18 Cognitive Processes in Speech Perception

JAMES M. McQUEEN AND ANNE CUTLER

1 Introduction

The recognition of spoken language involves the extraction of acoustic-phonetic information from the speech signal, and the mapping of this information onto cognitive representations. To develop accurate psycholinguistic models of this process, we need to know what information is extracted from the signal, and when and how it is integrated with stored knowledge.

An essential assumption in all models of spoken language understanding is the mental lexicon. The lexicon is certainly not a list of orthographic word forms. Instead, it is usually characterised as a dictionary in which a variety of information is contained in each entry: phonological and orthographic form information; morphological structure; syntactic, semantic and pragmatic information. A lexical entry may in fact not represent an individual word: groups of word forms (such as those corresponding to the different inflections of a verb) may share a lexical entry, and multi-word phrases such as idioms may have their own entries. Where we refer below to the accessing of words in the mental lexicon, we do so as a shorthand; we mean the accessing of lexical entries.

Whatever the exact structure and organisation of lexical entries, however, lexical access has to be central to any cognitive model of speech recognition. There is an infinite number of possible sentences that a listener might hear, but a finite number of words. Syntactic, semantic and pragmatic processes can therefore only operate for interpretation of utterances via the intermediary process of word recognition. An account of the cognitive processes involved in speech perception should thus be set in the context of lexical access and word recognition. In this chapter, we therefore ask what sources of information are involved in the pre-lexical processing of the speech signal, that is, in the processing which takes place prior to and for lexical access.

There are three types of information that might be involved in lexical access: acoustic-phonetic (information specifying segmental structure); lexical (information about the words of the input language); and prosodic (information specifying supra-segmental structure). Of these three, only acoustic-phonetic information has a mandatory role to play in pre-lexical processing. Since it is acoustic-phonetic information which primarily distinguishes words one from another, this has to be what provides the principal means of access to the lexicon. However, there has been considerable disagreement about how acoustic-phonetic information is used in the access process, as discussed in section 2. There are no *a priori* reasons, however, to assume that either lexical or prosodic information must be involved in the generation of a lexical access code. Both sources of information may only play a role after words have been accessed in the lexicon. In sections 3 and 4, we ask if these two sources of information do influence the lexical access process, and if so, how?

2 Acoustic-phonetic infonnation

2.1 Units of perception

There are two ways in which acoustic-phonetic information might be used in the process of lexical access. One possibility is that phonetic segments are extracted explicitly, in some pre-lexical level of representation, with a classification of the speech signal into "units of perception" (Healy and Cutting, 1976; McNeill and lindig, 1973). Units which have been postulated include acoustic-phonetic features (Eimas and Corbit, 1973; Marslen-Wilson, 1987; Marslen-Wilson and Warren, 1994; Stevens, 1988), phonemes (Foss and Blank, 1980; McClelland and Elman, 1986), context-sensitive allophones (Wickelgren, 1969), syllables (Cole and Scott, 1974; Mehler, 1981), and articulatory gestures (Liberman and Mattingley, 1985). Alternatively, acoustic-phonetic information could be used implicitly, in a direct mapping of the signal onto the lexicon with no explicit classification into pre-lexical units. Klatt (1980b, 1989) has suggested a template-matching process, where spectral information, as analysed by the peripheral auditory system, is mapped directly onto a lexicon of spectral templates of diphone sequences.

In some experiments, alternative perceptual units have been compared. Foss and Swinney (1973) and Segui, Frauenfelder and Mehler (1981) demonstrated that syllable monitoring was faster than phoneme monitoring. Listeners were quicker to detect a pre-specified target in a list of words and nonwords when that target was a syllable than when it was a phoneme. These findings suggested that the syllable functions as a basic perceptual unit. Norris and Cutler '1988), however, claimed that these results were artifactual. The subjects could detect syllables using only a partial analysis of the signal. In their experiment, Norris and Cutler (1988) included foil items, which contained phonetic near-matches of syllabic and phonemic targets. When subjects were thus required to analyse each stimulus fully, phoneme monitoring was faster than syllable monitoring. This finding, however, does not confirm that the phoneme is the unit of perception, only that phonemic information tends to become available, for a phonetic decision, more rapidly than syllabic information.

Other studies have provided evidence in favour of several different units. Selective adaptation effects have been taken as evidence of acoustic feature detectors (Eimas and Corbit, 1973, but these effects now appear to be due to general auditory adaptation: Kuhl and Miller, 1978; Sawusch and Jusczyk, 1981). Results from duplex perception experiments have been taken as evidence for the extraction of articulatory gestures during speech perception (Liberman and Mattingley, 1985). Data showing that listeners integrate coarticulatory information for segments over time can be taken as support for a phonemic unit of perception (Fowler, 1984). Subjects are faster to detect syllable targets when the targets match the syllabification of the target-bearing word than when they mismatch (Mehler, Dommergues, Frauenfelder and Segui, 1981), as would be predicted if the syllable were the unit of perception. Dupoux and Mehler (1990) found that phoneme monitoring to word-initial targets was faster in high- than in low-frequency monosyllables, but that there was no frequency effect in disyllables. These results suggest that pre-lexical processing involves extraction of syllabic information (the structural complexity of disyllables could delay lexical access, relative to that for monosyllables, and hence reduce the contribution of the lexicon, as characterized by the word frequency effect). But these results can also be explained by a pre-lexical process of rate normalization that does not involve syllabic information *per se*. Finally, features have been claimed to be the basic perceptual units on the grounds that lexical entries contain underspecified phonological representations rather than phoneme strings (Lahiri and Marslen-Wilson, 1991). If segments are not represented lexically there is no need to extract them pre-lexically; instead phonological features extracted from the signal could be mapped directly onto the lexicon.

