Psychological Review
2008, Vol. 115, No. 2, 357–395

0033-295X/08/$12.00 DOI: 10.1037/0033-295X.115.2.357

Shortlist B: A Bayesian Model of Continuous Speech Recognition

Dennis Norris
Medical Research Council, Cognition and Brain Sciences Unit

James M. McQueen
Max Planck Institute for Psycholinguistics

A Bayesian model of continuous speech recognition is presented. It is based on Shortlist (D. Norris, 1994; D. Norris, J. M. McQueen, A. Cutler, & S. Butterfield, 1997) and shares many of its key assumptions: parallel competitive evaluation of multiple lexical hypotheses, phonologically abstract prelexical and lexical representations, a feedforward architecture with no online feedback, and a lexical segmentation algorithm based on the viability of chunks of the input as possible words. Shortlist B is radically different from its predecessor in two respects. First, whereas Shortlist was a connectionist model based on interactive-activation principles, Shortlist B is based on Bayesian principles. Second, the input to Shortlist B is no longer a sequence of discrete phonemes; it is a sequence of multiple phoneme probabilities over 3 time slices per segment, derived from the performance of listeners in a large-scale gating study. Simulations are presented showing that the model can account for key findings: data on the segmentation of continuous speech, word frequency effects, the effects of mispronunciations on word recognition, and evidence on lexical involvement in phonemic decision making. The success of Shortlist B suggests that listeners make optimal Bayesian decisions during spoken-word recognition.

Keywords: Spoken-word recognition, Bayesian modeling, continuous speech

Sherlock Holmes seemed to know something about the power of Bayesian decision making when he said to Watson: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890, Ch. 6, p.93). We argue here that listeners know this and much more about Bayesian decision making. Specifically, we suggest that, in order to perceive continuous streams of speech as sequences of discrete words, listeners behave as optimal Bayesian recognizers. In support of this claim, we present a new computational model that gives a simple and elegant account of the main empirical findings on spoken-word recognition. This leads to a complete reconceptualization of the word recognition process. The model introduces new ways of thinking about word frequency, how words are matched to the perceptual input, lexical competition, and lexical activation.

In the literature on spoken-word recognition, there is almost universal acceptance that recognition involves a process whereby the perceptual input activates lexical representations, and these activated representations then compete with each other to determine an appropriate segmentation of the input into words (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; for reviews see Frauenfelder & Floccia, 1998; Gaskell & Marslen-Wilson, 2002; McQueen, 2007; McQueen, Norris, & Cutler, 1994). The Bayesian framework we advocate here leads us to abandon the concept of lexical activation.

Activation has been an extremely valuable metaphor in spoken-word recognition research. It is embodied in two of the most influential models, TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994), both of which are based on interactive-activation networks. In such networks each word (or lexical candidate) is represented by a single node, which is assigned an activation value. The activation of a lexical node increases as the node receives more perceptual input and decreases when subject to inhibition from other words. But what is the explanatory value of the concept of activation? Beyond the general notion that bigger is better, activation does not directly determine the behavior of these models. In particular, neither reaction time nor response probabilities can be derived directly from activation values without additional assumptions. Furthermore, as we will explain later, activation is not actually a core part of the theory motivating Shortlist. The interactive-activation network is simply a convenient mechanism for performing some of the computations required by Shortlist or, indeed, by any theory of spoken-word recognition. In this article, therefore, we replace the interactive-activation network in Shortlist with Bayesian computations that provide a more direct implementation of the theoretical principles underlying the model.

One consequence of adopting the Bayesian perspective is that activation is replaced by the concepts of likelihood and probability, both of which have a clear formal interpretation. In the case of probability, this can be linked directly to measures of behavior. Thus, although one goal of this article is to present a new version of Shortlist, the more general aim is to argue for the benefits of the Bayesian perspective in understanding spoken-word recognition.

The central theoretical claims embodied in Shortlist can be derived from a simple higher-level claim, namely that human

Dennis Norris, Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, United Kingdom; James M. McQueen, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands.

We thank Mark Pitt, Dom Massaro, Paul Luce, Sachiko Kinoshita, Delphine Dahan, and Anne Cutler for valuable comments on previous versions of this article, and we thank Mark Pitt for suggesting the name Shortlist B to us. We also thank Roel Smits for his help with this project, in particular for his work on the model's input, and Matthias Sjerps for his assistance with the figures.

Correspondence for this article should be addressed to Dennis Norris, MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, United Kingdom E-mail: dennis.norris@mrc-cbu.cam.ac.uk

listeners adopt a near optimal strategy for recognizing speech. Here we cast this in terms of the theoretical principle that human listeners approximate optimal Bayesian decision makers. We then ask how much of what we know about human speech perception can be explained purely on the basis of this simple premise. The answer, we will argue, is that the principle of optimality gives us a better explanation of a broader range of phenomena than any existing model. Importantly, this explanation follows automatically from this basic principle in such a straightforward way that the resulting computational model is much simpler than any of the competitor models.

The starting point for this work is an analysis of the computational problems that must be solved in order to complete the task of speech recognition successfully (Marr, 1982)—in Anderson's (1990) terms, a Rational Analysis. We have argued in detail elsewhere (Scharenborg, Norris, ten Bosch, & McQueen, 2005), that, for spoken-word recognition to be successful, the listener must solve a number of specific computational problems that arise from the nature of the acoustic speech signal and from the structure of the vocabulary. The original Shortlist model (Norris, 1994) and its developments (Norris, McQueen, Cutler & Butterfield, 1997; Scharenborg et al., 2005) offer algorithms for how these computational-level problems are solved. According to Anderson's (1990) Principle of Rationality, however, the overarching constraint on all information-processing theories is that the cognitive system should be optimized with respect to its goals. Spoken-word recognition, therefore, should be optimal in the sense that a listener's behavior should approach the best that it can be, given the constraints imposed both by the speech signal and by phonological and lexical knowledge. As we now show, a major attraction of a Bayesian model of spoken-word recognition is that its behavior is guaranteed to be optimal in exactly this way.

If speech consisted of strings of completely unambiguous isolated words, then optimal word recognition would simply entail the selection of the sequence of words that matched the current input. As we describe in more detail below, however, the speech signal is phonetically ambiguous, and it does not consist of a series of discrete words—instead, speech sounds unfold over time in a quasi-continuous stream. Bayesian inference allows word recognition to be optimal in the face of this ambiguity by combining the perceptual evidence that is available (no matter how ambiguous) with knowledge of the prior probabilities of words. Bayes's theorem ensures that the way these sources of information are combined is optimal. First, with completely unambiguous input the best-matching word will always be selected. Second, word recognition will also be optimal with ambiguous input. As the ambiguity of the input increases, the influence of the prior probability of the words will also increase: As perceptual uncertainty increases, the smart money goes on the events which are more likely to occur.

Our claim that human listeners approximate optimal Bayesian decision makers requires us to specify exactly what function listeners are attempting to optimize. This means we have to specify the listeners' task, as it is the task that determines the function that must be optimized. Norris (2006) provides an extensive discussion of optimality in tasks involving visual word recognition. In Norris's Bayesian Reader model, evidence is accumulated from the input by means of a noisy sampling process. In sequential sampling models, there are two standard definitions of optimality, depending on the task (see Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006, for a recent discussion of optimal decision making in sequential sampling models). In tasks requiring speeded decisions, optimality is defined as making the fastest decision possible while achieving a given level of accuracy (e.g., 95% correct). In tasks requiring a response based on a fixed amount of perceptual evidence (e.g., perceptual identification), optimality is defined as selecting the response (word) that is most probable, given the available input.

