Lexical Influence in Phonetic Decision Making: Evidence From Subcategorical Mismatches

James M. McQueen
Max Planck Institute for Psycholinguistics

Dennis Norris
Medical Research Council Cognition
and Brain Sciences Unit

Anne Cutler
Max Planck Institute for Psycholinguistics

In 5 experiments, listeners heard words and nonwords, some cross-spliced so that they contained acoustic–phonetic mismatches. Performance was worse on mismatching than on matching items. Words cross-spliced with words and words cross-spliced with nonwords produced parallel results. However, in lexical decision and 1 of 3 phonetic decision experiments, performance on nonwords cross-spliced with words was poorer than on nonwords cross-spliced with nonwords. A gating study confirmed that there were misleading coarticulatory cues in the cross-spliced items; a sixth experiment showed that the earlier results were not due to interitem differences in the strength of these cues. Three models of phonetic decision making (the Race model, the TRACE model, and a postlexical model) did not explain the data. A new bottom-up model is outlined that accounts for the findings in terms of lexical involvement at a dedicated decision-making stage.

How do listeners use lexical knowledge when they make phonetic decisions about spoken language, such as when they categorize phonemes or detect phonetic targets? Although there is no dispute that lexical information can, at least under certain circumstances, influence phonetic decision making, there is considerable theoretical disagreement about how this influence comes into play. At least three different models of phonetic decision making have been proposed; each makes different claims about the flow of information through the recognition system and the way in which listeners make phonetic decisions. In this article we discuss these three alternative accounts. Data from six experiments and from other published research lead us to conclude that all three of these models are inadequate. We conclude by outlining a new model of phonetic decision making (Norris, McQueen, & Cutler, in press), which is able to account for all the available data.

The first model we consider is the *Race model* (Cutler & Norris, 1979; Cutler, Mehler, Norris, & Seguí, 1987). Levels of processing in the Race model are autonomous. Lexical information is unable to influence the processes that operate

prelexically because the flow of information is only bottom-up. Phonetic decisions are based on a race between two procedures; one decision procedure is based on the prelexical analysis of the speech signal, whereas the other is based on phonological information stored in the lexicon. Whichever procedure completes first is responsible for the phonetic decision on that trial. This model is able to account for the lexical effects that have been observed in a number of phonetic judgment tasks. In phoneme monitoring, listeners are faster to detect targets in words than targets in nonwords (Cutler et al., 1987; Foss & Blank, 1980; Foss & Gernsbacher, 1983; Rubin, Turvey, & van Gelder, 1976). The Race model explains this as follows: Decisions to targets in words are based on two routes, whereas decisions to targets in nonwords are all based on the prelexical route. Because the mean completion time of two racing procedures with overlapping distributions of completion times is faster than the mean completion time of either one of those procedures, responses to targets in words tend to be faster than those to targets in nonwords.

The Race model is also consistent with a number of other findings from the phoneme monitoring task. Cutler et al. (1987) showed that the presence of lexical effects depended on task monotony. When all the fillers were monosyllabic, no lexical effects were observed, but when some of the fillers were bisyllabic, lexical effects were obtained. Eimas, Marcovitz Hornstein, and Payton (1990) further showed that lexical effects in a simple phonetic decision task were present only when listeners were also required to perform a secondary task that demanded lexical involvement (e.g., lexical decision). However, even this conditional lexical effect appears to depend on the nature of the context; it was obtained when the targets were in nonsense sentences but not when the targets were in meaningful sentences (Eimas &

Nygaard, 1992). The results of these studies suggest that lexical involvement in phonetic decision making depends heavily on task demands (see also Frauenfelder & Seguí, 1989; Marslen-Wilson, 1984; Pitt & Samuel, 1990; Seguí & Frauenfelder, 1986). When the experimental situation encourages listeners to use their lexical knowledge, lexical effects emerge in the phonetic task, but when use of lexical information is discouraged, the effects go away. This can be explained easily by the Race model: Listeners switch attention between the two routes (thus increasing or decreasing the extent of lexical involvement) according to task demands.

Frauenfelder, Seguí, and Dijkstra (1990) presented further experimental support for the Race model. They showed that lexical effects in phoneme monitoring for targets in long words depended on the position of the target phoneme in the word. There were no lexical effects for word-initial targets or for word-medial targets that appeared before the word's uniqueness point (i.e., the point at which, moving through the word, there is only one word the input could be), whereas there were lexical effects for targets that occurred after the uniqueness point. The Race model predicts that on longer words the lexical route will be unable to complete fast enough to contribute to reaction times (RTs) if the target phoneme is before the uniqueness point. Although Pitt and Samuel (1995) found that lexical effects in phoneme monitoring can be obtained on targets before the uniqueness point, the words in that study had higher frequency initial syllables than the matched nonwords. This result could thus be explained by the Race model if the prelexical level were sensitive to the sequential likelihoods of phonemes. Responses to targets in words would be faster than those to targets in nonwords not because of lexical involvement but because the prelexical level was sensitive to the fact that the targets in the words were more likely than those in the nonwords. Other research has shown that listeners are sensitive to the transitional probability of phonemes (McQueen & Pitt, 1996; Saffran, Newport, & Aslin, 1996) and that the mechanism underlying this sensitivity appears to have a prelexical locus (Pitt & McQueen, 1998; Vitevitch & Luce, 1998).

The Race model can also handle lexical effects on phonemic restoration (Samuel, 1981, 1987, 1996). Missing phonemes tend to be perceptually restored more frequently in words than in nonwords. In the Race model, this is because the lexical route can provide support for missing phonemes in words but not in nonwords. Furthermore, the Race model can account for lexical effects in the categorization of ambiguous phonemes (Burton, Baum, & Blumstein, 1989; Connine & Clifton, 1987; Fox, 1984; Ganong, 1980; McQueen, 1991; Miller & Dexter, 1988; Pitt, 1995; Pitt & Samuel, 1993). Ambiguous phonemes, midway along a continuum between two unambiguous endpoints (e.g., a voicing continuum between [b] and [p]) and placed in contexts in which one endpoint is a word and the other a nonword, tend to be categorized in a lexically consistent manner (e.g., more [b] responses to the continuum *beef–peef* and more [p] responses to the continuum *beace–peace*). In the Race model, this is because the proportion of lexically

consistent responses in a given continuum will increase every time the lexical route wins the race.

One result from phonetic categorization appeared problematic for the Race model. This was Elman and McClelland's (1988) demonstration of apparent lexical involvement in compensation for coarticulation. Listeners tend to label an ambiguous sound between [t] and [k] as [k] after [s] and as [t] after [ʃ] (Mann & Repp, 1981); the perceptual system appears to compensate for the acoustic consequences of the coarticulation of fricatives and stops. Elman and McClelland showed that this effect could be obtained both after unambiguous word-final fricatives (e.g., the [s] in *christmas* and the [ʃ] in *foolish*) and after an ambiguous word-final fricative [?] in the same contexts (listeners responded to the following stops as if they had heard [s] at the end of *christma?* and as if they had heard [ʃ] at the end of *fooli?*). Because one can assume that the mechanism responsible for compensation for coarticulation has a prelexical locus (it depends on low-level acoustic–phonetic information such as formant transition frequencies), this demonstration of apparent lexical involvement in the workings of a prelexical mechanism seems to be seriously problematic for the autonomous assumption in the Race model. This result might suggest that information does not flow only bottom-up but also top-down from the lexicon to prelexical levels of processing.

However, Elman and McClelland's (1988) materials contained a confound. As Cairns, Shillcock, Chater, and Levy (1995) found for a large corpus of English, [s] is more likely than [ʃ] after [ə] (as in all of the [s]-final words in Elman & McClelland, 1988), whereas [ʃ] is more likely than [s] after [ɪ] (as in all of the [ʃ]-final words). The results could thus be explained by an autonomous model (such as the Race model) if the prelexical level of processing in that model were sensitive to the transitional probabilities between phonemes (Cairns et al., 1995; Norris, 1993). The explanation for this effect is thus analogous to that offered for transitional probability effects found in other tasks. Consistent with this explanation, Pitt and McQueen (1998) recently found that the categorization of an ambiguous fricative between [s] and [ʃ] at the end of nonwords was influenced by the transitional probability of those fricatives and that the labeling of following stops was also influenced by the transitional likelihood of the fricatives. These findings suggest that the prelexical level of processing includes a mechanism for compensation for coarticulation and a mechanism that encodes the sequential likelihoods of phonemes.

Pitt and McQueen (1998) also found that when an ambiguous fricative was placed at the ends of words in which the transitional probabilities of [s] and [ʃ] were controlled, there were no lexical effects in compensation for coarticulation (i.e., there was no influence on categorization of following stops) even though there was a lexical effect in the labeling of the fricative. This result is exactly as would be predicted by the Race model, in which lexical knowledge can influence fricative decisions through the operation of the lexical route but cannot influence the prelexical compensation for coarticulation process because there is no top-down flow of information.

Although the Race model can therefore account for a wide range of data, there are nevertheless a number of recent findings that the model cannot explain. Connine, Titone, Deelman, and Blasko (1997) found that phoneme monitoring responses to targets in nonwords became faster as the nonwords became more wordlike. Thus, for example, responses to the [t] in *shuffinet* tended to be slower than those to the [t] in *mabinet* (which is more like *cabinet* but in which the initial [m] is maximally different from the [k]), and responses to the [t] in *mabinet* in turn tended to be slower than those to the [t] in *gabinet* (in which the [g] differs minimally from the [k] of *cabinet* only by one distinctive feature). Because in the Race model all phonetic decisions to nonwords must be made via the prelexical route, there should be no lexical involvement in those decisions.

The Race model is also challenged by phonetic decision data. Marslen-Wilson and Warren (1994) created nonwords by cross-splicing a word with a nonword and by cross-splicing two nonwords (see Table 1). For example, they created two cross-spliced versions of *smob* by splicing either the *smo* from the word *smog* or the *smo* from the nonword *smod* onto the *b* from *smob*. These nonwords both contained mismatching acoustic–phonetic information (i.e., *subcategorical mismatch;* Whalen, 1984), as the formant transitions in the vocalic portions signaled an upcoming [g] or [d] and thus were mismatched with the information in the final release burst, which signaled a [b]. Marslen-Wilson and Warren's study was based on the earlier finding (Whalen, 1984) that phonetic decisions to fricatives were slower (relative to conditions with no acoustic–phonetic mismatch) when vocalic formant transition information mismatched the frication noise and decisions to stops were slower when the vowel information mismatched release burst information.

Marslen-Wilson and Warren (1994) also observed this basic mismatch effect: Phonetic decisions to the final stops in cross-spliced items were harder than those to stops in items without mismatching information (e.g., a *smob* made by splicing the *smo* from *smob* onto the *b* from a different token of *smob*). Although the Race model can explain this mismatch effect, it would predict no differential effect of the lexical status of the cross-spliced portion on the phonetic decisions because all decisions to nonwords must, according to the model, be made via the prelexical route. Contrary to this prediction, Marslen-Wilson and Warren found that

performance on the nonwords made by cross-splicing with a word was poorer than that on nonwords made by cross-splicing with another nonword.

Both Connine et al. (1997) and Marslen-Wilson and Warren (1994) found lexical influences on phonetic decisions to nonwords. Newman, Sawusch, and Luce (1997) challenged the Race model in a similar way, but with data from the phonetic categorization task. In contrast to the studies described earlier in which word–nonword and nonword–word continua were compared, Newman et al. examined nonword–nonword continua, in which the nonwords at each continuum endpoint varied in their similarity to real words. Thus, for example, the continuum *gice–kice,* in which *gice* has more lexical neighbors than *kice,* was compared with the continuum *gipe–kipe,* in which the opposite endpoint, *kipe,* had the higher neighborhood density. Newman et al. found that there were more responses in the ambiguous region of the continuum consistent with the endpoint nonword with a denser lexical neighborhood (i.e., more [g] responses to *gice–kice* and more [k] responses to *gipe–kipe*). Again, the Race model predicts no lexical involvement in these nonword decisions.

Analyses of lexical neighborhood density by Frauenfelder, Baayen, Hellwig, and Schreuder (1993), however, found that transitional probability of phoneme sequences is strongly correlated with neighborhood density: Words with many neighbors contain more probable sequences. As a consequence, nonwords with many word neighbors will also be likely to contain more probable phoneme sequences. It is thus possible that the Newman et al. (1997) result reflects prelexical sensitivity to the transition probability of phoneme sequences rather than effects of lexical neighborhood density per se; Vitevitch and Luce (1998), as well as Pitt and McQueen (1998), found that phoneme transition probability does affect prelexical processing. The role of this factor in the Newman et al. result remains to be established. Nevertheless, apart from this finding, there is still a growing body of evidence against the Race model.

The second model we considered is the TRACE model (McClelland & Elman, 1986). TRACE differs from the Race model in many ways; three differences are central to the current argument. First, phonetic decisions are not made via a race. Instead, evidence is integrated at phoneme nodes and decisions are based on a comparison between the levels of

Table 1
Experimental Conditions

Item type	Notation	English example	Dutch example
Words		*job*	*sloop*
1. Word 1 + Word 1	W1W1	job + jo<u>b</u>	sloop + sloo<u>p</u>
2. Word 2 + Word 1	W2W1	j<u>og</u> + jo<u>b</u>	sloot + sloo<u>p</u>
3. Nonword 3 + Word 1	N3W1	j<u>od</u> + jo<u>b</u>	slook + sloo<u>p</u>
Nonwords		*smob*	*smep*
1. Nonword 1 + Nonword 1	N1N1	smob + smo<u>b</u>	smep + sme<u>p</u>
2. Word 2 + Nonword 1	W2N1	smog + smo<u>b</u>	smet + sme<u>p</u>
3. Nonword 3 + Nonword 1	N3N1	smod + smo<u>b</u>	smek + sme<u>p</u>

Note. Items were constructed by splicing together the underlined portions. The notations defined in this table are the same as those used by Marslen-Wilson and Warren (1994).

activation of those nodes. Second, these nodes provide the only means for making a phonetic decision (in contrast to the two routes in the Race model). Third, TRACE is an interactive model: Information flows both bottom-up from feature nodes to the phoneme nodes and then to word nodes and also top-down from the word nodes back to the phoneme nodes.

The top-down flow of information in TRACE from the lexicon to the phoneme nodes accounts for (a) faster responses to words than nonwords in phonetic decision and phoneme monitoring; (b) stronger restoration of missing phonemes in words than in nonwords; and (c) lexical biases in phonetic categorization. The model can therefore account for all the basic lexical effects. It can also account for some of the results that are problematic for the Race model. Lexical effects in nonwords such as those found by Connine et al. (1997) are easily accounted for: As a nonword becomes more wordlike, the degree of bottom-up activation of that word's lexical node will increase, as will the degree of top-down feedback to the phoneme node corresponding to the target. Thus responses to [t] in *gabinet,* for example, should, because of increased activation of the word node for *cabinet,* be faster than those to [t] in *mabinet* or *shuffinet.* TRACE could also account for the findings of Newman et al. (1997): Nonwords will partially activate their lexical neighbors, which in turn will send feedback to the phoneme nodes corresponding to their constituents, producing neighborhood density biases in the categorization of ambiguous phonemes.

Even though TRACE can account for many of the lexical effects in the literature, there are a number of results that are problematic for the model. One problem is the variability, attributable to task demands, of lexical effects in phoneme monitoring. As Cutler et al. (1987), Eimas et al. (1990), and Eimas and Nygaard (1992) argued, these effects can easily be explained by the Race model, in which attention, modulated by the demands of the experimental situation, shifts between the two routes. Those authors have pointed out that TRACE cannot account for these findings so parsimoniously; attention would have to act to switch on and off top-down flow of information. Because the top-down connections are intended to improve performance of the recognition system, why should it be possible to switch them off? Clearly TRACE can only account for these findings in an unsatisfactory manner.

Furthermore, Frauenfelder et al. (1990) tested and failed to confirm what they took to be a prediction of TRACE that phoneme monitoring should be harder on a target embedded in a nonword that differed from a real word on only that target (e.g., the [t] in the French nonword *vocabutaire,* close to the word *vocabulaire*) than on a control nonword (e.g., *socabutaire*). This prediction follows from the joint effects of top-down connections and inhibitory connections between phoneme nodes. The nonword *vocabutaire* should activate *vocabulaire,* resulting in top-down support for the phoneme [l], which would interfere with detection of the [t]. Wurm and Samuel (1997), in a replication of this study in English, also failed to find inhibition (comparing, e.g., responses to [t] in *vocabutary* and *socabutary*). Although the failure to find inhibition challenges the TRACE model, it

needs to be treated with some caution. First, it is a null effect. Second, despite the Frauenfelder et al. (1990) assumption, the prediction does in fact depend on the parameters used in TRACE. A model with exactly the same interactive architecture but different network parameters can account for these results (Norris et al., in press).