No definitive answer to the unit of perception question is available. Instead, the results suggest that there is in fact no one basic unit. Many different "units" can be used by listeners, depending on the demands of the listening situation (Pisoni and Luce, 1987). It is therefore perhaps impossible to establish whether any particular units are always constructed during normal comprehension. But even if there is no basic unit, this does not mean that lexical access must be a direct mapping of the unanalysed speech stream onto the lexicon. The assumption that acoustic-phonetic information is transformed into more abstract pre-lexical representations (whatever their exact form) is supported by the need for speech normalization.

2.2 Normalization

The acoustic cues to segments are far from invariant. They vary greatly depending on a large number of factors, including: coarticulation (the realization of segments depends upon both preceding and following phonological context/

Fowler, 1984; see also Farnetani, COARTICULATION AND CONNECTED SPEECH PROCESSES); speech rate (e.g. temporal cues such as Voice Onset Time [VOT] change depending on speed of articulation, requiring rate-dependent processing; Miller, 1981; Gordon, 1988); and variation between speakers due to differences in sex, age and dialect. Some authors have argued that this variation is dealt with by the extraction of acoustic cues which are invariant (Stevens and Blumstein, 1981); others that the variation is lawful, and can be exploited by the listener (Elman and McClelland, 1986). In either case, however, it is clear that the perceptual system must be able to deal with this variability. The same physical signal must be interpretable as different segments, and different signals must be interpretable as the same segment (Repp and Liberman, 1987). It seems clear that normalization should take place pre-lexically, prior to lexical access.

In a study of speech rate effects, for example, Miller, Green and Schermer (1984) demonstrated that subjects labelled more ambiguous consonants, midway between /b/ and /p/ and embedded in the continuum *bath-path,* in a contextually congruent manner (i.e. more *bath* responses in a bathing context), but only when subjects were explicitly told to attend to the sentence context. These effects were absent in a speeded response condition, which focused the subjects' attention on the target words. Speaking rate was also varied, resulting in shifts in the category boundary between /b/ and /p/, but the task demand manipulation did not influence this rate-dependent boundary placement. Miller and Dexter (1988) also used the phonetic categorization task to examine effects of lexical status and speaking rate. They found that under speeded response conditions, there was no tendency to label ambiguous initial consonants in a lexically-consistent manner (e.g. as /b/ in a *beef-peef* continuum and as /p/ in a *beece-peace* continuum). Listeners could not ignore the rate manipulation, however: even under speeded response instructions they based their decision on the early portion of the syllable, treating it as if it was physically short (the /b/ - /p/ boundary shifted to a smaller VOT for fast responses). These studies neatly demonstrate that rate normalization is a mandatory feature of pre-lexical processing. The analysis of acoustic information specifying speech rate appears to be essential for accurate lexical access (Miller, 1987).

Other research on normalization has explored effects of between-speaker variability. Mullennix, Pisoni and Martin (1989) showed that listeners could identify words more easily in lists spoken by a single speaker than when the same word-lists were spoken by 15 different speakers, and that this effect was more marked when the speech signal was physically degraded. Normalization across speakers thus appears to operate at a low level, like rate normalization. Mullennix and Pisoni (1990) have further shown that speaker normalization, again like rate normalization, is mandatory. Subjects could not ignore voice variability when categorizing unambiguous initial phonemes in lists of words spoken by one or several speakers, nor could they ignore the variability in the initial consonants when categorizing the words as being spoken by either a male or female speaker. Asymmetries in this interference suggested that extraction of phonetic and speaker information are independent but closely related processes: phonetic decisions appear to be at least partially contingent upon the process of speaker normalization.

In a related series of experiments (Goldinger, Pisoni and Logan, 1991; Martin, Mullennix, Pisoni and Summers, 1989; Palmeri, Goldinger and Pisoni, 1993), speaker variability has been shown to affect recall. This suggests that information about speakers' voices is retained in long-term memory. These authors argued that their results contradict the view that normalization entails a mapping of the input onto abstract linguistic representations with the consequent loss of information about the speaker's voice. Rather, this information appears to be preserved. But these data are not inconsistent with a process of linguistic abstraction: they suggest only that speaker information is not discarded during recognition. These results are nevertheless consistent with the suggestion that the mapping of the signal onto the lexicon should be viewed as a collection of interacting parallel processes, extracting different pieces of information, rather than as a unified process extracting a transcription composed of abstract units at a single level of representation.

2.3 Summary

Acoustic-phonetic information clearly plays a fundamental role in lexical access. Evidence from speech normalization suggests that there is an intermediate level of prelexical processing, mediating between the raw acoustic input and the mental lexicon. It has not been possible, however, to determine which linguistic unit, if any, is constructed at this level of processing. There is support for several different processing units. But the search for a single "unit of perception" is perhaps futile. Evidence that seems to support a particular unit is very often obtained when the subject is required to make responses based at that level of representation. Nevertheless, it is clear that pre-lexical processing entails the transformation of acoustic-phonetic information into more abstract representations. The form of these representations remains to be determined. One important constraint on pre-lexical representations is that they must have a "shared vocabulary of representation" (Connine and Clifton, 1987) with the lexicon (or at least there has to be a very direct mapping between these representations). The form of lexical representations may thus constrain the form of pre-lexical representations.

3 Lexical information

Is lexical information involved in pre-lexical processing? Discussion of this question is best done in the context of two competing theories. Interactive models hold that lexical information is brought to bear pre-lexically. We will focus on one particular model of this class, TRACE (McClelland and Elman, 1986; McClelland, 1991). This interactive activation model has three levels of processing units, corresponding to features, phonemes and words. Units within a level compete with each other via lateral inhibitory connections. Units at lower levels activate the units at higher levels with which they are consistent via facilitatory connections. Thus, during recognition, activation of a feature node leads to activation of consistent phoneme nodes, which in turn activate word nodes. Importantly, higher level nodes also facilitate lower level units. Activated word units boost the activation of their constituent phonemes: this top-down facilitation instantiates the claim that lexical information influences pre-lexical processing.