In both cases the primary requirement is to calculate the conditional probability of each word given the available input, and this conditional probability is exactly what Bayes's theorem allows us to calculate:

$$P(Hypothesis_i|Evidence) = \frac{P(Evidence|Hypothesis_i) \times P(Hypothesis_i)}{\sum_{j=1}^{j=n} P(Evidence|Hypothesis_j) \times P(Hypothesis_j)} \quad (1)$$

Bayes's theorem specifies how to revise or update beliefs in the light of new evidence. Given some initial belief about the probability of a hypothesis being true, *P(Hypothesis)*, Bayes's theorem tells us how to update this prior probability and compute the posterior probability of the hypothesis being true, given the evidence, *P(Hypothesis|Evidence)*. *P(Evidence|Hypothesis)* is the likelihood of the evidence given the hypothesis. When modeling word recognition, the hypotheses correspond to words:

$$P(Word_i|Evidence) = \frac{P(Evidence|Word_i) \times P(Word_i)}{\sum_{j=1}^{j=n} P(Evidence|Word_j) \times \mathrm{P}(Word_j)} \quad (2)$$

where n is the number of words in the lexicon. Bayes's theorem therefore gives us exactly the information we need: The conditional probability of each word, given the available evidence. This holds regardless of whether the input is spoken or written. It makes no difference whether the evidence is being accumulated by noisy sampling, as in the Bayesian Reader, or because spoken input is arriving over time, as in Shortlist B. Once these conditional probabilities have been computed, optimal decisions can be made either by selecting the most probable word at a given time or by making a response when the probability of a word exceeds a predetermined probability threshold. The focus of most of the simulations here will be on calculation of the relevant probabilities; but, in the final simulations, we will estimate actual reaction time (RT) measures in the tasks of lexical decision and phonetic categorization.

There is one important qualification to the claim for optimality we have just presented. The probabilities calculated will only be true probabilities to the extent that the listener's prior beliefs are a true reflection of real probabilities. For example, if a listener has a completely mistaken belief as to the probability of encountering a particular word in a particular context, the posterior probability he or she assigns to that word will no longer correspond to the actual probability. Listeners' erroneous beliefs may lead them to make inaccurate decisions. In most of the simulations presented here, we make the simplifying assumption that the appropriate prior probabilities can be estimated from standard measures of word frequency. However, we will also need to take account of the fact that

word frequency represents only a fraction of the knowledge that listeners have at their disposal when recognizing continuous speech.

A new, Bayesian version of Shortlist will therefore be presented: Shortlist B. In the original version of the model (henceforth, Shortlist A, for Activation) the output is a pattern of word activations over time. In contrast, Shortlist B operates as a Bayesian classifier and its output is a list of the posterior probabilities of words. Nevertheless, Shortlist B shares most of the key assumptions of its predecessor about, for example, the nature of prelexical and lexical representations and the processing architecture of the recognition system. Shortlist B therefore offers the same solutions as Shortlist A to the computational problems associated with spoken-word recognition but with the additional major benefit that the model makes optimal Bayesian decisions.

The Bayesian approach taken in Shortlist B is attractive not only because it instantiates the assumption that word recognition is optimal but also because almost all of the characteristics of the model follow directly from this assumption. As we discuss in greater detail below, the ways the model deals with the data on lexical competition, word frequency, perceptual match and mismatch, and the relation between lexical and sublexical information are all forced by the optimality assumption. This has two major benefits. First, the result is a much simpler theory than in the corresponding connectionist models. TRACE, Shortlist A, and Merge (Norris, McQueen, & Cutler, 2000b) all have a large number of free parameters. Changes to these parameters can produce quite large differences in model behavior (Pitt, Kim, Navarro, & Myung, 2006). In contrast, as will become apparent, Shortlist B has very few parameters, and the exact values of these parameters are not critical. Furthermore, the parameters that are used in the model are all designed to reflect expectations about the structure of the linguistic input or the task to be performed.

Second, the optimal Bayesian account of phenomena such as word frequency effects encourages a complete reassessment of core aspects of the word recognition process. Shortlist B generates insights into word recognition that cannot readily be derived from the traditional activation-based approach. For instance, the effect of mispronunciations on lexical access is usually cast in terms of perceptual similarity: The degree of activation of a lexical hypothesis varies simply as a function of the phonetic similarity of the mispronunciation to the base word. Probability must also be considered, however. For example, irrespective of phonetic similarity, some mispronunciations may be more likely renditions of a base word than others. As will be demonstrated, Shortlist B offers a formal account of how the likelihood of different realizations of words can modulate word recognition.

A further advantage of modeling speech recognition within a Bayesian framework is that it allows us to take advantage of some of the principles developed in the Bayesian Reader model of visual word recognition (Norris, 2006). As will be shown, this enables us to give a principled account of lexical decision and to simulate both RTs and error rates. Please note that further discussion of the benefits of Bayesian methods, which apply just as well to spoken as to visual word recognition, is to be found in Norris (2006).

The Bayesian approach has another important motivation. Unlike the activation values output by connectionist models, the posterior probabilities generated by Shortlist B have a clear formal interpretation. Activation rarely corresponds directly to behavioral observations such as speed, accuracy, or probability of responding. Following McClelland and Rumelhart (1981), the R.D. Luce (1959) choice rule is sometimes used to generate response probabilities from activation values (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, & Tanenhaus, 2001; Luce, Goldinger, Auer, & Vitevitch, 2000; McClelland & Elman, 1986). But this is largely a pragmatic procedure for generating probabilistic data from a deterministic model. It is therefore still unclear what activation values mean: In particular, do they reflect the probabilities that words will be recognized? Davis, Gaskell and Marslen-Wilson (1998) have suggested that activations in a recurrent network trained to identify words should be interpreted as the conditional probabilities of identifying those words given the input. It is not clear that this is a formal property of this model, however, or even that the model would produce such probabilities under conditions other than the specific circumstances of the simulations that Davis et al. report. Because the interpretation of activation values is not straightforward, we avoid any use of the activation metaphor here. Shortlist B works entirely within the probability domain; interpretation of the posterior probabilities output by the model is therefore completely unambiguous. A related reason to favor the Bayesian approach over connectionist models is that the use of such models does not guarantee optimal word recognition. It might be possible to build an interactive-activation network that computed the same Bayesian functions as Shortlist B, but it is also possible to build networks that compute other functions. The use of an interactive-activation framework therefore does not ensure that recognition will be optimal. In contrast, as we have already argued, the Bayesian framework guarantees optimality.

Continuous Speech Recognition

Previous models have taken a similar approach to the problem of speech recognition that we advocate here. One is the Neighborhood Activation Model (NAM; Luce, 1986; Luce & Pisoni, 1998); another is the Fuzzy Logical Model of Perception (FLMP; Massaro, 1987, 1989b; Oden & Massaro, 1978). As in Shortlist B, the NAM instantiates the critical assumption that words' prior probabilities (i.e., their frequency of occurrence) are combined with bottom-up evidence to determine word recognition. As we describe later in the section on NAM and Shortlist B, however, the way in which NAM operates is not strictly Bayesian. Furthermore, the NAM offers an account only of the recognition of isolated monosyllabic words. A major goal in developing Shortlist B was to provide an account of the recognition of words in the continuous speech stream and not just isolated words.

The FLMP foreshadowed Shortlist B in assuming that speech recognition should be optimal, that optimality should be achieved through the independent evaluation of different sources of evidence (Massaro, 1987; Massaro & Cohen, 1991), and that recognition can be conceived of as a Bayesian process (Massaro, 1987; Massaro & Friedman, 1990). But, as with the NAM, although there are strong formal similarities between FLMP and Bayes's theorem (Massaro & Friedman, 1990), the FLMP is not strictly Bayesian (see the section *FLMP and Shortlist B* below). In addition, although the FLMP has been applied to many types of perceptual data, as with the NAM, it has not been applied to large-vocabulary continuous speech recognition. Where the FLMP has been applied

to data on recognizing words in sentences (Massaro, 1978, 1989a), the critical words have always been treated exactly as they would in a model of isolated word recognition.

Any adequate model of spoken-word recognition must be able to recognize words in continuous speech. Because of the spaces between written words in a text like this, one can consider that recognition of written sentences entails the repeated application of the procedures used to recognize individual written words. The spaces provided by the writer tell the reader where one word ends and the next one begins. But speakers do not segment their utterances in this way for their listeners. Although there is a wide range of cues in the speech signal that are correlated with word boundaries (Church, 1987; Cutler & Carter, 1987; Nakatani & Dukes, 1977) and although listeners are sensitive to these cues and use them in segmentation (Cutler & Butterfield, 1992; Cutler & Norris, 1988; Davis, Marslen-Wilson, & Gaskell, 2002; Gow & Gordon, 1995; Mattys, White, & Melhorn, 2005; McQueen, 1998; Norris, McQueen, & Cutler, 1995; Salverda, Dahan, & McQueen, 2003; Shatzman & McQueen, 2006; Tabossi, Collina, Mazzetti, & Zoppello, 2000; Vroomen & de Gelder, 1995), none of these cues is completely reliable.[1] Word recognition therefore requires a solution to this segmentation problem: A mechanism that can work in the absence of any such signal-based cues.

