11

Speech Perception

JAMES M. MCQUEEN

The core process in speech perception is word recognition. If someone tells you *'Jim chose cucumber-green paper to wrap the present in'*, the only way you can come to understand this utterance is to map the auditory information in the speech signal onto your stored knowledge about the sound forms of the words of your language. Given that you usually do not know in advance what someone is going to say, and that every talker has an infinity of possible utterances to choose from, the only way you can decode any particular message is to recognize each of the words in that message (since the words do come from a finite set). This may sound trivial and, indeed, spoken word recognition seems to be pretty effortless. If someone says *'Jim chose cucumber-green paper to wrap the present in'*, no matter how unlikely that utterance may be, those are the words that you will perceive, and you will probably do so without any apparent difficulty.

But this is no mean feat. Spoken word recognition entails a complex decoding problem. The physical speech signal is a quasi-continuous stream of acoustic energy, varying over time in amplitude and spectral shape (i.e. consisting of different components at different frequencies). As shown in the spectrogram in Figure 11.1, for example, the first consonant [dʒ] is an affricate which consists of an abrupt onset of energy, followed by turbulent noise (aspiration and frication noise caused as air passes the partial constriction of the tongue on the roof of the mouth). This noise is spread mainly over relatively high-frequency bands above 1500 Hz. The first vowel [ɪ] consists of periodic energy (voicing due to vibration of the vocal folds), with energy concentrated in a number of frequency bands called formants (the resonant frequencies of the vocal tract). Each vowel has a characteristic formant pattern, depending on the shape of the vocal tract as that vowel is spoken (e.g. the position of the tongue, jaw and lips). The first three formants for this token of the vowel [ɪ] are centered roughly at 500, 1800 and 2300 Hz. This is the merest sketch of the acoustic parameters that code [dʒ] and [ɪ]. Ladefoged and Maddieson (1996) and Stevens (1998) provide detailed accounts of the acoustic structure of these and many other sounds of the world's languages, and of how the articulatory system generates those sounds.

To add to the complexity of the decoding problem, speech sounds (and therefore spoken words) are not invariant. As can be seen in Figure 11.1, for example, the acoustic realization of the [p] at the beginning of *paper* is not the same as the [p] at the end of *wrap*, nor the [p] in the middle of *paper*, nor the [p] at the beginning of *present*. This variability is due in part to the phonological context. Thus, syllable-initial and syllable-final stops are not the same (the initial [p] in *paper* has a release burst and has high-frequency aspiration energy; syllable-final stops do not have aspiration and may indeed not have a burst, as in the [p] of *wrap*). Furthermore, the two [p]'s in *paper* differ because of the stress pattern and syllabification of the word (the first [p] is syllable-initial in a stressed syllable; the second is ambisyllabic, at the boundary between a full and a reduced vowel). Similarly, the first [k] of *cucumber* is in a stressed syllable, and is therefore different from (e.g. longer than) the second [k], which is in an unstressed syllable.

Furthermore, speech is coarticulated: the shape of the vocal tract at any moment in time is usually determined by motor commands not only for the current speech sound, but also in part by those for preceding and following sounds. This means that

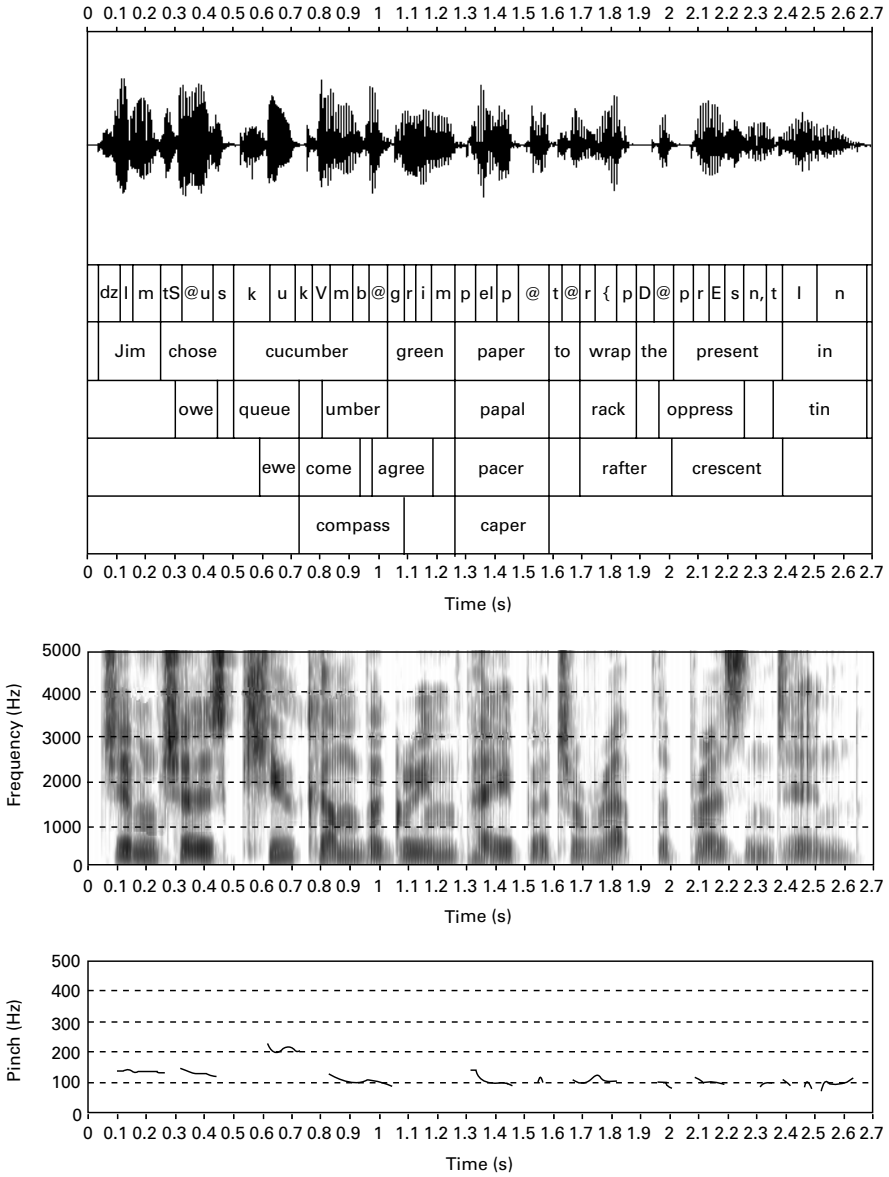

Figure 11.1 *A waveform, spectrogram and pitch track representing the sentence '*Jim chose cucumber-green paper to wrap the present in*', spoken by a male native speaker of British English. The waveform (top panel) shows the waveform as it varies in amplitude (on the vertical axis) over time (on the horizontal axis). Each sound in the utterance is transcribed using the Speech Assessment Methods Phonetic Alphabet (SAMPA, Wells, 1987; see Ladefoged, 2001, for an introduction to phonetic transcription and the International Phonetic Alphabet, IPA, which is used in the main text). Each sound is roughly aligned with the center of the information for that sound in the utterance. A selection of the words that are consistent with the different parts of the utterance is given under the phonetic transcription. The spectrogram (middle panel) shows how the different components of the speech signal are arrayed over a range of frequencies (on the vertical axis); these components change continuously over time. The relative darkness of the different components indicates their relative amplitude. The pitch track (bottom panel) shows how the speaker's fundamental frequency changes over the course of the utterance.*

each sound in the acoustic signal is influenced by its neighboring sounds, and thus that the realization of any given word is dependent, at least in part, on the sounds in neighboring words (e.g. the [ə] preceding the [p] and the following [r] in *the present* have consequences for the realization of the [p], thus making it different from the first [p] *in green paper*, which in turn is influenced by its neighboring sounds). Similarly, coarticulation has consequences on the realization of vowels, in particular their formant structure (i.e. as the mouth changes shape to create different consonantal constrictions the resonant frequencies of the vocal tract change; e.g. compare the slightly rising formants of the [ə] at the end of *paper*, before a [t], with the strongly falling formants of the [ə] of *to*, before a [r]).

The position of words and sounds in the prosodic structure of an utterance also has a strong effect on the way they are produced. Thus, for example, the production of *cucumber*, which the speaker accented, is marked by the lengthening of its segments, increased amplitude and pitch movement. The placement of this accent also has consequences for the realization of the rest of the utterance, as do the location of intonational boundaries (e.g. the [ɪ] of utterance-initial *Jim* is approximately 45 ms long, and thus much shorter than the [ɪ] of utterance-final *in*, which is approximately 124 ms long) and the type of utterance (e.g. whether it is a declarative or interrogative sentence).

To make matters even worse, the speech signal varies as a function of many other factors: the talker's sex, age and dialect; the idiosyncrasies of his or her vocal tract; his or her speaking rate and speaking style (e.g. formal, careful speech versus casual, colloquial speech); and the nature of any background noise. The utterance in Figure 11.1 would obviously look very different if it had been spoken by a female speaker of American English in a crowded restaurant to her partner, instead of by a male speaker of British English as a psycholinguistic example in a recording booth.

The listener is faced with yet other problems. The phonological space of possible words in a given language can be quite dense: many words share the same sequences of sounds, either because they begin in the same way (e.g. *wrap, rack, rafter, ramekin*, etc.), or because they rhyme (e.g. *present, crescent*), or because longer words often have shorter words embedded within them (e.g. *cucumber* contains *queue, ewe, come* and *umber*). It is therefore not easy to distinguish one word from all other words. The continuous nature of the speech signal (e.g. in Figure 11.1, the lack of a break in the signal between *to* and *wrap* in spite of the word boundary) adds to this problem: there is no way of knowing, in advance, how many words are in a given utterance, nor where they begin and end.