We will contrast the TRACE model with an autonomous model which holds that lexical information is not involved in pre-lexical processes. In the Race model (Cutler and Norris, 1979; Cutler, Mehler, Norris and Segui, 1987), phonetic decisions can be based either on pre-lexical processing (the pre-lexical route) or on phonological information, stored in the lexicon (the lexical route). The two routes race with each other when a phonetic decision is being made. Whichever route produces an output first wins the race. But processing is strictly bottom-up only, so the lexicon cannot influence pre-lexical processing. Below, we will describe how the TRACE and Race models, as instances of interactive and autonomous theories, account for lexical involvement in various tasks.

3.1 Lexical effects

3.1.1 *Monitoring* Phoneme monitoring is sensitive to phonetic factors. Foss and Gernsbacher (1983) have shown an effect of vowel length: the longer the vowel, the longer the reaction time (RT) on the preceding target consonant. Another factor is the phonological similarity of the target phonemes to preceding phonemes (Newman and Dell, 1978; Dell and Newman, 1980). Detection of target phonemes in sentences is slower when the word preceding the target-bearing word begins with a phoneme closely related to the target. Several studies, however, have failed to find lexical effects. Foss, Harwood and Blank (1980) found that monitoring was no faster to words than to nonwords, and that the frequency of occurrence of the target-bearing word did not influence RT. Segui, Frauenfelder and Mehler (1981) also failed to find an RT advantage for word responses over nonword responses, and Segui and Frauenfelder (1986) did not find a frequency effect when subjects were required to monitor only for word-initial phonemes ("standard" phoneme monitoring).

These results support the claim that phoneme monitoring is based on pre-lexical processing which is open to the influence of phonetic information but not lexical information. But there are some studies which have demonstrated lexical effects. Segui and Frauenfelder (1986), for instance, did obtain a word-frequency effect when subjects were required to monitor not just for word-initial targets, but for targets which could appear anywhere in the words ("generalized" phoneme monitoring). Rubin, Turvey and van Gelder (1976) also found a word-nonword effect: subjects were faster to detect e.g. / b / in *bat* than in *bal.*

Lexical effects appear to be present only in some experiments. Stemberger, Elman and Haden (1985) took this variability as support for interactive models like TRACE. Lexical influences are taken to result from top-down facilitation from word nodes increasing the level of activation of target phoneme nodes, thus speeding responses to targets in words relative to nonwords. Where there are no lexical effects, it is assumed that responses are being made from the phoneme-node level, without top-down facilitation from the lexical level.

Cutler et al. (1987) describe how these lexical effects support the Race model. When the lexical route is available, there are two competing routes for word responses. Thus phonetic decisions will tend to be faster to words than to nonwords because there will be a proportion of trials in which the lexical route wins the race. Cutler et al. examined word-nonword effects in a series of experiments, using only monosyllabic target-bearing items. Lexical effects were found to come and go. Responses to targets in words were faster than those to targets in nonwords only when task monotony was reduced. It was argued that task monotony determines attentional focus in the Race model: given a monotonous list, subjects attended to the signal, and based their responses on the output of the pre-lexical route; given a less monotonous list, subjects attended to lexical route output, producing lexical effects.

Both models can therefore account for the lexical effect, and its variability, in phoneme monitoring. In another task, rhyme monitoring, where subjects detect words and nonwords which rhyme with a prespecified cue, responses are faster to words than to nonwords, and responses are faster to high- than to low-frequency rhyming words (McQueen, 1993). Again both models can explain these lexical effects.

3.1.2 Phonemic restoration If the medial / s / of *legislatures* is replaced with a cough, listeners report hearing a cough and the complete *legislatures,* with the absent phoneme perceptually restored (Warren, 1970). Low-level factors influence the effect: if the replacing noise is acoustically similar to the removed phoneme, the illusion is more likely to occur (Samuel, 1981a, 1981b, Warren and Obusek, 1971); and there is more restoration for fricatives and stops (which are more noise-like) than for liquids, vowels and nasals (Samuel, 1981a, 1981b).

Samuel (1981a) found that several lexical factors influenced the extent of the illusion: there was more restoration for longer than for shorter words; there was a more reliable illusion in words than in phonologically legal nonwords, and presenting an intact version of the target word before the target word also increased restoration. Samuel (1987) found further that there was more restoration for items with several possible restorations (e.g. **egion: legion* or *region)* than for items with a unique restoration (e.g. **esion: lesion).* He also found that there was more phonemic restoration in words which become unique early, moving left-to-right through the word (e.g. *boysenb*rry)* than in words which became unique late (e.g. *indel*ble).* Samuel explained these results in terms of a partially interactive model, in which top-down expectations are confirmed by the bottom-up signal, lexical information being used to facilitate perceptual decisions made at lower levels.

An autonomous account of the data is however also possible. If the illusion is due to attention being focused on lexical information, then the lexical effects can be explained without recourse to top-down connections. In Race model terms, restoration occurs because subjects are using the lexical route. Just as with the monitoring tasks, the evidence for lexical involvement in phoneme restoration does not allow us to distinguish between the two models.

3.2.3 *Phonetic categorization* In the phonetic categorization task, with a continuum of sounds from / d / to / t / in the contexts *deep-teep* and *deach-teach,* for example, a lexical effect would be shown by an increased proportion of /d/ responses in the ambiguous region of the continuum when the voiced endpoint formed a word *{deep),* and an increased proportion of / t / responses when the unvoiced endpoint formed a word *(teach).* This effect was originally demonstrated by Ganong (1980). In the TRACE model, this effect is again accounted for by top-down connections. In the Race model, the effect again reflects the operation of the race between pre-lexical and lexical routes.

Fox (1984) replicated this effect, and showed that there were no lexical effects for fast categorization responses. Connine and Clifton (1987) found both a lexical shift and an RT advantage for word responses relative to nonword responses in the boundary region. They further showed that the lexical effect was not due to postperceptual bias: it was not equivalent to an effect obtained using monetary reward to bias subjects' responses. Lexical effects have also been reported by Burton, Baum and Blumstein (1989), who found that the categorization of a word-initial continuum depended on the acoustic-phonetic quality of the continuum, and by Miller and Dexter (1988), who showed that lexical involvement in categorization was not mandatory, in contrast to rate-normalization processes (see above).