If words were not alike, then the lack of reliable segmentation cues would not be a problem. Each word would be perceptually distinct and could be recognized on the basis of its unique material even if it were not segmented from other words. But words are alike. Because the words of a given language are made up from a limited inventory of speech sounds, they tend to be phonologically very similar to each other. Words begin in the same way as many other words, end in the same way, and often have other words embedded within them (Cutler, Norris, Mister, & Sebastian-Galles, 2004; Luce, 1986; McQueen, Cutler, Briscoe, & Norris, 1995). Given the continuous nature of the speech signal, it also often contains words that straddle word boundaries and hence yet more words that are not part of the sequence of words intended by the speaker. For example, the Italian sequence *visi tediati*, "faces bored," contains the spuriously embedded word *visite* "visits." Experiments by Tabossi, Burani and Scott (1995) suggest that such between-word embedded words are indeed considered by listeners during sentence comprehension.

These observations have led speech researchers to reject earlier accounts, such as the Cohort model (Marslen-Wilson & Welsh, 1978), in which words were identified in a strictly sequential order (see also Cole & Jakimik, 1978, 1980), and to believe that spoken-word recognition involves a process of competition between lexical candidates, so that candidates that overlap in the input compete with each other (McClelland & Elman, 1986; Norris, 1994). Words such as *visite* will be considered but will lose the fight with the competing words *visi* and *tediati* and will not be consciously perceived as being present in the utterance. This competition process segments the input (e.g., finds the boundary between *visi* and *tediati*) even in the absence of any cues signaling a lexical boundary.

Both of the connectionist models that have been most widely used to explain continuous speech recognition (TRACE and Shortlist A) implement this competition process in terms of interactive-activation networks, where nodes representing the activation of lexical candidates inhibit each other via reciprocal inhibitory connections. This, however, is certainly not the only solution to the problem of identifying the sequence of words in a stretch of continuous speech. Indeed, this is never the solution adopted in Automatic Speech Recognition (ASR) systems. In ASR, the task of discovering the sequence of candidate words that provides the best coverage of the input is usually thought of as a search problem. As we will explain below, candidate words can be encoded in a graph or lattice (see Figure 1) and the task is to find the best contiguous path through the lattice. The best path should correspond to the best fitting sequence of words. In Marr's (1982) terms, lattice-based search and competitive inhibition in a connectionist model are different algorithms for the same computational function, that of searching for the sequence of words that best matches the input.

In a Bayesian model, such as Shortlist B, words do not have activation values, and there is no direct inhibition among lexical hypotheses. The same assumptions are made in the FLMP (Massaro, 1987, 1989b; Massaro & Oden, 1995). But path probabilities can be computed and compared with each other in Shortlist B. This, then, is the solution to the segmentation problem instantiated in Shortlist B. Although there is no competitive inhibition in the new model, it still has a search algorithm that evaluates multiple lexical hypotheses simultaneously. The parallels between the two approaches are illustrated in Figure 1 (see also Scharenborg et al., 2005). The upper panel shows the standard representations of candidate words in an interactive-activation network like that in Shortlist A. Each candidate is connected to every overlapping candidate by inhibitory links. The lower panel represents the same candidates in terms of a word lattice, as in ASR systems and Shortlist B. The aim of the recognizer here is to search for the best path through the lattice. This example mainly shows paths that link contiguous word candidates. Standard ASR practice is to allow discontinuous paths too, but to penalize paths that leave some part of the input unaccounted for. Examples of discontinuous paths are also shown in Figure 1; we discuss below how such paths are processed in Shortlist B.

In practical ASR systems the number of alternative paths through a lattice can become very large indeed. The search process is usually simplified by using dynamic programming techniques, such as Viterbi search (Viterbi, 1967) or token passing (Young, Russell, & Thornton, 1989). The parallels between lattice-based search and inhibition in an interactive-activation network were recognized by Norris (1994), who suggested that dynamic programming techniques could be used as an alternative to the interactive-activation network in Shortlist. Although Shortlist B does not use dynamic programming, it does work by performing computations on paths.

The details on how path probabilities are computed are given below. Two points about how Shortlist B deals with continuous speech can already be made, however. The first is that, with respect to the segmentation problem, the high-level computation performed by the new model is the same as that in Shortlist A. The second is that the key computations involve path probabilities rather than word probabilities. A major goal in the development of Shortlist B was to examine whether Bayesian decisions based on

[1] Furthermore, we know of no evidence suggesting that any combination of these cues is completely reliable.

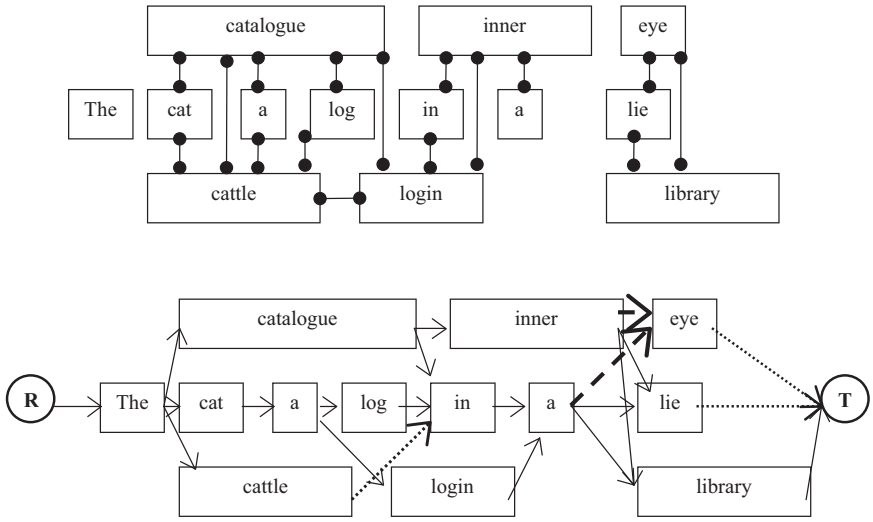


Figure 1. Recognition of the phrase "The catalogue in a library," as spoken by speaker of British English: /ðəkætəlɒgɪnəlaɪbrɪ]. The upper panel shows the competitive inhibition process that occurs among activated candidate words in an interactive-activation model, such as Shortlist A. Words that compete for the same stretch of input inhibit each other via direct, bidirectional inhibitory connections. Only a subset of the best-matching candidates is shown. The lower panel illustrates the path-based search through a word lattice used in automatic speech recognition and Shortlist B. Paths connect sequences of lexical hypotheses from a root node (R) to a terminal node (T); not all paths or words are shown. The dashed and dotted arrows are examples of connections between noncontiguous words (see text for details).

paths are successful in explaining the data on human continuous spoken-word recognition.

The Input to Shortlist B

A final motivation for the development of Shortlist B was the need to improve on the account of early phonetic analysis offered by the original Shortlist model. The input to Shortlist A is simply a string of phonemes. The representations of those phonemes have no internal structure, and all phonemes are treated equally. There is therefore nothing in the input to the word-recognition process to indicate that listeners find some phonemes more confusable than others. Furthermore, this kind of input to word recognition is discrete and categorical in two inappropriate ways. First, it is discrete in temporal terms. That is, there is no overlap of evidence for different speech sounds, as if, counterfactually, there were no effects of coarticulation in the speech signal. Second, this kind of input is discrete in informational terms: For any segmental position in the input there is 100% support for one and only one phoneme. There is, however, considerable evidence (reviewed in McQueen, 2007) to suggest that the word-recognition process is continuous in both the temporal and informational senses. Acoustic information modulates word recognition on a much finer time-scale than phoneme by phoneme, and that information concerns within-phoneme variability. The input to Shortlist A is therefore inadequate.

To date there have been three different approaches to producing more realistic input representations in models of spoken-word recognition. One option is to model the input noncategorically. The input in TRACE (TRACE II, to be more precise; McClelland and Elman, 1986), for example, consists of a vector of phonetic features that varies over time. Although this kind of input is more detailed, it still involves considerable oversimplification, particularly with respect to the time-course with which featural information becomes available. Critically, this approach depends on a largely untested set of assumptions about what evidence the listener can extract about different features (and hence phonemes) in any stretch of input.