The task faced by the cognitive psychologist seeking to understand speech perception, then, is to derive a theory of the mental processes and representations that are used by listeners as they map complex acoustic speech signals onto their knowledge of spoken words. In this chapter, I argue that spoken word perception involves the simultaneous evaluation of multiple lexical hypotheses and a process of inter-word competition among those hypotheses. I also argue that listeners perform a detailed phonetic analysis of the speech signal prior to lexical access, at a prelexical stage of processing. That is, they first derive an abstract representation of the sounds in the speech signal before lexical access, rather than try to map the speech signal directly onto the mental lexicon. However, information appears to flow in cascade from the signal to the prelexical level, and from there in cascade up to the lexical level. A controversial question has been whether information also flows from the lexical to the prelexical level. I shall argue that there is no benefit to be had from on-line feedback of information from the lexicon to lower levels of processing, and no evidence that requires this type of feedback. Thus, for example, the knowledge that *wrap* is a word does not interfere with the prelexical analysis of its component sounds [r], [æ] and [p].

The chapter concludes with a brief discussion of several models of speech perception. The focus will again be on models of the perception of words rather than on models of the perception of speech sounds. Although an understanding of how the individual sounds of speech are decoded is critical in any account of speech perception, as indeed is an understanding of how the prosodic structure of utterances is decoded, I suggest that these components must be viewed in the context of a theory about how we perceive spoken words. This is because word forms are the primary perceptual representations of spoken language. That is, when we listen to speech, the perceptual level of analysis on which we focus our attention is the lexical level, not the level of individual sounds. Furthermore, as I have already argued, word forms provide the key to unlocking the speech code: speech comprehension is only possible if speech sounds are linked up to meaning representations via the phonological knowledge stored in the mental lexicon.

Perception of Spoken Words

Multiple activation of lexical hypotheses

Word recognition involves the simultaneous evaluation of many candidate words. As an utterance like our example sentence unfolds over time, words like *cucumber, queue, cute, cube, come* and *umber* will be considered in parallel as *cucumber* is heard. A common way to describe this process is in terms of lexical activation: each word that is being considered

is assumed to be activated (e.g. like a node in a connectionist network). Evidence for multiple activation comes from cross-modal semantic priming experiments, where participants decide whether visually presented letter strings are real words or not, as they hear spoken words or sentences.[1] Facilitation (i.e. speeding up) of responses to semantic associates of particular words is taken to reflect activation of those words. Cross-modal semantic priming has shown that competitors beginning at the same time are activated (e.g. in Dutch, listeners responded more rapidly to associates of both *kapitaal*, capital, and *kapitein*, captain, when they heard [kæpɪt] than when they heard the beginning of an unrelated word: Zwitserlood, 1989; see also Moss, McCormick, & Tyler, 1997; Zwitserlood & Schriefers, 1995). Other evidence for multiple activation of words that begin in the same way comes from recognition memory experiments. Participants make false positive errors on words that have not been presented earlier in the experiment but that begin in the same way as words that have appeared earlier (Wallace, Stewart, & Malone, 1995; Wallace, Stewart, Shaffer, & Wilson, 1998; Wallace, Stewart, Sherman, & Mellor, 1995).

Words beginning at different points in the signal can also be activated. For example, in English, faster responses to an associate of *bone* were found when listeners heard *trombone* than when they heard an unrelated word (Shillcock, 1990; but see also Vroomen & de Gelder, 1997). Other evidence that words beginning at different points in the signal can be simultaneously activated comes from Italian and English. In Italian, responses to an associate of *visite*, visits, for example, were faster when listeners heard *visi tediati*, bored faces, than in a control condition (Tabossi, Burani, & Scott, 1995). In English, faster responses to associates of both *lips* and *tulips*, for example, were found when listeners heard *two lips* than in a control condition (Gow & Gordon, 1995). These latter two studies show that candidate words can be activated even if they span word boundaries in the input.

Words that end in the same way as the input speech material are also activated. Connine, Blasko, and Titone (1993), for example, found evidence in cross-modal semantic priming when nonword primes differed from the base words in only one or two articulatory features (*zervice* primed responses to *tennis*, presumably due to activation of *service*). Similar results have also been observed with intramodal (auditory-auditory) priming in Dutch (Marslen-Wilson, Moss, & van Halen, 1996). Connine et al. found no reliable priming effect, however, when the primes differed from the base words on more than two features (e.g. *gervice*; the [g] differs from the [s] of *service* on more featural dimensions than the [z] of *zervice* does; see, e.g. Jakobson, Fant, & Halle, 1952, for a feature-based analysis of speech sounds).

The importance of the degree of mismatch between the material in the speech signal and stored lexical knowledge has also been observed using the phoneme monitoring task. Connine, Titone, Deelman, and Blasko (1997) asked listeners to detect the phoneme /t/, for example, in *cabinet, gabinet* (one feature change on the initial phoneme), *mabinet* (many features changed) and *shuffinet* (control). Phoneme monitoring latencies were fastest for targets in the base words (*cabinet*), slower for targets in the minimally mismatching nonwords (*gabinet*), even slower for targets in the nonwords with greater mismatch (*mabinet*) and slowest for targets in control nonwords (*shuffinet*). As will be discussed in more detail later, phoneme monitoring latency reflects degree of lexical activation. These results thus suggest that the lexical representation for *cabinet* is more strongly activated when *gabinet* is heard than when *mabinet* is heard. Likewise, in another phoneme monitoring study, Frauenfelder, Scholten, and Content (2001) found evidence of lexical activation of French words when the initial phonemes of those words were distorted by a single feature (e.g. *vocabulaire*, vocabulary, produced as *focabulaire*). Monitoring latencies on target phonemes in these words were faster than on targets in control items, but only when the target phonemes were word-final. Frauenfelder et al. suggest that the lexical effect only emerged when enough time had elapsed for the matching information later in the word to overcome the effects of the initial mismatch.

It therefore appears that the word recognition process is relatively intolerant of mismatch between the speech signal and stored phonological knowledge. This is true for initial mismatch, as in the studies just cited, and for mismatch occurring later in the input. Zwitserlood (1989), for example, found evidence of activation of both *kapitaal* and *kapitein* when listeners had heard [kæpɪt], but of only the correct word when disambiguating information (the final vowel and consonant) had been heard. There is thus rapid selection of the intended word and rapid rejection of other candidates as soon as the signal mismatches with the incorrect candidates.

Results from phoneme monitoring support this suggestion. Frauenfelder et al. (2001) found no evidence of activation of, for example, *vocabulaire* given *vocabunaire* (i.e. responses to the final /r/ were no faster than in control nonwords). Soto-Faraco, Sebastián-Gallés, and Cutler (2001) carried out a series of cross-modal fragment-priming experiments (in Spanish) to examine the effects of noninitial mismatch. Lexical decision responses to visually presented words like *abandono*, abandonment, were faster, relative to a control condition, if listeners had just heard the matching fragment *aban*, and slower if they had just heard the mismatching fragment *abun*, the onset of *abundancia*,

abundance. Again, it appears that incorrect candidate words are rapidly rejected as soon as they mismatch with the signal.

In general, however, words that begin in the same way as the input and then diverge from it will tend to be activated earlier and more strongly than those that have initial mismatches but end in the same way as the input. This is because candidates with initial overlap will at least initially be as strongly supported by the input as the actual word in the input, while those with final overlap will never be as good a match to the input as the actual word. The difference due to overlap position has been observed in eye-tracking experiments, where participants' fixations to pictures on a computer screen are collected while they are auditorily instructed to click on one of those pictures. As the name of the target picture unfolds over time, participants tend to make more fixations to pictures with names that are compatible with the available spoken information than to unrelated pictures (e.g. looks to a picture of a beetle when the initial sounds of *beaker* are heard: Allopenna, Magnuson, & Tanenhaus, 1998). Somewhat later in time, and to a lesser extent, listeners tend to look at pictures with names that rhyme with the input (e.g. listeners look at a picture of a speaker when they hear *beaker*: Allopenna et al., 1998).

Modulation of lexical activation by fine-grained information

The activation of lexical hypotheses therefore varies continuously over time, in response to the goodness of match of each hypothesis to the current input. As *paper* is heard, for example, *papal* and *pacer* and *caper* (among others) will be activated to differing degrees at different moments in time (*papal* will be more active than *pacer*, and for longer, and will be as good a hypothesis as *paper* itself until evidence than there is no /l/ is heard; the activation of *caper* will increase as that for *pacer* drops, and so on). The results of Connine et al. (1993, 1997) suggest in addition that lexical activation is graded not only with respect to the number of matching and mismatching phonemes, but also in response to more fine-grained differences in the input: words that mismatch with the signal on fewer features are more strongly activated than those that mismatch on more features.

Many other studies have also shown that lexical activation varies in response to fine-grained differences in the speech signal (see, e.g. Warren & Marslen-Wilson, 1987, 1988). A series of studies on the effect of subphonemic mismatch in cross-spliced words and nonwords has examined this issue in detail (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Marslen-Wilson & Warren, 1994; McQueen, Norris, & Cutler, 1999; Streeter & Nigro, 1979; Warren & Marslen-Wilson, 1987; Whalen 1984, 1991). In these experiments, cross-spliced versions of monosyllabic words and nonwords were constructed by concatenating their initial portions (up to the vowel) with the final consonants of other tokens. For example, the word *job* could be made with the [dʒɒ] of *jog* and the [b] of *job*. The vocalic portion would thus contain formant-transition information consistent with a velar [g] (coarticulation, as described in the Introduction, results in changes in the formant structure of the vowel before the [g], relative to the same vowel before a [b]). The cues to an upcoming [g] would thus mismatch with the final bilabial stop release burst of the [b]. These experiments showed that the strength of the interference due to this kind of mismatch depends on lexical factors (whether the portions used for cross-splicing come from words or nonwords, and whether the resulting sequence is a word or a nonword). They thus show that subphonemic details filter through to the lexical level to influence lexical activation.