McQueen (1991a) and Pitt and Samuel (1993) have found lexical effects for Phonemes in word-final position (e.g. for an /f/-/s/ continuum in contexts such as *fish-fiss* and *kish-kiss).* McQueen (1991a) also replicated Burton et al.'s (1989) finding that lexical effects in the categorization task only appear when the materials are of poor acoustic quality. Pitt and Samuel (1993), however, have shown that poor stimulus quality is not a necessary condition for a lexical effect: lexical shifts were obtained with high-quality materials in both word-initial and word-final categorization. The basic lexical effect in this task is consistent with both types of model.

3.2 Test cases

Both models can account for lexical effects in several tasks. Are there any test cases which might allow us to distinguish between the two models? Can we establish whether or not lexical information is used in pre-lexical processing? Several attempts have been made to contrast divergent predictions of the TRACE and Race models.

3.2.1 Phoneme monitoring Frauenfelder, Segui and Dijkstra (1990) have presented evidence from the phoneme monitoring task which challenges the interactive position. TRACE predicts that activation of a lexical candidate will both boost the activation of its constituent phonemes by top-down facilitation and inhibit the activation of nonconstituent phonemes because of phoneme-to-phoneme inhibition. As this study showed, there are strong facilitatory effects on the detection of targets (such as /p/ in *olympiade),* which occur after the word becomes unique, relative to matched nonwords (e.g. *arimpiako).* In TRACE terms, this could be due to top-down facilitation of /p/ from the word node. If this were the case, detection of /t/ in *vocabutaire* should be inhibited relative to detection of /t/ in a matched nonword such as *socabutaire,* because of top-down facilitation of /l/ from the activated *vocabulaire* node followed by inhibition of other phoneme nodes by the /l/ node. No such inhibition was found, contrary to the TRACE account. This result is not problematic for the Race model, however. It predicts that performance on nonwords is insulated from lexical information because all nonword decisions have to be made via the pre-lexical route.

3.2.2 Word-final categorization McQueen (1991a) showed that the lexical effect in categorization of word-final ambiguous fricatives, in contexts such as *fish-fiss* and *kish-kiss,* was larger for faster responses. This finding has recently been replicated by Pitt and Samuel (1993). This reaction time effect is predicted by the Race model, where for word-final fricatives the lexical route can be assumed to be faster, on average, than the pre-lexical route. But the TRACE model assumes that lexical effects should build up gradually over time (McClelland and Elman, 1986), and thus predicts exactly the opposite pattern of results, that the lexical effect should be larger for slower responses.

3.2.3 *Compensation for coarticulation* One result appears to support interactive models. Mann and Repp (1981) showed that stops midway between /t/ and /k/ were more often categorized as /k/ after /s/, but as /t/ after /f/. The perceptual system appears to compensate for fricative-stop coarticulation. Elman and McClelland (1988) replicated this effect for ambiguous word-initial stops following fricative-final words such as *christmas* and *foolish,* and, most importantly, they showed that the effect occurred when the word-final fricatives were replaced with an ambiguous fricative. When the

/s/ in *christmas* was replaced with an ambiguous sound /?/, midway between /s/ and /f/, there were again more /k/ responses to the ambiguous stops. With *fooli?,* there were more /t/ responses.

Elman and McClelland claimed that this effect was strong evidence in favour of interactive models like TRACE. Lexical information appears to be influencing a compensation process that can be assumed to be operating pre-lexically. This seems to be direct evidence against the autonomous assumption that information flow is bottom-up. But autonomous models can account for Elman and McClelland's results. Norris (1993) and Chater, Shillcock, Cairns and Levy (submitted) have shown that connectionist recurrent net models, which are sensitive to temporal context but which have strictly bottom-up information flow, can simulate the compensation for coarticulation effect. During training, these networks learn to use contextual information, and thus become sensitive to sequential dependencies. They recognise /s/ in *christma?,* for example, because in training (on a large corpus of conversational English) /s/ was more likely after schwa than /f/. These models successfully simulate both Elman and McClelland's results, and those of McQueen (1991b). The effect is not, after all, a test case that might allow us to distinguish between interactive and autonomous models. Note that this demonstration also shows that effects which appear to be due to specific lexical entries may in fact be due to knowledge about regularities of the spoken language as a whole. Further research is required to establish which "lexical effects" are indeed due to the involvement of stored information about specific lexical entries.

3.3 *Attentional effects*

The previous section has shown that the little evidence there is which distinguishes between interactive and autonomous models supports the Race account. Lexical information does not appear to be involved in pre-lexical processing. Further evidence in support of this conclusion has come from an analysis of attentional effects. In addition to the task monotony effect in phoneme monitoring, that lexical effects tend to be absent when listeners hear °nly monosyllabic items (Cutler et al., 1987, see above), other attentional effects have been found with this task. In a comparison of standard and generalized phoneme monitoring, Frauenfelder and Segui (1989) found that responses were faster to target-bearing words that were preceded by associatively-related words than to those preceded by unrelated words, but only in generalized monitoring. In standard monitoring, subjects can attend to the initial sounds alone, but in generalized monitoring, where there was no prior cueing of the possible target position, subjects have to rely more heavily on lexical information.

Pitt and Samuel (1990a) explicitly examined attentional effects in phoneme monitoring. They manipulated the proportion of occurrences of targets in a particular position in disyllabic words in a generalized phoneme monitoring task: that is, they varied the probable location of the target phoneme. Subjects were faster and more accurate in detecting, for example, initial-position targets when they appeared in lists in which 75 per cent of the targets were in this position than when they appeared in an unbiased baseline condition; target detection in other (non-predicted) locations was worse than in the baseline condition. These authors argued that the manipulation of expected target location induced a strategy of attending to that location.