The TRACE model also faces three more serious problems, however. The first is that it is unable to account for the findings of Pitt and McQueen (1998). Although the original demonstration of apparent lexical involvement in compensation for coarticulation (Elman & McClelland, 1988) was taken to be strong evidence in favor of TRACE, the model cannot account for the more recent results. The model could account for transitional probability effects via top-down connections (i.e., the lexicon as a whole acting to indicate which segments are most likely in a given context). The model can therefore account for the finding that the transitional probabilities of [s] and [ʃ] in nonwords influenced both fricative identification, and, via compensation for coarticulation, stop identification. It can also account for the lexical effect in word-final fricative identification. As Pitt and McQueen (1998) argued, however, the model is unable to account for two dissociations in their data. One dissociation is the presence of lexical influence on word-final fricative identification when there was no lexical influence on the categorization of the following stops. If top-down connections are acting to bias fricative node activation, that should also trigger the compensation for coarticulation process. The second dissociation is that compensation for coarticulation on stops after ambiguous fricatives did not appear when the fricatives were on the end of words (without transition bias) but did appear when the fricatives were on the end of nonwords that were transitionally biased. Again, the top-down flow of information should produce both lexical and transition bias effects, not only one of them.

The second serious problem for TRACE is that it is unable to account for the distribution of lexical effects over time in word-final phonetic categorization (McQueen, 1991; Pitt & Samuel, 1993). The tendency toward lexically consistent responses in word final categorization (e.g., making more [ʃ] responses to ambiguous fricatives in the continuum *fish–fiss* and more [s] responses in the continuum *kish–kiss*) is largest in listeners' fastest responses and tends to disappear in their slower responses. This is problematic for TRACE because the model predicts that lexical effects should build up over time: As word nodes become more activated, the amount of top-down activation of phoneme nodes should also increase.

The third problem for TRACE is posed by the data of Marslen-Wilson and Warren (1994). Those authors showed, using computer simulations, that TRACE failed to account for their data in four different ways. First, the model predicts that phoneme targets in nonwords cross-spliced with words (e.g., the [b] in *smob* made from *smog* and *smob*) should barely become active (because of top-down activation of the lexically consistent phoneme, such as the [g] in this example). The model thus predicts much too strong a lexical effect (i.e., a very high error rate on the nonwords cross-spliced with words). Marslen-Wilson and Warren also used

their materials in a lexical decision task and observed a similar pattern in the cross-spliced nonwords to that which was observed in their phonetic decision data. TRACE's second problem with these data is that it again predicts much too strong a lexical effect on the nonwords cross-spliced with words: contrary to the data, the model predicts a high false-alarm rate on these nonwords (i.e., responding "yes," as if smob were the word smog).

The remaining two problems with Marslen-Wilson and Warren's (1994) data for TRACE concern its predictions for cross-spliced words. In addition to cross-splicing nonwords, Marslen-Wilson and Warren also cross-spliced words, again both with another word (e.g., job made from the initial consonant and vowel of jog plus the [b] from job) and with a nonword (e.g., job made from the initial portion of jod plus the [b] from job; see Table 1). These items were also presented for both lexical and phonetic decisions. In contrast to the nonwords, there were no differential effects of cross-splicing words with either words or nonwords. Performance on both tasks was equivalent for words cross-spliced with words and words cross-spliced with nonwords. TRACE, however, predicts that performance should be poorer in lexical decisions to the words cross-spliced with other words (because of activation of that other word) and that performance should be poorer in phonetic decisions to targets in these words. Note, however, that TRACE's problems with these data depend again on the network parameters of the model. Norris et al. (in press) found that an interactive model with the same architecture as TRACE but different parameters can provide a better simulation of Marslen-Wilson and Warren's (1994) data than the standard version of TRACE.

It would therefore appear that both TRACE and the Race model can be rejected because both models are unable to account for several different findings. The third model we consider is that offered by Marslen-Wilson and Warren (1994). They proposed that all phonetic decisions are based on output from the lexicon. Decisions to targets in words are assumed to depend on recognition of those words, whereas decisions to targets in nonwords are made by analogy to words activated by those nonwords. In this "postlexical model," there is thus no processing route by which a phonetic decision can be made that does not involve the lexicon. Given that all phonetic decisions are made via a lexical route, responses to targets in words should be faster than responses to targets in nonwords (the necessary phonological information can be retrieved faster from a single word, which would generate a stable representation, than from a more unstable representation based on a collection of words similar in sound to a nonword). The postlexical model can thus account for basic lexical effects in phoneme monitoring. It could offer a similar explanation for lexical effects in phonemic restoration and phonetic categorization.

In contrast to both TRACE and the Race model, the postlexical model, not surprisingly, can in principle also account for Marslen-Wilson and Warren's (1994) phonetic decision data (described above). According to Marslen-Wilson and Warren (1994, p. 672), there are no effects of the lexical status of the cross-spliced portions of words because there is no lexical competition in the model: Only the correct word will be recognized and made available for the phonetic decision regardless of whether the word is cross-spliced with another word or with a nonword. However, nonwords cross-spliced with words should be more difficult than those cross-spliced with other nonwords because in the former case the lexicon will provide support for the incorrect phoneme (such as for [g] given smob made from smog and smob).

The postlexical model might also be able to account for the results of Connine et al. (1997). As a nonword becomes more like a word, the lexicon will tend to provide more support for the target phoneme and performance should improve. Similarly, the postlexical model could account for the findings of Newman et al. (1997): Postlexical phoneme decisions to ambiguous phonemes could well be biased toward phonemes in nonwords with larger lexical neighborhoods. Although further specification of exactly how phonetic decisions to segments in nonwords in this model would need to be made, it is clear that the model could account for effects such as these without any radical changes to its basic architecture.

However, like both the Race and TRACE models, the postlexical model is not without problems. The model is challenged by the variability of lexical effects observed in phonetic tasks (Cutler et al., 1987; Eimas et al., 1990; Eimas & Nygaard, 1992; Frauenfelder & Seguí, 1989; Seguí & Frauenfelder, 1986). If all phonetic decisions are based on output from the lexicon, then it should not be possible to observe phonetic decision making in the absence of lexical involvement. Likewise, the model appears to predict, like TRACE, that there should be inhibitory lexical effects, such that, for example, monitoring for [t] should be harder in vocabutaire than in socabutaire. Such effects have not been found (Frauenfelder et al., 1990; Wurm & Samuel, 1997).

Furthermore, it is unclear how the postlexical model would account for the fact that lexical influence in phonetic categorization of word-final segments diminishes in listeners' slower responses (McQueen, 1991; Pitt & Samuel, 1993). If decisions are based on lexical output, lexical involvement in categorization judgments should not diminish over time. It is also unclear how the model would account for the data on compensation for coarticulation (Elman & McClelland, 1988; Pitt & McQueen, 1998). Because the compensation effect operates between words (e.g., an influence of christma? on tapes), and perceptual decisions are postlexical in the model, it would appear that the compensation for coarticulation process would have to be located postlexically so that effects with ambiguous fricatives could be explained. The compensation process, which depends on detailed acoustic–phonetic information that may not be available in more abstract postlexical representations, does not have such a late locus (Mann, 1986). Furthermore, even if the process did operate postlexically, the model could not account for the dissociation between lexical and transitional probability biases observed by Pitt and McQueen (1998): If the lexicon were acting to produce both a lexical and a transitional probability bias in fricative identification, and the probability bias were also influencing stop identification via the compensation process,

the model would predict that the lexical bias would also influence stop identification, but this effect was not observed.

This review of lexical effects in phonetic decision making suggests that none of the models that are currently available provides a satisfactory account of the data. Although there is no one result that is problematic for all models, no model can account for all the data. In the present research, we first attempted to clarify the pattern of results in experiments modeled on those of Marslen-Wilson and Warren (1994).

The data presented by Marslen-Wilson and Warren (1994) are particularly important because they simultaneously challenge both of the older models and directly support the postlexical model. This latter model, however, is itself strongly challenged by data showing the variability of lexical effects in phonetic decision making (e.g., Cutler et al., 1987; Eimas et al., 1990; Eimas & Nygaard, 1992). In the experiments we report here, we examined whether the lexical effects observed by Marslen-Wilson and Warren (1994), like the other lexical effects, would also be modulated by task demands. It is possible that the inhibitory effects caused by subcategorical mismatch actually behave differently from the facilitatory effects that were manipulated in earlier studies. Therefore, in our first experiment, we sought to replicate Marslen-Wilson and Warren's findings. We used the same phonetic decision task in a close repetition of their study, albeit with Dutch materials and Dutch listeners.

Experiment 1

Method

Participants. Forty-two volunteers from the Max Planck Institute subject pool were paid to take part in the experiment. They were all native speakers of Dutch and had no known speech or hearing disorders.

Materials and design. Fifteen triplets based on real words and 15 triplets based on nonwords were selected. These items are listed in Appendix A. All the items were monosyllabic; all ended with unvoiced stops; and within each triplet one item ended with [p], one ended with [t], and one ended with [k]. Some items began with single consonants and some with consonant clusters, but all had only a single consonant (the unvoiced stop) after the vowel. In addition to triplets based on unvoiced stops, Marslen-Wilson and Warren (1994) also used triplets based on final voiced stops in their phonetic decision experiment. Voiced stop triplets cannot be constructed in Dutch because Dutch has only two voiced stops ([b] and [d]; [g] is not in the native inventory) and because voiced stops do not appear in word-final position in Dutch.

The word triplets consisted of a base word (e.g., *sloop*, pillowcase), plus a matched word, identical to the base word except that it ended with a different unvoiced stop (e.g., *sloot*, ditch), and a matched nonword, also beginning in exactly the same way but ending with the third unvoiced stop (e.g., *slook*, which is not a word in Dutch). Five base words ended with [p], 5 with [t], and 5 with [k]. A number of constraints applied in the selection of these triplets. The most obvious constraint, of course, was that for a given word ending in a stop, there had to be one other word ending in a stop with a different place of articulation, but only one other word. This ruled out many words. In addition, word frequency constraints were applied. Words with a frequency of less than 10 counts per 42

million words (from the CELEX corpus; Burnage, 1990) were not considered as possible base words or their possible matched words. Pairs of words were also matched as closely as possible in frequency. Overall, the base words had a mean frequency of 17 per million, and the matched words had a mean frequency of 14 per million. In addition, triplets were ruled out if there were any monosyllabic words, identical to the other triplet members except for the final consonant, that were of higher frequency than either the base word or the matched word. These constraints generally matched those of Marslen-Wilson and Warren (1994) and, as in their English materials, left a relatively small number of acceptable triplets. All triplets in the language that could be found given these constraints were used.

The nonword triplets were constructed in a similar manner. Each consisted of a base nonword (e.g., *smep*), plus a matched word, identical to the base nonword except that it ended with a different stop (e.g., *smet*, stain), and a matched nonword, also beginning in exactly the same way but ending with the third stop (e.g., *smek*, which is not a word in Dutch). Five base nonwords ended with [p], 5 with [t], and 5 with [k]. The primary constraint in generating these triplets was that for a given word ending in an unvoiced stop, a change in that stop to either of the other two unvoiced stops would result in a nonword in both cases. Only words with these constraints and with a frequency greater than 10 per 42 million words were considered. The mean frequency of the matched words in the nonword triplets was 10 per million. Triplets were again ruled out if there was any other monosyllabic word of a higher frequency than the matched word, beginning in the same way as the triplet members but ending with a consonant other than a stop.

These triplets of items were used to construct the materials used in the experiment. Listeners heard only the base words and the base nonwords, but there were three different versions of each item. The closure and burst of the final stop of each base word was spliced onto the initial portion (up to the end of the vowel) of another token of the base word (*Word 1 + Word 1*, W1W1; e.g., *sloo* from one token of *sloop* plus the closure and burst of [p] from another token of *sloop*). The same stop information was also spliced onto the initial portion of the matched word (*Word 2 + Word 1*, W2W1; e.g., *sloo* from *sloot* plus the [p] information from *sloop*) and onto the initial portion of the matched nonword (*Nonword 3 + Word 1*, N3W1; e.g., *sloo* from *slook* plus the [p] information). The same procedure was used for the base nonwords. The closure and burst portions of their final stops were spliced onto another token (*Nonword 1 + Nonword 1*, N1N1; e.g., *sme* from one token of *smep* plus the [p] information from a second token), onto the matched word (*Word 2 + Nonword 1*, W2N1; e.g., *sme* from *smet* plus the [p] information), and onto the matched nonword (*Nonword 3 + Nonword 1*, N3N1; e.g., *sme* from *smek* plus the [p] information).

The resulting six conditions are shown in Table 1; they match those of Marslen-Wilson and Warren (1994). In all conditions the resulting tokens sounded like the base word or nonword, but in the nonidentity conditions there were conflicting cues to the place of articulation of the final stop, those present in the presplice segment (principally formant-transition information), and those present in the postsplice segment (release burst information). There were six different types of phonetic conflict (bilabial [p] plus either alveolar [t] or velar [k], etc.) that varied across and within the nonidentity conditions. Crucially, what was held constant within each of these conditions was the lexical status of the tokens used in the splicing. The information mismatching with the final stop was consistent with either a word or a nonword, and the resulting token was either a word or a nonword.

Three lists of 30 items were constructed. Each list contained one version of each of the 15 words and one version of each of the 15

nonwords, in random order. Splice conditions were counterbalanced over lists, such that each list contained five of each of the six conditions. Target phonemes were also counterbalanced over lists, such that each list contained 10 items ending with [p], 10 ending with [t], and 10 ending with [k]. These three lists were in turn used to construct three experimental sequences, each of which consisted of all three lists but in different orders. The first third of the experiment (Block 1; i.e., the first list of 30 items in each sequence) therefore contained all 90 items, counterbalanced across the three sequences. Because the second and third versions of each item in a sequence sounded like repetitions of the first version, it was possible that repetition effects would interact with splicing effects (Whalen, 1991, observed practice effects in a lexical-decision experiment with cross-spliced items, although in his study these effects did not interact with any other factor, including mismatch). The present design allowed for an analysis of the first third of the experiment alone, which contained no repetitions. Each sequence was preceded by a short practice list consisting of 18 monosyllables ending in unvoiced stops (9 words and 9 nonwords, spliced in the same six conditions as in the main experiment, in random order).

Procedure. Two tokens of each of the items listed in Appendix A, plus those to be used in the practice list, were recorded by a female native speaker of Dutch. She was instructed to speak clearly; all her productions contained released final stops. The recording was made in an acoustically isolated booth onto digital audio tape (sampling at 48 kHz, with 16-bit resolution). The materials were then transferred to the hard disk of a computer (down-sampling to 16 kHz) and edited using the Entropics Xwaves speech editor. Splicing was performed following the procedure used by Marslen-Wilson and Warren (1994). In each case, the splice was made at the offset of periodic energy associated with the vowel (i.e., at the onset of the stop closure). The splices produced no discontinuities that could be heard.

The materials were then copied onto the hard disk of a personal computer (up-sampling to 20 kHz). Items were played to listeners over closed-ear headphones, with digital-to-audio conversion direct from this hard disk. Experiment control and data collection were carried out using the Max Planck Institute for Psycholinguistic's NESU software. Each trial began with the auditory presentation of a warning tone, coupled with the visual presentation of the two response alternatives on a computer screen. There were three possible pairs of alternatives: P T; P K; and T K. These alternatives were presented as uppercase letters, one on the left of the screen and one on the right. Consistency in the response mapping was held as constant as possible: the P always appeared on the left side, and the K always appeared on the right; only the position of the T varied.

For the items with conflicting cues, the two letters corresponded

to the two cued segments. Thus, for example, the letters P and T appeared with the W2W1 version of *sloop* (which was made from *sloot* and *sloop*). A response on these trials was taken to be correct when it corresponded with the place of articulation of the stop release after the splice. For the items with no conflicting cues (W1W1 and N1N1), one letter corresponded to the correct stop and one to the stop consistent with the matched word in that triplet.

