I feel like I'm going to burst. I've got to CONFIDE (Karen's secret to someone / my secret to someone), AVOID seeing Clare until tonight.

12. D/G As a speaker, Richard has his faults. I think PRIDE (can be a bad thing / makes you sound pompous), CRUDE jokes should be avoided.

13. D/G You won't get very far with that. These days a QUID (can hardly cover a cup of tea / buys you next to nothing), ROD needs live bait to catch the big fish.

14. D/G What about your daily travel. Do you RESIDE (close to the centre / more than a mile from the centre), REGARD the bus system as adequate.

15. N/M The kitchen looks more like a bomb site. We got the BASIN (plastered with rubbish / coated with rubbish), BROKEN chairs out of the room.

16. N/M Intent on making a good impression, Terry took a CLEAN (pullover from the drawer / gown from the wardrobe), FAWN tie out of the wardrobe.

17. N/M They had done this routine so many times. First the CLOWN (pretended to faint / called for a volunteer), HORN would hoot twice.

18. N/M As they watched from their hideout in the marsh, the CRANE (picked up a large twig / gripped a twig with its claws), STONE fell of the cliff and into the nest.

19. N/M For months we felt lost without Felix. He was a DIVINE (person / creature), GOLDEN example to us all.

20. N/M We inspected the postmark. It looked like a FOREIGN (postcard / card), GERMAN letter.

21. N/M As it was the middle of the wet season, the GRAIN (became mouldy in just three weeks / came at the wrong time for the villagers), TAN didn't have time to develop.

22. N/M At four o'clock the lesson began. The teacher took a GREEN (book from the shelf / cookery book from the shelf), PEN out of his drawer.

23. N/M The garage is notorious. It's run by that tall Geordie and his INSANE (brother / grandfather), COMMON law wife.

24. N/M I'm afraid I can't stand the subject. This LATIN (book is ridiculous / course is ridiculous), FASHION and art course is rubbish.

25. N/M We have a houseful of fussy eaters. Sandra will only eat LEAN (bacon / gammon), BROWN loaves.

26. N/M Alison ran around the garden. She was pretending to be a MARTIAN (pilot / girl), MOUNTAIN explorer.

27. N/M It's hard to get any peace in this village. The PARSON (pesters us all the time / comes round all the time), MILKMAN wakes us up at six in the morning.

28. N/M The kitchen was very well looked after. Along one wall stood a PINE (bench / cupboard), FINE sideboard.

29. N/M Sue didn't remain upset for long. The PUN (brought a wry look to her face / crept into her mind as she read the paper), TUNE helped to pick her up.

30. N/M We can't use the stove tonight, the PYLON (broke in the wind last night / cracked in the wind last night), ROTTEN thing's gone wrong again.

31. N/M We won't have enough time to see it. I think the QUEEN (broadcasts at three / catches the plane at three), PLANE takes off at two.

32. N/M The pitch had been carefully prepared, but the RAIN (put the groundsman in a tricky situation / gave the groundsman a trick problem), VAN had left massive tracks in the grass.

33. N/M I don't think you'll make it in time, the TRAIN (bypasses Royston / gets in at four), PLAN is to leave at five.

34. T/K We finally came to a halt at six. The BOAT (grounded in the thick mud / beached in the thick mud), SKIRT of the hovercraft was punctured.

35. T/K What is the current rate? I want to CONVERT (Greek to British currency / British to Greek currency), BUDGET for my holiday abroad.

36. T/K Some areas weren't affected for a while. The DROUGHT (gradually spread to the south / mainly affected the south), CHART showed that the South would be alright.

37. T/K The animals were all very different. The GOAT (grew up so quickly / behaved like a clown), FAT one just sat around all day.

38. T/K The bailiffs couldn't break in. The MALLET (glanced off the oak door / bounced back off the oak door), RACKET they made woke up the dog.

39. T/K I think they were well received. They were asked to give a REPEAT (golf display / performance), CONCERT in New York.

40. T/K Are you sure about the wine? I think TROUT (goes with something lighter / belongs with something lighter), FRUIT salad needs something lighter.

41. T/P I'd know that face anywhere. I've always had an ACUTE (memory for faces / camera-like memory), EXPERT knowledge in this area.

42. T/P I don't know how I remained sane. I used to COMMUTE (by train / quite frequently), FORGET all my appointments.

43. T/P Everyone brought something along. Andrew made a lovely DATE (bake / concoction), STOUT ale.

44. T/P The afternoon was a disaster. Philip thought the KITE (belonged to him / caught the branches of a tree), SKATE must have been left at the rink.

45. T/P The magician called for silence. The PLATE (began to roll across the table / glimmered in the darkness), SWEAT glistened on her forehead.

46. T/P It's not as easy as it looks. You should SHOOT (before the target stops / carefully or you'll miss), LET me show you first.

47. T/P At first everything went smoothly but unfortunately the TART (burned in the oven / collapsed in the oven), VOTE went the wrong way.

48. T/P I think it's early closing day today. You'll have to TROT (back to the shops before four / quickly if you want to get there in time), WAIT until tomorrow.