Studies by Andruski, Blumstein, and Burton (1994) and by Utman, Blumstein, and Burton (2000) also show that lexical activation varies as a function of subphonemic differences in the speech signal. Andruski et al., for example, present evidence that the degree of activation of words beginning with unvoiced stop consonants varies with the Voice Onset Time (VOT) of those stops (VOT is the time from the release burst to the onset of voicing, approximately 90 ms for the initial [k] of *cucumber* and approximately 16 ms for the [g] of *green* in Figure 11.1; VOT is one acoustic cue to the distinction between unvoiced and voiced stops, such as [k] and [g], or [p] and [b]). The [p] of *pear*, for example, was presented by Andruski et al. in unedited form, or with the VOT reduced by either one-third or two-thirds. All three forms primed responses to an associate word (*fruit*), but responses were faster after the unedited prime than after the less extremely edited prime, which in turn were faster than responses after the more extremely edited prime.

Information about the syllabic structure of utterances also influences the activation levels of words. Tabossi, Collina, Mazzetti, and Zoppello (2000) have shown in a fragment-priming study in Italian that words that match the syllable structure of the fragments are more strongly activated than words that mismatch the fragments. For example, *si.lenzio*, silence, appeared to receive more support than *sil.vestre*, silvan, when the input was the fragment [si.l], taken from *si.lenzio*. The opposite was true when the fragment was [sil] from *sil.vestre*. Small durational differences between the vowels in these fragments appear to have signaled the difference in syllabic structure. Spinelli, McQueen, & Cutler (2003) have recently shown in a cross-modal

identity-priming experiment that, even though French sequences such as *dernier oignon* (last onion) and *dernier rognon* (last kidney) are phonemically identical, French speakers signal the difference between these two sequences to listeners. Only responses to the words that the speaker intended were significantly primed. The duration of the medial [ʁ] in *dernier oignon*, for example (where it was the final sound of *dernier*, and was produced because of the phonological process of liaison), was slightly shorter than the equivalent [ʁ] in *dernier rognon* (where it was the first sound of *rognon*), and this difference influenced the degree of activation of these words.

Gow and Gordon (1995) also found that fine-grained durational differences in the speech signal influence lexical activation. They found evidence of activation of, for example, both *tulips* and *lips* when listeners heard *two lips* (where the medial [l] was longer) but of only *tulips* when listeners heard *tulips* (where the [l] was shorter). Likewise, subtle durational differences between productions of an ambiguous sequence (e.g. [kæp]), which can either be a monosyllabic word (*cap*) or as the onset of a longer word (*captain*), bias lexical activation in favor of the speaker's intentions (Davis, Marslen-Wilson, & Gaskell, 2002; Salverda, Dahan, & McQueen, 2003).

The picture that therefore emerges is that lexical activation is fine-tuned to the acoustic-phonetic detail in the speech signal. Small differences in the duration of segments can influence the amount of support for particular candidate words. Two other sources of variation in the signal also influence lexical activation. First, the phonological context can determine the effect of mismatching information. Spoken in isolation, the sequence [grim] is not a good match to the word *green*. But this is how this word was produced in the utterance shown in Figure 11.1, where the following word (*paper*) begins with the bilabial stop [p]. In English, word-final coronal consonants (like [n]) can take on the place of articulation of the following consonant ([n] can thus become the bilabial [m]). Research on this assimilation process has shown that a form such as [grim] can be recognized as a token of the word *green*, but only if the change is contextually appropriate (Coenen, Zwitserlood, & Boelte 2001; Gaskell & Marslen-Wilson, 1996, 1998, 2001; Gow, 2001; Marslen-Wilson, Nix, & Gaskell, 1995). Gow (2002) has recently shown that there are also subphonemic cues to assimilation (e.g. the [raɪp] in *right berries*, where the final /t/ of *right* is assimilated to a [p], is not identical to the [raɪp] in *ripe berries*), and that this information, like the other subphonemic cues mentioned above, appears to influence lexical activation.

Second, lexical stress information appears to influence the degree of activation of words (Cooper, Cutler, & Wales, 2002; Cutler & van Donselaar, 2001; Soto-Faraco et al., 2001; see also the review in Cutler, Dahan, & van Donselaar, 1997). For example, Soto-Faraco et al. showed that mismatching stress information inhibits responses in Spanish cross-modal fragment priming: responses to the visual target *principio* (beginning), which is stressed on the second syllable, were slower after listeners heard the fragment *PRINci-*, the beginning of *príncipe* (prince), which is stressed on the first syllable, than after they heard a control fragment. The use of stress information appears to be limited to languages where this information is useful for lexical disambiguation. In French, for example, where words do not contrast in stress, this information does not appear to influence lexical activation (Dupoux, Pallier, Sebastián-Gallés, & Mehler, 1997).

At any one moment in the unfolding of a speech signal, therefore, multiple lexical hypotheses are being considered. The activation process is fairly intolerant of mismatching information, such that, as soon as one word has sufficient support, the system settles on that candidate and the other words will drop out of the running. Furthermore, lexical representations are activated in a graded fashion, in response to their goodness of fit to the available input. Even very subtle acoustic details in the speech signal have an impact on lexical activation levels.

Competition between lexical hypotheses

How then is a choice made from among the set of activated lexical hypotheses? When *paper* is heard, how is *paper* selected and how are *papal, caper* and so on rejected? One possibility is that the selection is made purely on the basis of bottom-up goodness of fit. The word that best matches the input is selected, and all other words are then rejected. Another possibility, however, is that, in addition, the lexical hypotheses compete with each other. Competition between candidate words would act to sharpen the distinctions in activation levels made on the basis of goodness of match.

There are many different sources of evidence suggesting that there is inter-word competition between candidate words. Listeners in a word-spotting task find it harder to spot words embedded at the beginning of longer words (like *sack* in [sækrəf], the beginning of *sacrifice*) than in matched sequences that are not word onsets (like [sækrəf]: McQueen, Norris, & Cutler, 1994). This effect presumably reflects competition between the shorter and the longer word. Competition also occurs in word spotting when the shorter and the longer word begin at different points (e.g. spotting *mess* in [dəmɛs], the beginning of *domestic*, is harder than in the nonword onset [nəmɛs]: McQueen et al., 1994). Furthermore, the number of competitors beginning at a different point in the signal from the target word influences

how easy it is to recognize that target. Recognizing a word embedded in a longer nonsense word is harder when the longer word contains a sequence consistent with many other words than when it contains a sequence consistent with fewer other words (Norris, McQueen, & Cutler, 1995; Vroomen & de Gelder, 1995).

A comparison between the studies of Soto-Faraco et al. (2001) and Cutler and van Donselaar (2001) offers further support for lexical competition. As mentioned earlier, Soto-Faraco et al. showed that responses to the Spanish target word *principio* (stressed on the second syllable, *prinCIpio*), were slower after listeners heard the mismatching fragment *PRINci-*, the beginning of *principe* (stressed on the first syllable, *PRINcipe*), than after they heard a control fragment. In a similar fragment-priming study in Dutch, however, no such inhibition was observed (e.g. the prime *MUzee*, a fragment mismatching with the target *museum*, which, as in English, is stressed on the second syllable, *muSEum*, did not inhibit responses to the target: Cutler & van Donselaar, 2001). The difference between the two studies is that the mismatching fragments in Spanish were consistent with other words (e.g. *principe*, prince) while those in Dutch were not (e.g. *MUzee* is not the beginning of any Dutch word). The inhibitory effect in Spanish thus appears to be due to the conjoint effects of bottom-up mismatch and lexical competition.

Lexical competition can also be inferred from the effects of manipulating the lexical neighborhood density of words (the number and frequency of similar-sounding words). It should be harder to recognize a word in a dense neighborhood than in a sparse neighborhood because of stronger inter-word competition in the denser neighborhood. Inhibitory effects of high neighborhood density have indeed been found (Cluff & Luce, 1990; Luce, 1986; Luce & Large, 2001; Vitevitch, 2002; Vitevitch & Luce, 1998, 1999). In priming paradigms, responses to target words tend to be slower when they are preceded by phonologically related prime words than when they are preceded by unrelated words, suggesting that the target words were activated when the related primes were heard, and that they then lost the competition process, making them harder to recognize (Goldinger, Luce, Pisoni, & Marcario, 1992; Luce, Goldinger, Auer, & Vitevitch, 2000; Monsell & Hirsh, 1998; Slowiaczek & Hamburger, 1992).

Yet another source of evidence for lexical competition comes from experiments examining the effects of subphonemic mismatch on the activation of monosyllabic words. A number of studies (Andruski et al., 1994; Gow, 2001; Marslen-Wilson et al., 1996; Marslen-Wilson & Warren, 1994; McQueen et al., 1999) have all shown that degree of activation of short mismatching words depends on the lexical competitor environment. Monosyllabic words that mismatch with the input tend to be activated more strongly when there are no other phonologically close candidate words (i.e. when their activation is not reduced because of strong lexical competition).

Segmentation of continuous speech into words

Speech decoding thus involves the multiple graded activation of candidate words, and competition between those words. Lexical competition acts to exaggerate the differences in activation of candidate words as signaled by the (sometimes very fine-grained) acoustic detail in the speech signal, and thus makes word recognition more efficient. Lexical competition can also help in segmenting the continuous signal into individual words (McClelland & Elman, 1986; McQueen, Cutler, Briscoe, & Norris, 1995; McQueen et al., 1994; Norris, 1994). As the evidence reviewed earlier suggests, candidate words compete with each other when they begin at the same point in the signal (e.g. *cucumber* and *queue*) and when they begin at different points (e.g. *cucumber* and *agree*, which, though their onsets are well separated, are still fighting over one vowel in the input, see Figure 11.1; note that this is true in Standard Southern British English, but not in American English). This competition process will thus help to select the best-matching sequence of words for the entire utterance. That is, not only will competition help in the selection of, for example, *cucumber*, but it will also help in segmenting the signal at word boundaries (e.g. as *cucumber* and *green* win, candidates spanning the boundary between these words, such as *agree*, will lose). The correct segmentation will thus emerge out of the competition process, even when no word boundary is marked in the signal at that point.