Listeners were tested singly or in groups of 2, 3, or 4 in a room containing three acoustically isolated booths and one other testing carrel; they were unable to hear each other's keypress responses. They were told that they would be presented with a sequence of trials and that on each trial they would be required to press either the left or the right button on a console. They were also told that on each trial they would see a pair of letters on the computer screen in front of them, one letter positioned above the left button and one positioned above the right button, and that they would then hear either a real word or a nonsense word over the headphones. They were asked to decide as quickly as possible which sound the word ended with and then to press the button below the appropriate letter. RTs were measured from the acoustic onset of each item. These RTs were then adjusted, to measure from the splice point, by subtraction of the duration of the appropriate presplice portion. Listeners heard the practice list, followed by one of the three experimental sequences. Each of the three sequences was heard by 14 participants. Short breaks were given between each list.

Results and Discussion

The results are shown in Table 2. Analyses of variance (ANOVAs) were carried out on both the RTs and the errors, with participants ($F1$) and items ($F2$) as random factors. In the RT analysis, all outlying RTs (those outside the range of 150–1,500 ms) were excluded (this constraint excluded only 0.4% of the data). Missing data points (errors or outlying RTs) for each listener were replaced with the mean of the available data points for that listener in that condition; in the items analyses, missing data points were likewise replaced with the mean of the available RTs for that item in that condition.

RT analyses. There was a main effect of lexical status: Responses to words were faster than responses to nonwords, but this was significant only by participants, $F1(1, 39) = 16.34$, $p < .001$; $F2 < 1$. There was a highly significant effect of splice: Responses to identity-spliced items (W1W1 and N1N1) were faster than responses to the corresponding cross-spliced items, $F1(2, 78) = 222.19$, $p < .001$; $F2(2,$

Table 2
Mean Phonetic Decision Latencies (in Milliseconds, Measured From Phoneme Target Onsets) and Mean Error Rates (Percentage of Incorrect Responses) in Experiment 1

	Words						Nonwords					
	Word 1 + Word 1 (W1W1)		Word 2 + Word 1 (W2W1)		Nonword 3 + Word 1 (N3W1)		Nonword 1 + Nonword 1 (N1N1)		Word 2 + Nonword 1 (W2N1)		Nonword 3 + Nonword 1 (N3N1)	
Block	M	%	M	%	M	%	M	%	M	%	M	%
1	544	3	651	8	640	5	589	2	665	5	648	7
2	549	3	671	4	646	5	559	4	659	6	663	3
3	543	2	654	4	635	2	579	3	651	3	654	5
Overall	545	3	658	5	640	4	576	3	658	5	655	5

56) $= 60.72, p < .001$. There was also an interaction between lexical status and splicing: The splicing effect was somewhat larger for words than for nonwords, but, again, this was only significant by participants, $F1(2, 78) = 4.35, p < .05; F2(2, 56) = 1.24, p = .3$.

The effect of splicing was examined in planned comparisons between each splicing condition within words and nonwords. In the words analysis, responses to identity-spliced words (W1W1) were reliably faster than those to words cross-spliced with other words (W2W1; 113 ms on average), $t1(41) = 14.81, p < .001; t2(14) = 6.77, p < .001$, and to words cross-spliced with nonwords (N3W1; 95 ms on average), $t1(41) = 12.70, p < .001; t2(14) = 7.95, p < .001$. The difference between the two cross-spliced words (responses to W2W1 words were, on average, 18 ms slower than those to N3W1 words) was significant only by participants, $t1(41) = 2.71, p = .01; t2(14) = 0.94, p > .3$. In the nonwords analysis, responses to identity-spliced nonwords (N1N1) were reliably faster than those to nonwords cross-spliced with words (W2N1; 82 ms on average), $t1(41) = 11.10, p < .001; t2(14) = 7.92, p < .001$, and to nonwords cross-spliced with other nonwords (N3N1; 79 ms on average), $t1(41) = 9.56, p < .001; t2(14) = 7.41, p < .001$. The difference between the two cross-spliced nonwords was not significant.

The RT results are clear: For both words and nonwords, there were strong effects of acoustic–phonetic mismatch. Identity-spliced items were responded to more rapidly than were cross-spliced items. Furthermore, for both words and nonwords, there was no significant effect of the lexical status of the cross-spliced portion. Cross-splicing with words was equivalent to cross-splicing with nonwords. The effect for words replicated Marslen-Wilson and Warren (1994); the effect for nonwords did not.

Although all listeners heard all three versions of each item, within the first third of the experiment there were no repetitions of items. Even though there was no main effect of block in the overall analysis, and this factor did not interact with either lexical status or splicing, a complete parallel analysis of Block 1 was also carried out to test the effects of subcategorical mismatch in the absence of any repetition effects. Analyses of the Block 1 data were also carried out in Experiments 2, 3, and 4. We report only outcomes from the analyses that differed from those of the overall analyses. The sole difference in the Experiment 1 RTs was that the interaction between splicing and lexical status was not significant in Block 1.

Error analyses. The only significant effect was one of splice, and this was significant only by participants, $F1(2, 78) = 4.45, p < .05; F2(2, 56) = 2.29, p = .11$. There was a weak tendency for performance to be more accurate on the identity-spliced items than on the cross-spliced items. Planned pairwise comparisons between the splicing conditions within words and nonwords did not reveal any significant differences. In the error analyses of Block 1 alone, the effect of splice was fully reliable across participants and items. The planned comparisons within Block 1 showed one significant difference for the words and two for the nonwords. For the words, performance on identity-spliced words was reliably

more accurate than that on words cross-spliced with other words, $t1(41) = 2.23, p < .05; t2(14) = 2.20, p < .05$. For the nonwords, performance on identity-spliced nonwords was better than that on nonwords cross-spliced with words (significant by participants and marginal by items), $t1(41) = 2.24, p < .05; t2(14) = 2.09, p = .06$, and better than that on nonwords cross-spliced with other nonwords (significant only by participants), $t1(41) = 2.89, p < .01; t2(14) = 1.62, p = .13$. In none of the error analyses was there any indication of differences between the two cross-spliced item types. As in the RT analyses, performance on cross-spliced words was equivalent regardless of whether the cross-splice involved another word or a nonword. Likewise, speed and accuracy on cross-spliced nonwords did not depend on whether the cross-splice involved a word or another nonword.

Experiment 2

The results of Experiment 1 clearly fail to replicate the findings of Marslen-Wilson and Warren (1994). Although the words showed the same pattern as in the earlier study, the crucial difference observed in that study between nonwords cross-spliced with words (W2N1) and nonwords cross-spliced with other nonwords (N3N1) was absent here. There was no indication that the lexical status of the cross-spliced portion influenced the speed or accuracy of performance. This suggests that lexical knowledge was not being used to any significant extent by the listeners in Experiment 1. Consistent with this claim is the finding that there was no fully reliable main effect of lexical status. Although words were responded to more rapidly than nonwords, this difference was significant only by participants, and there was no lexical effect in the error analysis.

In Experiment 2 we used the phoneme-monitoring task. The goal was to establish whether the pattern of results obtained in Experiment 1 would generalize to another, simpler phonetic judgment task or whether the pattern found by Marslen-Wilson and Warren (1994) would be obtained with this task. If listeners failed to use their lexical knowledge in phonetic decision because the task did not demand lexical processing (listeners could, e.g., focus attention on the item-final target location and ignore the contextual information), it would be likely that they would also not use lexical knowledge in the phoneme-monitoring task because, as described earlier, it is known that this task can be performed without the use of the lexicon (Cutler et al., 1987). On the other hand, the absence of lexical effects in Experiment 1 could have been due to the choice of task; a relatively difficult forced-choice decision could encourage listeners to focus attention on the final bursts. If this were true, the simpler monitoring task could reveal greater use of the contextual information in the vocalic portions of the materials and hence possibly lexical effects.

Method

Participants. Forty-two volunteers from the Max Planck Institute subject pool were paid to take part in the experiment. None of

these volunteers had participated in Experiment 1. They were all native speakers of Dutch and had no known speech or hearing disorders.

Materials and design. The materials were based on those of Experiment 1, with the addition of filler items needed for the phoneme-monitoring task. Each of the three lists of 30 items was split into three sublists, one for each of the three target phonemes ([p], [t], or [k]). Each sublist contained 5 words and 5 nonwords ending with the target phoneme and 24 fillers that did not end with the target phoneme. The fillers were all monosyllabic and ended either with an unvoiced fricative [f,s,x] or a nasal consonant [m,n,ŋ]. The experimental items and fillers within each sublist were mixed in pseudorandom order.

These fillers were constructed in the same way as the experimental materials (i.e., by splicing the initial sequences of words and nonwords, up to and including the vowel, onto the final consonants of other words and nonwords). Thus, within each set of 24, there were 12 words (6 identity-spliced, 3 cross-spliced with other words, and 3 cross-spliced with nonwords) and 12 nonwords (6 identity-spliced, 3 cross-spliced with words, and 3 cross-spliced with other nonwords). Unlike the experimental items, the fillers did not form triplets (three versions of the same word or nonword spliced in three ways). Instead, there was only one version of each filler constructed from two appropriate words, two appropriate nonwords, or one word and one nonword. As with the experimental items, only words with a frequency of greater than 10 counts per 42 million words were used. The proportions of fricative- and nasal-final items in each sublist were approximately equal, and we attempted to balance the number of different fricative and nasal pairings (e.g., the number of items based on cross-splicing [f]- and [s]-final items, the number of items based on identity-spliced [f]-final items, etc.). The fillers were thus highly similar to the experimental items, except that they did not end with unvoiced stops.

The design of Experiment 1 was used here. As before, there were three experimental sequences, such that all listeners heard all three versions of the experimental items (and all fillers) in three different orders of presentation and such that the first third (Block 1) was itself a fully balanced experiment. Within each list, however, there were three sublists, one for each target phoneme. The order of presentation of these sublists was rotated within lists in a fully counterbalanced manner. (The presentation order of target sublists within lists was in fact ignored in the analysis of results.) There was also a practice list consisting of 9 words and 9 nonwords in random order (spliced in the same six conditions as in the main experiment). Six of these items ended with the target phoneme [s].

Procedure. The fillers and practice items were recorded at the same time as the items used in Experiment 1. They were digitized and edited using the same procedure. Participants were instructed that they would be given target sounds, presented as letters on a computer screen (P, T, K, or S in the practice), and that they would then hear lists of words and nonwords presented over headphones. They were also told that their task was to press the button on the console in front of them as fast as possible if the word or nonword they heard ended with the target sound. A warning tone preceded each auditory item. As in Experiment 1, RTs were measured from the acoustic onset of each experimental item. They were again adjusted before the analysis so as to measure from the splice point. Participants were again tested either singly or in groups of up to 4 in the same room as in Experiment 1. They heard the practice list followed by nine lists (three lists corresponding to the lists in Experiment 1, with each divided into three sublists with different target phonemes, that were presented on the computer screen before the onset of each sublist). There were short breaks between the sublists.

Results and Discussion

The results are shown in Table 3. The same pattern of responses as that found in Experiment 1 was found here. As in Experiment 1, we conducted ANOVAs on the RTs and the errors. In the RT analysis, all outlying RTs (again, those outside the range of 150–1,500 ms) were excluded (this constraint again excluded only 0.4% of the data). Missing data points were replaced in the same way as in Experiment 1. The results of 3 listeners were excluded from the analysis, leaving three groups of 13 listeners who heard each of the three sequences of items. One of the listeners whose data were dropped had failed to detect at least 50% of the target phonemes. The data from the other 2 listeners were dropped to maintain counterbalancing across groups; they were in fact the only other listeners to miss more than 5% of the targets.

RT analyses. There was again a main effect of lexical status: Responses to words were faster than responses to nonwords, but, as in Experiment 1, this was significant only by participants, $F1(1, 36) = 5.19, p < .05; F2 < 1$. The highly significant effect of splice was replicated: Responses to identity-spliced items (W1W1 and N1N1) were faster than responses to the corresponding cross-spliced items, $F1(2, 72) = 189.23, p < .001; F2(2, 56) = 76.27, p < .001$. There was also an interaction between lexical status and splicing, $F1(2, 72) = 8.88, p < .001; F2(2, 56) = 3.17, p <$

Table 3

Mean Phonetic-Monitoring Latencies (in Milliseconds, Measured From Phoneme Target Onsets) and Mean Error Rates (Percentage of Missed Responses) in Experiment 2

	Words						Nonwords					
	Word 1 + Word 1 (W1W1)		Word 2 + Word 1 (W2W1)		Nonword 3 + Word 1 (N3W1)		Nonword 1 + Nonword 1 (N1N1)		Word 2 + Nonword 1 (W2N1)		Nonword 3 + Nonword 1 (N3N1)	
Block	M	%	M	%	M	%	M	%	M	%	M	%
1	418	1	543	3	542	2	462	1	537	1	532	1
2	428	2	520	2	511	1	446	1	509	0	501	1
3	412	1	509	2	501	2	441	0	515	3	505	0
Overall	419	1	524	2	518	2	450	0	521	1	513	1

.05. In contrast to in Experiment 1, there was also a main effect of list: Responses became faster moving from the first to the second to the third list, $F1(2, 72) = 7.24, p < .005$; $F2(2, 56) = 14.17, p < .001$. This factor did not reliably interact with any other factors.

The effect of splicing was again examined in planned comparisons between each splicing condition within words and nonwords. For words, responses to identity-spliced words (W1W1) were faster than those to words cross-spliced with other words (W2W1; 105 ms on average), $t1(38) = 15.26, p < .001; t2(14) = 8.30, p < .001$, and to words cross-spliced with nonwords (N3W1; 99 ms on average), $t1(38) = 11.83, p < .001; t2(14) = 9.70, p < .001$. There was no difference between the two cross-spliced words. For nonwords, responses to identity-spliced nonwords (N1N1) were faster than those to nonwords cross-spliced with words (W2N1; a mean difference of 71 ms), $t1(38) = 12.72, p < .001; t2(14) = 7.58, p < .001$, and to nonwords cross-spliced with other nonwords (N3N1; a mean difference of 63 ms on average), $t1(38) = 8.65, p < .001$; $t2(14) = 8.08, p < .001$. The difference between the two cross-spliced nonwords was again not significant. Exactly the same splicing effects were therefore found for words and nonwords; the interaction between lexical status and splicing indicates simply that the splicing effect was somewhat larger for words than for nonwords.

Error analyses. There were no reliable effects in the analyses of missed responses.

In summary, the results replicate those of Experiment 1. For both words and nonwords, there were strong mismatch effects (targets in identity-spliced items were detected more rapidly than those in cross-spliced items), and there were no effects of the lexical status of the cross-spliced portion. The failure to find a lexical effect in the cross-spliced nonwords again contradicts the results of Marslen-Wilson and Warren (1994).

Experiment 3

The combined findings of Experiments 1 and 2 indicate that the lexical effect on nonwords observed by Marslen-Wilson and Warren (1994) is not easily replicable. Like other effects of lexical status in phonetic judgment tasks (Cutler et al., 1987; Eimas et al., 1990; Eimas & Nygaard, 1992), the effect observed by Marslen-Wilson and Warren may come and go according to changes in task demands. The model proposed by Marslen-Wilson and Warren cannot explain the absence of lexical effects in phonetic decision making. If all phonetic decisions to both words and nonwords are based on the output of lexical processes, then it should not be possible to observe performance in a phonetic task that does not reveal lexical involvement, nor should it be possible for lexical effects to come and go in different experimental situations. However, why would the effect have occurred in Marslen-Wilson and Warren's experiment but not in either of our Experiments 1 or 2? The answer to this question, based on the earlier literature on the variability of lexical effects, is that in some circumstances, task demands discourage the use of lexical knowledge, whereas

in other circumstances, the use of lexical knowledge is encouraged. According to this view, it would appear that some factor in both the present phonetic decision task (Experiment 1) and phoneme-monitoring task (Experiment 2) caused listeners not to use their lexical knowledge to any significant degree.

It was possible, however, that the materials in Experiments 1 and 2 were somehow at fault and that the manipulation of lexical status was not powerful enough for lexical effects to be detected. To test this possibility, we used the lexical-decision task in Experiment 3. This task has been used in previous studies of the effect of mismatching phonetic information on lexical access (Streeter & Nigro, 1979; Whalen, 1991). The results of these studies showed that there were differential effects of mismatch in words and nonwords. Streeter and Nigro (1979) found that, relative to unspliced items, mismatch between vowel–consonant and consonant–vowel transitions in the middle of bisyllabic words delayed "yes" responses to words but did not delay "no" responses to nonwords. Whalen (1991) found somewhat stronger effects of mismatch between vowel transitions and final fricatives in words than in nonwords (although this difference was not statistically reliable). In neither of those studies, however, were there within-items comparisons of the effects of cross-splicing words and nonwords with both words and nonwords, as in Marslen-Wilson and Warren's (1994) study.