Nusbaum, Walley, Carrell and Ressler (1982) and Samuel and Ressler (1986) have examined the role of attention in the phoneme restoration effect. The lexical status of stimuli could cause subjects to attend to word-level representations rather than to individual phonemes. But if subjects could be trained on potential targets and cued as to the identity and location of the critical phoneme then the illusion could perhaps be removed. Samuel and Ressler (1986) found that training and cueing did indeed inhibit the restoration illusion.

Eimas, Marcovitz Hornstein and Payton (1990) asked subjects to identify unambiguous item-initial stops in words and nonwords, which were presented at the end of a neutral phrase. There were no lexical effects. However, when subjects also performed a secondary task which required lexical knowledge (such as lexical decision) on the target items subsequent to the identification decision, reliable lexical effects emerged. The concurrent task, which focused attention on the lexicon, appeared to produce lexically-mediated performance in the phonetic task. Eimas and Nygaard (1992) showed further that lexical effects in phoneme monitoring did not emerge in sentential contexts, even with a secondary lexical task. They occurred only in random word strings, and then only when a secondary task was being performed. Eimas and Nygaard argued that when subjects were given meaningful contexts, the secondary decisions were based on higher-level representations, and the phonetic decisions were based on pre-lexical processing. It is only in neutral contexts, where listeners must use lexical information to perform the secondary task, that attention is directed to the lexical level for phoneme monitoring.

These attentional effects show that, in contrast to rate or speaker information, lexical information plays no mandatory role in pre-lexical processing Attentional manipulations can encourage or discourage listeners from using lexical information during phonetic decision-making. But listeners cannot avoid using rate and speaker information, even when they are asked to ignore this information. As Cutler et al. (1987), Eimas et al. (1990) and Eimas and Nygaard (1992) have argued, these results are more in keeping with the Race model than with TRACE. An attentional mechanism fits more parsimoniously into the Race model framework, requiring only that attention, in response to task demands, can be focused on the output of one or other route. Shifting attention to one or other level of representation naturally entails shifting attention to one or other route. In the TRACE framework, however, attentional modulation of lexical involvement has to be modelled in terms of a gain control increasing or decreasing the amount of top-down facilitation. Adjusting the gain in this way does not naturally follow from a shift of attention from one level of representation to the other.

3.4 Summary

As claimed by the Race model, lexical information is not used to constrain the process of lexical access. Although many results can be handled by both models, a few findings are problematic for the TRACE account. Furthermore, attentional modification of lexical involvement, though not contradicting either model, is more consistent with the Race account. It might be argued that no Race model explanation of compensation for coarticulation has been offered. Lexical involvement in compensation for coarticulation was accounted for by autonomous recurrent networks. But the Race model has recently been considered to be part of a fuller account of spoken-word recognition which includes the recurrent network responsible for compensation effects: the Shortlist model (Norris, 1994).

Shortlist is a two-stage model of spoken-word recognition (for a fuller description, see Norris, 1994). In the first stage, lexical hypotheses, consistent with the segmental information in the input, are activated. This stage can be modelled using the type of recurrent net which can explain the compensation for coarticulation results. This stage encapsulates pre-lexical processing in the model: it is capable of rate normalization (Norris, 1990), and, in accordance with the empirical evidence, the lexical access code is based on acoustic-phonetic information specifying segmental structure. Since phonetic decisions can be based on the output of this stage, it instantiates the pre-lexical route of the Race model. In the second stage, the activated lexical hypotheses (constituting a "shortlist" of candidate words) compete with each other via a process of lateral inhibition, until one word (or a series of words in the case of continuous speech recognition) dominates the activation pattern, and can then be recognised. This stage instantiates the lexical route of the Race model. The Shortlist model thus provides an account both of the compensation for coarticulation results and the lexical effects that can be explained by the Race model. Information flows only bottom-up in Shortlist, and thus the model embodies the claim that lexical information does not influence pre-lexical processing.

Note that we are not claiming that lexical information is not involved in lexical access, just that it is not involved in the processing which takes place Prior to lexical access. The lexical information specifying the structure of the vocabulary and the form-based relationships between words has an important role to play in word recognition. There is growing evidence that multiple lexical hypotheses are activated during the recognition process (Marslen-Wilson, 1987,1990; Shillcock, 1990; Swinney, 1981; Zwitserlood, 1989). Recent evidence suggests that once words consistent with the input speech have been activated, they then compete (Goldinger, Luce and Pisoni, 1989; Goldinger, Luce, Pisoni and Marcario, 1992; McQueen, Norris and Cutler, 1994; Norris, McQueen and Cutler, 1995; Slowiaczek and Hamburger, 1992; Vroomen and de Gelder, 1995). Accessed words compete with each other until one word dominates the others: this one word can then be recognised. This competition process is instantiated, in different ways, in several models of spoken-word recognition: the Neighborhood Activation Model (Luce, 1986), TRACE, and, as described above, Shortlist. Lexical information, such as the extent to which one word overlaps with other words, and the nature and number of words embedded within other words, is thus crucially important in the process of word recognition, but only after lexical access has taken place.

4 Prosodic information

The third type of information whose role in lexical access we will investigate is prosodic information. By prosody we mean variations in fundamental frequency, timing structure and intensity across an utterance, and we confine ourselves here to prosodic variation which is not solely the direct consequence of a segmental decision.

By this restriction we in fact exclude a very large amount of prosodic variation. Speech is realised as sound, and sounds must perforce be uttered with a certain fundamental frequency (f_0), a certain amplitude, and a certain duration. Phonetic segments may vary in intrinsic f_0, and listeners certainly exploit this information (see Silverman, 1987, for a review). Likewise segments may be longer or shorter in one context than in another simply due to properties of that immediate context: effects of adjacent segments, for example, or of the position of occurrence within a word; again, we exclude this source of prosodic variation.