Appendix B — Materials for Experiment 3

The 48 items included in the pretest for Experiment 3 are listed below. Items 21, 26, 29, 30, 33, 38, 44 and 47 were excluded after analysis of the pre-test scores. Item 6 was excluded from the analysis of Experiment 3 because it contained two tokens of the target segment. The format of the list is: target / phonological change, preceding context, (real word carrier / nonword carrier), viable context / unviable context.

1. D/B First the champagne, then the boat (SLID / BLID) prettily out to sea / gracefully out to sea.
2. D/B A little hankie won't stop the (BLOOD / ZUD) pouring from the cut / gushing from the cut.
3. D/B I remember the (KID / TID) pulling at my sleeve / grabbing my arm.
4. D/B The film shows a (TOAD / GROAD) pouncing on a fly / gulping at a fly.
5. D/B Steve will regret it as soon as the (WORD / KWORD) passes his lips / comes out.
6. D/B Don't let the (BLADE / SKADE) pierce your skin / cut your finger.
7. D/B The starter kit (MADE / TADE) plenty of beer / gallons of beer.
8. D/B With the extra money we (COULD / FOULD) paint the kitchen / get a new TV.
9. D/G Sam thought about setting up to (BREED / PREED) cats / parrots.
10. D/G After plenty of watering, the (POD / THOD) came apart / burst open.
11. D/G Sally honestly saw the (BOARD / SLOARD) consigning her to the scrap heap / promoting her to manager.
12. D/G Graham saw them (PLOD / STOD) carefully back to the house / back to the house.
13. D/G Jane has never felt safe since the (SHED / TWED) caught fire / burnt last year.
14. D/G When no-one was looking, Tony (FED / VED) carrots to the goat / biscuits to the goat.
15. D/G Nancy shares out sweets when they are (GOOD / LOOD) quiet boys / boys.
16. D/G The earlier version was meant to (FADE / CHADE) calmly away / peacefully away.
17. N/M Tim could (TUNE / STUNE) bass guitars fairly well / guitars fairly well.
18. N/M Due to the storm, they had (FLOWN / SPOAN) below cloud level all the way / gradually up through the cloud.
19. N/M After all the hot weather, the (CORN / RORN) became over-ripe / grew quickly.
20. N/M The city got two awards for its (CLEAN / THREAN) parks / guest houses.
21. N/M The dish is best made with (LEAN / GEAN) pork / goose.
22. N/M Rachael thought there were about (TEN / SKEN) packets left / kilos left.
23. N/M Zoe stood up quickly so the (SWAN / FLON) bolted off / glided off.
24. N/M The spotlight (SHONE / GRON) brightly across his face / grimly across his face.
25. N/M The moss from the roof (SOON / FOON) plugged the gutter / clogged up the gutter.
26. N/M The forward was reputed to be a real (BONE / YOAN) breaker / crusher.
27. N/M This year it should be a (FUN / VUN) party / camp.
28. N/M The receipt was at the bottom of his (PEN / CHEN) pocket / cup.
29. N/M Eddie was a (KEEN / BLEEN) basketball player / collector.

30. N/M By six thirty Mary had (GROWN / TROAN) preoccupied with her book / crotchety.
31. N/M We ought to start (WHEN / DREN) Paul arrives / Kate arrives.
32. N/M At the art class, Julie decided to draw a (MOON / POON) buggy / crater.
33. T/P Are you sure we have a (NUT / DRUT) bake for Christmas / cracker for Christmas.
34. T/P Simon made sure the door was (SHUT / SHRUT) behind him / cleanly.
35. T/P In the morning we will (SHOOT / VOOT) bigger things / clay pigeons.
36. T/P Leave some space, we've (GOT / ZOT) boxes coming as well / cake coming up.
37. T/P The day's walk made her (FEET / YEET) bloody and grimy / grimy and bloody.
38. T/P When she was angry, she would (POUT / JOUT) brusquely / coolly.
39. T/P Luckily, the ship was only a (FREIGHT / PRAYT) bearer / carrier.
40. T/P Cool the oven down or you'll have the (PLATE / CLAYT) baking / cracking.
41. T/K Anne was such a (FAT / SWAT) girl / baby.
42. T/K If there's no room you could always (PUT / SKOOT) glasses in the fridge / beer in the fridge.
43. T/K In the shop, Richard (BOUGHT / FLORT) gravy powder / pencils for design lessons.
44. T/K The news said the union will (BOOT / CHEWTE) grocers and butchers / police officers.
45. T/K The new manager seems a (BIT / THIT) grumpy / babyish.
46. T/K The hill-side (HUT / BLUT) gave them refuge from the wind / provided refuge from the wind.
47. T/K While we were there, we saw the (BRUTE / SMUTE) grip her hand / punch the chef.
48. T/K While you're there you should (JOT / THOT) guesses down / problems down.

Appendix C — Materials for Experiment 4

The 48 items included in the pretest for Experiment 4 are listed below. Items 1, 17, 18, 21, 25, 29, 30, 33, 36, 38, 44 and 47 were excluded from the main experiment. The format of the list is: target, word-final segment (unchanged / viable / unviable), Preceding context, carrier (real word / nonword), following context. The target is the initial segment of the following context.