Marslen-Wilson and Warren (1994) observed an inhibitory lexical effect in lexical decisions analogous to what they had found in phonetic decisions: "No" decisions to nonwords cross-spliced with words (e.g., *smob* made from *smog* and *smob*) were slower than decisions to nonwords cross-spliced with other nonwords (e.g., *smob* made from *smod* and *smob*) because of interference from the word (*smog*) only in the former case. They also found that responses to nonwords cross-spliced with nonwords were no slower than responses to identity-spliced nonwords (in contrast to the mismatch effect found with these items in phonetic tasks, both in Marslen-Wilson and Warren and in our Experiments 1 and 2). This, they argued, was because cross-splicing two nonwords does not create a mismatch at the lexical level; both parts of the resulting nonword are consistent with a "no" response. It is only when the mismatch involves information consistent with a word that a mismatch effect should be observed in lexical decision. Furthermore, Marslen-Wilson and Warren found in lexical decision to words the same pattern as they had found in the phonetic decision task. There was no lexical effect in their comparison of lexical decisions to words cross-spliced with words and words cross-spliced with nonwords, but both types of cross-spliced item were responded to more slowly than were identity-spliced items. Marslen-Wilson and Warren argued that this was because, in contrast to cross-splices in nonwords, any kind of cross-splice in a word, whether involving another word or a nonword, should interfere with recognition of the word and thus slow lexical decisions to that word.

In Experiment 3 we tested for similar effects in lexical decision using the Dutch materials from Experiments 1 and 2. Lexical involvement in decision making should be seen with lexical decision if it is to be found at all. When the

decision concerns the lexical status of an item, the lexical status of the components of that item ought to influence that decision. In Experiment 3 we thus tested whether the failure to observe inhibitory lexical effects in Experiments 1 and 2 was due to a problem with the items, in which case the effects would not be found even with the lexical-decision task. Alternatively, the failure to observe lexical effects in the earlier experiments could be due to the task demands of the phonetic tasks, which could have discouraged lexical involvement, in which case the effects ought to be found with the lexical-decision task, which does encourage lexical involvement.

Method

Participants. Forty-five volunteers from the Max Planck Institute subject pool took part. They were paid for their participation. All were native speakers of Dutch and had no known speech or hearing disorders, and none had participated in either Experiments 1 or 2. Seventeen participants heard one of the experimental sequences, whereas 14 participants heard each of the other two sequences.

Materials and procedure. The materials from Experiment 1 were used in a lexical-decision task. Exactly the same auditory materials were used, which were again divided into three lists and presented in the same three running orders to three different groups of participants. The only changes were in the instructions and procedure. Participants were asked to listen to each item and to decide whether it was a real word in Dutch. They were asked to respond as quickly as possible by pressing either the button labeled "Ja" (yes) or the button labeled "Nee" (no) on the console in front of them. As in Marslen-Wilson and Warren's (1994) lexical-decision experiment and our Experiment 2, only a warning tone preceded each item (i.e., there was no visual presentation of response alternatives on a computer screen, as in Experiment 1). RTs were measured from the acoustic onset of each item. Before the analysis these were adjusted by subtraction of the duration of the appropriate item so as to measure from the acoustic offset of that item. Participants were again tested singly or in groups of up to 4, in the same room as for Experiments 1 and 2.

Results and Discussion

The results are shown in Table 4. ANOVAs were again carried out on both the RTs and the errors. In the RT analysis, all outlying RTs (those slower than 1,500 ms) were excluded. (This constraint again excluded only 0.4% of the

data; responses faster than 150 ms were, in contrast to in Experiments 1 and 2, not excluded here. There were several fast responses, particularly to identity-spliced words, made soon after the acoustic offset of words.) Missing data points were replaced in the same way as in the other experiments.

RT analyses. There was a main effect of splice: Responses to identity-spliced items were considerably faster than those to cross-spliced items, $F1(2, 84) = 184.42, p < .001; F2(2, 56) = 29.86, p < .001$. There was also a main effect of list: Listeners responded more rapidly in the later lists, $F1(2, 84) = 17.75, p < .001; F2(2, 56) = 30.97, p < .001$. Listeners were faster in making "yes" responses than "no" responses, but this difference was significant only by participants, $F1(1, 42) = 5.05, p < .05; F2 < 1$. Importantly, there was a significant interaction between splice and lexical status, $F1(2, 84) = 40.79, p < .001; F2(2, 56) = 9.59, p < .001$.

Pairwise *t* tests showed that the splice effect within words was the same as in the earlier two experiments. Responses to identity-spliced words (W1W1) were faster than those to words cross-spliced with other words (W2W1), $t1(44) = 20.06, p < .001; t2(14) = 6.41, p < .001$, and those to words cross-spliced with nonwords (N3W1), $t1(44) = 16.41, p < .001; t2(14) = 7.53, p < .001$. There was no reliable difference between the two types of cross-spliced words.

There was a different pattern in the nonwords. RTs to identity-spliced nonwords (N1N1) were faster than those to nonwords cross-spliced with words (W2N1), $t1(44) = 6.94, p < .001; t2(14) = 4.60, p < .001$, but they were not reliably faster than those to nonwords cross-spliced with other nonwords (N3N1), $t1(44) = 2.89, p < .01; t2(14) = 1.70, p = .11$. Crucially, the difference between the two cross-spliced nonwords was also significant, $t1(44) = 3.01, p < .005; t2(14) = 2.47, p < .05$. Nonwords cross-spliced with words (W2N1) were thus reliably harder to process than both other types of nonword, which did not reliably differ from each other. In the Block 1 analyses, exactly the same pattern was observed, except that in the items analyses none of the differences was statistically significant.

For the first time, therefore, the pattern of results was not the same for words and nonwords. The word responses showed the same pattern as in the earlier experiments and as in Marslen-Wilson and Warren's (1994) lexical-decision experiment. The nonwords, however, showed a different

Table 4
Mean Lexical-Decision Latencies (in Milliseconds, Measured From Item Offsets) and Mean Error Rates (Percentage of Incorrect Responses) in Experiment 3

	Words						Nonwords					
	Word 1 + Word 1 (W1W1)		Word 2 + Word 1 (W2W1)		Nonword 3 + Word 1 (N3W1)		Nonword 1 + Nonword 1 (N1N1)		Word 2 + Nonword 1 (W2N1)		Nonword 3 + Nonword 1 (N3N1)	
Block	M	%	M	%	M	%	M	%	M	%	M	%
1	375	9	513	16	504	22	457	2	500	4	473	3
2	331	5	468	12	464	11	420	1	482	2	465	3
3	316	8	454	11	442	10	399	2	446	3	416	3
Overall	340	7	478	13	470	14	425	1	476	3	451	3

pattern. As in the earlier lexical-decision study, there was an inhibitory lexical effect on the processing of nonwords cross-spliced with words and no significant difference between identity-spliced nonwords and nonwords cross-spliced with nonwords.

Error analyses. In the analysis of proportion of incorrect responses, there was a significant effect of lexical status: Overall, responses to nonwords were more accurate than responses to words, $F1(1, 42) = 56.83, p < .001; F2(1, 28) = 9.08, p < .01$. Listeners, particularly in the first list, appeared to have a bias toward making "no" responses. More important, there was also an effect of splice, $F1(2, 84) = 11.50, p < .001; F2(2, 56) = 7.32, p < .005$. The splice effects were not equivalent for words and nonwords, $F1(2, 84) = 4.83, p < .05; F2(2, 56) = 3.04, p = .06$. There was also an effect of list: Error rates dropped in the later lists, $F1(2, 84) = 12.87, p < .001; F2(2, 56) = 4.58, p < .05$. This effect interacted with lexical status—the disadvantage for words over nonwords lessened in the later lists, $F1(2, 84) = 6.32, p < .005; F2(2, 56) = 3.15, p = .05$—but did not interact with any other effects.

Within the words, t tests showed that identity-spliced words (W1W1) were responded to more accurately than either words cross-spliced with other words (W2W1), $t1(44) = 3.95, p < .001; t2(14) = 2.57, p < .05$, or words cross-spliced with nonwords (N3W1), $t1(44) = 3.90, p < .001; t2(14) = 4.10, p < .005$. The difference between the two types of cross-spliced words was not significant. There were no significant differences in the pairwise comparisons between nonwords.

The error analyses were thus consistent with the RT analyses. Within the words, there were no differences between the two types of cross-spliced items in either speed or accuracy. On both measures, however, performance on identity-spliced words was better than that on the cross-spliced words. This was a consistent pattern observed on these items in lexical and phonetic decision (both here and in Marslen-Wilson & Warren, 1994) and in phoneme monitoring (Experiment 2). Within the nonwords, there were no reliable splicing effects in the error analyses. Nevertheless, in RT, performance was poorer on the items cross-spliced with words than on both identity-spliced items and items cross-spliced with other nonwords. This is the pattern observed for similar English items in lexical decision in Marslen-Wilson and Warren's (1994) study.

It would thus appear that the Dutch materials used in the present experiments are capable of revealing inhibitory lexical effects in nonword decisions. If so, though, why were these effects not observed in Experiments 1 and 2? As discussed earlier, it is possible that the task demands of these experiments discouraged lexical involvement. If that were the case, however, why did Marslen-Wilson and Warren (1994) find equivalent inhibitory effects in nonwords cross-spliced with words in phonetic and lexical decision? In Experiment 4 we sought to examine further the discrepancy between the present experiments and those of Marslen-Wilson and Warren.

Experiment 4

Careful comparison of Experiment 1 and the phonetic decision experiment in Marslen-Wilson and Warren (1994) revealed two subtle differences. The first concerned the response mapping. In Experiment 1, the mapping of response to button was kept consistent: [p] Responses were always made with the left button; [k] responses were always made with the right button; and only [t] responses varied, on the right when the alternative was a [p] and on the left when the alternative was a [k]. This appeared to be the simplest setup for listeners to learn. Marslen-Wilson and Warren (1994), however, made the task somewhat more difficult: "Targets were assigned pseudorandomly to left and right positions on the screen" (p. 665). Although it seems unlikely that such a difference could account for the discrepancy between Experiment 1 and Marslen-Wilson and Warren's phonetic decision experiment, target position on the screen was made to vary pseudorandomly in Experiment 4.

The second difference was that responses in Experiment 1 were limited to the set of unvoiced stops [p,t,k], whereas in Marslen-Wilson and Warren (1994) there were both unvoiced stops and the voiced stops [b,d,g]. As with the varied response mapping, increasing the set of possible responses and mixing those response alternatives from trial to trial would act to make the task more difficult. Experiment 4 therefore included additional sets of response pairs. It was impossible to include triplets of items ending in voiced stops in Dutch, both because [g] is not a native phoneme and because the voiced stops [b] and [d] do not occur word-finally (they are devoiced in this position). However, it was possible to include triplets ending with the nasal consonants [m,n,ŋ] and triplets ending with the fricatives [f,s,x].

Fricative-final items were used in Marslen-Wilson and Warren's (1994) lexical-decision experiment but were dropped from the phonetic decision task because they had failed to produce clear effects of subcategorical mismatch (in contrast to the clear mismatch effects observed with fricatives in both phonetic and lexical decision by Whalen, 1984, 1991). Marslen-Wilson and Warren (1994) avoided nasals (p. 657, Footnote 8) because of difficulties in cross-splicing such items. The fricative- and nasal-final items in Experiment 4, however, were added not to test for effects of mismatch on those items but simply to make the task more difficult and thus more compatible with Marslen-Wilson and Warren's experiment. The prediction was straightforward: If lexical influence in phonetic tasks depends on task demands (as the results of Cutler et al., 1987; Eimas et al., 1990, and Eimas & Nygaard, 1992, suggest), then increasing the difficulty of the phonetic decision task by both varying the targets and the response mappings of those targets would cause lexical effects to emerge. It was possible that the easier version of phonetic decision used in Experiment 1 and the simple low-level focus of the phoneme-monitoring task used in Experiment 2 allowed listeners to focus attention on the acoustic–phonetic information and to ignore lexical information.

Method

Participants. Thirty-six members of the Max Planck Institute subject pool were paid to take part. They were all native speakers of Dutch and had no known speech or hearing disorders, and none of them had participated in any of the earlier experiments. Twelve participants heard each of the three experimental sequences.

Materials and design. The experiment was based closely on Experiment 1. The experimental materials were identical to those used in all the earlier experiments, and they were again split over three lists. The only change from Experiments 1 and 3 was that these lists also contained filler items ending in fricatives [f,s,x] and nasal consonants [m,n,ŋ], to which the participants also had to respond. These fillers were selected from the fillers used in Experiment 2, which, like the experimental items, were constructed by cross-splicing. In addition to the 30 experimental items in each list, there were 24 fillers (6 fricative-final words, 2 identity-spliced, 2 cross-spliced with other words, and 2 cross-spliced with non-words; 6 fricative-final nonwords, 2 identity-spliced, 2 cross-spliced with words, and 2 cross-spliced with other nonwords; and a similar set of 12 nasal-final items). Note that the fillers were not based on triplets of items (unlike the experimental items); there was thus only one version of each filler word or nonword (rather than three differently spliced versions). Nine fricative-final items (three of each splice type) and 9 nasal-final items (three of each type) were added to the practice list from Experiment 1. All lists consisted of a pseudorandom mixture of stop-, fricative-, and nasal-final items, such that items of all types were evenly distributed over each list.

Procedure. This was identical to Experiment 1, except in the presentation of response alternatives on the computer screen. In addition to the target specifications for the experimental items (P, T, and K), there were also fricative targets (F, S, and G for [x]) and nasal targets (M, N, and NG for ŋ). Instead of a fixed response mapping, as in Experiment 1, target positions varied on the screen. Half the presentations of each response pairing had one alternative on the left of the screen, whereas for the other half that alternative was on the right of the screen. Target positions thus varied from trial to trial.

Results and Discussion

The results for responses to stop-final items are shown in Table 5. ANOVAs were again carried out on both the RTs and the errors. As in Experiment 1, all outlying RTs (those outside the range of 150–1,500 ms) were excluded (1.1% of

the data), and missing data were replaced in the same way as in the other experiments.

RT analyses. There was a significant effect of splice, $F1(2, 66) = 269.95, p < .001; F2(2, 56) = 71.69, p < .001$. This effect interacted with lexical status in the participants analysis, $F1(2, 66) = 11.56, p < .001; F2(2, 56) = 1.56, p = .22$. Responses to targets in words were faster than those to targets in nonwords, but this was significant only by participants, $F1(1, 33) = 9.55, p < .005; F2 < 1$. There were no other significant effects.

The planned pairwise comparisons within words revealed exactly the same pattern as was observed on the words throughout: an effect of cross-splicing but no difference between the two types of cross-spliced words. Responses to targets in identity-spliced words (W1W1) were reliably faster than those to targets in words cross-spliced with other words (W2W1; a difference of 136 ms on average), $t1(35) = 16.66, p < .001; t2(14) = 7.41, p < .001$, and those to targets in words cross-spliced with nonwords (N3W1; a difference of 134 ms on average), $t1(35) = 18.31, p < .001; t2(14) = 7.10, p < .001$. The difference of only 2 ms between the two types of cross-spliced word was not significant.

The nonwords showed a pattern different from that obtained in Experiments 1 and 2. As in Experiment 3 (lexical decision), responses to targets in nonwords cross-spliced with words (W2N1) were slower than those to targets in nonwords cross-spliced with nonwords (N3N1; a difference of 27 ms on average), significant only in the participants analysis, $t1(35) = 4.05, p < .001; t2(14) = 1.81, p < .1$. This difference was fully reliable in Block 1 (a mean difference of 41 ms), $t1(35) = 2.27, p < .05; t2(14) = 3.06, p < .01$. Responses were faster to identity-spliced nonwords (N1N1) than to either type of cross-spliced nonwords. For N1N1 versus W2N1, there was a difference of 115 ms on average, $t1(35) = 15.00, p < .001; t2(14) = 10.09, p < .001$. For N1N1 versus N3N1, there was a difference of 88 ms on average, $t1(35) = 11.04, p < .001; t2(14) = 6.67, p < .001$.

Error analyses. There were no significant effects in the analyses of incorrect responses.