Furthermore, in many languages inter-segmental contrasts are realised solely in duration (vowels have three levels of duration in Estonian, for example)- Analogous to this, we argue, is the case of lexical tone in languages such as Mandarin, Vietnamese or Yoruba, in which two morphemes consisting of the same sequence of segments are distinguished by f_0 variation. In many languages, tone realisation can be in part contextually dependent (which is of course also true of coarticulated segments); but it remains the case that, for example, a CV syllable with a rise-fall tone can be a *different lexical item* from the same CV with a level tone. And although tone is best conceptualised as a property of syllables, it is actually realised to all intents and purposes on syllabic nuclei, i.e. vowels; thus it is conceivable that a recogniser could process a given vowel with a rise-fall tone and the same vowel with a level tone simply as different segments, irrespective of context. There is as yet little experimental evidence available on spoken-word recognition in tone languages, but what evidence there is supports a parallelism between segmental processing and the processing of tonal information.

For instance, lexical information can affect tone categorisation in just the same way as it affects segment categorisation. In the segment categorisation study of Ganong (1980), listeners' category boundaries between /t/ and /d/ shifted to produce more /d/ responses preceding /i:p/ but more /t/ responses preceding /i:tf/; similarly, in a tone categorisation study by Fox and Unkefer (1985), listeners' category boundaries between two tones of Mandarin Chinese shifted as a function of which endpoint tone produced a real word given the syllable the tone was produced on. (This was of course only true when the listeners were Mandarin speakers; English listeners showed no such shift.) Lexical priming studies in Cantonese also suggest that the role of a syllable's tone in word recognition is analogous to the role of the vowel (Cutler and Chen, 1995). Nevertheless, there is evidence (Cutler and Chen, in press) that tonal information is not processed in conjunction with vocalic information; listeners can detect the difference between two CV syllables with the same onset and the same tone but a different vowel more rapidly than they can detect the difference between two CV syllables with the same onset and the same vowel but a different tone.

The above types of prosodic information we will exclude from further consideration; the principal question at issue here concerns prosodic information which is not segmentally constrained, and the use of which is, therefore, in principle optional. We will consider two aspects of prosodic structure: lexical stress and sentence rhythm. These neither exhaust the canon of prosodic variation, nor, more importantly, can they lay claim to universal validity: for example, many languages do not have lexical stress. They are chosen chiefly because they have both been the subject of considerable research effort.

4.1 *Lexical stress*

The term *lexical stress* itself suggests that stress pattern can have a lexically distinctive function. Indeed, pairs of unrelated words differing only in stress pattern do exist, although in the world's lexical stress languages such pairs are extremely rare. In English, for instance, although stress oppositions between verb and noun forms of the same stem *(decrease, conduct)* are common, there are very few such pairs which are lexically clearly distinct (such as *forbear, or insight/incite).*

Due to the greater acoustic reliability of stressed syllables, stress can affect recognition: stressed syllables are more readily identified than unstressed syllables when cut out of their original context (Lieberman, 1963), and distortions of the speech signal are more likely to be detected in stressed than in stressed syllables (Cole and Jakimik, 1980; Browman, 1978; Bond and Games, 1980). Detection of word-initial target phonemes can also be faster on stressed than unstressed syllables, although only when acoustic differences are relatively large, as for instance in spontaneous speech; such differences do not arise with laboratory-read materials (Mehta and Cutler, 1988).

Studies of English vocabulary structure show that stress pattern information could be of use in word recognition. A partial phonetic transcription which includes stress pattern information applies to a smaller candidate set of words than one which does not (Aull, 1984; Waibel, 1988). An automatic recognition algorithm operating at this level of phonetic specification performs significantly better with stress pattern information than without (Port, Reilly and Maki, 1988). But stress information does not facilitate human word recognition: neither visual nor auditory lexical decision is facilitated by prior specification of stress pattern, nor does whether or not a bisyllabic word conforms to the canonical English word class pattern (initial stress for nouns, final stress for verbs) affect how rapidly its grammatical category is judged (Cutler and Clifton, 1984).

Mis-stressing, to be sure, inhibits word recognition. English listeners presented with English spoken by Indian speakers, including stress patterns unorthodox by British English standards, tend to interpret the input in conformity with the stress pattern, often in conflict with the segmental information (Bansal, 1966). Puns are unsuccessful if they require a stress shift (Lagerquist, 1980). Deliberately mis-stressed words are responded to more slowly in recognition tasks than correctly stressed words (Bond and Small, 1983; Cutler and Clifton, 1984). The mis-stressing used in such studies, however, was not simply a prosodic manipulation. Pairs of English words with stress-pattern opposition usually also differ vocalically. Thus *OBject* and *obJECT, CONtent* and *conTENT* have quite different vowels in their first syllables - the stressed syllables have a full vowel, while the unstressed syllables have schwa. Just as the vowel difference in *cot* and *cut* is lexically significant, so may observed effects of stress simply reflect the lexical significance of different vowels resulting from stress differences. To English listeners, the vowel quality distinction is indeed more crucial than the purely prosodic distinction; cross-splicing vowels with different stress patterns produces unacceptable results only if vowel quality is changed (Fear, Cutler and Butterfield, 1995). In Fear et al.'s study, listeners heard tokens of, say, *autumn,* which has primary stress on the initial vowel, and *audition,* which has an unstressed but unreduced vowel, with the initial vowels exchanged; they rated these tokens as insignificantly different from the original, unspliced, tokens.