1.	P	D/B/G	First the champagne, then the boat (SLID / BLID) prettily out to sea.
2.	P	D/B/G	A little hankie won't stop the (BLOOD / ZUD) pouring from the cut.
3.	P	D/B/G	I remember the (KID / TID) pulling at my sleeve.
4.	P	D/B/G	The film shows a (TOAD / GROAD) pouncing on a fly.
5.	P	D/B/G	Steve will regret it as soon as the (WORD / KWORD) passes his lips.
6.	P	D/B/G	Don't let the (BLADE / SKADE) pierce your skin.
7.	P	D/B/G	The starter kit (MADE / TADE) plenty of beer.
8.	P	D/B/G	With the extra money we (COULD / FOULD) paint the kitchen.
9.	K	D/G/B	Sam thought about setting up to (BREED / PREED) cats.
10.	K	D/G/B	After plenty of watering, the (POD / THOD) came apart.
11.	K	D/G/B	Sally honestly saw the (BOARD / SLOARD) consigning her to the scrap heap.
12.	K	D/G/B	Graham saw them (PLOD / STOD) carefully back to the house.
13.	K	D/G/B	Jane has never felt safe since the (SHED / TWED) caught fire.
14.	K	D/G/B	When no-one was looking, Tony (FED / VED) carrots to the goat.
15.	K	D/G/B	Nancy shares out sweets when they are (GOOD / LOOD) quiet boys.
16.	K	D/G/B	The earlier version was meant to (FADE / CHADE) calmly away.
17.	B	N/M/NG	Tim could (TUNE / STUNE) bass guitars fairly well.
18.	B	N/M/NG	Due to the storm, they had (FLOWN / SPOAN) below cloud level all the way.
19.	B	N/M/NG	After all the hot weather, the (CORN / RORN) became over-ripe.
20.	P	N/M/NG	The city got two awards for its (CLEAN / THREAN) parks.
21.	P	N/M/NG	The dish is best made with (LEAN / GEAN) pork.
22.	P	N/M/NG	Rachael thought there were about (TEN / SKEN) packets left.
23.	B	N/M/NG	Zoe stood up quickly so the (SWAN / FLON) bolted off.
24.	B	N/M/NG	The spotlight (SHONE / GRON) brightly across his face.
25.	K	N/NG/M	The moss from the roof (SOON / FOON) clogged up the gutter.
26.	K	N/NG/M	The forward was reputed to be a real (BONE / YOAN) crusher.
27.	K	N/NG/M	This year it should be a (FUN / VUN) camp.
28.	K	N/NG/M	The receipt was at the bottom of his (PEN / CHEN) cup.
29.	K	N/NG/M	Eddie was a (KEEN / BLEEN) collector.
30.	K	N/NG/M	By six thirty Mary had (GROWN / TROAN) crotchety.
31.	K	N/NG/M	We ought to start (WHEN / DREN) Kate arrives.
32.	B	N/M/NG	At the art class, Julie decided to draw a (MOON / POON) buggy.
33.	B	T/P/K	Are you sure we have a (NUT / DRUT) bake for Christmas.

34.	B	T/P/K	Simon made sure the door was (SHUT / SHRUT) behind him.
35.	B	T/P/K	In the morning we will (SHOOT / VEWT) bigger things.
36.	B	T/P/K	Leave some space, we've (GOT / ZOT) boxes coming as well.
37.	B	T/P/K	The day's walk made her (FEET / YEET) bloody and grimy.
38.	B	T/P/K	When she was angry, she would (POUT / JOUT) brusquely.
39.	B	T/P/K	Luckily, the ship was only a (FREIGHT / PRAYT) bearer.
40.	B	T/P/K	Cool the oven down or you'll have the (PLATE / CLAYT) baking.
41.	G	T/K/P	Anne was such a (FAT / SWAT) girl.
42.	G	T/K/P	You could always (PUT / SKOOT) glasses in the fridge.
43.	G	T/K/P	In the shop, Richard (BOUGHT / FLORT) gravy powder.
44.	G	T/K/P	The news said the union will (BOOT / CHEWTE) grocers and butchers.
45.	G	T/K/P	The new manager seems a (BIT / THIT) grumpy.
46.	G	T/K/P	The hill-side (HUT / BLUT) gave them refuge from the wind.
47.	G	T/K/P	While we were there, we saw the (BRUTE / SMUTE) grip her hand.
48.	G	T/K/P	While you're there you should (JOT / THOT) guesses down.

Appendix D — Simulation Materials

D.1 TRACE Simulations

Stimuli for TRACE simulations 1 & 2, with their close competitor environment. The words are presented using the ASCII phonetic transcription of TRACE. Bracketed stimuli are not included in the TRACE lexicon (i.e. they are nonwords).

CATEGORY A

Bisyllabic, early uniqueness point, nonword mismatch.

Original	Mismatch	Fragment	Competitor Environment			
bat^S	(bat^t)	(bat^)	bagis	bapub	badal	bakar
d^lub	(d^luk)	(d^lu)	d^par	d^gis	d^dal	d^kar
sipar	(sipal)	(sipa)	sig^t	sigis	sid^S	sikiS

CATEGORY B

Bisyllabic, late uniqueness point, nonword mismatch.