A second option is to construct a model that takes the raw acoustic waveform as its input. Both TRACE I (Elman and McClelland, 1986; McClelland & Elman, 1986) and SpeM (Scharenborg et al., 2005) take this approach. A limitation of this method, once again, is that there is little reason to believe that there will be a close mapping between the acoustic-phonetic processes and representations in these models and those used by human listeners. Scharenborg et al., for example, derive phonemic representations with a conventional hidden Markov model phone recognizer, as used in ASR systems. To the extent that this recognizer deviates from human behavior, the results of the SpeM model as a whole could be misleading.

A third alternative is to accept that it may be premature to expect to produce a well-motivated model of the early stages of speech recognition and, instead, to try to simulate these processes using data from human phoneme or word confusions (e.g., Luce & Pisoni, 1998). Even though this approach sidesteps the question of how the early stages of recognition operate, it enables one to present later stages of a model with input that corresponds more closely to the input that would be received from the human perceptual system. For example, if listeners have more difficulty discriminating one pair of phonemes than another, then the input to the model should reflect that difference. Luce and Pisoni (1998) have used this procedure to great effect in the NAM to explain a

wide range of data on lexical neighborhoods and word frequency in tasks such as perceptual identification and lexical decision.

The confusion data driving the NAM is derived entirely from errors that listeners make in identifying words in noise (see also Benki, 2003; Miller & Nicely, 1955; Pickett, 1957; Wang & Bilger, 1973). This is a significant limitation because most of the psycholinguistic data one would wish to model are collected under relatively noise-free listening conditions. These confusion data also provide no information about how listeners accumulate perceptual evidence as the acoustic waveform arrives over time. This is again a serious problem because many aspects of the word-recognition data to be modeled concern the time-course of lexical processing and the speed of responding in RT tasks.

The input to Shortlist B is similar in spirit to that in NAM, in the sense that it too is based on perceptual confusion data. But, unlike the data used by NAM, those data are not based on identification of words in noise and do provide information about how confusions change over time. Specifically, they are derived from a gating task in Dutch (Smits, Warner, McQueen, & Cutler, 2003; Warner, Smits, McQueen, & Cutler, 2005). These gating data provide fine-grained information about how listeners accumulate perceptual information from the speech signal over time. These data are also very extensive in that they cover confusions about almost all possible diphones in Dutch.

Given the Bayesian principles of Shortlist B, its input should consist of probability values rather than, for example, phoneme activations. The confusion data from the gating task are ideal in this regard. As will be described in more detail later, it is straightforward to derive, from the responses of the listeners in the gating task, a sequence of multiple phoneme probabilities over three time slices per segment. This forms the input to the word-recognition process in the new model.

Shortlist B

In summary, Shortlist B makes two significant advances over its predecessor. First, the new model is based on Bayesian principles rather than on interactive activation. These principles, based on path probabilities rather than simple word probabilities, are applied to the problem of word recognition in continuous speech. Second, the input to the new model is based on phonetic confusion data, derived from a large-scale gating study. Thus, in contrast to Shortlist A, Shortlist B has a much more realistic input.

We will show that the new model can simulate key findings in spoken-word recognition. These results establish the viability of Shortlist B and, perhaps more importantly, show how a Bayesian perspective can offer valuable insights into the problems of speech recognition. For example, Shortlist B offers new and principled accounts of word frequency effects and the effects of perceptual match and mismatch. First, however, we present the model itself.

Bayesian Assumptions

Basic Equations

The most important equation is Equation 2, which specifies how to compute the conditional probability of each word given the evidence. In Equation 2, $P(Word_i)$ represents the listener's prior belief, before any new perceptual evidence has been accumulated, that $Word_i$ will be present in the input. In all of the simulations reported here we assume that $P(Word_i)$ can be approximated by the word's frequency of occurrence in the language. However, $P(Word_i)$ will also be influenced by factors outside the scope of the present model, such as semantic or syntactic context.

$P(Evidence|Word_i)$ is calculated from the evidence for sublexical units of representation. Spoken-word recognition appears to be mediated by the recognition of phonologically abstract sublexical units at a prelexical level of processing (Healy & Cutting, 1976; Massaro, 1975; McNeill & Lindig, 1973; McQueen, Cutler, & Norris, 2006; Scharenborg et al., 2005). A number of units could serve this function, including (bundles of) features, phonemes, and position-specific allophones. It remains to be determined which of these alternatives is the most plausible. In Shortlist B, as indeed in Shortlist A, we make the assumption that these units are phonemes. It is important to stress, however, that this is only a working assumption; we have no reason to commit to phonemes as the prelexical unit of representation. The following arguments, and the implementation of Shortlist B, do not depend on the phonemic status of the prelexical representations, only that there is a prelexical stage of processing involving abstract sublexical representations of phonological form that mediates between the speech signal and the mental lexicon. The choice of phonemes as units is also constrained by the choice of the diphone database as input to the model.

Given the assumption of prelexical phonemes, therefore, $P(Evidence|Word_i)$ is derived from phoneme probabilities that, in turn, using Bayes's theorem, are derived from phoneme likelihoods. First:

$$P(Phoneme_i|Evidence) = \frac{P(Evidence|Phoneme_i) \times P(Phoneme_i)}{\sum_{j=1}^{j=m} P(Evidence|Phoneme_j) \times P(Phoneme_j)} \quad (3)$$

where m is the number of phonemes in the language. The likelihood of $Word_i$ is then given by the product of the probabilities of the phonemes in that word, $P(PhonemeString_i)$:

$$P(Evidence|Word_i) = P(PhonemeString_i) = \prod_{j=1}^{l} P(Phoneme_j|Evidence) \quad (4)$$

where l is the length of the word. $P(Word|Evidence)$ is then computed as follows:

$$P(Word_i|Evidence) = \frac{P(PhonemeString_i) \times P(Word_i)}{\sum_{j=1}^{j=n} P(PhonemeString_j) \times P(Word_j)} \quad (5)$$

Note that these computations do not take directly into account any statistical dependencies among phonemes (e.g., differences in their transition probabilities). In the case of word recognition, however, these dependencies are built into the lexicon (words with common sequences will tend to have many lexical neighbors). Sequential dependencies will thus modulate word recognition as a function of

the influence of similar-sounding words on *P(Word | Evidence)*; see Equation 5. As we will see later, Shortlist B can indeed simulate the negative effects of large and dense lexical neighborhoods on word recognition (Luce & Pisoni, 1998; Vitevitch & Luce, 1998, 1999). But if the task is phoneme identification or the input does not consist of a word, one might also want to take account of sequencing constraints. It has indeed been shown that listeners are sensitive to phonotactic constraints (Massaro & Cohen, 1983b) and phoneme transition probabilities (Pitt & McQueen, 1998) in phonetic categorization and that transition probability effects are different for words and nonwords (Vitevitch & Luce, 1998, 1999). A more complete model would therefore include modulation of the computation of *P(PhonemeString)* as a function of transition probabilities. In the current version of Shortlist B, however, the probability of each phoneme in an input string is computed independently of all other phonemes in that string.

Phoneme Likelihoods

Phoneme likelihoods—that is, *P(Evidence|Phoneme)*; see Equation 3—are an essential component of the Bayesian theory underlying Shortlist B. Implicit in the use of Bayes's theorem is the idea that a particular input signal might possibly have been generated by more than one phoneme, that is, that there is some ambiguity in the input. If the input were unambiguous, it would correspond to a sequence of phonemes, each of which having a probability of 1.0. A single word would therefore also have a probability of 1.0, and, at least in the case of isolated word recognition, successful word recognition would be a rather trivial consequence of phoneme recognition. The speech signal, however, is inherently ambiguous. First, there is variability in the way phonemes are realized. A given acoustic signal can be a possible realization of more than one phoneme (e.g., Sawusch & Jusczyk, 1981). Second, ambiguities can be introduced by noise in the environment. Additional ambiguities could arise as a consequence of noise within the perceptual system itself.