There is, however, more to segmentation than competition. Although the speech signal is continuous, with coarticulation of sounds both within and between words, and there are no fully reliable cues to word boundaries (Lehiste, 1972; Nakatani & Dukes, 1977), there are nonetheless a variety of cues to likely word boundaries in the speech signal. When those cues are available, listeners appear to use them (Norris, McQueen, Cutler, & Butterfield, 1997). For example, some sequences of speech sounds do not co-occur within syllables (e.g. [mr] in English); such sequences signal the location of a likely word boundary (e.g. between the [m] and the [r]). Listeners use such phonotactic constraints for lexical segmentation (Dumay, Frauenfelder, & Content, 2002; McQueen, 1998; Weber, 2001). Likewise, sound sequence constraints that are probabilistic rather than absolute (some sound sequences are more likely to be associated with a

word boundary than other sequences) are also used by listeners for segmentation (van der Lugt, 2001).

The rhythmic structure of speech is also used by listeners for lexical segmentation. English and Dutch listeners use metrical information based on stress (Cutler & Butterfield, 1992; Cutler & Norris, 1988; Norris et al., 1995; Vroomen & de Gelder, 1995; Vroomen, van Zon, & de Gelder, 1996). French, Catalan and Spanish listeners use metrical information based on the syllable (Cutler, Mehler, Norris, & Seguí, 1986; Pallier, Sebastián-Gallés, Felguera, Christophe, & Mehler, 1993; Sebastián-Gallés, Dupoux, Seguí, & Mehler, 1992; but see also Content, Meunier, Kearns, & Frauenfelder, 2001). Japanese listeners use metrical information based on a sub-syllabic structure, called the mora (Cutler & Otake, 1994; McQueen, Otake, & Cutler, 2001; Otake, Hatano, Cutler, & Mehler, 1993). The differences between the results of these studies reflect differences in the rhythmic structures of the languages that were tested. In each case, however, the edge of a rhythmic unit (a strong syllable in English, a syllable in French, a mora in Japanese) can be considered to be a likely word boundary in the speech signal (see Norris et al., 1997, for further discussion).

In addition to phonotactic and metrical cues to word boundaries, other sources of evidence in the speech signal indicate where words are likely to begin and end. Allophonic cues such as the aspiration of word-initial stops (such as the first [k] of *cucumber* in Figure 11.1) could be used in segmentation (Church, 1987; Lehiste, 1960; Nakatani & Dukes, 1977). In addition to allophonic cues, other acoustic differences (such as in the duration of segments) signal the correct segmentation of ambiguous sequences such as *night rate/nitrate* or *two lips/tulips*, and appear to be used by listeners for this purpose (Christie, 1974; Gow & Gordon, 1995; Lehiste, 1972; Nakatani & Dukes, 1977; Oller, 1973; Quené, 1992, 1993; Turk & Shattuck-Hufnagel, 2000).

How then can all of these cues influence the activation of candidate words, the competition among them, and hence the segmentation of an utterance? Norris et al. (1997) have proposed that a penalty is applied to the activation levels of candidate words that are misaligned with likely word boundaries (irrespective of how any particular boundary might be cued). They argued that a candidate word would count as misaligned if the stretch of speech between that word and the likely word boundary did not contain a vowel. Such stretches of speech (single consonants, or consonant sequences) are themselves not possible words, so any segmentation of an utterance into a potential sequence of words (a 'lexical parse') involving these consonantal chunks is highly implausible. For example, the lexical parse *queue come b agree m paper* is not going to be what the speaker of the utterance in Figure 11.1 intended. The Possible Word Constraint (PWC: Norris et al., 1997) thus penalizes *agree*, for example, because the [m] between the end of *agree* and the boundary at the onset of *paper* (cued in this case by the fact that the first syllable of this word is strong [Cutler & Norris, 1988] and/or by the fact that the [p] is aspirated [Church, 1987] is not a possible English word. Penalization of misaligned words will help the correct words to win the competition process.

Norris et al. (1997) provided empirical support for the PWC. English listeners found it harder to spot words like *apple* in nonsense sequences like *fapple* (the [f] between the onset of *apple* and the preceding silence, a very likely word boundary, is not a possible English word) than in sequences like *vuffapple* (*vuff* is not an English word, but could be). Further support for the PWC has since been found in English (Norris, McQueen, Cutler, Butterfield, & Kearns, 2001) and a number of other languages (Dutch: McQueen, 1998, McQueen & Cutler, 1998a; Japanese: McQueen, Otake, & Cutler, 2001; and Sesotho: Cutler, Demuth & McQueen, 2002). In spite of differences between these languages in what constitutes a well-formed word (how long or complex a sequence of sounds has to be to form a word in a particular language), this research suggests that the PWC operates according to a simple, and possibly language-universal, principle: namely, that a stretch of speech without a vowel is not a viable chunk in the lexical parse of an utterance.

The PWC is thus an important mechanism that enriches the competition process. It acts to rule out spurious candidate words, and thus helps in the segmentation of continuous speech into the sequence of words that the speaker of that utterance intended. While the PWC may be the means by which many cues to word boundaries modulate the activation and competition process, it may not be the only mechanism. In some cases, the activation of candidate words appears to be influenced by cues to word-boundary locations even when those words are not misaligned with those locations. As was discussed earlier, subtle durational differences in the input can influence lexical activation (Davis et al., 2002; Gow & Gordon, 1995; Salverda et al., 2003). Salverda et al., for instance, found in a study in Dutch that the word *ham* (id.) was more strongly activated when listeners heard a token of *hamster* (id.) in which the first syllable [ham] had been taken from an utterance where the speaker intended the word *ham* (and was longer in duration), than when it had been taken from another token of the word *hamster* (and was shorter).

One way to interpret findings such as this is to assume that the durational information signals a word boundary (e.g. a longer [ham] indicates that there is a boundary after the [m]). Many authors have indeed argued that segmental lengthening

signals the edges of prosodic domains (e.g. utterance and phrase boundaries, but also word boundaries: Beckman & Edwards, 1990; Cho & Keating, 2001; Fougeron, 2001; Fougeron & Keating, 1997; Turk & Shattuck-Hufnagel, 2000). While the PWC could penalize candidate words that are misaligned with boundaries that have been cued by segmental lengthening, it will not penalize an aligned word such as *hamster*. The results of Davis et al. (2002) and Salverda et al. (2003) thus suggest that, in addition to the PWC, there may be another segmentation process, one that would boost the activation of words that are aligned with prosodic boundaries. It will be important in future research to establish whether a single mechanism can account for the modulation of competition-based segmentation by word-boundary cues, and, if not, to establish the relative roles of processes that boost the activation of words that are aligned with possible word boundaries and processes that penalize misaligned words.

Summary

Spoken word perception emerges from competition among candidate lexical hypotheses. The activation of any given word is modulated by three crucial factors: its own goodness of fit to the current signal; the number of other words that are currently active; and their goodness of fit. The quality of match between any candidate and the signal is in turn determined by a number of factors: how well the signal and the stored lexical knowledge match in terms of segmental material (individual speech sounds); how well they match in terms of suprasegmental material (e.g. stress pattern); the appropriateness of the word's form in the phonological context (e.g. as in the case of place assimilation); and its (mis)alignment with likely word boundaries in the signal. We have seen that lexical activation is finely tuned to the signal, such that even very subtle acoustic details in the input can influence the pattern of activation at the lexical level. We have also seen that the segmentation of a continuous utterance into individual words is a product of this activation and competition process.

There is of course much more to speech comprehension than this. Word perception is not language understanding. The listener has much more to do than recognize the words of an utterance if he or she wants to comprehend the speaker's message. An interpretation must be built on the basis of the meanings of the words, the syntactic structure of the utterance, and the discourse and situational context of the utterance. It is beyond the scope of this chapter to review the evidence on these higher levels of speech comprehension. A review of research on the morphological structure of words (inflectional structures like *leap* + *ing*; derivational structures like *sweet* + *ly*; compounds like *tree* + *stump*) and the role of morphology in speech processing can be found in McQueen and Cutler (1998b). A recent discussion of the representation of word semantics is provided by Rodd, Gaskell, and Marslen-Wilson (2002). In speech comprehension (as opposed to reading), the prosodic and intonational structure of sentences plays a key role in syntactic and discourse processing. Reviews on these issues are to be found in Cutler et al. (1997) and in Cutler and Clifton (1999).

The next part of the chapter will instead focus on the way in which the speech signal is mapped onto the mental lexicon, and on the perception of speech sounds. It is important to note, however, that this perceptual process appears to be uninfluenced by higher-level factors. That is, sentential context does not affect the early stages of the perceptual analysis of speech (van Alphen & McQueen, 2001; Connine, 1987; Connine, Blasko, & Hall, 1991; Miller, Green, & Schermer, 1984; Samuel, 1981). With respect to the effects of context on lexical activation, it appears that context does not influence which words are activated, but does influence the process of word selection (Marslen-Wilson, 1987; Zwitserlood, 1989). Recent evidence suggests that context can start to influence lexical interpretation very early, that is, before information in the signal is available to distinguish the intended word from its competitors (van den Brink, Brown, & Hagoort, 2001; Van Petten, Coulson, Rubin, Plante, & Parks, 1999). But it has not yet been determined whether context influences the activation of form representations of words, or meaning representations, or the integration of lexical information into the interpretation of the utterance. What is clear, however, is that the core perceptual process, that is, the generation of lexical hypotheses, is driven by the speech signal alone.

Perception of Speech Sounds

Intermediate representations

One possible account of speech perception is that acoustic-phonetic information is mapped directly onto the mental lexicon, with no intermediate stage of processing. On this view, the output of the peripheral auditory system would be the input to the lexicon, and knowledge about the sound-form of words would have to be stored in each lexical representation in the form of auditory primitives (Klatt, 1979). The alternative is that there is a prelexical level of representation (or indeed, more than one such level), which mediates between the output of the auditory system and the lexicon. There are several arguments that can be made in support of a prelexical level of processing.

First, intermediate representations could remove considerable redundancy that otherwise would have

to exist at the lexical level. If knowledge about the acoustic form of a particular speech sound, [p] for example, can be stored prelexically, then it does not need to be stored multiple times at the lexical level, in the representations of every word with one or more [p]'s. Instead, only the symbolic phonological information would need to be stored lexically. Recoding the acoustic signal in terms of a relatively small number of speech sounds (most languages have fewer than 50 phonemes: Maddieson, 1984), or in terms of other abstract phonological representations (e.g. syllables: Mehler, 1981), would achieve considerable information reduction and would thus make word recognition much more efficient.