Original	Mismatch	Fragment	Competitor Environment			
puriS	(purit)	(puri)	purip	purik	pur^d	puras
laS^b	(laS^k)	(laS^)	laS^d	laS^p	laSit	laSis
t^gar	(t^gal)	(t^ga)	t^gak	t^gap	t^g^d	t^gis

CATEGORY C

Bisyllabic, late uniqueness point, real word mismatch.

Original	Mismatch	Fragment	Competitor Environment			
giruS	girut	(giru)	girut	girul	girak	gir^s
kudab	kudak	(kuda)	kudak	kudat	kud^t	kudas
Sabir	Sabil	(Sabi)	Sabil	Sabit	Sab^k	Sabas

CATEGORY D

Trisyllabic, early uniqueness point, nonword mismatch.

Original	Mismatch	Fragment	Competitor Environment			
r^sakiS	(r^sakit)	(r^saki)	r^g^d	r^pub	r^dal	r^kar
bulip^b	(bulip^k)	(bulip^)	bukal	bupub	budar	bukar
lap^gir	(lap^gil)	(lap^gi)	laduk	lagis	ladub	lakar

D.2 Simulations 1 and 2

Test and context words for Simulations 1 and 2, transcribed using the CSTR machine readable phonemic alphabet for English.

TEST WORDS

D/B/G	sauund, mensh@nd, r@uud, @uuld, frend, staatid
N/M/NG	viktoori@n, iiv@n, dauun, wuhn, sod@n, mein
T/P/K	kwait, ekstent, muhst, aakitekt, fit, rait

CONTEXT WORDS

pei, bai, mei, g@uu, koz

D.3 Simulation 3

TEST WORDS

All Test words were presented with a non-coronal final target segment (i.e., the +coronal bias words were presented assimilated). The context words were as in Section D.2.

	+Coronal Bias	-Coronal Bias
1.	bilt	bangk
2.	deivid	duhr@m
3.	eniwuhn	eksplooring
4.	fain	flaiing
5.	glad	gruup
6.	h@@d	him
7.	iiv@n	jhob
8.	kwait	kuhm
9.	lis@n	luk
10.	mait	maak
11.	nain	neim
12.	peintid	pleigruup
13.	ruup@t	ruum
14.	striit	sist@m
15.	triitid	taim
16.	wikid	wiik
17.	yuunit	yuhng
18.	thoot	v@@b
19.	@bauut	@long
20.	@said	@syuum
21.	bauut	biiing
22.	deit	desk
23.	eitiin	enithing
24.	fiild	fruhm
25.	had	geting
26.	hap@n	hauusing
27.	int@estid	iivning
28.	kan	kingdom
29.	lant@n	laik
30.	meid	meik
31.	rait	riid@ship
32.	sev@n	sm@uuk
33.	th@n	teip
34.	wuhrid	wili@m
35.	@gein	@k@m
36.	@piild	athletik
37.	auut	aask
38.	bet	bleim
39.	chest	chom

40.	eit	duuing
41.	greit	griik
42.	lit	luk
43.	wot	taip
44.	yet	uhp

D.4 Simulations 5 and 6

191 word randomly selected from the training corpus.

leg	peintid	waiz	kooind	s@spektid
menii	truh	ch@n	wiilz	w@@d
kanz	p@z	luh	shuu@r	spiik
suhmwe@r	laik	nev@	buhnd@l	chaps
kauuns@l	frend	thing	distrikt	wood@n
gluuing	auu@z	@r	nii@n	klii@
koot	d@	truuth	jhyuu	fol@wing
s@@t@nlii	puu@	ikspii@rii@ns	we	ron@ld
damn	fig@	mait	dis@@teish@n	nait
meid	swiit	ingglish	f@@dh@r	hi
men	woond	@uunlii	thrii	wuhrii
laif	soos@z	ileksh@nz	probablii	suhmthing
m@uust	athletik	glaasiz	regyul@	aim
brekf@st	auu	kuhd@l	w@@king	freiz@l
grandpe@rents	suhm	re	si	uh
meik	wel	chest	hai	of
iriteitid	@uuld	had@nt	@chiivd	piian@uu
@@r	s@p@uuz	th@@tiin	ho	kr@uuz
k@neksh@n	miinz	wud@nt	@v	sit
dhe@z	fainansh@lii	emptii	l@uuk@l	s@@t@n
difr@nt	@k@m	sta	ei	bl@m
eksaktlii	yuuz	od	dei	sivilaizd
thik	litr@ch@	deit	an	freiz
biznis	jhon	@d	miiting	huum
giv	duhz	huuaa	miin	yooself
aidh@	k@rekt	raadh@	redii	chomli
tik	komp@nii	plezans	duu	bikeim
similaritiz	pul	kanongeit	p@	en
jhanyuuarii	eniwei	yuul	gram@	eniwuhn
fiild	jhent@lm@n	deiz	h@uum	huhzb@nd
hiz	richm@nd	bi	memb@	fiftii
rooial	hap@n	wei	e@rii@	mikeila
gosh	pluhs	kidii	presiidents	aiv
twentii	siin	mensh@nd	wuhn	dhat
kuhm	peip@z	dyuubi@s	shud	tid
fluhdz	hel	sto	amyuuzd	p@uust
duhb@l	p@@fektlii	luhki	and	pozitiv
duhr@m	leib@	ireiz	laivlii	ekspansh@n
@uuldz				

D.5 Simulation 7

TEST WORDS

All test words shown as presented to the network (i.e. in assimilated form for the +coronal bias words). The context words are as in Section D.2.