In summary, the results of Experiment 4 replicate the phonetic decision results of Marslen-Wilson and Warren (1994). The task was made more difficult by varying the

Table 5
Mean Phonetic Decision Latencies (in Milliseconds, Measured From Phoneme Target Onsets) and Mean Error Rates (Percentage of Incorrect Responses) in Experiment 4

| | Words | | | | | | Nonwords | | | | | |
| | Word 1 + Word 1 (W1W1) | | Word 2 + Word 1 (W2W1) | | Nonword 3 + Word 1 (N3W1) | | Nonword 1 + Nonword 1 (N1N1) | | Word 2 + Nonword 1 (W2N1) | | Nonword 3 + Nonword 1 (N3N1) | |
Block	M	%	M	%	M	%	M	%	M	%	M	%
1	679	4	804	3	801	6	712	6	838	6	797	6
2	665	1	790	8	800	4	700	4	797	7	801	4
3	661	4	817	6	805	4	706	2	828	6	784	7
Overall	668	3	804	5	802	5	706	4	821	6	794	6

response mapping and including nasal and fricative targets; this made the experimental situation highly similar to that of the earlier study. It would appear that under these circumstances, listeners were more dependent on lexical knowledge, such that inhibitory lexical effects were observed in the nonwords that were cross-spliced with words. The results were not the same as those in Experiments 1 and 2, even though in those experiments we used exactly the same experimental materials. The earlier experiments were somewhat easier (as can be seen from the slower RTs in Experiment 4), however. Listeners appeared to be able to perform their tasks in the earlier experiments without the use of lexical knowledge. No differential effects on the nonwords cross-spliced with words and those cross-spliced with nonwords were therefore observed.

As we argue later, the pattern of results across the four experiments presented so far strongly constrains models of how and when the lexicon is involved in phonetic decision making. Our argument, however, depends on an important assumption about the nature of the mismatch effects that we have not yet tested. This is the assumption that the acoustic–phonetic information in the vocalic portions of the cross-spliced items provides support for the final consonants that the speaker originally produced in the tokens used for cross-splicing. In other words, is the mismatch effect in, for example, the W2N1 item smep due to evidence specifically cuing a [t] (as in the matched word smet), or is it a general mismatch effect due to acoustic information that simply mismatches the [p] without providing any particular support for the [t]? To be able to argue that the mismatch effects we found in the cross-spliced items were due to the influence of specific lexical entries, we will need to show that the mismatching information was consistent with those words.

To show that the presplice portions of the cross-spliced tokens actually contained these specific coarticulatory cues, Marslen-Wilson and Warren (1994) conducted a gating experiment. Listeners in that study were presented with the materials in gated fragments of ever-increasing length and were asked to decide from what word or nonword they thought the fragments came. The crucial responses were to the fragments that did not include any of the final release bursts. These responses indicated the extent to which there were coarticulatory cues in the presplice portions to the final stops produced in the tokens from which those portions were excised. There was clear evidence in all four cross-spliced conditions (i.e., words cross-spliced with other words or with nonwords, nonwords cross-spliced with words or with other nonwords) that the presplice portions contained strong cues to the stops that had been produced in the original utterances. In fragments shorter than the full presplice portion of a W2W1 word such as job, for example, there were many responses consistent with a velar place of articulation (i.e., consistent with the [g] of the W2 word jog that had been used in the cross-splicing).

Experiment 5

Experiment 5 was therefore carried out to confirm that the mismatch effects observed in the earlier experiments were

due to coarticulatory cues that specifically signaled the place of articulation of the stops with which the original W2 words and N3 nonwords had ended. It was a gating experiment modeled as closely as possible on the gating study in Marslen-Wilson and Warren (1994). Experiment 5 also had a second purpose: to retest, using a different task, for lexical effects. We have suggested that lexical involvement in phonetic decision making (Experiments 1, 2, and 4) depends on task demands. It was only in the more difficult conditions of Experiment 4 that lexical effects were observed. These effects were similar to those obtained in lexical decision (Experiment 3), that is, in a task that requires the use of lexical knowledge. Although listeners do not have to use lexical knowledge to respond in a gating task, it is likely that they will do so. If the information in a speech fragment can be continued to form either a word or a nonword, then listeners are likely to prefer to give the word as a response.

We therefore predicted strong lexical biases akin to those found in Marslen-Wilson and Warren's (1994) gating experiment. In the nonwords condition, we predicted that listeners hearing fragments of the presplice portion of a W2N1 item would be more likely to give that W2 word as a response (e.g., smet to the presplice portion of W2N1 smep) than they would be to give an N3 nonword response to fragments of the presplice portion of an N3N1 item (e.g., smek to the presplice portion of N3N1 smep). In addition to this lexical bias, however, we predicted that if there were mismatching information in these cross-spliced items, both of them would produce a radically different pattern of responses to that for the matching N1N1 items. Responses to an N1N1 item like smep, for example, should be consistent with a [p] even for short fragments, whereas responses to the cross-spliced versions of this item should be inconsistent with a [p] (and consistent with the presplice portions) except in the longest fragments (i.e., those that include the [p] release burst).

In the words condition, we predicted a similar mismatch effect. We predicted that responses consistent with the final stop in W1W1 items would begin early in the shorter fragments (e.g., in W1W1 sloop, responses consistent with the [p]) but that responses consistent with the final stop in W2W1 and N3W1 would appear reliably only in the final gates. However, a lexical bias could encourage listeners to produce responses consistent with the final stop also in the shortest fragments of cross-spliced words (i.e., before there was enough coarticulatory information to specify any particular place of articulation). This is because, unlike in the nonwords, responses consistent with the final stop in words would be word responses (i.e., sloop responses are word responses irrespective of how the word was made). Nevertheless, we again predicted a mismatch effect, namely that the majority of the responses, particularly in the longer presplice portions of the cross-spliced words, would be inconsistent with the final stops and instead would be consistent with the stops cued by the presplice portions. A lexical effect, as in the nonwords, was predicted here, too: Listeners hearing the shorter fragments of a W2W1 word should be more likely to give the W2 word as a response (e.g., sloot to W2W1 sloop) than to give an N3 nonword response to the shorter fragments of an N3W1 word (e.g., slook to N3W1 sloop).

Finally, we predicted that this lexical bias would be smaller in the words condition (because there was a strong alternative lexical candidate that could be chosen instead of the W2 word; i.e., the W1 word) than in the nonwords condition (where the only lexical candidate ending in a stop was the W2 word).

Method

Participants. Thirty-six volunteers from the Max Planck Institute subject pool were paid to take part. None of them had taken part in any of the earlier experiments. They were all native speakers of Dutch with no known speech or hearing disorders. Twelve participants heard each of the three test sequences.

Materials and design. The materials consisted of gated versions of all 90 experimental items used in the earlier experiments and gated versions of 18 filler items from Experiment 4. Nine of the fillers ended with fricatives, and 9 ended with nasal consonants. There were three fillers of each type of item (words or nonwords constructed by identity splicing or by cross-splicing with a word or a nonword).

All the items were prepared such that they could be presented to participants in 12 fragments that increased in duration in 25-ms increments. The fragments were aligned with the splice point, such that for each item the shortest fragment ended 125 ms before the splice point, the 6th fragment ended at the splice point, and the 11th fragment ended 125 ms after the splice point. The 12th fragment was the complete item (so the increment between the 11th and 12th item varied in length). The final 2.5 ms of each fragment (except for the complete item) was linearly ramped to silence, such that there were no offset discontinuities.

Three test sequences were made, with each containing, in random order, one version of each of the 15 experimental words, one version of each of the 15 experimental nonwords, and all 18 fillers. Item types were counterbalanced over these sequences, such that each sequence contained five tokens in each of the six experimental conditions. Each sequence was preceded by a short practice session comprising three words and three nonwords. These items were gated in exactly the same way as the other items and were selected from the fricative- and nasal-final fillers used in Experiment 4, one in each of the six experimental conditions.

Procedure. Participants were tested in separate carrels in a quiet room, with as many as 4 participants tested in a single session. They sat in front of a computer screen and keyboard. They were told that they would hear fragments of spoken items presented in portions of ever-increasing length. They were then asked to decide how they thought the fragment they had heard could be completed. To discourage listeners from adopting the strategy of providing only word responses, we emphasized that nonword responses were just as acceptable as word responses and that they should respond on the basis of what they had heard, not only on the basis of what possible lexical continuations they could think of. A host computer played the fragments out directly from hard disk to the participants over closed-ear headphones. This computer also controlled the operation of slave computers linked to each listener's keyboard and screen. After each fragment was presented, listeners were prompted by an arrow on their screens to type in their responses. After each response, they were further prompted to rate, on a scale from 1 to 10, how confident they were about the accuracy of their response. The slave computers logged both of these responses. After all listeners had responded or after 12 s if any listener had failed to respond within that time window, the host computer moved on to the next trial.

Results and Discussion

We adopted the same scoring procedure as Marslen-Wilson and Warren (1994). Each response to each fragment was placed in one of three categories: consistent with the presplice portion, consistent with the postsplice portion, or any other response. If the response had the correct initial consonants and the correct vowel, then the following consonant was categorized as being consistent with the pre- or postsplice portion if it had the same place of articulation as that portion, irrespective of voicing or manner of articulation. Thus, for example, the response *sloong* to *sloop* made from *slook* would be scored as being consistent with the presplice portion, whereas the response *sloom* would be scored as being consistent with the postsplice portion. Before this categorization, obvious typographic errors in the listeners' responses (such as a double *p* at the end of "sloop") were corrected, but only if there was no doubt about what listeners had intended.

Responses to words are plotted in Figure 1 and those to nonwords in Figure 2. In each case, the average percentage of responses at each gate consistent with both pre- and postsplice portions is plotted for identity-spliced items (W1W1 and N1N1), for items made by cross-splicing with words (W2W1 and W2N1), and for items made by cross-splicing with nonwords (N3W1 and N3N1). Note that for the identity-spliced items, responses consistent with the presplice portions are also responses consistent with the postsplice portions (e.g., in *sloop* made from two versions of *sloop*, bilabial responses are consistent with both the pre- and postsplice portions).

There was a strong similarity between the data for the cross-spliced items and those presented by Marslen-Wilson and Warren (1994, p. 663, Figures 3 and 4). As in the earlier study, there were three phases in the responses. The first corresponded to approximately the first six gates (i.e., up to the splice point). Overall, there was a gradual increase in the proportion of responses consistent with the presplice portions: As more information built up, particularly in the transitions out of the vowel, about the place of articulation of the following stops, the proportion of responses consistent with those stops increased. The second phase consisted of a plateau in which responses remained stable. For both word and nonword stimuli, in both the early phase and the plateau phase, there were more responses consistent with the presplice portions when those portions were made from words than when they were made from nonwords. The third phase began between 75 and 100 ms after the splice point. It was at this time that information about the final burst began to appear. There was a rapid crossover in the responses, with a sudden buildup of responses being consistent with the postsplice portions. For the complete items, almost all responses were consistent with the postsplice cues.

The overall pattern for the identity-spliced items varied between words and nonwords (note that Marslen-Wilson & Warren, 1994, did not report the results from their identity-spliced conditions). For words (W1W1), the pattern of responses matched that for the words cross-spliced with other words (W2W1) in the early and plateau phases. Then,
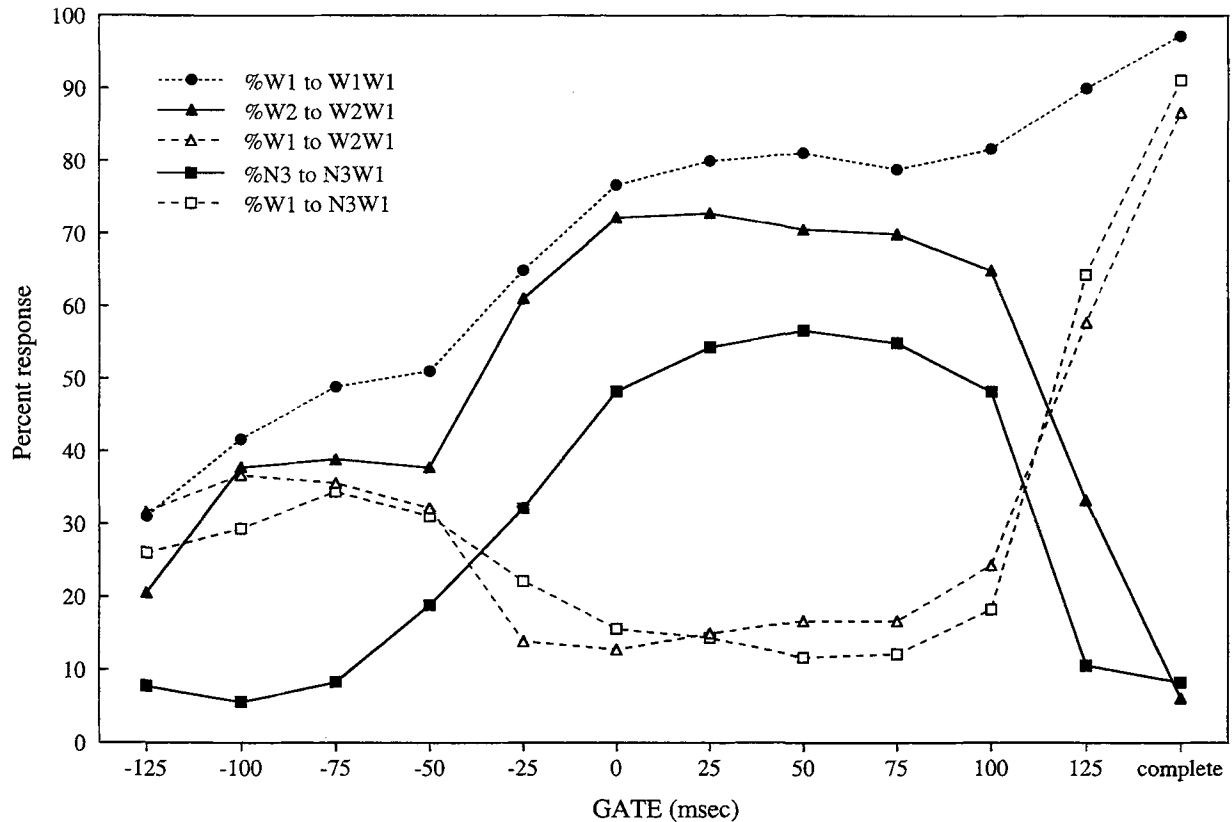
Figure 1. Percentage of responses consistent with the place of articulation cued by the presplice portions of the word stimuli (Word 1 + Word 1 [%W1 to W1W1], Word 2 + Word 1 [%W2 to W2W1], and Nonword 3 + Word 1 [%N3 to N3W1]), and those consistent with the postsplice portions (%W1 to W2W1 and %W1 to N3W1), plotted by gate (0 = splice point; complete = the entire word). For the W1W1 words, only one function is plotted because the place of articulation cued by the presplice portion in these items was the same as that cued by the postsplice portion.

as information about the final burst arrived in the last few gates, the proportion of consistent responses simply continued to increase to ceiling because there was no mismatch between the pre- and postsplice portions in these items. For nonwords (N1N1), the responses in the first and second phases matched those made to nonwords cross-spliced with other nonwords (N3N1). Then, as the final burst information arrived, the proportion of consistent responses to N1N1 nonwords again simply continued to increase to ceiling.

In the overall patterns, then, it is clear that there was a great deal of mismatching information in the cross-spliced items and that this information specifically signaled the place of articulation of the "missing" final stops (i.e., the stops that were originally produced at the ends of the words and nonwords used in cross-splicing). In the word condition, a comparison of the proportion of postsplice consistent responses in W1W1 items with the equivalent proportions in the cross-spliced items shows that in the earliest gates, there is little difference between the three functions. As more information in the vowel arrives, the functions diverge, such that the proportion of postsplice responses in the cross-spliced items drops toward zero, whereas the proportion of responses in the identity-spliced items rises to about 80%. It

is only after all of the final burst has been heard, at the last gate, that the functions again converge. It is clear why there were few post–splice-consistent responses to cross-spliced items in the plateau region: Listeners were instead producing responses consistent with the presplice portions. Statistical analyses showed that the differences between post–splice-consistent responses to fragments from W1W1 items were reliably different from those to fragments from the cross-spliced words. Details of these tests, and of all other tests conducted on Experiment 5's data, are presented in Appendix B.