Clearly, given the variability and complexity of the speech signal, as sketched in the Introduction (e.g. the different [p] sounds in the utterance in Figure 11.1 are all different), it is far from trivial to recode the signal in terms of abstract phonological representations. The variability problem raises many issues about the nature of prelexical representation (e.g. should there be position-specific, 'allophonic' units that code the major differences within phoneme categories, such as those for syllable-initial versus syllable-final [p]?). Nevertheless, variability does not invalidate the benefits of abstract prelexical recoding. Word recognition would benefit if at least part of the speech code could be cracked prelexically.

Speech-sound perception entails the integration of many different acoustic cues (Diehl & Kluender, 1987; Fitch, Halwes, Erickson, & Liberman, 1980; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Repp, 1982; Repp & Liberman, 1987; Repp, Liberman, Eccardt, & Pesetsky, 1978). The prelexical level provides a stage of processing at which this integration can take place. The prelexical level also provides a way of coding how those speech sounds are contextually conditioned. Indeed, coarticulation and other connected speech phenomena make it clear that speech sounds must always be interpreted in their phonetic context (Fowler, 1984). As Repp and Liberman (1987) point out, different acoustic signals need to be interpretable, depending on context, as the same sound, and it must also be possible to interpret the same acoustic signal in different contexts as different sounds. One kind of context is that provided by neighboring speech sounds. The identity of fricative sounds like [s] and [ʃ], for example, is dependent in large part on the frication noise itself, but also on neighboring vowels (Mann & Repp, 1980; Nearey, 1992; Smits, 2001a, 2001b; Whalen, 1981). Similarly, fricative sounds affect the perception of following stop consonants (Mann & Repp, 1981). It is usually assumed that the mechanisms that are responsible for these perceptual compensations for coarticulation have a prelexical locus (Elman & McClelland, 1986, 1988). One source of evidence in support of this assumption is that compensatory effects can dissociate from lexical effects on phonetic perception (Pitt & McQueen, 1998; see below). If the mechanism responsible for compensation for coarticulation acted at the lexical level, compensatory and lexical effects should not dissociate.

Another kind of contextual influence on speech sound perception is that due to speech rate (Miller, 1981). Miller and Liberman (1979) showed that the duration of the formant transitions from the consonants to the vowels in [ba] and [wa] is a cue to the distinction between [b] and [w], but also that this durational cue depended on the speaking rate. Changes in the duration of the steady-state vowel (as a cue to speaking rate) modified the interpretation of the formant transition duration. As with the adjustments for coarticulation, those responsible for rate normalization appear to dissociate from higher-level biases in phonetic perception (Miller et al., 1984; Miller & Dexter, 1988). It is therefore reasonable to assume that the mechanisms responsible for speaking rate normalization should also have a prelexical rather than a lexical locus.

There are thus good grounds to assume that there is an intermediate level of processing in the speech perception system, between auditory processing and lexical processing. This stage of processing (or perhaps separate sub-stages) acts to normalize the signal and to abstract some kind of phonetic code that can then be used for lexical access. One line of research that is potentially problematic for the view that speech perception involves abstract prelexical representations is that on talker variability. This research has shown, for example, that listeners cannot ignore talker variability, that the processing of talker information and phonetic information are not entirely independent, and that talker-specific knowledge is stored in long-term memory (Church & Schacter, 1994; Goldinger, 1996; Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Palmeri, Goldinger, & Pisoni, 1993; Schacter & Church, 1992). Such data have been taken as evidence for episodic theories of spoken word recognition, in which the lexicon consists of detailed episodic traces of the words that the listener has heard (Goldinger, 1998), and in which there are no abstract prelexical representations. These results certainly challenge pure abstractionist accounts, in which all speaker-specific information is thrown away at a prelexical normalization stage. But there is considerable scope for hybrid models in which talker-specific information (and other indexical information) is preserved, but in which there are also abstract, symbolic representations (Schacter & Church, 1992). Such models could account for the evidence on memory for talker details and at the same time could account for the evidence supporting abstract intermediate representations.

Even among theories that include a prelexical level, however, there is no consensus on the size of

the representational units. Many possibilities have been proposed, including phonemes (Foss & Blank, 1980; Nearey, 2001; Norris, 1994), syllables (Massaro, 1987; Mehler, 1981), acoustic-phonetic features (Lahiri & Marslen-Wilson, 1991; Marslen-Wilson & Warren, 1994; Stevens, 2002), a combination of a featural level followed by a phonemic level (McClelland & Elman, 1986), articulatory gestures (Liberman & Mattingly, 1985) and context-sensitive allophones (Luce et al., 2000; Wickelgren, 1969). Space limits preclude a detailed review of the arguments for and against each of these representational alternatives. The short version of the story is this: the issue has not been resolved. One recurring issue in this debate has been whether the objects of speech perception are fundamentally acoustic in nature (Kluender, 1994), or are gestural in nature (Fowler, 1986, 1996; Liberman & Mattingly, 1985), or are the product of pattern-recognition processes (Massaro, 1987; Nearey, 1997). Another, related issue has been whether speech perception calls on special processes (Liberman & Mattingly, 1985), or depends on general auditory processes that are also used in the perception of other complex sounds (Pastore, 1981; see, e.g. the debate between Fowler, Brown, & Mann, 2000, and Lotto & Kluender, 1998, and Lotto, Kluender, & Holt, 1997).

One reason why an answer to the 'unit of perception' question has eluded speech scientists is that such units are mental representations to which, during normal communicative speech processing, the listener does not attend (instead, as I suggested earlier, the listener focuses on words and on building an interpretation of the current utterance). There is thus no direct way to examine the contents of the prelexical level. Representations at the prelexical level of processing must be inferred from experiments that probe lexical-level processing, or from experiments that require decisions about speech sounds (e.g. the classification of a sound as one of two phonetic categories, or the detection of target sounds or sound sequences). Either way, it is possible that such experiments may call not only on the processes and representations that are used in normal speech comprehension, but also on processes that are required for the experimental task but that are not part of normal processing. It is thus possible that results that might be taken as support for a particular form of prelexical representation instead reflect lexical-level representations (if the task involves word responses) or decision-level representations (if the task involves phonological decision-making). Another reason for the stalemate in the 'unit of perception' debate is that results are often consistent with a number of representational alternatives (see, e.g. McQueen et al., 1999).

While it remains very difficult to tie down the size of the prelexical 'unit of perception', one aspect about the way the prelexical level operates does now seem to be very clear. This is that it operates in cascade. Activation flows through this level of processing in a continuous fashion. The evidence showing subphonemic effects on lexical activation reviewed earlier in this chapter supports this claim. Since fine-grained information in the speech signal modulates the activation of words, it cannot be the case that the prelexical level acts in a serial and categorical way. If phonemic units, for example, reached some threshold of activation and then only passed on discrete signals to the lexicon that particular phonemes had been detected, then subphonemic variation would be unable to influence lexical activation. Instead, prelexical representations must be activated in proportion to their match with the speech signal, and in turn pass activation continuously to word representations. Units larger than the phoneme (e.g. syllabic representations) or smaller (e.g. featural representations) could also pass fine-grained information on to the lexical level in cascade, just like phonemic representations.

Note that cascaded processing in speech perception is not limited to the interface between the prelexical and lexical stages. The evidence that was presented earlier on the activation of word meanings before the acoustic offset of those words (e.g. Allopenna et al., 1998; van den Brink et al., 2001; Zwitserlood, 1989) suggests that activation also cascades from word-form representations to meaning representations. Thus, for example, it is not the case that categorical decisions about each of the sounds of *cucumber* are made before the lexical representation of *cucumber* is activated. Instead, as soon as some degree of support has accumulated at the prelexical level for [k], for example, this information cascades up to the lexical level, influencing the activation of words beginning with [k], and the activation of their meanings.

Flow of information in the recognition system

Information must flow from the prelexical level to the lexical level, but need not flow from the lexicon back to the lower level. There is in fact no benefit for word recognition to be had from lexical–prelexical feedback (Norris, McQueen, & Cutler, 2000a). The best that the lexical level can do is select the words that best match the information in the current utterance. Feedback offers no improvement on this performance, since all it does is cause the prelexical level to agree with the decision that has already been reached at the lexical level.

There is, however, a considerable body of evidence suggesting that listeners use lexical knowledge in tasks that require explicit phonemic judgments. For example, in phonetic categorization tasks, where listeners have to identify ambiguous phonemes, there are lexical biases: listeners are

more likely to identify a sound between [b] and [p] as [b] in a *beef–peef* context than in a *beace–peace* context (Ganong, 1980). Although such effects are variable (i.e. are sometimes present and sometimes absent), they have now been replicated many times under a variety of different conditions (Burton, Baum, & Blumstein, 1989; Connine, 1990; Connine & Clifton, 1987; Fox, 1984; McQueen, 1991; Miller & Dexter, 1988; Pitt, 1995; Pitt & Samuel, 1993). The phonemic restoration illusion (Warren, 1970) also shows lexical biases (it is harder to distinguish between a sound that has had noise added to it and a sound that has been replaced with a noise when the sound is in a real word than when it is in a nonsense word: Samuel, 1981, 1987). These lexical biases are also variable (Samuel, 1996). Furthermore, target phonemes can be identified faster in real words than in nonwords, but these effects again come and go (Cutler, Mehler, Norris, & Seguí, 1987; Eimas, Marcovitz Hornstein, & Payton, 1990; Eimas & Nygaard, 1992; Frauenfelder, Seguí, & Dijkstra, 1990; Pitt & Samuel, 1995; Rubin, Turvey, & Van Gelder, 1976; Wurm & Samuel, 1997). Lexical biases in phonemic tasks can also be found in nonwords (Connine et al., 1997; Frauenfelder et al., 2001; Marslen-Wilson & Warren, 1994; McQueen et al., 1999; Newman, Sawusch, & Luce, 1997). As was mentioned earlier in the chapter, for example, Connine et al. found that phoneme monitoring latencies were fastest for targets in real words (e.g. /t/ in *cabinet*), slower for targets in the nonwords that were very similar to the real words (*gabinet*), and slower still for targets in nonwords with greater mismatch (*mabinet*).