	+Coronal Bias	**-Coronal Bias**		
	Real Word	Nonword	Real Word	Nonword
1.	ruup@p	chuup@p	taip	maip
2.	chesp	lesp	@k@m	ok@m
3.	wuhrib	guhrib	sist@m	kist@m
4.	kam	tham	pleigruup	freigruup
5.	@saib	isaib	jhob	rob
6.	deivib	seivib	fruhm	gruhm
7.	faing	waing	geting	teting
8.	meig	deig	eksplooring	aksplooring
9.	hap@ng	nap@ng	meik	seik
10.	greik	kleik	griik	briik
11.	maik	shaik	athletik	uuthletik
12.	wikig	bikig	@long	ulong

References

Aaronson, D., & Watts, B. (1987). Extensions to Grier's computational formulae for A' and B" to below chance performance. *Psychological Bulletin*, **102**(3), 439-442.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, **9**, 147-69.

Amit, D. J. (1989). *Modeling Brain Function: The World of Attractor Neural Networks*. New York: Cambridge University Press.

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

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: lexical access and/or production? *Journal of Memory and Language*, **24**, 89-106.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics*, **44**, 395-408.

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

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the Mind*. Oxford: Blackwell.

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory and Cognition*, **8**, 493-511.

Berko, J. (1958). The child's learning of English morphology. *Word*, **14**, 150-177.

Blosfeld, M. E., & Bradley, D. C. (1981). Visual and auditory word recognition: Effects of frequency and syllabicity. Paper presented at the Third Australian Language and Speech Conference, Melbourne.

Bradley, D. C., & Forster, K. I. (1987). A reader's view of listening. *Cognition*, **25**, 103-134.

Briscoe, E. J. (1989). Lexical access in connected speech recognition. In *Proceedings of the 27th Congress, Association for Computational Linguistics* (pp. 84-90). Vancouver.

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: CUP.

Brown, R. (1973). *A First Language*. Cambridge, MA: Harvard University Press.

Bruner, J. (1983). *Child's talk*. Cambridge, MA: Harvard University Press.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (submitted). Bootstrapping word boundaries: a bottom-up corpus based approach to speech segmentation.

Campbell, C., Sherrington, D., & Wong, K. Y. M. (1989). Statistical mechanics and neural networks. In I. Aleksander (Ed.), *Neural Computing Architectures* (pp. 239-57). London: North Oxford Academic.

Caplan, D. (1972). Clause boundaries and recognition latencies for words in sentences. *Perception and Psychophysics*, **12**, 73-76.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, **25**, 975-979.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Academic Press.

Church, K. W. (1987). Phonological parsing and lexical retrieval. *Cognition*, **25**, 53-70.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, **1**, 372-381.

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics*, **13**, 153-156.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and Production of Fluent Speech.* Hillsdale, NJ: Erlbaum.

Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *Journal of the Acoustical Society of America*, **64**, 44-56.

Cole, R. A., & Perfetti, C. A. (1980). Listening for mispronunciations in a children's story: The use of context by children and adults. *Journal of Verbal Learning and Verbal Behavior*, **19**, 297-315.

Coltheart, M. (1980). Deep dyslexia: a review of the syndrome. In M. Coltheart, K. Patterson, & J. C. Marshall (Eds.), *Deep Dyslexia.* London: Routledge.

Coltheart, M., Curtis, B., & Atkins, P. (in press). Models of reading aloud: dual-route and parallel-distributed-processing approaches. *Psychological Review*.

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.

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception and Psychophysics*, **5**(6), 365-373.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, **31**, 218-236.

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

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, **25**, 385-400.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, **19**, 141-177.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, **24**, 381-410.

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.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**(1), 113-121.

Daugherty, K. G., MacDonald, M. C., Petersen, A. S., & Seidenberg, M. S. (1993). Why no mere mortal has ever flown out to center field but people often say they do. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* Hillsdale, NJ: Lawrence Erlbaum.

Dell, G. S., & Newman, J. E. (1980). Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behavior*, **19**, 608-623.

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

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, **7**, 195-225.

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

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360-385). Hillsdale, NJ: Erlbaum.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, **27**, 143-165.

Feldmann, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, **6**, 205-54.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A., Bever, T. G., & Garrett, M. (1974). *The Psychology of Language*. New York: McGraw-Hill.

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

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time. *Quarterly Journal of Experimental Psychology*, **33A**, 465-496.

Forster, K. I. (1989). Basic issues in lexical processing. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press.

Foss, D. J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior*, **8**, 457-462.

Foss, D. J. (1970). Some effects of ambiguity upon sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, **9**, 699-706.

Foss, D. J., & Gernsbacher, M. A. (1983). Cracking the dual code: Towards a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, **22**, 609-632.