To investigate *prosodic* effects on word recognition in a lexical-stress language, it is necessary to control for vowel quality. Although most unstressed syllables in English have a neutral (schwa) vowel, a reasonably large class or polysyllabic words with exclusively full vowels does exist. *Nutmeg* and *typhoon* are two such words. In their mispronunciation experiment, Cutler and Clifton (1984) explicitly compared bisyllabic words in which the unstressed syllable contained schwa *(wisdom, deceit)* with words like *nutmeg* and *typhoon-* Word recognition was clearly inhibited by mis-stressing for the former group- The words with full vowels, however, were only harder to recognize when mis-stressed if their citation form pronunciation had initial stress. That is, *nutMEG* was much harder to recognize than *NUTmeg;* but *TYphoon* was no significantly more difficult than *tyPHOON*. (In English, the demands of sentence rhythm can cause stress to shift in words like *typhoon;* they are in practice encountered sufficiently often in initially-stressed form for this form perhaps to have achieved the lexical status of an optional pronunciation.) Similarly Taft (1984), using a monitoring task, found that SW (strong, weak) words produced slower responses when mis-stressed *(cacTUS),* but mis-stressing of WS words *(SUSpense)* actually led to response times which were somewhat faster than those with the correctly stressed words. Slowiaczek (1990) also demonstrated increased recognition difficulty for pronunciations like *nutMEG.*

The process of word recognition includes several subsidiary operations, however, and there are at least two ways in which prosody could be relevant in recognition. These correspond to the commonly drawn distinction between lexical access and lexical retrieval. On the one hand, lexical prosody, i.e. stress marking, could be an essential part of the access code by which lexical entries are located; on the other, it could be part of the phonological code listed for a word in the lexicon and consulted only in retrieval, i.e. once access has been achieved. The mis-stressing results do not distinguish between these two possibilities. If prosodic information is present in the lexical access code, *nutMEG* could be hard to recognize because the initial access attempt will encounter no match, and successful access will only be achieved after the code has been recomputed. If prosody does not play a role in access, however, *nutMEG* could be hard to recognize because the complete phonological form in the accessed lexical entry fails to match the input.

If prosody participates in lexical access, in much the same way that segmental identity does, then minimal stress pairs, i.e. words with identical segments but different prosody, should generate distinct lexical access codes, and be, in practice, not confusable. In fact, the rarity of minimal stress pairs itself suggests that lexical stress may hardly ever exercise such constraint. Experimental evidence supports this, by suggesting that the lexical access code does not draw on the information available from English word prosody. Using the cross-modal priming paradigm (Swinney, 1979), in which listeners hear a sentence and at some point during the sentence perform a visual lexical decision, Cutler (1986) showed that strings like *forbear* are functionally homophonous: *both FORbear and forBEAR* facilitate recognition of words related to *each* of them (e.g. *ancestor, tolerate).* In other words, listeners did not distinguish between these two word forms in initially achieving access to the lexicon.

Where English listeners *can* use stress information, they do use it; so when Connine, Clifton and Cutler (1987) asked listeners to categorize an ambiguous initial consonant in either *DIgress-TIgress* (in which *tigress* is a real word) or *tiGRESS-tiGRESS* (in which *digress* is a real word), they reported /t/ more often for the initial-stress items, /d/ more often for the final-stress items. In other words, they were using the stress information (both that in the signal and that in their stored representations of these words) to resolve ambiguity in a difficult perceptual situation. As we pointed out in section 3.3, however, evidence that a given source of information is used in a phonetic categorization task does not constitute evidence that it is used pre-lexically; this holds as true for prosodic correlates of stress (although they could in principle be processed in a bottom-up way) as for information about lexical identity (which implies top-down flow of information). The listeners in Connine et al.'s study had the opportunity to consult the phonological code listed in two lexical entries, and to use the prosodic information contained therein to motivate their phonetic categorisation decision.

In fact, it may well make good sense for the listener not to make early use of lexical stress for lexical access. In order to know the stress pattern of a word, the listener's word recognition system must know how many syllables the word has; in effect, therefore, it could not begin the process of lexical access until the end or nearly the end of the word if it were to need stress pattern information before access could be attempted. Perhaps, therefore, lexical prosody does not participate in the *pre-lexical* access code simply because the information it provides cannot outweigh the disadvantage of delayed initiation of access.

4.2 Sentence rhythm

The characteristic rhythm of English - in which language most of the relevant experimental evidence is again to be found - is based on stress. Rhythmic patterns are accompanied by segmental variations: by far the majority of unstressed syllables in English contain weak (reduced) vowels. And again, rhythmic effects in the recognition of spoken English largely reduce to effects which can be interpreted in terms of segmental processing. Consider, for instance, the effects observed in speech segmentation. In continuous speech, word boundaries are rarely reliably marked, and listeners in practice adopt explicit procedures which assist with the location of points at which lexical access attempts should most usefully be commenced; in English, and in rhythmically similar languages such as Dutch, the procedure is based on the assumption that strong syllables are most likely to be word-initial. The evidence for this comes partly from studies of word boundary misperceptions, in which listeners most commonly err by assuming strong syllables to be word-initial and weak syllables to be non-initial (Cutler and Butterfield, 1992; Vroomen, van Zon and de Gelder, in press), and word-spotting studies, in which real words embedded in nonsense bisyllables are harder to detect if detecting them requires processing segments from two consecutive strong syllables, i.e. across the canonical point of speech segmentation (Cutler and Norris, 1988; MeQueen et al, 1994; Norris et al., 1995; see also Vroomen and de Gelder, 1995; Vroomen et al., in press).

In all of these studies the effective parameter was vowel quality rather than prosodic stress *per se*. In Cutler and Norris' (1988) original word-spotting study/ for example, prosodic stress was not varied. Thus detection of the word *mint* was compared in *mintayf* and *mintef;* both bisyllables had the same stress pattern (initial stress). The difference was solely in the vowel which occurred in the second syllable, and it was this vowel which affected listeners' responses: *mint* was much harder to detect when the second vowel was strong, as in *mintayf*.