In all of these cases a particular signal presented to the word-recognition process might have been generated by more than one phoneme. Figure 2 illustrates this by considering an idealized case where there are only two phonemes, A and B, that differ along a single perceptual dimension, *I*. *I* is a continuously valued variable whose probability distribution for a particular phoneme is given by the density function *f(I|Phoneme)*. Figure 2 shows the probability density functions (pdfs) of the values of tokens of the two phonemes on that dimension. The broader distribution for phoneme B indicates that the realization of phoneme B is much more variable than the realization of phoneme A. Given the input I_x, the probability that the input is $Phoneme_A$ is given by Equation 6:

$$P(Phoneme_A|I_x) = \frac{f(I_x|Phoneme_A) \times P(Phoneme_A)}{\sum_{i=1}^{i=n} f(I_x|Phoneme_i) \times P(Phoneme_i)} \quad (6)$$

where $f(I_x|Phoneme_x)$ corresponds to the height of the pdf at I_x and n is the total number of phonemes. $f(I_x|Phoneme_i)$ is called the likelihood function of $Phoneme_i$. When different phonemes are being compared on the basis of I_x, it is the ratio of the phoneme likelihoods that influences the revision of the prior probabilities (as shown in Equation 6). In general, of course, different speech

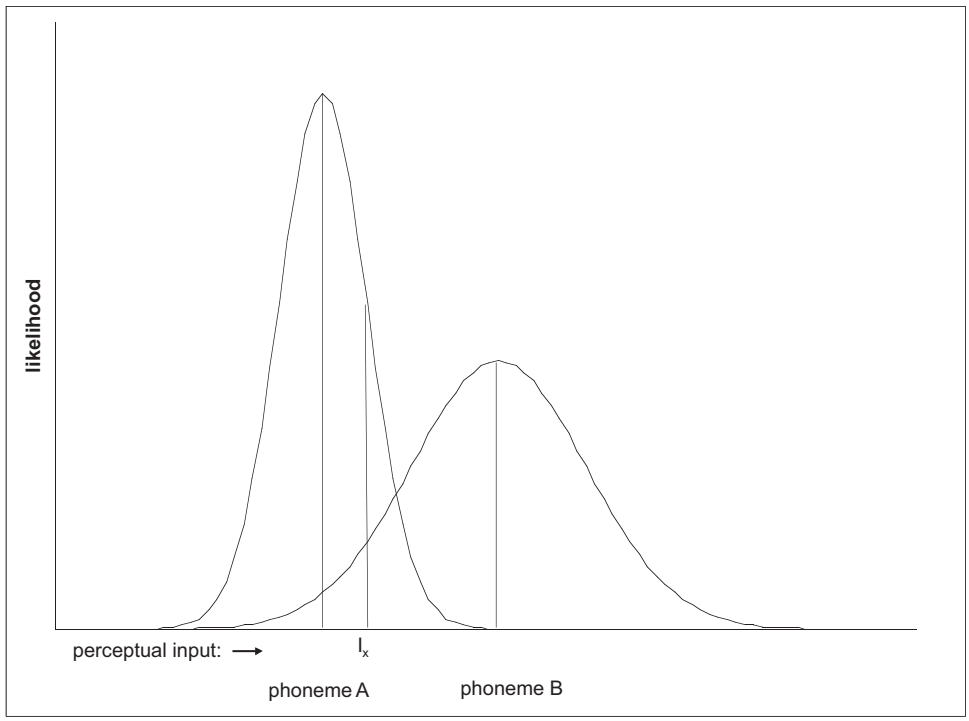


Figure 2. Illustration of possible probability density functions of two phonemes on a perceptual dimension I.

sounds are likely to vary on multiple physical or perceptual dimensions. In this case the variance of the likelihood functions in the different dimensions will determine how the evidence from the different dimensions is weighted. Because they are more diagnostic, dimensions with small variance will have a greater influence than more variable dimensions.

The underlying assumption here, therefore, is that listeners are able to learn the likelihood functions for recurring units (e.g., phonemes) in their language. This is exactly the same assumption that underlies the use of Hidden Markov Models in ASR and is similar to the proposals for episodic theories of speech recognition presented by Johnson (1997a; 1997b) and Pierrehumbert (2002), although these theories are not expressed in terms of Bayes's theorem. There are several ways that listeners could learn to characterize the likelihood functions of phonemes. For example, each phoneme might be described in terms of Gaussian distributions over each perceptual dimension. Whatever way this learning may be implemented, the central assumption is that listeners have knowledge about the likelihood that speech-sound categories are associated with particular perceptual events. The foundations of this learning are put in place in the first year of life (Maye, Werker, & Gerken, 2002; Werker & Tees, 1999). However, the listener's estimate of the likelihood function should not simply represent the aggregate of past experience but should be updated in the light of new experience. As we will see later, adjustments can still be made even in adulthood.

A complete implementation of a Bayesian model of spoken word recognition would compute phoneme probabilities in the manner described above. Unfortunately, however, we have no direct access to the representations of the likelihood functions that listeners have acquired and, therefore, cannot estimate *f(Evidence|Phoneme)*, especially not across the multitude of perceptual dimensions along which speech sounds vary. Indeed, we cannot even be sure what all the relevant perceptual dimensions may be. The only practical solution available, therefore, was to find a way of estimating phoneme probabilities, while still preserving the key theoretical assumption that listeners acquire knowledge of phoneme likelihoods. The solution in the implementation in Shortlist B was to take advantage of the perceptual confusion data from Smits et al. (2003). The listeners' identification responses are used to estimate directly $P(Phoneme_i|Evidence)$, and hence $P(Evidence|Word_i)$; see Equation 4:

$$P(Evidence|Word_i) = \prod_{j=1}^{l} P(RespondPhoneme_j|StimulusPhoneme_j) \quad (7)$$

where $P(RespondPhoneme_j \mid StimulusPhoneme_j)$ is the probability that the listeners identified the j_{th} phoneme in the input as the j_{th} phoneme in the word. We specify in more detail below how phoneme probabilities are derived from the perceptual confusion data.

NAM and Shortlist B

Equation 7 has the same form as the Stimulus-Word-Probability equation of Luce and Pisoni (1998), and inserting Equation 7 into Equation 2 makes the latter take on the same form as Luce and Pisoni's Frequency-Weighted Neighborhood Probability Rule (FWNPR). However, despite the superficial similarities, the interpretation of these equations is different in NAM and Shortlist B. First, in Shortlist B all of these probabilities depend on the assumption that listeners are able to compute the likelihood *f(Evidence|Phoneme).* This is central to our claim that listeners are behaving as optimal Bayesian recognizers. As we will show later, this has important implications for our analysis of perceptual match and mismatch. Thus, although our use of the confusion data from the diphone database is similar in spirit to the use of confusion data in the NAM and this leads also to a computation based on response probabilities (Equation 7), it is important to stress that these similarities concern the way Shortlist B has been implemented and not the underlying theory. In other words, if we did have access to listeners' likelihood functions, there would be less similarity between the models. The core assumption that listeners compute phoneme likelihoods is not made in the NAM.

Second, in Shortlist B the left-hand term in Equation 5 is a posterior probability. In contrast, Luce and Pisoni (1998) interpret their corresponding equation as being an application of the R. D. Luce (1959) choice rule. That is, the FWNPR estimates *P(ID)*, the probability of correctly identifying a stimulus word. A response probability is not the same as the posterior probability of a hypothesis given the evidence. For example, in the NAM, if $P(ID) = 0.95$, this implies that the listener will respond with the stimulus word 95% of the time. Shortlist B, however, employs the optimal Bayesian decision rule; and thus, if $P(Word_x|Evidence) = 0.95$, Shortlist B will always respond with $Word_x$ because $Word_x$ has the maximum posterior probability.

The NAM and Shortlist B therefore have important similarities—in particular the use of perceptual confusion data, weighted by word frequency, and the use of a relative evaluation metric. Both models are based on the key idea that optimal word recognition depends on the combination of bottom-up evidence and prior lexical probabilities. Finally, as we will see later, contextual information is considered to influence lexical priors in Shortlist B just as word frequency does. This assumption was also prefigured by the NAM. But the NAM and Shortlist B also have fundamental differences—only Shortlist B is strictly Bayesian, and only Shortlist B is designed to recognize words in continuous speech.

Continuous Speech

In themselves, Equations 2 and 7 are not sufficient to assign probabilities to words in continuous speech. Even if a word matches the input extremely well, it will not be recognized if it overlaps with competitors or is not, with other words, part of a path through the input which fully accounts for that input. Furthermore, we cannot simply try to calculate probabilities by comparing a word only with other words that it overlaps with. Those words may, in turn, overlap with yet other words that influence their probabilities. What we need is a measure of word probability that takes account of whether or not the word is on a high or low probability path. For that we first need to calculate path probabilities.