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.

Foss, D. J., & Lynch, R. H. J. (1969). Decision processes during sentence comprehension: Effects of surface structure on decision time. *Perception and Psychophysics*, **5**, 145-148.

Fowler, C. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics*, **36**(4), 359-368.

Frauenfelder, U. H., & Peeters, G. (1990). Lexical segmentation in TRACE: An exercise in simulation. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing.* Cambridge, MA: MIT Press.

Frauenfelder, U. H., & Peeters, G. (1992). *Simulating the time course of word recognition: An analysis of lexical competition in TRACE*. MIT Press.

Frauenfelder, U. H., Segui, J., & Dijkstra, T. (1990). Lexical effects in phonemic processing: Facilitory or inhibitory. *Journal of Experimental Psychology: Human Perception and Performance*, **16**(1), 77-91.

Frauenfelder, U. H., & Tyler, L. K. (1987). The process of spoken word recognition: an introduction. *Cognition*, **25**, 1-20.

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

Garnham, A. (1985). *Psycholinguistics: Central Topics*. New York: Methuen.

Gasser, M., & Lee, C. (1989). Networks that learn phonology (Technical report No. TR300). Indiana University.

Gentner, D. (1982). Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In S. A. Kuczaj (Eds.), *Language, Thought and Culture* Hillsdale, NJ: Erlbaum.

Giachin, E. P., Rosenberg, A. E., & Lee, C. (1991). Word juncture modeling using phonological rules for HMM-based continuous speech recognition. *Computer Speech and Language*, **5**, 155-168.

Goldsmith, J. (1976). An overview of autosegmental phonology. *Linguistic Analysis*, **2**, 23-68.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: computing formulas. *Psychological Bulletin*, **75**(6), 424-429.

Grosjean, F. (1980). Spoken word recognition and the gating paradigm. *Perception and Psychophysics*, **28**, 267-283.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, **38**(4), 299-310.

Grosjean, F., & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition*, **25**, 135-156.

Hakes, D. T. (1971). Decision processes during sentence comprehension: Effects of surface structure reconsidered. *Perception and Psychophysics*, **8**, 229-232.

Hare, M., & Elman, J. L. (1992). A connectionist account of inflectional morphology: evidence from language change. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society.* Princeton, NJ: Erlbaum.

Hare, M., & Elman, J. L. (1993). From weared to wore: A connectionist account of language change. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Harrington, J., Johnson, I., & Cooper, M. (1987). The application of phoneme sequence constraints to word boundary identification in automatic, continuous speech recognition. In J. Laver & M. Jack (Eds.), European Conference on Speech Technology, 1 (pp. 163-166).

Harrington, J., & Johnstone, A. (1987). The effects of word boundary ambiguity in continuous speech recognition. In *Proceedings of XIth International Congress of Phonetic Sciences*. Tallin, Estonia.

Harrington, J., Watson, G., & Cooper, M. (1989). Word boundary detection in broad class and phoneme strings. *Computer Speech and Language*, **3**, 367-382.

Harris, C. (1991). Parallel Distributed Processing Models and Metaphors for Language and Development. PhD Thesis, University of California, San Diego.

Hawkins, P. (1984). *Introducing Phonology*. London: Hutchinson.

Hayes, B. (1992). Comments on F Nolan, 'The descriptive role of segments: evidence from assimilation'. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody.* Cambridge: CUP.

Hebb, D. O. (1949). *The Organisation of Behaviour*. New York: John Wiley & Sons.

Hillis, A., & Carramazza, A. (1990). Category-specific naming and comprehension impairment: a double dissociation. (Report No. 6). Cognitive Neuropsychology Laboratory, John Hopkins University.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations.* Cambridge, MA: MIT Press/Bradford Books.

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

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

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science USA*, **79**, 2554-8.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science USA*, **81**, 3088-92.

Howes, D., & Solomon, R. I. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, **41**, 401-410.

Hura, S. L., Lindblom, B., & Diehl, R. (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech*, **35**(1, 2), 59-72.

Huttenlocher, D. P., & Zue, V. W. (1984). A model of lexical access based on partial phonetic information. In Proceedings ICASSP, (pp. 26.4.1-26.4.4).

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

Jarvella, R. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, **10**, 409-416.

Johansson, S., & Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar*. Oxford: 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: Lawrence Erlbaum.

Kallman, H. J., & Massaro, D. W. (1979). Similarity effects in backward recognition masking. *Journal of Experimental Psychology: Human Perception and Performance*, **5**(1), 110-128.

Kaye, J. (1989). *Phonology: A Cognitive View*. Hillsdale NJ: Erlbaum.

Kaye, J. (1993). *The phonology morphology interface*. Paper presented at Birkbeck College, London.

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**.

Kiparsky, P. (1982). Lexical morphology and phonology. In I.-S. Yang (Ed.), *Linguistics in the Morning Calm.* Seoul: Hanshin.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimisation by simulated annealing. *Science*, **220**(671-80).

Klatt, D. H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), *Perception and Production of Fluent Speech.* Hillsdale, NJ: Erlbaum.