In the nonword condition, the proportion of post–splice-consistent responses also differed reliably between the identity-spliced items (N1N1) and the two cross-spliced items. After overlap at the earliest gates, with few post–splice-consistent responses in any condition, only the proportion of postsplice responses to the N1N1 nonwords increased during the gates near the end of the presplice portions. Listeners did not produce post–splice-consistent responses to the cross-spliced nonwords because, as their responses reflect, there was phonetic information in the vowels consistent with the presplice place of articulation. Again, it is only
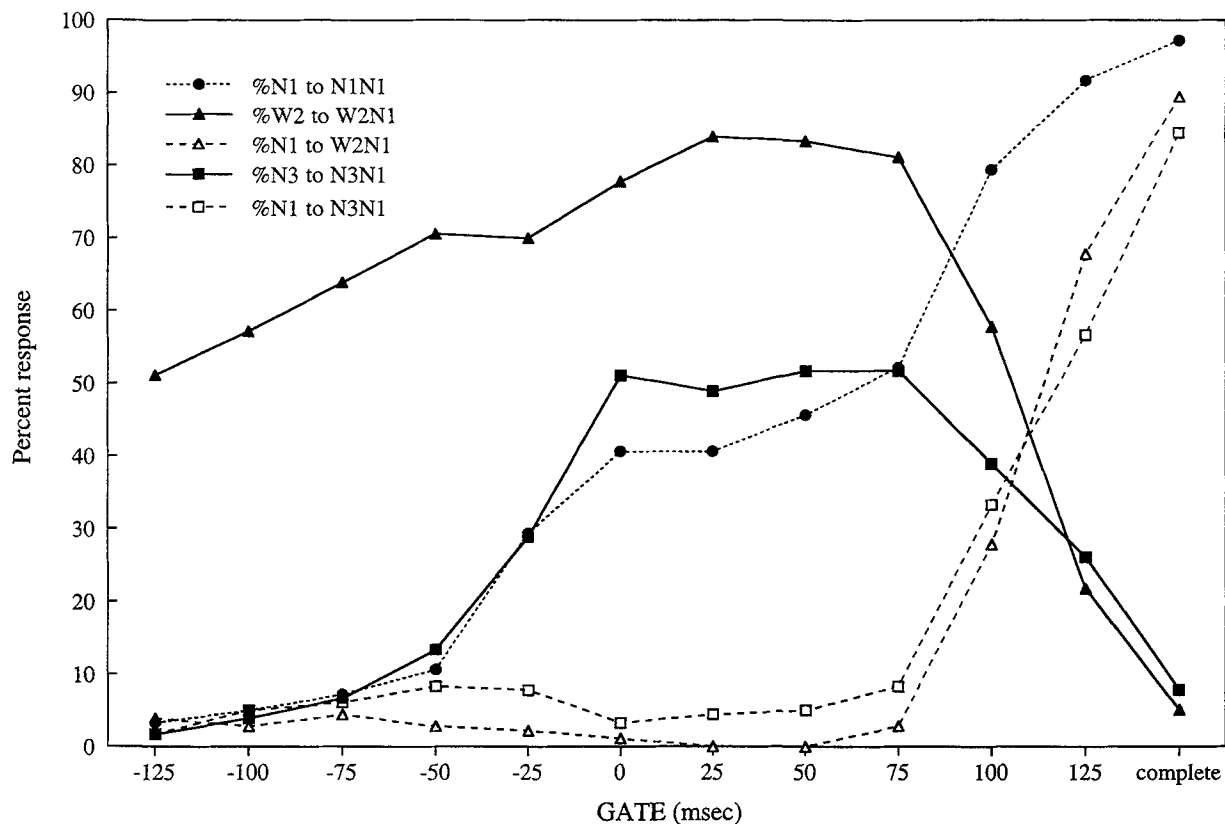
Figure 2. Percentage of responses consistent with the place of articulation cued by the presplice portions of the nonword stimuli (Nonword 1 + Nonword 1 [%N1 to N1N1], Word 2 + Nonword 1 [%W2 to W2N1], and Nonword 3 + Nonword 1 [%N3 to N3N1]), and those consistent with the postsplice portions (%N1 to W2N1 and %N1 to N3N1), plotted by gate (0 = splice point; complete = the entire nonword). For the N1N1 nonwords, only one function is plotted because the place of articulation cued by the presplice portion in these items was the same as that cued by the postsplice portion.

at the final gate that the three functions for the postsplice consistent responses converge.

As predicted, Experiment 5 produced strong lexical effects in both the word and the nonword conditions. In the postsplice responses, this can be seen by comparing the cross-spliced word conditions with the cross-spliced non-word conditions. Most markedly in the first four gates, but also continuing through to the end of the plateau phase, listeners were more likely to produce responses consistent with the postsplice stop in the word condition (where such responses would tend to be a word, the W1 itself) than to the nonword condition (where such responses would tend to be the nonword N1). Indeed, until the 10th gate, there were fewer than 10% post–splice-consistent responses at any one gate in the nonword condition. As predicted, particularly at the earliest gates in words, before sufficient information was available to signal an alternative place of articulation, listeners were biased to produce those words as responses (the same bias can be seen in Marslen-Wilson & Warren's, 1994, data).

Lexical effects in the pre–splice-consistent responses can be seen most clearly at the shortest gates both in the words

and, most obviously, in the nonwords. In the early portions of the word *sloop,* for example, there were two stop-final words consistent with the initial consonant-consonant-vowel sequence: One was *sloop* itself, and the other was the word used for cross-splicing (W2 *sloot*). In the first four gates, there were 20%–40% of responses consistent with the W2 word (and, as already discussed, similar proportions in all three conditions consistent with the W1 word). However, there were fewer than 10% of responses at any of these gates consistent with the presplice portion in N3W1 because such responses, if they were stop-final, would by definition have to be nonword responses. Note that effects such as these motivate the scoring scheme that we adopted. If only stop responses were scored as consistent, then this lexical bias would appear even more strongly.

In the nonword condition, there was only one stop-final word consistent with the early portion (i.e., the word used to make the W2N1 materials). This produced a large lexical bias in the short fragments. Even at the earliest gate, more than 50% of responses were consistent with the presplice portion of W2N1 items (e.g., *smet* responses to *smep* made from *smet*). This is in contrast to the W2W1 word items, in

which there were two stop-final lexical alternatives available and hence fewer pre–splice-consistent responses in the early gates. It is also in contrast to the N3N1 nonwords, in which in the early gates there were few pre–splice-consistent responses (again, all stop-final responses would have to be nonwords). Note that the lexical effect was much stronger in the early nonword gates here than in Marslen-Wilson and Warren's (1994) study. It would appear that the Dutch listeners, in spite of being encouraged to produce nonword responses, had a stronger lexical bias than their English counterparts.

In summary, the results of Experiment 5 show that the coarticulatory cues in the presplice portions of the cross-spliced items that were responsible for the mismatch effects found in Experiments 1–4 were as one would predict them to be: They cued the place of articulation of the stops that were excised during initial construction of the materials. Thus, the mismatch effects were due to the availability of perceptual evidence supporting specific stop consonants and were not general mismatch effects in which the perceptual evidence was simply inconsistent with the final stop without being supportive of any particular alternative.

The results of Experiment 5 also show that there was strong lexical involvement in gating performance. These lexical effects were related to the mismatch effects observed earlier, most particularly those found in Experiment 4. In the nonwords condition, as in all the previous experiments, performance on the cross-spliced items was different from that on the identity-spliced items because of the information in the presplice portions. Furthermore, there was a stronger effect on the nonwords cross-spliced with words than on nonwords cross-spliced with other nonwords, as in Experiments 3 and 4 because of lexical involvement (the influence of W2 words). Note also that responses to the N3N1 items showed the same pattern as in Experiment 4, that is, they were different from those to N1N1 items. They did not show the same pattern as in Experiment 3, where they were not reliably different from those to N1N1 items. In the shorter gates, the presplice coarticulatory cues in N3N1 items influenced performance, as they did in the phonetic judgment tasks. In lexical decision, however, because these cues had no consequences at the lexical level (both N3 and N1 were nonwords), no reliable difference was found. In the words condition, there was also a difference between the identity- and cross-spliced conditions, as observed in all four earlier experiments. However, in contrast to those experiments, there was also a lexical effect within the cross-spliced words. Performance on words cross-spliced with words was different from that to those cross-spliced with nonwords in gating, but not in either lexical or phonetic decision. This was because the gating task required that decisions be made to incomplete items, where there was more bottom-up support for W2 than W1, and thus the influence of W2 could be seen. In the complete items used in Experiments 1–4, however, there was more bottom-up support for W1, so the influence of W2 was then no longer seen. We return to this issue in the General Discussion section.

The results of Experiment 5 therefore support one of the central claims of this article: that lexical involvement in perceptual decision making depends on task demands. In lexical decision (Experiment 3), listeners were required to use lexical knowledge. In gating, although the use of the lexicon was not mandatory, listeners tended to use lexical knowledge to assist in their decision making. With exactly the same materials, however, no evidence of lexical involvement was observed in the phonetic tasks reported in Experiments 1 and 2. Nevertheless, when the task demands of Experiment 4 encouraged lexical involvement, listeners did tend to use lexical knowledge in phonetic decision making. In addition to being able to account for the detailed pattern of lexical effects, an adequate model of phonetic decision making must also be able to explain the variability of these effects. We outline such a model in the General Discussion section.

Experiment 6

There is one other question that might be raised regarding our materials. A key aspect of the results was the difference between the two types of the cross-spliced nonword (W2N1 and N3N1), a difference that appeared in phonetic decision (Experiment 4), in lexical decision (Experiment 3), and in gating (Experiment 5). We argue later that this effect was due to the influence of lexical entries. There is an alternative explanation, however: The presplice portions made from words (W2 in W2N1) may have contained stronger coarticulatory cues to the following stops than those presplice portions made from nonwords (N3 in N3N1). If this were the case, then the stronger mismatch effects in the nonwords cross-spliced with words may have nothing to do with lexical status per se but could instead be due to the presence of stronger acoustic cues created when the speaker produced the original word and nonword tokens. It is possible that the speaker found it easier to produce the words than the nonwords and thus tended to say them more rapidly and with more coarticulation.

There are two problems with this alternative explanation. The first is that it makes the wrong prediction for the cross-spliced *words*. If there were more coarticulatory cues in words than in nonwords, then there also ought to have been stronger mismatch effects in the words made by cross-splicing words than in the words made by cross-splicing with nonwords. No such effect was obtained in any of our first four experiments or in the lexical and phonetic decision studies reported by Marslen-Wilson and Warren (1994). The second problem is that there was no evidence from the durations of our materials that our speaker tended to produce words more rapidly, and thus with more coarticulation, than nonwords ($Ms = 595$ and 598 ms, respectively). Within the cross-spliced nonwords, those items made from cross-spliced words were in fact slightly longer (608 ms on average) than those made from cross-spliced nonwords (592 ms on average). Within the cross-spliced words, there was no difference between those items made from cross-spliced words and those made from cross-spliced nonwords ($Ms = 602$ and 603 ms, respectively).

One possibility remains, however: that there happened to be a difference in the amount of coarticulatory information

in the words and nonwords that were used to cross-splice with nonwords, but not between the words and nonwords that were used to cross-splice with words (and that these differences did not result in durational differences). Such an asymmetry (as far-fetched as it may seem) could result in the pattern of performance observed in Experiment 4. If the two types of cross-spliced words (W2W1 and N3W1) had equivalent amounts of mismatching information, although the nonwords made by cross-splicing with words (W2N1) had more mismatching information than those made by cross-splicing with nonwords (N3N1), then one would predict exactly what was observed: A difference in the nonwords but not in the words.

There is a problem even with this proposal, however. Why would different amounts of mismatch between W2N1 and N3N1 sometimes have an effect in phonetic decision (Experiment 4) but not always (Experiments 1 and 2)? This is particularly difficult to explain when, in all three of these experiments, there were main effects of mismatch (comparing the cross-spliced with the identity-spliced nonwords). Nevertheless, any interpretation of the different pattern of results between Experiments 1 and 2 and Experiment 4 depends on establishing whether there were indeed differential amounts of coarticulatory information in the cross-spliced materials. We therefore sought to address this concern in Experiment 6.

A simple forced-choice phonetic decision experiment was chosen to test whether there would be any differences in the strength of the coarticulatory cues in the cross-spliced items. Listeners heard only the vowels excised from each version of each experimental word and nonword. Thus, for example, they heard all three tokens of [o] from *sloop,* and all three tokens of [ε] from *smep,* and were required to decide, on the basis of only the information in the vowels, what the following segments had been. The response alternatives in each condition were chosen to remove lexical biases. If the vowel plus the correct following consonant formed a word (as with the [o] from N3 *slook* because of *ook,* also), then a response alternative was selected that formed another word (in this case, [m] was selected, consistent with the word *oom,* uncle). If, on the other hand, the vowel plus the correct following consonant formed a nonword, then a response alternative was selected that also formed a nonword (thus, e.g., the alternatives [p] and [t] could be used both with the [o] from W1 *sloop* and the [o] from W2 *sloot* because neither [op] nor [ot] are Dutch words). By equating response alternatives on lexical status, we hoped to obtain measures of the strength of coarticulatory cues in the vowels that were uncontaminated by lexical biases (unlike the gating data from Experiment 5).

If there were no differences in the amount of coarticulation in the different vocalic portions used for cross-splicing, then the proportion of correct responses to the vowels from all four cross-splice conditions (W2W1, N3W1, W2N1, and N3N1) should be the same and should be no different from the proportion of correct responses to the vowels from the two identity-splice conditions (W1W1 and N1N1). On the other hand, if there were stronger coarticulatory cues in the vowels taken from words than in the vowels taken from

nonwords, then there should be more correct responses to the vowels from the three word conditions (W1W1, W2W1, and W2N1) than to the vowels from the three nonword conditions (N1N1, N3W1, and N3N1). Finally, if the amount of coarticulation in the words and nonwords used in the word condition was equivalent but in the nonword condition there was more coarticulation in the vowels from words than in the vowels from nonwords, then there should be an interaction: a difference in the proportions of correct responses to vowels from W2N1 and N3N1 items, but no difference in the proportions of correct responses to vowels from W2W1 and N3W1 items.

Method

Participants. Thirty volunteers from the Max Planck Institute subject pool were paid for their participation. They were all native speakers of Dutch and had no known speech or hearing disorders. None of them had taken part in any of the earlier experiments.

Materials and design. The experiment was based closely on Experiment 1. There was only one change in the materials. Instead of the complete items used in Experiment 1, we used only the vowels from each item. All items, including practice materials, began at vowel onset (at zero crossings closest to the onset of the periodic energy associated with the vowel) and ended at vowel offset (i.e., at the splice point). There was only one change in the design, which was that each participant heard each vowel three times. The fully counterbalanced design of Experiment 1 was repeated three times within participants, such that each participant heard each of the 90 vowels (three repetitions of each of the three vowels associated with each word and nonword) in each of the three parts of the complete experiment.

Procedure. Participants were tested in acoustically isolated booths. The procedure was the same as in Experiment 1, except that participants were told that they would hear vowels excised from longer stretches of speech and that they were to listen carefully to each vowel and to decide on the basis of what they had heard what they thought the following sound had been. As in Experiment 1, two response alternatives appeared on the screen before each fragment was heard. Participants were required to choose which of these two alternatives was the most likely following sound.

There were nine different pairs of consonant alternatives. They were selected so that the lexical status of the vowel plus each consonant was matched. If the vowel plus the correct response [p,t,k] formed a word, then an alternative consonant was selected that also formed a word with the vowel; if the vowel plus the correct response formed a nonword, then the second consonant was selected to also form a nonword. Over the set of 90 vowels, 41 formed words with the correct following consonant and 49 formed nonwords. In addition to satisfying this constraint, the response pairs were selected such that they also met three other constraints. The first was that as few pairs as possible were used (given the limitations of the Dutch vocabulary to make alternative words or nonwords). The second was such that they were used more or less equally often. The third was such that the place of articulation of the incorrect consonant was different from that of the correct consonant (a constraint that was met in all cases bar one, the matching of [t] with [l], which was unavoidable because the only word alternative to pair with *uit* [out], which was the correct continuation in four items, was *uil* [owl]). The pairings, with the number of occasions they occurred, were P T (12); P K (10); P L (10); P S (10); T K (12); T L (10); M K (7); N K (10); and S K (9). As in Experiment 1, response mappings were held constant: P

always appeared on the left, as did M and N; K always appeared on the right, as did L; and the positions of T and S varied.