Given any possible path (string of phonemes), the probability of observing that path will largely be determined by the product of the probabilities of the phonemes on the path. But because paths are also sequences of words, we must also take into account the

fact that some sequences of phonemes (those consisting of concatenations of higher frequency words) will be more likely than others. As shown in Equation 8, therefore, path likelihoods are based on word likelihoods. The likelihood of a path is given by:

$$P(Evidence|Path_i) = \prod_{j=1}^{w} P(Evidence|Word_j) \times P(Word_j) \quad (8)$$

where w is the number of words in the path. The probability of each path is then given by normalizing over the sum of the path likelihoods:

$$P(Path_i|Evidence) = \frac{P(Evidence|Path_i)}{\sum_{j=1}^{j=p} P(Evidence|Path_j)} \quad (9)$$

where p is the number of paths through the lattice. $P(Word_i|Evidence)$ is then given by summing the $P(Path_t|Evidence)$ for all paths in which that word occurs in that position:

$$P(Word_i|Evidence) = \sum_{j=1}^{j=n} P(Path_j|Evidence) \quad (10)$$

where n is the number of paths the word lies on. This means, for example, that if there were two paths with the same probabilities, both of which contained the same word in the same position, the word probability would be twice what it would be if it appeared on only one of those paths.

We can now summarize the chain of probability estimations in Shortlist B that lead to the estimates of the probability of individual words given a continuous speech input, that is, $P(Word_i|Evidence)$. $P(Evidence|Word_i)$ is derived from the diphone database using Equation 7, and $P(Word_i)$ is derived from the frequency of occurrence counts in the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). These terms, across multiple words and paths, influence $P(Path|Evidence)$, as in Equations 8 and 9. Finally, $P(Path|Evidence)$ determines $P(Word_i|Evidence)$, via Equation 10.

Consequences of Bayesian Assumptions

Several things follow from this Bayesian path-based approach. First, if there is only one path with a nonzero likelihood, $P(Path|Evidence)$ will have a probability of 1.0, and all words on that path will have a probability of 1.0. This follows from the fact that what we are doing here is deriving the probability of words, given that the input really is a sequence of real words. If a path is the only possible one that is consistent with the input, then its probability, and the probability of all words on that path, must be 1.0, regardless of how well the words fit the input. The model thus follows the advice of Sherlock Holmes with which we began this article: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890, Ch. 6, p. 93). One might object that the word probability derived from path probabilities takes no account of the fact that listeners obviously can judge the goodness of a particular token of a word, even when there is no doubt as to which word is presented. The fact that listeners can make judgments about goodness of fit does not imply, however, that these judgments are based on exactly the same information as that used to determine recognition. For example, goodness of fit might be based on likelihoods ($P(Input|Word)$). In effect, when a listener is presented with a word that is poorly articulated, they might be in a position to be completely certain what the word must be but, at the same time, be certain that this is an unusual exemplar of that word.

Second, although all words on a path can have a probability of 1.0, this does not mean that the model always behaves in a winner-takes-all fashion. For example, if paths differ only in terms of two words with nonzero probabilities, the final probability of the better matching word will be reduced because of the presence of its competitor. This illustrates the fact that although the model does not have direct inhibition between alternative lexical candidates as in the original Shortlist model (and TRACE), there is, nevertheless, a form of lexical competition. The more probable one word is, the less probable overlapping words will be.

Third, a strictly Bayesian approach requires the computation of exact probabilities. In order to assign exact probabilities[2] to words we would need to calculate all possible paths through the lattice, because the denominator in Equation 9 corresponds to the sum of the probabilities of all paths. In fact, the denominator is equivalent to $P(Evidence)$, the probability of observing that particular input. In ASR systems, it is generally impractical to calculate all paths, so only the best few paths are computed. Because the main requirement is simply to identify the best path, there is little need to assign meaningful probability values to either the paths or the words. This is fortunate because the total number of possible paths can be enormous, in part because of limitations in the performance of the phoneme or word recognition techniques used. Another reason for the large number of possible paths is that it is often necessary in ASR to compare paths containing overlapping tokens of the same word or phoneme beginning at slightly different times, and this leads both to more paths and to problems in pooling evidence from the same words on different paths. (Note that these problems do not arise here, because, in Shortlist B, all phoneme and word candidates are aligned with fixed phoneme boundaries.) ASR systems, therefore, are usually designed to discover the path (sequence of words) that is most likely to have produced the observed input, that is, they use maximum likelihood methods that are designed to find the path that maximizes $P(Evidence|Path)$. For this, only the best few paths need to be computed. In some ASR applications, however, it is helpful to be able to assign confidence measures to individual words, and various techniques have been developed to approximate Bayesian word probabilities (cf. Bouwman, Boves, & Koolwaaij, 2000; Wessel, Schlueter, Macherey, & Ney, 2001).

[2] There is a long-running philosophical debate among Bayesians concerning whether probabilities are subjective or objective. People must necessarily operate on the basis of subjective probabilities; but, for the purposes of exposition, we also assume that the subjective probabilities approximate objective probabilities that could be calculated from empirical data. However, we acknowledge that in the context of natural language, it is unlikely to be possible to compute exact objective probabilities.

We adopt the same pragmatic approach in Shortlist B. In practice there is no need to consider words and paths with very low probabilities, as these will not make a significant contribution to the probabilities of more likely candidates. In fact the simulations to be described later are set up so that we can limit the number of candidate words starting at each position, exactly the constraint on which the name of the original Shortlist model was based. These hard limits on the candidate set in Shortlist B are purely for practical convenience, however: They allow the simulations to run more rapidly. Nevertheless, the idea that only a shortlist of candidates is considered at any point in time is more strongly motivated in the new model than in Shortlist A. Even if there were no practical limit imposed on set size, the Bayesian computations would guarantee that low probability words (and paths) effectively remove themselves from the running, so that only the shortlist of best candidates influences recognition performance.

A fourth consequence of the Bayesian approach in Shortlist B is that unknown words require special treatment. As described so far, an assumption behind our use of Bayes's theorem is that the input consists of a sequence of known words. For the posterior word probabilities in the model to have a direct interpretation as real probabilities, the input must be interpreted as a sequence of known words in the lexicon, where the probability of each input word is the same as its prior probability (i.e., the frequency as indicated in the lexicon). But clearly there will be occasions when the input (or part of it) will not consist of words in the lexicon. The input may contain, for example, a genuinely unknown word (such as a foreign name) or a word that is so badly mispronounced as to be unrecognizable. Under these conditions the product of the candidate word probabilities on all paths will be zero. Consequently, a single unrecognizable word could prevent recognition of any word in the utterance.

The problem here is that the model's prior belief is wrong: the input will not always be a sequence of known words, properly pronounced. As noted earlier, a Bayesian system (listener or model) will only make optimal decisions to the extent that prior beliefs accurately reflect the structure of the real world. The set of hypotheses under consideration must therefore be extended beyond the set of known words in the language. The model has to consider the hypothesis that the input is not a known word. We will refer to such hypotheses as dummy words. The dummy word serves a similar function to the garbage model used in ASR systems. A dummy word is a hypothesis that matches any stretch of the input to some extent. Dummy words can therefore fill in the gaps in incomplete paths (e.g., the paths including dotted and dashed arrows in the lower panel of Figure 1). This means that there will always be at least one complete path with a finite probability. Now consider what will happen when a nonword is inserted into a sentence. Assume that all paths pass through the nonword. The raw accumulated path scores of all paths would therefore be multiplied by the probability of the dummy word. Actual path probabilities depend on the ratio of a given path score to the sum of the path scores for all paths (Equation 9). As both the numerator and denominator of Equation 9 will be multiplied by the probability of the dummy word (cf. Equation 8), the effect of the dummy probability will cancel out, and the final word probabilities in any given path will therefore be unaffected by the presence of the nonword. In practice, it would be almost impossible to insert a nonword into a sentence without creating additional paths, but this example does illustrate how the use of a dummy word can make a path-based system behave robustly when faced with unknown words. Without the dummy word, Shortlist B would have the wrong prior beliefs and would make the wrong decisions.

The dummy word has two other important functions. First, as we discuss in detail when we present the simulations, the dummy plays a critical role in word segmentation. Second, it can be used to perform lexical decision. If a listener in a lexical-decision experiment is presented with a nonword, the only fully spanning path will contain a dummy word. The listener can therefore judge whether a stimulus is a word or not simply by determining whether the best path contains a dummy word (or is a dummy word). If a path containing a dummy word is much more probable than paths consisting only of words, the stimulus is a nonword; otherwise it is a word. The general procedure for performing lexical decisions using Bayesian techniques is discussed in more detail in Norris (2006). Although that article deals only with the case of visual word recognition, the principles apply equally to spoken word recognition.