Klatt, D. H. (1986). The problem of variability in speech recognition and in models of speech perception. In J. Perkell & D. Klatt (Eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Erlbaum.

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

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.

Koster, C. J. (1987). Word recognition in foreign and native language. PhD thesis, Rijksuniversiteit te Utrecht.

Kucera, H., & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence RI: Brown University Press.

Ladefoged, P. (1982). *A Course in Phonetics*. New York: Harcourt Brace Jovanovich.

Lahiri, A., & Marslen-Wilson, W. (1992). Lexical processing and phonological representation. In G. J. Docherty & D. R. Ladd (Eds.), *Laboratory Phonology II: Gesture Segment Prosody* (pp. 229-254). NY: CUP.

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

Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in SOAR: the anatomy of a general learning mechanism. *Machine Learning*, **1**(1).

Lamel, L., & Zue, V. W. (1984). Properties of consonant sequences within words and across word boundaries. In Proceedings ICASSP, (pp. 42.3.1-42.3.4).

Liberman, A. M., Cooper, F. S., Shankweiler, D. S., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

Liberman, A. M., & Mattingly, I. G. (1986). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

Luce, R. D. (1959). *Individual Choice Behaviour*. New York: Wiley.

Mack, M., & Blumstein, S. E. (1983). Further evidence of acoustic invariance in speech production: The stop-glide contrast. *Journal of the Acoustical Society of America*, **73**, 1739-1750.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, **69**, 548-558.

Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., Woest, A., & Pinker, S. (1993). German inflection: the exception that proves the rule. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum.

Marslen-Wilson, W. D. (1973). Speech shadowing and speech perception. PhD thesis, Department of Psychology, MIT.

Marslen-Wilson, W. (1984). Function and process in spoken word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X: Control of Language Processes*. Hillsdale, NJ: Erlbaum.

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

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

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

Marslen-Wilson, W., Brown, C. M., & Tyler, L. K. (1988). Lexical representation in spoken language recognition. *Language and Cognitive Processes*, **3**(1), 1-16.

Marslen-Wilson, W. D., & Gaskell, G. (1992). Match and mismatch in lexical access. Paper presented at the XXV International Congress of Psychology, Brussels.

Marslen-Wilson, W., Moss, H. E., & Halen, S. van. (submitted). Perceptual distance and competition in lexical access.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8**, 1-71.

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., & Warren, P. (submitted). Levels of representation and process in lexical access.

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

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

Martin, J. G., & Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, **8**(3), 473-488.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological enquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language,* **27**, 213-234.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, **21**, 398-421.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. In *Proceedings of the Third Annual Conference of the Cognitive Science Society* (pp. 170-2). Hillsdale, NJ: Erlbaum.

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. Part 1: An account of basic findings. *Psychological Review*, **88**, 375-407.

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.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas imminent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115-33.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: spotting words in other words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **20**(3), 621-638..

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, **20**(3), 298-305.

Minsky, M. A., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

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

Morton, J., & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, **15**, 43-51.

Nelson, K. (1974). Concept, word and sentence: interrelations in development. *Psychological Review*, **81**, 267-85.

Nelson, K., Hampson, J., & Shaw, L. K. (1993). Nouns in early lexicons: evidence, explanations and implications. *Journal of Child Language*, **20**, 61-84.

Newell, A., & Simon, H. A. (1963). GPS, a program that simulates human thought. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and Thought.* New York: McGraw-Hill.

Newman, J. E., & Dell, G. S. (1978). The phonological nature of phoneme monitoring: A critique of some ambiguity studies. *Journal of Verbal Learning and Verbal Behavior*, **17**, 359-374.

Ninio, A., & Bruner, J. (1978). The achievements and antecedents of labelling. *Journal of Child Language*, **5**, 1-16.

Nix, A., Gaskell, G., & Marslen-Wilson, W. D. (1993). Phonological variation and mismatch in lexical access. In *Proceedings of Eurospeech 1993*.

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

Norris, D. (1982). Autonomous processes in comprehension: A reply to Marslen-Wilson and Tyler. *Cognition*, **11**, 714-719.

Norris, D. (1986). Word recognition: context effects without priming. *Cognition*, **22**, 93-136.

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

Norris, D. (1991). Rewiring lexical networks on the fly. In *Proceedings of Eurospeech, 1991*.

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: Erlbaum.

Norris, D. (1993). Bottom up connectionist models of 'interaction'. In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Language Processes: Second Sperlonga Meeting*. Hove: Erlbaum.

Norris, D. G. (submitted). SHORTLIST: a hybrid connectionist model of continuous speech recognition.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

Ohala, J. J. (1984). Prosodic phonology and phonetics. *Phonology Yearbook*, **1**, 113-127.

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: Cambridge University Press.

Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory and Cognition*, **9**, 225-236.

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

Pearlmutter, B. A. (1990). Dynamic recurrent neural networks (Technical report No. CMU-CS-90-196). Carnegie Mellon University.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, **28**, 73-193.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, **25**, 21-52.

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

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

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, **38**, 43-102.

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

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, **8**(1), 1-56.

Prince, A., & Smolensky, P. (in press). *Optimality Theory*.