Results and Discussion

The mean percentages of correct responses in each of the six conditions are given in Table 6. We scored as correct those responses that corresponded to the consonant that the speaker had originally produced at the end of each item used to make the presplice portion in each condition. Missing responses made up only 1.5% of the data.

There are two important results. First, Experiment 6 again showed that the mismatch effects in Experiments 1–4 were due to coarticulatory cues that signaled the place of articulation of the stop that the speaker originally produced in each presplice portion. In other words, those earlier effects were not simply general mismatch effects. An overall accuracy of 80% suggests that there were strong cues to the following stops in the vowels, a finding consistent with the strong tendencies to produce responses consistent with the place of articulation of those stops that were observed in responses to the presplice gates in Experiment 5.

Second, Experiment 6 showed that there were no differences in the strength of the cues to subsequent stops in the vocalic portions of the experimental materials that could account for the pattern of data observed in Experiments 1–5. In particular, the present results are inconsistent with the possibility that the difference in mismatch effects found for cross-spliced nonwords but not for cross-spliced words in Experiments 3 and 4 might have been due to the different strengths of coarticulatory cues in the vowels used to make the cross-spliced items. ANOVAs showed that the only effect that was significant by both participants and items was a repetition effect, $F1(2, 54) = 11.99, p < .001; F2(2, 168) = 37.62, p < .001$. Accuracy improved by about 8% over the three parts of the experiment. Overall, performance was slightly more accurate on the vowels taken from the word condition ($M = 82\%$ correct) than on the vowels taken from the nonword condition ($M = 79\%$). This difference was significant only by participants, $F1(1, 27) = 7.02, p < .05; F2(1, 84) = 1.18, p > .1$. Similarly, the effect of splice condition was significant by participants but not items, $F1(2, 54) = 4.11, p < .05; F2 < 1$, and the pattern of differences between splice conditions was not the same for the vowels taken from the nonword conditions and those taken from the

word conditions, although, again, the interaction was significant only by participants, $F1(2, 54) = 11.88, p < .001; F2(2, 84) = 2.25, p > .1$. The interaction between splicing condition and experimental part was marginally significant, $F1(4, 108) = 2.24, p = .07; F2(4, 168) = 2.32, p = .06$. Planned comparisons within each condition showed that the differences among the vowels taken from the nonword condition were not significant, whereas in the word condition the only difference significant by both participants and items was that between the vowels from the W1W1 items and the vowels from the W2W1 items, $t1(29) = 4.48, p < .001; t2(28) = 2.40, p < .05$.

An analysis was also performed on the first part of the experiment alone (i.e., in which participants had no repetitions; they heard all 90 vowels once within that part). In the ANOVAs there were no differences significant by both participants and items. In the planned pairwise comparisons between the vowels from the nonword conditions, there were again no significant differences. For the vowels from the word conditions, however, both differences involving the vowels from the W2W1 items were now significant: for W1W1, $t1(29) = 3.25, p < .005, t2(28) = 2.34, p < .05$; for N3W1, $t1(29) = -2.54, p < .05, t2(28) = -2.06, p < .05$. The difference between vowels from W1W1 and N3W1 items was not significant.

If the different mismatch effects in the cross-spliced nonwords in Experiments 3 and 4 had been due to stronger coarticulatory cues in the W2N1 nonwords than in the N3N1 nonwords, there should have been better performance on the vowels from the W2N1 nonwords than on the vowels from the N3N1 nonwords. No significant difference was found. Although in the overall data there was a small trend in this direction, the trend reversed in the part of the experiment free of any repetition effects. Furthermore, for the explanation based on differing degrees of coarticulatory information to be able to explain the absence of a difference between the cross-spliced words in Experiments 1–4, there would have to be no difference between performance on the vowels from the cross-spliced words (W2W1 and N3W1). Although in the overall analysis no significant difference was found, in the first part of the experiment performance on the vowels from W2W1 items was reliably worse than that on the vowels from N3W1 items.

The failure to find support for this alternative explanation,

Table 6

Mean Percentage of Correct Identifications in Experiment 6 of the Consonants That Had Originally Followed the Excised Vowels

	Words			Nonwords		
Part	Word 1 + Word 1 (W1W1)	Word 2 + Word 1 (W2W1)	Nonword 3 + Word 1 (N3W1)	Nonword 1 + Nonword 1 (N1N1)	Word 2 + Nonword 1 (W2N1)	Nonword 3 + Nonword 1 (N3N1)
1	80	70	79	74	74	77
2	88	78	84	77	83	79
3	88	83	86	84	85	80
Overall	86	77	83	78	81	78

Note. Labels refer to the conditions in which the vowels had appeared in Experiments 1–5.

coupled with the fact that such an explanation would have difficulty explaining the different pattern of performance on the nonword items in Experiments 1 and 2 versus 3 and 4, led us to conclude that the results of Experiments 1–4 were due to the varying involvement of the lexicon, not to differences in the strength of coarticulatory cues between the cross-spliced conditions. We also concluded that the differences observed for both cross-spliced words and cross-spliced nonwords in the gating responses in Experiment 5 (more pre–splice-consistent responses to W2 than to N3 fragments in both words and nonwords, but to a greater extent in the nonwords than in the words) were not due to differences in amount of coarticulatory information between the words and nonwords used to cross-splice with nonwords versus the words and nonwords used to cross-splice with words. Instead, they were, as we argued earlier, due simply to lexical biases.

General Discussion

The results of our set of experiments are straightforward. Performance in Experiments 1–4 on identity-spliced words and nonwords was better than performance on cross-spliced items. This reflects the recognition system's sensitivity to fine-grained acoustic–phonetic information. When there was a mismatch between formant-transition information signaling one place of articulation and release burst information signaling another place of articulation, perceptual decisions became more difficult than when there was no mismatching information. Within the cross-spliced words in the first four experiments, performance was equivalent on words cross-spliced with other words and words cross-spliced with nonwords. Within the cross-spliced nonwords, however, there was in one case no mismatch effect: Nonwords cross-spliced with nonwords were not reliably harder to reject in lexical decision than identity-spliced nonwords. More important, in lexical decision (Experiment 3) and the version of phonetic decision modeled most closely on that used by Marslen-Wilson and Warren (1994; current Experiment 4), performance was worse on nonwords cross-spliced with words than on nonwords cross-spliced with other nonwords. There was no such difference when the same materials were used either in a simpler phonetic decision task (Experiment 1) or in a phoneme-monitoring task (Experiment 2). Control experiments showed that the coarticulatory information necessary to induce mismatch effects was clearly present in the presplice portions of the stimuli (Experiments 5 and 6) and was not differentially available in portions originally produced in words versus nonwords (Experiment 6). Together, these results support the argument that the lexicon acts to influence perceptual decisions to nonwords, albeit only under certain task conditions.

As we described in the introduction, the Race model (Cutler & Norris, 1979) has difficulty accounting for any lexical involvement in phonetic decisions to nonwords. Because all responses to nonwords should be made via the prelexical route, the lexical status of the cross-spliced portion in a nonword should not be able to influence those responses. The nonword data from Experiment 4, like the

equivalent phonetic decision data from Marslen-Wilson and Warren (1994), challenge the Race model. The model has also been challenged by other recent results (Connine et al., 1997; Newman et al., 1997).

As we also described earlier, the standard version of the TRACE model (McClelland & Elman, 1986) is unable to account for the absence of differences between the cross-spliced words found repeatedly both here and in Marslen-Wilson and Warren (1994). As simulations reported in Marslen-Wilson and Warren show, TRACE predicts that competition between the two words involved in words cross-spliced with other words (W2W1; such as that between *job* and *jog* in the example in Table 1) will make both lexical and phonetic decisions on such items harder than on words cross-spliced with nonwords (N3W1), in which there is no strong lexical competition. TRACE also predicts stronger inhibitory lexical effects on the cross-spliced nonwords than were obtained. It predicts a high false-alarm rate in lexical decisions to nonwords cross-spliced with words (e.g., saying "yes" to W2N1 *smob,* as if it were *smog,* almost as often as saying "yes" to words) and a very high error rate in phonetic decisions to these items (e.g., deciding that the final phoneme of W2N1 *smob* was [g] almost as often as making correct responses to [b] in N1N1 *smob*). The model is thus challenged by the results of all the experiments presented here and in Marslen-Wilson and Warren (1994). As discussed in the introduction, the model has also been challenged by other results (Frauenfelder et al., 1990; McQueen, 1991; Pitt & McQueen, 1998; Pitt & Samuel, 1993).

The third model considered in the introduction was the postlexical model proposed by Marslen-Wilson and Warren (1994) to account for their data. Although this model can account for the results of Experiments 3 and 4, it cannot account for the absence of inhibitory lexical effects on the nonwords cross-spliced with words in Experiments 1 and 2. If, as claimed in the model, all phonetic decisions are postlexical and decisions to nonwords are made on the basis of the activation of similar-sounding words, then it should simply not be possible to observe phonetic decision making without lexical involvement.

The postlexical model is similarly challenged by the variability of lexical influence seen in other studies (Cutler et al., 1987; Eimas et al., 1990; Eimas & Nygaard, 1992; Frauenfelder & Seguí, 1989; Seguí & Frauenfelder, 1986). The standard view on this variability is that task demands shift listeners' attention, causing lexical effects to come and go. When the experimental situation encourages the listener to use lexical knowledge, lexical effects emerge in the phonetic judgments. When the situation discourages the use of lexical knowledge, however, there are no lexical effects. In the present research, lexical influence on phonetic decisions on nonwords was observed only when the task was made more difficult (and more similar to Marslen-Wilson & Warren's, 1994, task in which lexical effects were also observed). The simpler tasks used in Experiments 1 and 2 appear to have discouraged lexical involvement. Listeners appear to have been able to focus attention on the acoustic–phonetic information (as the strong mismatch effects attest)

and ignore lexical information. This should not be possible in a model in which all phonetic decisions are based on the lexicon, however. This model thus cannot account for the present data or for other results showing variability of lexical effects in phonetic decision making or, as was pointed out in the introduction, for other recent data (Frauenfelder et al., 1990; McQueen, 1991; Pitt & Mc-Queen, 1998; Pitt & Samuel, 1993; Wurm & Samuel, 1997).

It would therefore appear that all three models that we have discussed should be rejected. Although it is clear that the lexicon can be involved in phonetic decision making, none of the previous accounts of how this involvement occurs is able to accommodate all the available evidence. What is required is a model that can deal with the variability of lexical influences on phonetic decision making and one that in particular can account for the constrained influence that the lexicon has on perceptual decisions made to nonwords.

The architecture that seems to us warranted for this task is one in which information from prelexical processing and information from the lexicon are independently made available to a decision-making stage, which can base its output on a merging of these two streams of information. We now briefly outline such a model—*the Merge model*—described in greater detail by Norris et al. (in press). This model, as we show, is consistent with all the available data.

Like the Race model, the Merge model is autonomous: Information flows bottom-up from the prelexical level of processing to the lexical level, but not top-down from the lexicon back to prelexical levels. Also as in the Race model, there are two routes through which information can flow for a phonetic decision: a prelexical route, via which acoustic–phonetic information extracted from the speech signal influences phonetic decisions, and a lexical route, via which phonological information retrieved from the lexicon influences phonetic decisions. Thus far, the Merge model is similar to the Race model. However, it differs in one fundamental way. Although in the Race model there is a simple first-past-the-post race, with independent procedures at the prelexical and lexical levels racing to be first to produce an output, there is no race in the Merge model. Instead, information from the two routes is integrated at a phonetic decision level. The model is implemented as a simple competitive network model (Norris et al., in press) in which activation from both prelexical and lexical representations is integrated at phoneme decision nodes. These nodes inhibit each other, and decisions can be taken when the activation of one node dominates that of all other nodes in the competition.

The principles on which the architecture of Merge is based have been instantiated most generally in the Fuzzy Logical Model of Perception (FLMP; Massaro, 1987, 1989, 1997; Massaro & Cohen, 1991; Massaro & Oden, 1995; Oden & Massaro, 1978). In FLMP, different sources of perceptual information are evaluated independently and then integrated for perceptual decisions; the model has been applied to performance on a variety of perceptual tasks. We believe that this architecture is the one that best accounts for

the available evidence on phonetic decision making. Note, however, that the Merge model is not therefore an instantiation of FLMP for phonetic decision making. On the one hand, although a central assumption of FLMP is that the same processes operate in many different perceptual domains, this architecture is postulated in Merge solely for explicit phonetic decision making. Merge therefore should not be seen as just a part of a general FLMP framework. On the other hand, FLMP offers no account of spoken word recognition as such, whereas Merge is designed to fit into the theory of spoken word recognition encapsulated in the Shortlist model (Norris, 1994). Indeed, as we discuss, it is the combination of the assumptions of integrative phonetic decisions and of word recognition based on lexical competition (as instantiated in Shortlist) that allows Merge to account for the available data (Norris et al., in press). In contrast, FLMP has no competition mechanism, or indeed any specification of the process of lexical access and word recognition, and thus cannot explain the observed competition effects. In brief, although Merge and FLMP are conceptually isomorphic with respect to a general architecture, they differ in that FLMP provides a broad-ranging account of the integration of different sources of information and Merge provides a detailed account of the processes of phonetic decision making based on the principles of competition and autonomy.

We further emphasize that Merge is not simply a part of Shortlist. The shortlist model is a model of the basic process of spoken word recognition; Merge is a model specifically of the process of making phonetic decisions. Word recognition does not necessarily entail phonetic decision making. Illiterate individuals (Liberman, Shankweiler, Liberman, Fowler, & Fisher, 1977; Morais, Bertelson, Cary, & Alegria, 1986; Morais, Cary, Alegria, & Bertelson, 1979), or those not taught an alphabetic script (Read, Zhang, Nie, & Ding, 1986), cannot perform tasks such as phoneme monitoring. They can, of course, recognize spoken language, but their ability to perform metalinguistic tasks such as phoneme identification does not follow as an automatic consequence. In other words, they have "Shortlist," but not "Merge."

Three features of Shortlist are crucial for the compatibility of the Merge and Shortlist models. First and foremost, both Shortlist and Merge are autonomous; in both accounts, the lexicon is unable to influence prelexical processing. Second, a level of prelexical processing is assumed in Shortlist that generates an abstract representation of the input that is then used for lexical access. These abstract prelexical representations (phonemes in current implementations of Shortlist) can also provide the prelexical input to the phoneme decision nodes in Merge. Third, word recognition in Shortlist is based on competition between activated candidate words. These lexical representations provide the lexical input to the decision nodes in Merge. The fact that lexical candidates compete via lateral inhibition proves crucial in the Merge model's account of the data presented here and in Marslen-Wilson and Warren (1994).

Merge can account for all the basic lexical effects in phonetic decision tasks. Responses to targets in words will tend to be faster than those to targets in nonwords because lexical-level activation will boost the phoneme decision nodes of targets in words. Lexical activation should also produce lexical biases in the categorization of ambiguous fricatives and should produce more phonemic restoration in words than in nonwords. Although it is autonomous, the Merge model, in contrast to the Race model, can also explain lexical effects in nonwords. As a nonword becomes more like a word, lexical-level activation will increase, thus speeding responses to targets in more wordlike nonwords (as in the results of Connine et al., 1997).

Merge can also account for the results of Experiment 4 and the phonetic decision results of Marslen-Wilson and Warren (1994). The model predicts the lexical inhibition observed in nonwords cross-spliced with words, such as *smob* made from *smog* and *smob:* Activation of *smog* will increase the activation of [g], which will in turn inhibit the activation of the [b], making decisions harder than when the nonword is made from two nonwords. The model predicts the lack of a lexical effect in the cross-spliced words because of the competition operating at the lexical level. Competition between, for example, *job* and *jog* given a version of *job* made by cross-splicing those two words will rapidly settle on *job* and the activation of *jog* will drop. The final activation for *job* is thus unaffected by the presence of the competitor *jog*. The lexical level will therefore support the activation of the [b] phoneme decision node both with words cross-spliced with words and words cross-spliced with nonwords. Simulations of the phonetic and lexical-decision data with the implementation of Merge are presented in Norris et al. (in press).