At this point, we should note that the dummy words, and lexical candidates in general, are all tokens representing phonological forms. A lexical candidate represents the hypothesis that the input corresponds to at least one word in the lexicon with that phonological form. If the phonological form matches more than one lexical entry (i.e., homophonous words), higher-level syntactic, semantic, or contextual information would have to be brought to bear to determine the appropriate interpretation of that phonological form. A dummy word represents the hypothesis that a particular sequence of segments does not correspond to any word in the lexicon.

Model Input: The Diphone Database

The input to Shortlist B is derived from a database of perceptual confusions collected in a gating task (for discussion of the gating task, see Grosjean, 1996). Smits et al. (2003) presented Dutch listeners with 2,294 diphone sequences in Dutch—effectively all possible diphones in the language (1,179 different diphones sequences, most recorded in multiple stress contexts). Each diphone was presented at six gates, corresponding roughly to each third of each of the two phonemes in the diphone. On each trial, listeners were presented with one of the six possible gates of one of the diphones and were required to identify both phonemes of the diphone as members of the standard inventory of 38 Dutch speech sounds. All stimuli were presented in pseudorandom order such that listeners never heard the same diphone on successive trials. That is, listeners were not exposed to each diphone gated incrementally. A detailed statistical analysis of the pattern of confusions obtained is presented in Warner et al. (2005).

The data from this gating task tell us, for each input phoneme, in each diphone context, what the probability is that that phoneme will be identified as each possible Dutch phoneme. This database therefore has three major advantages over other confusion data. First, it gives information about how confusions change over time. Second, it tells us how confusions for a particular phoneme vary according to their phonetic context (i.e., as a function of the other phoneme of the diphone). Finally, at gates corresponding to the first phoneme in the diphone (Gates 1–3) it tells us how coarticu-

latory information present during that phoneme can start to specify the identity of the upcoming phoneme.

Computing Probabilities from the Diphone Database

There are 37 phonemes in Shortlist B, 22 consonants and 15 vowels. There were 16 vowels in the stimulus and response sets in the diphone identification experiment, but, as discussed by Smits et al. (2003), the vowels /ə/ and /ʉ/] effectively formed a single category in the listeners' responses. One vowel, /}/, thus represents this compound category both in the diphone database and in Shortlist B. The phoneme inventory of the model is listed in Appendix A. The input for any simulation consists of a sequence of these phonemes (and, optionally, a silence marker, "[").Unless noted otherwise, examples throughout this article of input to the model and output from it will be based on the machine-readable (DIstinct Single Character; DISC) transcriptions used by the model rather than IPA transcriptions.

We assume that the average responses of the listeners in the gating experiment correspond to the output of prelexical processing. Any phonotactically legal sequence of phonemes can be represented as a concatenation of diphones. We can therefore use the database to estimate the similarity of phoneme sequences of arbitrary length to any word in Dutch. For each phoneme at a particular gate and in a particular diphone context, we know the probability that listeners will identify that phoneme as each possible phoneme, that is, $P(RespondPhoneme_j \mid StimulusPhoneme_j)$, as required for Equation 7.

The prelexical stage of processing is therefore simulated by retrieving these conditional probabilities from the diphone database. These probabilities then form the input to the word-recognition component of the model. Posterior word probabilities are derived from this input (and from the prior word probabilities) using Equations 7–10. These computations are carried out cyclically for each of the three gates within each phoneme.

For all but the first and last phoneme in a sequence, the diphone database provides two estimates of phoneme confusability. Consider, for example, the case where a single consonant-vowel-consonant (CVC) word, p1p2p3, is presented. The word can be represented by the overlapping diphones /p1,p2/ and /p2,p3/, with each diphone corresponding to six gates. For the vowel p2 (and, more generally, any nonterminal phoneme) we therefore have responses to overlapping gates contributed from both diphones. The probabilities used in the calculations in Shortlist B are derived by taking the maximum of the probabilities from each pair of overlapping gates, and then renormalizing those values so that they sum to 1.0. Using another procedure, such as taking the average probability or always using the probability for either the first or second diphones, would make little difference to the behavior of the model. The most reliable evidence for a given phoneme usually comes from the response at the last gate of the diphone where that phoneme is the first half of that diphone. As the final phoneme of a word is necessarily the final phoneme in a diphone, however, there is no option for word-final phonemes but to use the probability derived from the last gate of the final diphone (i.e., where that phoneme is the second half of that diphone).

As a concrete example, the probabilities of correctly identifying the phonemes of the input /b}s/ (the Dutch word meaning "bus") are shown in Table 1. The sharp increase in the probability of correctly identifying the /}/ between gates 3 and 4 (and the /s/ between gates 6 and 7) is typical of the database (see Figures 1–3 in Smits et al., 2003). This corresponds to the point where the input changes from providing information based on anticipatory coarticulation alone to providing information from the segment itself (e.g., gate 3 is the last gate of the /b/, containing the stop release burst, with relatively little information about the upcoming vowel, while gate 4, in contrast, is the first third of the vocalic portion of the vowel itself). This example also illustrates that the probability of the correct phoneme can sometimes go down as well as up. Note also that once the diphone containing a particular phoneme has finished (e.g., gate 6 for /b/), the probability of the phoneme remains fixed at its final value throughout the rest of the input.

Table 1
The Probabilities of Identifying /b/, /}/, and /s/ Across the Nine Gates (Three per Phoneme) of the Input /b}s/; There Are No Values for P(s) During the First Three Gates as the Data There Come Only From the Diphone /b}/

Gate	P(b)	P(})	P(s)
1	0.53	0.12	
2	0.70	0.07	
3	0.94	0.19	
4	0.83	1.00	0.30
5	0.94	0.92	0.08
6	0.86	0.92	0.06
7	0.86	0.92	0.94
8	0.86	0.92	0.97
9	0.86	0.97	0.97

P(Evidence|Word) for the word /b}s/ is given by multiplying the phoneme probabilities over the row corresponding to the current gate. At Gates 1–3, therefore, the probability is determined by only the first two phonemes. If the input really is the word /b}s/ then, after two phonemes, *P(Evidence|Word)* is given entirely by the probabilities of those two phonemes.

Limitations of the Diphone Database

Although using the diphone database to generate the input to a model of speech recognition represents a considerable advance over Shortlist A, it still has a number of limitations. The first is that the procedure for collecting the diphone confusions allowed participants as much time as they liked to make responses (performance on the task was self-paced, and participants were not put under any time pressure). This means that the data indicate how much information listeners can potentially extract from a stimulus of a particular duration, but this will not necessarily be an exact reflection of how much information they have extracted at the time the stimulus ends. It seems reasonable to assume that there must be some lag between presentation of the input and complete perceptual processing of the input. If this lag were constant for all stimuli, it would not really be a source of concern. It seems more than likely, however, that some speech sounds will take longer to process than others. Because this is not captured in the diphone data, there is almost certainly an undesirable source of error in the simulations.

Another element of noise in the diphone data is that, although listeners in the gating study performed extremely well, they were

not perfect (they were approximately 90% correct on final gates overall; Smits et al., 2003). There are many reasons for this, including errors arising from ambiguities in the speech materials, errors due to perceptual noise, and random errors in participants' responses. It is impossible to distinguish among these types of error or, especially at early gates, between errors and the perceptual confusions we are interested in here. The high overall accuracy rates, and the orderliness in the diphone data (Smits et al., 2003; Warner et al., 2005) indicate that there are relatively few such errors, but those that are in the database cannot be avoided; they can only be passed on to the model.

There is also no noise or moment-by-moment variability in processing in Shortlist B. This makes it hard to simulate both RT and error rate together. Later we will present simulations using a variant of the model based on the stimulus sampling mechanism used in the Bayesian Reader (Norris, 2006). In this version of the model, the same input can generate different response probabilities on different trials. These simulations show that it is possible to give a principled account of how a Bayesian model should operate with noise in perceptual processing. Unfortunately it is not yet practical to do this with the diphone input. For these later simulations we will therefore have to use hand-crafted input.