Pulman, S. G., & Hepple, M. R. (1993). A feature based formalism for two-level phonology: a description and an implementation. *Computer Speech and Language*, **7**, 333-358.

Reilly, R. (1993). Boundary effects in the linguistic representations of simple recurrent networks. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, (pp. 854-859). Hillsdale, NJ: Erlbaum.

Repp, B. H. (1978). Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. *Perception and Psychophysics*, **24**, 471-485.

Repp, B. H. (1983). Bidirectional contrast effects in the perception of VC-CV sequences. *Perception and Psychophysics*, **33**(2), 147-155.

Rosenblatt, F. (1962). *The Principles of Neurodynamics*. New York: Spartan.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, **9**, 487-494.

Rubin, P., Turvey, M. T., & Gelder, P. V. (1976). Initial phonemes are detected faster in spoken words than non-words. *Perception and Psychophysics*, **19**, 394-398.

Rubin, P. E. (1975). Semantic influences on phonetic identification and lexical decision. PhD thesis, University of Conneticut.

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

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations*. Cambridge, MA: MIT Press/Bradford Books.

Salasoo, A., & Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word recognition. *Journal of Memory and Language*, **24**, 210-231.

Samuel, A. G. (1981a). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, **110**, 474-494.

Samuel, A. G. (1981b). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 1124-1131.

Sartori, G., Miozzo, M., & Job, R. (1993). Category-specific naming impairments? Yes. *Quarterly Journal of Experimental Psychology*, **46A**(3), 489-504.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America*, **35**, 200-206.

Segui, J., & Frauenfelder, U. (1986). The effect of lexical constraints upon speech perception. In F. Klix & H. Hagendorf (Eds.), *Human Memory and Cognitive Capabilities: Mechanisms and Performances.* Amsterdam: North-Holland.

Segui, J., Frauenfelder, U., & Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, **72**, 471-477.

Seidenberg, M. S. (1993). Connectionism without tears. In S. Davis (Eds.), *Connectionism: Theory and Practice* New York: Oxford University Press.

Seidenberg, M. S., & Bruck, M. (1990). Consistency effects in the generation of past tense morphology. Paper presented at the Psychonomics Society, New Orleans.

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

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, **1**, 145-68.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Symposium on the Mechanisation of Thought Processes* London: HMSO.

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

Seybold, J. L. C. (1992). An Attractor Neural Network Model of Spoken Word Recognition. PhD thesis, University of Oxford.

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: 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, Indiana: Erlbaum.

Shillcock, R. C. (1990). Lexical hypotheses in continuous speech. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing* Cambridge, MA: MIT Press.

Slowiaczek, L. M., & Hamburger, M. (1992). Prelexical facilitation and lexical interference in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18**(6), 1239-1250.

Slowiaczek, L. M., Nusbaum, H. C., & Pisoni, D. B. (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **13**(1), 64-75.

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

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Elmas & J. L. Miller (Eds.), *Perspectives on the study of speech.* Hillsdale, NJ: Erlbaum.

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

Suomi, K. (1993). An outline of a developmental model of adult phonological organization and behaviour. *Journal of Phonetics*, **21**, 29-60.

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

Swinney, D., Onifer, W., Prather, P., & Hirshkowitz, M. (1978). Semantic facilitation across sensory modalities in the processing of individual words and sentences. *Memory and Cognition*, **7**, 165-195.

Tabossi, P. (1993). Connections, competitions and cohorts: comments on the chapters by Marslen-Wilson; Norris; and Bard and Shillcock. In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing: Second Sperlonga Meeting* Hove: Erlbaum.

Tanenhaus, M. K., & Lucas, M. M. (1987). Context effects in lexical processing. *Cognition*, **25**, 213-234.

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception and Psychophysics*, **34**, 409-420.

Tyler, L. K., & Wessels, J. (1985). Is gating an on-line task: Evidence from naming latency data. *Perception and Psychophysics*, **38**, 217-222.

Waibel, A. (1989). Modular construction of time-delay neural networks for speech recognition. *Neural Computation*, **1**, 39-46.

Waibel, A., & Hampshire, J. (1989, August). Building blocks for speech. *BYTE*, p. 235-42.

Waibel, A., Sawai, H., & Shikano, K. (1989). Modularity and scaling in large phonemic neural networks. *IEEE Transactions on Acoustics, Speech and Signual Processing*, **37**, 1888-99.

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

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

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

West, R. F., & Stanovich, K. E. (1982). Source of inhibition in experiments on the effect of sentence context on word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **8**, 385-399.

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

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

Whalen, D. H. (1991). Subcategorical phonetic mismatches and lexical access. *Perception and Psychophysics*, **50**, 351-360.

Whalen, D. N. (1983). The influence of subcategorical mismatches on lexical access. In *Haskin Laboratories Status Report on Speech Research* (pp. 1-15).

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In 1960 IRE WESCON Convention Record, (pp. 96-104). New York: IRE.

Williams, J. N. (1988). Constraints upon semantic activation during sentence comprehension. *Language and Cognitive Processes*, **3**(3), 165-206.

Wright, B., & Garrett, M. (1984). Lexical decision in sentences. *Memory and Cognition*, **12**, 31-45.

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