Note that this prediction differs from that made by TRACE for cross-spliced words. In standard versions of TRACE lexical competition tends not to produce a single dominant candidate word as rapidly as in Shortlist and Merge. Thus, although both models instantiate lexical competition, inhibitory lexical effects in cross-spliced words are predicted by TRACE, but not by Merge. However, as pointed out earlier, an interactive model with TRACE architecture but different network parameters from the standard model can account for the present data and those of Marslen-Wilson and Warren (1994; see Norris et al., in press). The failure of TRACE to capture the word results led Marslen-Wilson and Warren to reject lexical competition. They argued that the effects of lexical competitors (e.g., *jog* given the W2W1 *job*) ought to be observed if there is an active process of competition. However, Merge correctly predicts the absence of a differential effect in the cross-spliced words precisely because of active competition. It is because of the competition between, for example, *job* and *jog* that *job* can rapidly dominate the activation pattern, leaving no trace of that competition process in the responses made to the cross-spliced words. It is only if competition were less efficient that effects of the activation of *jog* given W2W1 *job* would be observed. Note that the effects of W2 word activation were indeed observed in gating

(Experiment 5), where with shorter fragments the proportion of responses consistent with the presplice place of articulation was greater for W2W1 items than for N3W1 items (see Figure 1). This is because for the shorter fragments there was more perceptual evidence for the W2 words than for the W1 words. The gating task thus allowed us to observe the activation of W2s before the point at which they were later overwhelmed by W1s in the competition process.

Merge explains the variability of the inhibitory lexical effect in the cross-spliced nonwords (and indeed the variability of lexical effects in general) in much the same way as the Race model. When task demands discourage the use of lexical knowledge, phonetic decisions are made solely on the basis of the prelexical route; lexical activation does not contribute to decision node activation. This might appear to be similar to the mechanism adopted in the TRACE model, in which top-down connections are switched on and off according to task demands. These mechanisms are not equivalent, however. In TRACE, the attentional mechanism causes permanent connections that are part of the normal word recognition system to be turned on and off. In Merge, on the other hand, the connections to the phoneme decision nodes are not part of the normal word recognition system, and they are only built on the fly when the listener is required to make phonetic decisions. When task demands encourage the use of lexical knowledge, as in Experiment 4, connections are built from both prelexical and lexical levels. When the use of lexical knowledge is discouraged, as in Experiments 1 and 2, only connections from the prelexical level are constructed.

Another recent result that can be explained by Merge is that reported by Pitt and McQueen (1998). On the assumptions that the prelexical processing level is sensitive to transitional probabilities and the process responsible for compensation for coarticulation has its locus at that level, Merge can account for Pitt and McQueen's results in a straightforward way. Transitional probability can bias fricative activation at the prelexical level and therefore also modulate the compensation process. The outcome of these processes can be fed via the prelexical route to the phoneme decision nodes. The lexicon can also influence fricative decisions via the lexical route to the decision nodes, but, as observed, this lexical bias will not have any effect on stop identification because the prelexical level is immune to lexical influences.

The focus of this research has been on models of phonetic decision making and on the flow of information through the normal speech recognition system. We have had little to say about the nature of prelexical representations. In contrast, Marslen-Wilson and Warren (1994) argued not only that all phonetic decisions are based on output from the lexicon but also that lexical access is based on featural rather than phonemic representations. In their view, acoustic–phonetic features are extracted prelexically, with direct lexical access from these representations without any intervening phonemic level. Our view is that the available data do not allow any strong conclusions to be drawn about the nature of

prelexical representations. The Merge model is accordingly neutral with respect to this issue. Although its current instantiation has phonemic prelexical representations (which serve as input both to the lexicon and directly to the phoneme decision units), Merge could work just as well with prelexical featural representations.

One line of argument in support of direct featural access was presented by Streeter and Nigro (1979), in their study of acoustic–phonetic mismatch. They found effects of mismatch (slower lexical-decision responses) in spliced versus unspliced words, but not in spliced versus unspliced nonwords. They argued that if featural information were integrated prelexically into phonemic representations, delays in that integration caused by mismatching information should be observed regardless of whatever the lexical status of the item. Whalen (1991) reached exactly the opposite conclusion: that featural mismatches are resolved at a prelexical phonemic level. This conclusion was based on his finding that both words and nonwords showed mismatch effects. The effects were larger for words than nonwords in lexical decision, but the relevant interaction was not significant. In auditory naming, at least for slow responses, mismatch effects were larger for nonwords than for words, but, again, the interaction was not significant.

The present results, and those of Marslen-Wilson and Warren (1994), offer an explanation for this mixed pattern for nonwords observed in the earlier studies. They show that in lexical decision there were reliable mismatch effects only for nonwords cross-spliced with other words (the differences between identity-spliced nonwords and nonwords cross-spliced with other nonwords were not significant in either study). Marslen-Wilson and Warren (1994), consistent with Streeter and Nigro (1979), argued that the absence of a mismatch effect in lexical decisions to the nonwords cross-spliced with other nonwords supported their view that lexical access is based on features, with no intermediate phonemic representations. However, Merge simulates this data pattern, even though it has phonemic prelexical representations. This is because, according to Merge, all mismatch effects observed in lexical decision are due to the activation of lexical nodes (or the lack of such activation) and so do not depend on the nature of prelexical representations. The presence of an activated lexical hypothesis (i.e., in the W2N1 case) is crucial for interference because of the mismatching information to be observed in a task tapping into the lexical level. In the case of nonwords cross-spliced with other nonwords, there is no strongly activated lexical node, so responses are as fast as with identity-spliced nonwords (i.e., all parts of an N3N1 item are as consistent with a "no" response as are all parts of an N1N1 item).

The absence of mismatch effects in lexical decisions to nonwords cross-spliced with other nonwords thus does not support the view that there is direct lexical access based on features. The only class of model of prelexical representation that is ruled out is one in which classification into phonemic units is completed before lexical access. In

such a model, lexical access must wait until phonetic classification is complete, and thus mismatch effects in lexical decision should occur whenever there is subcategorical mismatch, irrespective of the lexical status of the components of cross-spliced items. However, there is continuous mapping of information onto the lexicon in Merge, so this problem does not occur. In the current instantiation of Merge, this is continuous mapping of phonemic information, but the data do not demand that this mapping must be phonemic.

In the phonetic judgment tasks in our research and that of Marslen-Wilson and Warren (1994), in contrast, there were always reliable mismatch effects for nonwords cross-spliced with other nonwords. Merge also simulates this data pattern. This is because decisions in the lexical-decision task are modeled on the basis of the activation of lexical hypotheses, whereas decisions in phonetic tasks are modeled in the Merge on the basis of the activation of the phoneme decision nodes. The activation of these nodes is sensitive to any mismatching acoustic–phonetic information in the signal, even if it is not consistent with any particular lexical hypothesis. Again, these results do not speak to the issue of prelexical representations. The effects are considered in Merge to be due to the operation of phoneme decision nodes, but no strong claim can be made about the representations that serve as input to those nodes. Again, as long as Merge allows for the continuous mapping of information from the prelexical level onto the phoneme nodes, the model can account for the results, irrespective of the nature of the prelexical representations.

We have argued that the Merge model can account for the present data and a range of other phenomena. The other models of phonetic decision making that have been discussed, on the other hand, cannot account for the present data, and are each challenged by a number of other recent findings. If researchers are to understand the process of spoken language comprehension, then they need accurate models of performance in laboratory tasks that are used to tap into that process. Merge offers a new account of performance in tasks that require phonetic decisions; it should be of value in the development of the understanding of speech recognition. Furthermore, the Merge model is consistent with the account of spoken word recognition offered by the Shortlist model. Both approaches embody the claim that the recognition system has an autonomous architecture, through which information flows only bottom-up.

References

Burnage, G. (1990). CELEX: A guide for users. Nijmegen, The Netherlands: CELEX.

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.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1995). Bottom-up connectionist modelling of speech. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds.), Connectionist

models of memory and language (pp. 289–310). London: UCL Press.

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.

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

Cutler, A., & Norris, D. (1979). Monitoring sentence comprehension. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 113–134). Hillsdale, NJ: Erlbaum.

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. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27,* 143–165.

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

Foss, D. J., & Gernsbacher, M. A. (1983). Cracking the dual code: Toward a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior, 22,* 609–632.

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., Baayen, R. H., Hellwig, F. M., & Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language, 32,* 781–804.

Frauenfelder, U. H., & Seguí, J. (1989). Phoneme monitoring and lexical processing: Evidence for associative context effects. *Memory & Cognition, 17,* 134–140.

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.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fisher, F. W. (1977). Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & D. L. Scarborough (Eds.), *Towards a psychology of reading* (pp. 207–226). Hillsdale, NJ: Erlbaum.

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r." *Cognition, 24,* 169–196.

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. (1984). Function and process in spoken word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 125–149). Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual

representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101,* 653–675.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological enquiry.* 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., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology, 23,* 558–614.

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., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 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., & Pitt, M. A. (1996). Transitional probability and phoneme monitoring. *Proceedings of the Fourth International Conference on Spoken Language Processing, 4,* 2502–2505.

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.

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

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7,* 323–331.

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. G. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Norris, D. G., McQueen, J. M., & Cutler, A. (in press). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences.*

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172–191.

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. (1990). Attentional allocation during speech perception: How fine is the focus? *Journal of Memory and Language, 29,* 611–632.

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.

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

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35,* 606–621.

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

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.

Seguí, J., & Frauenfelder, U. (1986). The effect of lexical constraints upon speech perception. In F. Klix & H. Hagendorf (Eds.), *Human memory and cognitive capabilities: Mechanisms and performances* (pp. 795–808). Amsterdam: North-Holland.

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.

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

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

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

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.

Appendix A

Experimental Materials, With Phonetic Transcriptions of the Base Words and Nonwords

Words		
Base words (W1)	Matched words (W2)	Matched nonwords (N3)
knaap (lad) [knap]	knaak (Nfl 2.50 coin)	knaat
knoop (knot) [knop]	knook (bone)	knoot
sloop (pillow-case) [slop]	sloot (ditch)	slook
strip (cartoon) [strɪp]	strik (bow)	strit
strop (rope) [strɔp]	strot (throat)	strok
breed (broad) [bret]	break (break)	breep
buit (loot) [bœyt]	buik (belly)	buip
kreet (scream) [kret]	kreek (creek)	kreep
schut (lock) [sxʉt]	schub (scale)	shuk
smaad (indignity) [smat]	smaak (taste)	smaap
draak (dragon) [drak]	draad (thread)	draap
prak (mash) [prɑk]	prat (pride)	prap
sik (goatee) [sɪk]	sip (glum)	sit
truck (truck) [trʉk]	trut (frump)	trup
vlok (flake) [vlɔk]	vlot (fluent)	vlop

Nonwords		
Base nonwords (N1)	Matched words (W2)	Matched nonwords (N3)
fuip [fœyp]	fuik (net)	fuit
jeup [jøp]	jeuk (itch)	jeut
kleep [klep]	kleed (garment)	kleek
kwiep [kʋip]	kwiek (brisk)	kwiet
smep [smɛp]	smet (stain)	smek
get [xɛt]	gek (mad)	gep
spraat [sprat]	spraak (speech)	spraap
spreut [sprøt]	spreuk (proverb)	spreup
struit [strœyt]	struik (shrub)	struip
troet [trut]	troep (mess)	troek
blook [blok]	bloot (naked)	bloop
bruuk [bryk]	bruut (brute)	bruup
fruik [frœyk]	fruit (fruit)	fruip
krijk [krɛik]	krijt (chalk)	krijp
sloek [sluk]	sloep (boat)	sloet

Appendix B

Inferential Statistics for Experiment 5

1. *Identity- Versus Cross-Spliced Words.* In analyses of variance (ANOVAs) we compared the proportion of responses consistent with the postsplice stop in the Word 1 + Word 1 (W1W1) condition with the mean of these proportions in the two cross-spliced conditions over all 12 gates. There was a highly significant effect of gate, $F1(11, 363) = 139.78, p < .001; F2(11, 154) = 29.16, p < .001$, and a highly significant effect of condition, $F1(1, 33) = 496.58, p < .001; F2(1, 14) = 80.05, p < .001$, with more post–splice-consistent responses in the W1W1 condition. There was also an interaction between gate and condition, $F1(11, 363) = 76.07, p < .001; F2(11, 154) = 26.55, p < .001$. Pairwise comparisons at each gate (two-tailed t tests, as in all subsequent pairwise comparisons reported in this experiment) showed that the difference between the identity- and cross-spliced conditions was reliable ($p < .05$ by both participants and items, and in all subsequent analyses) at all gates except the first two gates and the last.

2. *Identity- Versus Cross-Spliced Nonwords.* The proportion of responses consistent with the postsplice stop in the Nonword 1 + Nonword 1 (N1N1) condition was compared with the mean of these proportions in the two cross-spliced conditions. Again, there were highly significant effects of gate, $F1(11, 363) = 540.67, p < .001; F2(11, 154) = 63.11, p < .001$, and condition, $F1(1, 33) = 245.93, p < .001; F2(1, 14) = 36.38, p < .001$, with more post–splice-consistent responses in the N1N1 condition. There was also an interaction between gate and condition, $F1(11, 363) = 57.03, p < .001; F2(11, 154) = 13.87, p < .001$. Pairwise comparisons at each gate showed that the difference between the identity- and cross-spliced nonwords was reliable at all gates except the first four.

3. *Cross-Spliced Words Versus Nonwords: Post–Splice-Consistent Responses.* There was a lexical effect: There were more post–splice-consistent responses in the word condition than in the nonword condition, $F1(1, 33) = 163.28, p < .001; F2(1, 28) = 10.99, p < .001$. There was no difference between items cross-spliced with words and those cross-spliced with nonwords, nor did this factor interact with the lexical effect. There was, however, a significant effect of gate, $F1(11, 363) = 584.59, p < .001; F2(11, 308) = 75.52, p < .001$, and an interaction between the lexical effect and gate, $F1(11, 363) = 32.91, p < .001; F2(11, 308) = 4.81, p < .001$. Pairwise comparisons of the mean proportion at each gate of post–splice-consistent responses in the word and nonword conditions (collapsing over the two types of cross-spliced items in each condition) showed that the lexical effect was reliable for the first eight gates.

4. *Cross-Spliced Words Versus Nonwords: Pre–Splice-Consistent Responses.* There was a lexical effect in the pre-splice-consistent responses, and this was larger in the nonwords than in the words. The responses consistent with the presplice portions for the first six gates (i.e., up to the splice point) were included in the ANOVAs. Because these were compatible with analyses performed by Marslen-Wilson and Warren (1994), we used the same factor labels. Factors were gate, lexical status (whether the presplice portions were words or nonwords), and competitor (whether the fragments came from the word set, in which there were two word competitors, or from the nonword set, in which there was only one lexical competitor). There was an effect of gate, $F1(5, 165) = 163.64, p < .001; F2(5, 140) = 49.94, p < .001$, an effect of lexical status, $F1(1, 33) = 211.20, p < .001; F2(1, 28) = 44.97, p < .001$, and a competitor effect, $F1(1, 33) = 25.83, p < .001; F2(1, 28) = 5.55, p < .05$. The lexical effect interacted with gate, $F1(5, 165) = 8.87, p < .001; F2(5, 140) = 2.80, p < .05$, but it also interacted with competitor, $F1(1, 33) = 32.71, p < .001; F2(1, 28) = 4.57, p < .05$. That is, the difference between Word 2 (W2) and Nonword 3 (N3) responses was larger for nonwords than for words.

Pairwise comparisons showed that over these six gates, the difference between W2 and N3 was reliable in the nonwords at all gates and at Gates 2, 3, and 5 in the words (the differences at Gates 1, 4, and 6 were significant at the .05 level by participants and at the .1 level by items). Listeners were more likely to produce pre–splice-consistent responses in the W2 conditions than in the N3 conditions, but this tendency was more marked when there was only one stop-final word consistent with the fragment than when there were two. Note also that the proportion of responses consistent with N3 nonwords was equivalent in the word (Nonword 3 + Word 1, N3W1) and nonword (Nonword 3 + Nonword 1, N3N1) conditions (no reliable pairwise comparisons at any of the six gates), whereas the proportion of responses consistent with W2 words was reliably higher in the nonword (Word 2 + Nonword 1, W2N1) than in the word (Word 2 + Word 1, W2W1) conditions at Gates 1, 3, and 4 and significant by participants but marginal by items ($p < .1$) at Gate 2. In the N3 conditions, there was no bias to produce N3 responses more reliably in one condition than the other, but in the W2 conditions, the lexical bias was stronger when there was only one stop-final word consistent with the fragment (i.e., a W2 word in the nonword condition such as *smet*) than when there were two stop-final words (i.e., the W2 words in the word condition such as *sloot* have to compete with the W1 words such as *sloop*).