One way to explain these results is to postulate feedback from the lexicon to the prelexical level (McClelland & Elman, 1986). On an interactive account such as this, decisions about speech sounds are based on the activation of prelexical representations. Feedback from the lexicon would boost the activation of lexically consistent sounds in words, thus speeding responses to those sounds relative to sounds in nonwords. The other lexical effects in phonemic decision-making could be explained in the same way, including the lexical biases in word-like nonwords (where partially activated word representations feed activation back to prelexical representations). An alternative account, however, is that phoneme decisions are based on the activation of decision units, which receive input from both the prelexical and lexical levels (Norris et al., 2000a). According to this modular account, there is no feedback: information flows forward from the prelexical level to the lexical level and forward to the decision level, and from the lexical level forward to the decision level. The activation of units at the decision level would nonetheless be biased by lexical activation when words or word-like nonwords are heard, in much the same way as in the interactive account, but without feedback.

It is difficult to distinguish between these two alternative accounts (see Norris et al., 2000a, 2000b, and accompanying commentaries for extended discussion). Nevertheless, there are several reasons to prefer the feedforward account. The first reason has already been mentioned: feedback offers no benefit to spoken word recognition. Feedback would therefore be postulated simply to explain the data on lexical involvement in phonemic decision-making. But since these data can be explained without feedback, there is no need to include feedback in an account of speech perception. One might want to argue, however, that since phoneme decision units are not a necessary part of spoken word recognition, their addition is also only motivated by a need to explain the data on lexical effects. In other words, is the postulation of decision units any more parsimonious than the postulation of feedback? But, since listeners can make explicit phonemic judgments, the machinery for this ability must be included in any model of speech perception, whether it has feedback or not. As Norris et al. (2000b) argued, a model with feedback requires all of the components for phonemic decision-making that a model without feedback has, with the addition of the feedback connections themselves. A model without feedback, especially since feedback has no benefit to offer in normal word recognition, should therefore be preferred.

There is also empirical evidence that challenges the feedback account. In the earlier discussion of compensation for coarticulation, I pointed out that Pitt and McQueen (1998) have shown that lexical and compensatory effects can dissociate. Listeners tend to label more sounds on a [t]–[k] stop continuum as [t] when the sounds are presented after a word ending with the fricative [ʃ], such as *foolish*, than after a [s]-final word such as *christmas* (Elman & McClelland, 1988; Mann & Repp, 1981; Pitt & McQueen 1998). As argued above, this compensation for coarticulation process has a prelexical locus. Demonstrations of this compensation effect when word-final fricatives were replaced with an ambiguous sound (e.g. *foolish* becoming [fulɪ?]; *christmas* becoming [krɪsmə?]) have thus been taken as evidence for feedback (the lexicon providing the missing fricative information, and thus inducing the prelexical compensation process: Elman & McClelland, 1988).

Pitt and McQueen (1998) have shown, however, that there was a confound in the Elman and McClelland (1988) study between lexical bias and the phoneme transition probability from the final vowels in the context words to the final fricatives. When these transition probabilities were controlled, lexical biases still determined decisions about the ambiguous fricatives, but those decisions had no further consequences for the processing of the following stops. The dissociation between the lexical effect and the compensation for coarticulation

effect suggests that they have different loci. If the lexical bias were due to feedback, it should have caused a compensation effect (a shift in the identification of the following stop). As mentioned above, dissociations between lexical biases and another low-level adjustment process, that of speech rate normalization, have also been observed (Miller & Dexter, 1988; Miller et al., 1984). These dissociations challenge the feedback account, and support the feedforward account, where the perceptual effects have a prelexical locus, and the lexical effects are due to biases acting at the decision stage.

It would be misleading to suggest that the debate on feedback in speech perception is resolved. Recent papers have presented evidence that appears to challenge the feedforward account of lexical effects (Samuel, 1997, 2001). Interestingly, however, these experiments involved a learning component (i.e. selective adaptation effects, in which listeners' perception of speech sounds was changed over time as a result of repeated exposure to particular sounds: Eimas & Corbit, 1973; Samuel, 1986). It is possible that lexical knowledge could influence perceptual learning over time, without having any effect on the on-line processing of any particular utterance. It has in fact recently been shown that listeners do use their lexical knowledge to adjust their perceptual categories over time as they listen to unfamiliar speech sounds (McQueen, Norris, & Cutler, 2001; Norris, McQueen, & Cutler, 2003). The lexicon may thus influence perceptual learning at the prelexical level but not on-line prelexical processing.

Summary

Speech-sound perception is in many ways more controversial than spoken word perception. While there is considerable consensus on many aspects of the word recognition process, there is considerable disagreement about how speech sounds are perceived. Thus, while I have argued that there is an intermediate level of processing between the output of the auditory system and the mental lexicon, I have also had to point out that there is no agreement about the nature of the representations at this level of processing, nor indeed agreement that such a level of processing must exist. It is clear, however, that if there is a prelexical stage of processing, activation must cascade continuously through this stage, from the auditory level to the lexicon. I have also argued that there is no feedback of information from the lexicon to the prelexical stage, except perhaps where feedback acts as an error-correcting signal in perceptual learning. The claim that there is no on-line feedback in the perceptual system comes, however, with an important additional claim: that listeners' judgments about speech sounds are not based directly on the output of the prelexical processing stage, but instead on separate decision units.

Models of Speech Decoding

In conclusion, I turn to models of speech decoding. Table 11.1 lists a number of these models, together with their major features. Note, however, that there may be important differences even between models that appear to share a particular assumption. For example, although both TRACE and ARTWORD have both featural and phonemic representations that mediate between the speech signal and the lexicon, the nature of processing in these two models is radically different. The table thus omits a large number of fundamental differences between the models. I make no attempt to provide full descriptions of the models (many, but not all, of which have computational implementations). The references on each model provide those details. Nor do I offer a detailed evaluation of each model. Instead, my aim is to provide a recapitulation of the major themes of this chapter in the context of these models.

In the listing of word-form representations, the term 'logogen' is used. This term derives from Morton's (1969) seminal logogen model of word recognition. A logogen is a mental representation that accrues evidence, coded in terms of an activation value, for a particular word. All of the models, with the exception of the Lexical Access From Spectra (LAFS) model, MINERVA 2, the Distributed Cohort Model (DCM) and the Lexical Access From Features (LAFF) model, could thus be said to have logogens of one kind or another, since they all have abstract localist word representations with activation levels that code support for candidate words. The lexical representations in the Cohort and Shortlist models are not listed as logogens, however, since additional claims have been made about the phonological content of word representations in these two models.

As can be seen from Table 11.1, all the models assume that multiple words are simultaneously activated when a speech signal is heard, but only four assume that there is direct competition between those candidate words. In each of these four models, there are inhibitory connections between word representations, and thus direct competition between words. In two of the other models, however, competition is included as a decision-level process (i.e. in the Cohort model and the Neighborhood Activation Model, NAM). In these models, decisions about which word has been heard are based on computations of the degree of support for a given word relative to the degree of support for other candidates. Thus, while there is no inter-word competition in these two models, the number of alternative candidate words, and their degree of support, influences lexical identification. Furthermore, although there are no inter-word connections in the DCM (note that there can be no such connections in a network with fully distributed

TABLE 11.1 *Models of spoken word perception*

Model	Primary references	Prelexical representations	Word-form representations	Multiple activation of lexical hypotheses	Direct inter-word competition	Feedforward prelexical–lexical cascade	On-line lexical–prelexical feedback
Lexical Access From Spectra (LAFS)	Klatt (1979, 1989)	None	Spectral templates	Yes	No	Not applicable	Not applicable
MINERVA 2	Hintzman (1986); Goldinger (1998)	None	Episodic traces	Yes	No	Not applicable	Not applicable
Lexical Access From Features (LAFF)	Stevens (2002)	Features and segments	Features and segments	Yes	No	No	No
Cohort	Marslen-Wilson & Welsh (1978); Marslen-Wilson (1987, 1993); Lahiri & Marslen-Wilson (1991)	Features	Underspecified phonological structures	Yes	No	Yes	No
Distributed Cohort Model (DCM)	Gaskell & Marslen-Wilson (1997, 1999)	Features	Fully distributed vectors in a recurrent network	Yes	No	Yes	No
Neighborhood Activation Model (NAM)	Luce (1986); Luce & Pisoni (1998)	Acoustic-phonetic patterns	Logogens	Yes	No	Yes	No
PARSYN	Luce, Goldinger, Auer, & Vitevitch (2000)	Allophones	Logogens	Yes	Yes	Yes	Yes
TRACE	Elman & McClelland (1986); McClelland & Elman (1986); McClelland (1991)	Features and phonemes	Logogens	Yes	Yes	Yes	Yes
Shortlist	Norris (1994); Norris, McQueen, Cutler, & Butterfield (1997)	Phonemes	Phoneme strings	Yes	Yes	No	No
ARTWORD	Grossberg, Boardman, & Cohen (1997); Grossberg & Myers (2000)	Features and phonemes	Logogens	Yes	Yes	Yes	Yes

representations), there is nonetheless a form of competition between candidate words as the network settles into a stable state (i.e. the performance of the model is sensitive to the number of words activated by a particular input, and the similarities among those words). Although lexical competition has not been addressed directly in the remaining models (LAFS, MINERVA 2, LAFF), the lexical selection processes in these models are also likely to be affected by the number and similarity of activated lexical hypotheses.

Thus, while only some models have direct inter-word competition, all models have some form of competitive process. Direct inter-word competition may be preferred because it provides an efficient means by which continuous speech can be segmented into words (Grossberg & Myers, 2000; McQueen et al., 1995). Note also that, of these ten models, only Shortlist includes a mechanism that uses cues to likely word boundaries in continuous speech to modulate the competition process and thus improve lexical segmentation (Norris et al., 1997).