Although we interpret these rhythmic effects as reflecting exploitation of vowel quality, i.e. segmental information, it is also relevant that the strong/ weak vowel difference in English is the manifestation of language rhythm. In other languages with different rhythmic patterns, segmentation procedures also exploit rhythmic structure (see, e.g. Cutler, Mehler, Norris and Segui, 1992; Cutler and Otake, 1994; Otake, Hatano, Cutler and Mehler, 1993). In fact, rhythm allows a single, universally valid description of the different segmentation procedures used across languages. Important for the present discussion is that these effects thus indicate a way in which the lexical access process is indeed affected by prosodic structure (although in English the segmental reflections of rhythmic structure make a purely segmentally-based segmentation procedure feasible). The heuristic procedures which listeners use to facilitate word boundary location both in English and in other languages amount to a direction of attention to certain portions of the input rather than to other portions.

It is not only for speech segmentation purposes that such effects may be observed. A series of studies using the phoneme monitoring task provide evidence that listeners also use the overall prosodic contour of a sentence to direct attention to words bearing sentence accent. Response time to detect the initial phoneme of an acoustically constant word token is faster when the word occurs in a prosodic context consistent with sentence accent falling at that point than when it occurs in a context consistent with lack of accent (Cutler, 1976; Cutler and Darwin, 1981). Listeners can derive sufficient information to perform this attentional focus even when f_0 variation has been removed (Cutler and Darwin, 1981), although when dimensions of prosodic information conflict - such that, for example, the rhythm predicts accent where the f_0 contour predicts lack of accent - listeners refrain from deriving predictive information from prosody at all (Cutler, 1987). Similarly, the initial phonemes of nonsense words are detected more rapidly in contexts in which sentence rhythm predicts that the syllables containing the target will be accented (Shields, McHugh and Martin, 1974). Pitt and Samuel (1990b) presented acoustically constant versions of disyllabic minimal stress pairs at the ends of auditory lists in which all the disyllabic items had the same stress pattern; detection of a phoneme in these words was again faster when the syllable containing the target phoneme was predicted to be stressed, suggesting that listeners used the predictive information to attend selectively to stressed syllables.

The utility of prosodic information in human speech processing seems, therefore, to depend on the type of prosodic information involved. Information about word identity directly encoded in the stress pattern of a word is, as we saw above, not used in the computation of the initial lexical access code. Information about the general prosodic pattern-class of words, on the other hand, is used, and it is used in such a way as effectively to guide initiation of the access process. In a language such as English, this latter decision process can in practice be effected via vowel identity, i.e. via the segmental information alone, and thus does not depend on prosodic information. But this decision process is nevertheless highly similar to the process which exploits predictive sentence-level prosody. Rhythmic information directs attention to strong syllables (i.e. effectively to syllables containing full vowels) for the purpose of lexical segmentation; strong syllables are the most likely locations of word onsets in English. Predictive prosodic information directs attention to accented words or syllables, and effectively speeds processing and results in faster responses to the presence of target phonemes. In the final section we will consider the implications of this pattern of findings for the architecture of the human speech recognition system.

5 Conclusions

Each type of information we have discussed plays a different role in the processing which takes place for lexical access. Acoustic-phonetic information has a central and mandatory role in pre-lexical processing. The recognition system must deal with the variability of acoustic-phonetic information in the speech signal; and normalization processes such as those dealing with speech rate and speaker variability appear to operate always, outside of attentional control. These features suggest that pre-lexical processing entails the abstraction of segmental information from the speech signal. Although it is unclear precisely which representational units are abstracted pre-lexically, it seems clear that they constitute the basis of the lexical access code.

Lexical information, on the other hand, appears to play no role in pre-lexical processing, contrary to the claims of interactive models such as TRACE (McClelland and Elman, 1986), but consistent with autonomous models like Shortlist (Norris, 1994). Top-down connections from lexical to pre-lexical levels of processing are not required. Lexical effects which do occur in tasks requiring phonetic decisions can be explained either by a race between pre-lexical and lexical procedures (as in the Race model, Cutler et al., 1987) or by a pre-lexical mechanism which is sensitive to the sequential dependencies in speech (as in the recurrent net explanation of apparent lexical involvement in compensation for coarticulation). Both these mechanisms are instantiated in the Shortlist model. Furthermore, lexical effects in phonetic tasks are influenced by attentional manipulations. The modulation of lexical involvement by attention is more parsimoniously explained by autonomous than by interactive models.

Prosodic information, finally, plays an intermediate role. Information specifying the rhythmic structure of the language is used to constrain the process of lexical segmentation, and information about sentence-level accent can be used predictively, to benefit the processing of a word in an accented position. Prosodic information about lexical stress, however, does not appear to be used pre-lexically.

The direction of attention to some parts of the signal rather than others, which is how prosody is exploited pre-lexically, may seem analogous to potential lexical constraints on the pre-lexical access code. But the two are fundamentally different. The prosodic information which is used is present in the speech signal (information specifying vowel quality, sentence accent pattern, and so on), while lexical information has to be constructed from the signal via contact with a higher-level representation. In other words, prosodic information can be used bottom-up, while lexical information can only be used top-down. This means that prosodic information, like acoustic-phonetic information, can be used immediately, to benefit on-line processing. Top-down processing, on the other hand, is likely to be time-delayed, since lexical representations have to be accessed before they can influence processing. (Note that the prosodic information which cannot be computed on-line without delaying access, that is, lexical stress information, is precisely the type of prosodic information which appears not to be used pre-lexically.)

Furthermore, as Massaro (1989) has pointed out, lexical information, if used top-down, can act to distort the acoustic-phonetic information available in the signal. In the TRACE model, Massaro argues, top-down activation can eventually obliterate bottom-up evidence. The recognition system would surely be better designed if bottom-up information were not lost, and remained available to be used in perceptual decisions. Prosodic information available bottom-up, however, cannot distort the other information available in the signal.

In conclusion, it appears that human word recognition is an autonomous process. Information that is available in the speech signal is used to generate the lexical access code. This is largely acoustic-phonetic information specifying segmental structure, but includes information specifying the rhythmic structure of the input language and sentence-level prosodic information. Lexical information, however, does not appear to influence pre-lexical processing.