Of the eight models with prelexical representations, only the LAFF model and Shortlist have no cascaded processing from the prelexical level to the lexical level. Little has been said about the temporal dynamics of lexical access in the LAFF model (Stevens, 2002); whether the model can account for subphonemic effects on lexical activation will depend on specification of the time course of the spread of information from the prelexical featural representations to the lexicon. In the Norris (1994) implementation of Shortlist, processing evolves on a phoneme-by-phoneme basis, with categorical phonemic representations at the prelexical level acting as the input to lexical processing. It is important to note, however, that this implementation was considered to be an approximation of a continuous lexical access process (Norris, 1994). Indeed, more recent simulations of lexical access in Shortlist involve cascaded processing (Norris et al., 2000a). Shortlist, like the other models with prelexical–lexical cascade, may therefore be able to account for the evidence of continuous lexical access, including at least some of the recent results showing the influence of fine-grained acoustic detail on lexical activation.

Three models (TRACE, PARSYN and ARTWORD) have on-line feedback from the lexicon to prior stages of processing. The challenge posed by Norris et al. (2000a) to theorists postulating this kind of feedback was that the value of feedback for speech perception needs to be shown. As far as I am aware, this challenge has not yet been met. It is possible, however, that lexical feedback could be required for perceptual learning (Grossberg, 2000; Norris, 1993; Norris et al., 2000b, 2003). This kind of feedback would act off-line, to alter prelexical representations over time, but would in principle not influence on-line perception of a particular utterance. It is possible, however, that feedback for perceptual learning could produce effects that might appear to be the result of on-line feedback. It will be important to establish (not only in this domain but also in other areas of speech perception) the extent to which particular effects reflect learning (the dynamic adjustments of the perceptual system over time) and the extent to which effects reflect the steady-state operation of the system.

This review has provided a brief introduction to models of spoken word perception by setting them in the context of the issues that were raised in the main part of this chapter. A more thorough evaluation of these models would require a much more detailed analysis of the detailed assumptions of each model. Nevertheless, it should be clear that while there is considerable agreement on several of the major assumptions of the models, there are also many fundamental disagreements. There is therefore still plenty work to be done. Progress will be made in this field, as in other fields of cognitive science, through the continuing interplay between the development of computationally explicit models and simulation results on the one hand, and the results of experimental investigations with human listeners on the other.

One final point is that Table 11.1 is incomplete. The models listed in the table have all explicitly addressed spoken word perception. There are many other theories of speech perception, which have been concerned primarily with speech-sound perception. Such accounts include: the motor theory of speech perception (Liberman & Mattingly, 1985); a direct-realist theory of speech perception (Fowler, 1986, 1996); pattern-classification models (Nearey, 1992, 1997; Smits, 2001b); the Merge model of phonemic decision-making (Norris et al., 2000a); and the Fuzzy Logical Model of Perception (FLMP: Massaro, 1987, 1989, 1997). Lexical factors do play a role in some of these models (e.g. Merge is explicitly linked to the Shortlist account of word recognition; FLMP has been used to examine lexical effects in phonemic decision-making: Massaro & Oden, 1995), but even in these cases the emphasis has been on the perception of speech sounds. In most of these theories, no account of word perception is provided.

These accounts of speech-sound perception have been, and will continue to be, critical to the development of our understanding of speech decoding (e.g. in the debate on 'units of perception', or in specifying how the acoustic-phonetic information that specifies particular sounds is integrated). In my view, however, such accounts will be incomplete until they specify how spoken words are perceived. In particular, it will be necessary to make explicit the extent to which the processes that are postulated for speech-sound perception and those for word perception are shared or distinct, and, if distinct, how they are related. This complaint can of course

also apply to models of word perception, where the details of prelexical processing and speech-sound perception have often been evaded. The ideal, of course, is a model that accounts for the perception of both the words and the individual sounds of speech (and indeed the prosodic and intonational structure of spoken utterances). I believe that word perception will nevertheless remain as the central element in such a model of speech decoding, both because words are the primary objects of speech perception and because words are the key to the speech code.

Acknowledgements

I thank my colleagues Anne Cutler, Delphine Dahan, Dennis Norris and Roel Smits for comments on a previous version of this chapter, and Tau van Dijck for assistance with Figure 11.1.

Note

1 Many experimental tasks used in speech perception and spoken word recognition are described in a special issue of the journal *Language and Cognitive Processes, 11* (1996), which also appears as Grosjean and Frauenfelder (1996). Please see those descriptions for details on the tasks described in this chapter.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

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

Beckman, M. E., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 152–178). Cambridge: Cambridge University Press.

Burton, M. W., Baum, S. R., & Blumstein, S. E. (1989). Lexical effects on the phonetic categorization of speech: the role of acoustic structure. *Journal of Experimental Psychology*: *Human Perception and Performance, 15*, 567–575.

Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, *29*, 155–190.

Christie, W. M. (1974). Some cues for syllable juncture perception in English. *Journal of the Acoustical Society of America*, *55*, 819–821.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 521–533.

Church, K. W. (1987). Phonological parsing and lexical retrieval. *Cognition*, *25*, 53–69.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 551–563.

Coenen, E., Zwitserlood, P., & Boelte, J. (2001). Variation and assimilation in German: consequences for lexical access and representation. *Language and Cognitive Processes*, *16*, 535–564.

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: the role of sentence context. *Journal of Memory and Language*, *26*, 527–538.

Connine, C. M. (1990). Effects of sentence context and lexical knowledge during speech processing. In G. T. M. Altmann (ed.), *Cognitive models of speech processing: psycholinguistic and computational perspectives* (pp. 281–294). Cambridge, MA: MIT Press.

Connine, C. M., Blasko, D., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: temporal and linguistic constraints. *Journal of Memory and Language*, *30*, 234–250.

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.

Connine, C. M., & Clifton, C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 291–299.

Connine, C. M., Titone, D., Deelman, T., & Blasko, D. (1997). Similarity mapping in spoken word recognition. *Journal of Memory and Language*, *37*, 463–480.

Content, A., Meunier, C., Kearns, R. K., & Frauenfelder, U. H. (2001). Sequence detection in pseudowords in French: where is the syllable effect? *Language and Cognitive Processes*, *16*, 609–636.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: evidence from native and nonnative listeners. *Language and Speech*, *45*, 207–228.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: evidence from juncture misperception. *Journal of Memory and Language*, *31*, 218–236.

Cutler, A., & Clifton, C. (1999). Comprehending spoken language: a blueprint of the listener. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 123–166). Oxford: Oxford University Press.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech*, *40*, 141–201.

Cutler, A., Demuth, K., & McQueen, J. M. (2002). Universality versus language-specificity in listening to running speech. *Psychological Science*, *13*, 258–262.

Cutler, A., & van Donselaar, W. (2001). Voornaam is not a homophone: lexical prosody and lexical access in Dutch. *Language and Speech*, *44*, 171–195.

Cutler, A., Mehler, J., Norris, D., & Seguí, 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., & Seguí, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, *19*, 141–177.

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

Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, *33*, 824–844.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: evidence for lexical competition. *Language and Cognitive Processes*, *16*, 507–534.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden-path: segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 218–244.

Diehl, R. L., & Kluender, K. R. (1987). On the categorization of speech sounds. In S. R. Harnad (ed.), *Categorical perception* (pp. 226–253). Cambridge: Cambridge University Press.

Dumay, N., Frauenfelder, U. H., & Content, A. (2002). The role of the syllable in lexical segmentation in French: word-spotting data. *Brain and Language*, *81*, 144–161.

Dupoux, E., Pallier, C., Sebastián-Gallés, N., & Mehler, J. (1997). A destressing deafness in French. *Journal of Memory and Language*, *36*, 399–421.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*, 99–109.

Eimas, P. D., Marcovitz Hornstein, S. B., & Payton, P. (1990). Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language*, *29*, 160–180.

Eimas, P. D., & Nygaard, L. C. (1992). Contextual coherence and attention in phoneme monitoring. *Journal of Memory and Language*, *31*, 375–395.

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 of speech processes* (pp. 360–380). 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.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception and Psychophysics*, *27*, 343–350.

Foss, D. J., & Blank, M. A. (1980). Identifying the speech codes. *Cognitive Psychology*, *12*, 1–31.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, *29*, 109–135.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728–3740.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics*, *36*, 359–368.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3–28.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1730–1741.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 877–888.

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 526–540.

Frauenfelder, U. H., Scholten, M., & Content, A. (2001). Bottom-up inhibition in lexical selection: phonological mismatch effects in spoken word recognition. *Language and Cognitive Processes*, *16*, 583–607.

Frauenfelder, U. H., Seguí, J., & Dijkstra, T. (1990). Lexical effects in phonemic processing: facilitatory or inhibitory? *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 77–91.

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

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 144–158.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613–656.

Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 380–396.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science*, *23*, 439–462.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical ambiguity resolution and spoken word recognition: bridging the gap. *Journal of Memory and Language*, *44*, 325–349.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., & Marcario, J. K. (1992). Form-based priming in spoken word recognition: the roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1211–1238.

Gow, D. W. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*, 133–159.

Gow, D. W. (2002). Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 163–179.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 344–359.

Grosjean, F., & Frauenfelder, U. H. (Eds.) (1996). *A guide to spoken word recognition paradigms*. Hove, UK: Psychology Press.

Grossberg, S. (2000). Brain feedback and adaptive resonance in speech perception. *Behavioral and Brain Sciences*, *23*, 332–333.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 481–503.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: interword integration and duration-dependent backward effects. *Psychological Review*, *107*, 735–767.

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Jakobson, R., Fant, C. G. M., & Halle, M. (1952). *Preliminaries to speech analysis: the distinctive features and their correlates*. Cambridge, MA: MIT Press.

Klatt, D. H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. D. Marslen-Wilson (ed.), *Lexical representation and process* (pp. 169–226). Cambridge, MA: MIT Press.

Kluender, K. (1994). Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher (ed.), *Handbook of psycholinguistics* (pp. 173–214). San Diego: Academic Press.

Ladefoged, P. (2001). *A course in phonetics* (4th ed.). New York: Harcourt Brace Jovanovich.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford: Blackwell.

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

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica*, *5* (Suppl. 5), 1–54.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, *51*, 2018–2024.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

Lotto, A., & Kluender, K. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Perception and Psychophysics*, *60*, 602–619.

Lotto, A., Kluender, K., & Holt, L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America*, *102*, 1134–1140.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon (PhD dissertation, Indiana University). In Research on Speech Perception, Technical Report No. 6, Speech Research Laboratory, Department of Psychology, Indiana University.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception and Psychophysics*, *62*, 615–625.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, *16*, 565–581.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the Neighborhood Activation Model. *Ear and Hearing*, *19*, 1–36.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [S]-[s] distinction. *Perception and Psychophysics*, *28*, 213–228.

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.

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

Marslen-Wilson, W. D. (1993). Issues of process and representation in lexical access. In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The Second Sperlonga Meeting* (pp. 187–210). Hillsdale, NJ: Erlbaum.

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

Marslen-Wilson, W. D., Nix, A., & Gaskell, M. G. (1995). Phonological variation in lexical access: abstractness, inference and English place assimilation. *Language and Cognitive Processes*, *10*, 285–308.

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

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

Massaro, D. W. (1987). *Speech perception by ear and eye: a paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989). Testing between the TRACE model and the Fuzzy Logical Model of Speech Perception. *Cognitive Psychology*, *21*, 398–421.

Massaro, D. W. (1997). *Perceiving talking faces: from speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Oden, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *21*, 1053–1064.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, *23*, 1–44.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *10*, 1–86.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 433–443.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*, 21–46.

McQueen, J. M., & Cutler, A. (1998a). Spotting (different types of) words in (different types of) context. *Proceedings of the 5th International Conference on Spoken Language Processing* (Vol. 6, pp. 2791–2794). Sydney: Australian Speech Science and Technology Association.

McQueen, J. M., & Cutler, A. (1998b). Morphology in word recognition. In A. Spencer & A. M. Zwicky (Eds.), *The handbook of morphology* (pp. 406–427). Oxford: Blackwell.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes*, *10*, 309–331.

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

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1363–1389.

McQueen, J. M., Norris, D., & Cutler, A. (2001). Can lexical knowledge modulate prelexical representations over time? In R. Smits, J. Kingston, T. M. Nearey, & R. Zondervan (Eds.), *Proceedings of the SPRAAC Workshop* (pp. 9–14). Nijmegen: MPI for Psycholinguistics.

McQueen, J. M., Otake, T., & Cutler, A. (2001). Rhythmic cues and possible-word constraints in Japanese speech segmentation. *Journal of Memory and Language*, *45*, 103–132.

Mehler, J. (1981). The role of syllables in speech processing: infant and adult data. *Philosophical Transactions of the Royal Society of London B*, *295*, 333–352.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39–74). Hillsdale, NJ: Erlbaum.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 369–378.

Miller, J. L., Green, K., & Schermer, T. (1984). On the distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception and Psychophysics*, *36*, 329–337.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics*, *25*, 457–465.

Monsell, S., & Hirsh, K. W. (1998). Competitor priming in spoken word recognition. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *24*, 1495–1520.

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

Moss, H. E., McCormick, S. F., & Tyler, L. K. (1997). The time course of activation of semantic information during spoken word recognition. *Language and Cognitive Processes*, *10*, 121–136.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, *47*, 379–390.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*, 365–378.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, *62*, 714–719.

Nearey, T. M. (1992). Context effects in a double-weak theory of speech perception. *Language and Speech*, *35*, 153–171.

Nearey, T. M. (1997). Speech perception as pattern classification. *Journal of the Acoustical Society of America*, *101*, 3241–3254.

Nearey, T. M. (2001). Phoneme-like units and speech perception. *Language and Cognitive Processes*, *16*, 673–681.

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 873–889.

Norris, D. (1993). Bottom-up connectionist models of 'interaction'. In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The Second Sperlonga Meeting* (pp. 211–234). Hillsdale, NJ: Erlbaum.

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

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *21*, 1209–1228.

Norris, D., McQueen, J. M., & Cutler, A. (2000a). Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–325.

Norris, D., McQueen, J. M., & Cutler, A. (2000b). Feedback on feedback on feedback: it's feedforward. *Behavioral and Brain Sciences*, *23*, 352–370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, *34*, 191–243.

Norris, D., McQueen, J. M., Cutler, A., Butterfield, S., & Kearns, R. (2001). Language-universal constraints on speech segmentation. *Language and Cognitive Processes*, *16*, 637–660.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, *54*, 1235–1247.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, *32*, 358–378.

Pallier, C., Sebastián-Gallés, N., Felguera, T., Christophe, A., & Mehler, J. (1993). Attentional allocation within the syllable structure of spoken words. *Journal of Memory and Language*, *32*, 373–389.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *19*, 309–328.

Pastore, R. E. (1981). Possible acoustic factors in speech perception. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 165–205). Hillsdale, NJ: Erlbaum.

Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *21*, 1037–1052.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, *39*, 347–370.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 699–725.

Pitt, M. A., & Samuel, A. G. (1995). Lexical and sublexical feedback in auditory word recognition. *Cognitive Psychology*, *29*, 149–188.

Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics*, *20*, 331–350.

Quené, H. (1993). Segment durations and accent as cues to word segmentation in Dutch. *Journal of the Acoustical Society of America*, *94*, 2027–2035.

Repp, B. H. (1982). Phonetic trading relations and context effects: new evidence for a phonetic mode of perception. *Psychological Bulletin*, *92*, 81–110.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. R. Harnad (ed.), *Categorical perception* (pp. 89–112). Cambridge: Cambridge University Press.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 621–637.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: semantic competition in lexical access. *Journal of Memory and Language*, *46*, 245–266.

Rubin, P., Turvey, M. T., & Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in non-words. *Perception and Psychophysics*, *19*, 394–398.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*, 51–89.

Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General*, *110*, 474–494.

Samuel, A. G. (1986). Red herring detectors and speech perception: in defense of selective adaptation. *Cognitive Psychology*, *18*, 452–499.

Samuel, A. G. (1987). Lexical uniqueness effects on phonemic restoration. *Journal of Memory and Language*, *26*, 36–56.

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General*, *125*, 28–51.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology*, *32*, 97–127.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*, 348–351.

Schacter, D. L., & Church, B. A. (1992). Auditory priming: implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *18*, 521–533.

Sebastián-Gallés, N., Dupoux, E., Seguí, J., & Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language*, *31*, 18–32.

Shillcock, R. C. (1990). Lexical hypotheses in continuous speech. In G. T. M. Altmann (ed.), *Cognitive models of speech processing: psycholinguistic and computational perspectives* (pp. 24–49). Cambridge, MA: MIT Press.

Slowiaczek, L. M., & Hamburger, M. B. (1992). Prelexical facilitation and lexical interference in auditory word recognition. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *18*, 1239–1250.

Smits, R. (2001a). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 1145–1162.

Smits, R. (2001b). Hierarchical categorization of coarticulated phonemes: a theoretical analysis. *Perception and Psychophysics*, *63*, 1109–1139.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in

lexical access. *Journal of Memory and Language*, *45*, 412–432.

Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language*, *48*, 233–254.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891.

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

Tabossi, P., Burani, C., & Scott, D. (1995). Word identification in fluent speech. *Journal of Memory and Language*, *34*, 440–467.

Tabossi, P., Collina, S., Mazzetti, M., & Zoppello, M. (2000). Syllables in the processing of spoken Italian. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 758–775.

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, *28*, 397–440.

Utman, J. A., Blumstein, S. E., & Burton, M. W. (2000). Effects of subphonetic and syllable structure variation on word recognition. *Perception and Psychophysics*, *62*, 1297–1311.

van Alphen, P., & McQueen, J. M. (2001). The time-limited influence of sentential context on function word identification. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 1057–1071.

van den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience*, *13*, 967–985.

van der Lugt, A. H. (2001). The use of sequential probabilities in the segmentation of speech. *Perception and Psychophysics*, *63*, 811–823.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *25*, 394–417.

Vitevitch, M. S. (2002). Influence of onset density on spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 270–278.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: levels of processing in spoken word recognition. *Psychological Science*, *9*, 325–329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374–408.

Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 98–108.

Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 710–720.

Vroomen, J., van Zon, M., & de Gelder, B. (1996). Cues to speech segmentation: evidence from juncture misperceptions and word spotting. *Memory and Cognition*, *24*, 744–755.

Wallace, W. P., Stewart, M. T., & Malone, C. P. (1995). Recognition memory errors produced by implicit activation of word candidates during the processing of spoken words. *Journal of Memory and Language*, *34*, 417–439.

Wallace, W. P., Stewart, M. T., Shaffer, T. R., & Wilson, J. A. (1998). Are false recognitions influenced by prerecognition processing? *Journal of Experimental Psychology: Learning*, *Memory*, *and Cognition*, *24*, 299–315.

Wallace, W. P., Stewart, M. T., Sherman, H. L., & Mellor, M. (1995). False positives in recognition memory produced by cohort activation. *Cognition*, *55*, 85–113.

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

Warren, P., & Marslen-Wilson, W. (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.

Weber, A. (2001). Language-specific listening: the case of phonetic sequences. Doctoral dissertation, University of Nijmegen (MPI Series in Psycholinguistics, 16).

Wells, J. C. (1987). Computer-aided phonetic transcription. *Journal of the International Phonetic Association*, *17*, 94–114.

Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. *Journal of the Acoustical Society of America*, *69*, 275–282.

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

Whalen, D. H. (1991). Subcategorical phonetic mismatches and lexical access. *Perception and Psychophysics*, *50*, 351–360.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, *76*, 1–15.

Wurm, L. H., & Samuel, A. G. (1997). Lexical inhibition and attentional allocation during speech perception: evidence from phoneme monitoring. *Journal of Memory and Language*, *36*, 165–187.

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

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