Model Overview

The processing architecture of Shortlist B is the same as that of Shortlist A. Specifically, there is in both models a prelexical and a lexical level of processing, and information flows forward from the prelexical to the lexical level but not back from the lexical to the prelexical level during on-line processing. The two models also agree with respect to key representational assumptions. In Shortlist A, prelexical representations are phonemes, whereas in Shortlist B the prelexical level outputs phoneme probabilities. Furthermore, lexical representations in both models are phonologically abstract (they are strings of phonemes in both cases). Importantly, both models also make a distinction between type representations of words (lexical representations stored in long-term memory) and token representations of words (those standing for current hypotheses about what is being heard). Thus, in common with the token word units that are wired on the fly into the lexical network of Shortlist A, the word nodes in the word lattice of Shortlist B are also all temporary token representations. Token representations are necessary so that multiple versions of the same word can simultaneously stand for hypotheses that instances of that word appear at different locations in an utterance (see Norris, Cutler, McQueen, & Butterfield, 2006).

Word recognition proceeds phoneme by phoneme and within each phoneme using the data from the diphone database, gate by gate. At each gate, the model performs the following sequence of operations to compute word probabilities:

1. Derive phoneme probabilities from the diphone responses corresponding to that gate.
2. For every segment, calculate *P(Evidence|Word)* for all words beginning at that segment according to Equation 7.
3. Construct or update the word lattice.
4. Calculate path probabilities according to Equations 8 and 9.
5. Sum word probabilities over paths to compute P(Word|Evidence) for all words, as in Equation 10.
6. Input next gate and return to Step 1.

Some additional housekeeping during the simulations is also required. The number of word candidates considered at any one phoneme (candidates always start on a phoneme boundary) is limited to 50. When this limit is exceeded, the lowest scoring candidate words are eliminated. Similarly, there is a limit of 500 paths, so low-scoring paths are pruned when this limit is passed. These limits are quite generous and the model performs almost identically with much smaller or larger numbers of phonemes, words, or paths.[3] These limits are therefore not free parameters, strictly speaking. In its basic form, the model is parameter-free, and its behavior follows directly from Bayesian principles. However, as will be discussed below, the complete model requires five free parameters, three that influence the model's segmentation and lexical decision performance, one that influences how it deals with mispronunciations, and one that deals with aligning the model to eye-tracking data. In principle these should not be free parameters at all. The first four are all probabilities that should accurately reflect the statistical properties of the input. In effect, they represent the model's prior beliefs about the input and the task that is being performed. These parameters have fixed values across the simulations we report; it was thus unnecessary to adjust these values for each specific simulation nor, indeed to, choose the particular values we use here (the model is stable across a range of values). The fifth parameter is required to link the model to the eye-tracking behavior. It simply shifts model output relative to the data by a fixed amount, as motivated by the literature (see below). It too is therefore not truly a free parameter.

Simulations

Shortlist B will now be evaluated via five sets of simulations. In most simulations the model was run on an actual or possible set of items from a listening experiment, and the model's performance was averaged over those full item sets. In a few cases, however, the model was run on a single input in order to explicate an aspect of its operation. The simulations are designed to reflect, as transparently as possible, the underlying principles of the model and the way it is influenced by the properties of the diphone database. Therefore, we have not added additional parameters or modifications to improve the fit of the model to particular data-sets.

All of the Shortlist B simulations use a lexicon of 20,250 Dutch words. These word-forms, together with their frequency of occurrence, were extracted from the CELEX database (Baayen et al., 1995). Phonemic transcriptions were adjusted where necessary so that all words could be described in terms of the 37 phonemes in the model's inventory (see Appendix A). This involved collapsing the vowels /ʉ/ and /ə/ to /}/ (see above), and the voiced and voiceless velar fricatives /x/ and /ɣ/ to one voiceless category /x/ (the voicing distinction for velar fricatives is preserved in CELEX, but many Dutch speakers now neutralize the distinction (Gussenhoven, 1992); only the voiceless variant was therefore included in the Smits et al., 2003, study). The lexicon consists of the 20,000

[3] For example, in the first simulation we report, increasing the number of candidates to 100, and the number of paths to 10,000, does not change any of the calculated word probabilities by more than .02. Even reducing the number of word candidates (excluding dummy words) and paths both to 10 does not change the final probabilities of the critical words (plotted in Figure 3b) by more than .0001.

most frequent words in CELEX, plus 250 additional words that were added as required for specific simulations (e.g., if an experiment used a word which was not in the top 20,000). Like its predecessor, Shortlist B is therefore able to simulate listeners' performance across a range of actual items in an experiment, using a realistically large lexicon.[4]

Recognizing Continuous Speech

The first critical test of Shortlist B is whether it can indeed recognize words in continuous speech. As we have already argued, any plausible model of spoken-word recognition must have a means of segmenting the speech stream into words in the absence of any segmentation cues. In interactive-activation models, such as TRACE and Shortlist A, inhibitory competition among activated lexical hypotheses has this role. We have suggested that path-based Bayesian evaluation can serve the same function. Can Shortlist B therefore recognize words in continuous speech when no lexical boundaries are marked?

A related question concerns the model's ability to revise interpretations on the basis of following context. A listener's interpretation of an utterance can certainly change in the light of information arriving later in time (Bard, Shillcock, & Altmann, 1988; Connine, Blasko, & Hall, 1991; Grosjean, 1985). To varying degrees, the interactive activation networks in TRACE and Shortlist A allow them to account for these retroactive effects. In Norris (1994), this was illustrated by presenting Shortlist A with the input "shippingquiry." This input first activates *ship*, but then *shipping* dominates the activation landscape. When *inquiry* becomes strongly activated, however, it competes with *shipping*, and the interpretation of the input is finally revised to the correct segmentation *ship inquiry*. That is, a word boundary is finally postulated in the right place, even though the input to the simulation did not mark that boundary in any way. An example like this thus serves not only to test whether subsequent context can modulate word recognition but also, more fundamentally, whether the model can segment continuous speech.

Norris (1994) pointed out that the optimal interpretation of an input like *ship inquiry* could be obtained by resetting the network activations in Shortlist A after every phoneme. When activations are reset, the network recomputes a near optimal interpretation of the input taking both new and old information into account. If activations are not reset, the network can lock into a state where the high activations of words favored by the initial interpretation of the input prevent that interpretation being revised by the later context. Norris et al. (1995, 1997) showed that simulations of a number of effects were more accurate when Shortlist A employed an activation reset. Shortlist B also needs to recompute word probabilities as path probabilities change over time. For example, if a path reaches an impasse and its probability drops, then the probabilities of all words on that path need to be modified, and this will alter the final probabilities of all of the words on all other paths. The need to recompute probabilities is an inevitable consequence of the need to revise interpretations. The only alternative (in any model) is to wait until the end of the utterance and then perform a single computation. Given the strong evidence on the continuity of spoken-word recognition (see McQueen, 2007, for review), this alternative is very implausible.

In the first simulation, therefore, we tested Shortlist B on a Dutch version of *ship inquiry.* We also took the opportunity to compare the new model with the old. Furthermore, to facilitate direct comparison between the new and old models and to highlight the effect of word frequency on the new model, we also ran simulations using the new model, but with the effect of frequency disabled.

The input *kar personen*, cart people, was therefore presented to Shortlist B and, using the same Dutch lexicon, to a version of the 1994 Shortlist A model. Like *shipping* in *ship inquiry*, /kArp}rson}/ contains *karper*, carp. The results are shown in Figure 3. For clarity, only a selection of candidates is plotted. In both models many other words are considered.

At the broadest level, the two models produce similar output: the same candidates are considered at the same times. The most obvious difference between the two is that Shortlist B reaches a completely unambiguous interpretation of the input: at the offset of the input only the intended words have a probability greater than 0.01. In contrast, although the two intended words are activated most strongly in Shortlist A, the unintended word /kArp}r/ is almost as strongly activated as the intended word /kAr/ at the end of the input. This has two causes. In Shortlist A, the activations are partly determined by the amount of bottom-up evidence. The longer word has more evidence than the shorter word, and this partly compensates for the fact that /kArp}r/ is being inhibited by both /kAr/ and /p}rson}/. A second factor is that the final interpretation in Shortlist B is strongly influenced by frequency, whereas there is no frequency effect in Shortlist A. The effect of frequency can be seen most clearly in the comparison between Figures 3b and 3c. /p}rson}/ (*personen*, people) is much higher in frequency than /p}rson/ (*persoon*, person), and /kAr/ is higher in frequency than /kArp}r/. The difference in probability between the words on the /kAr/ /p}rson}/ parsing and their competitors is therefore more extreme in the simulation incorporating frequency than in the simulation without frequency.

Shortlist B can thus segment continuous speech successfully and use following context to revise earlier interpretations.

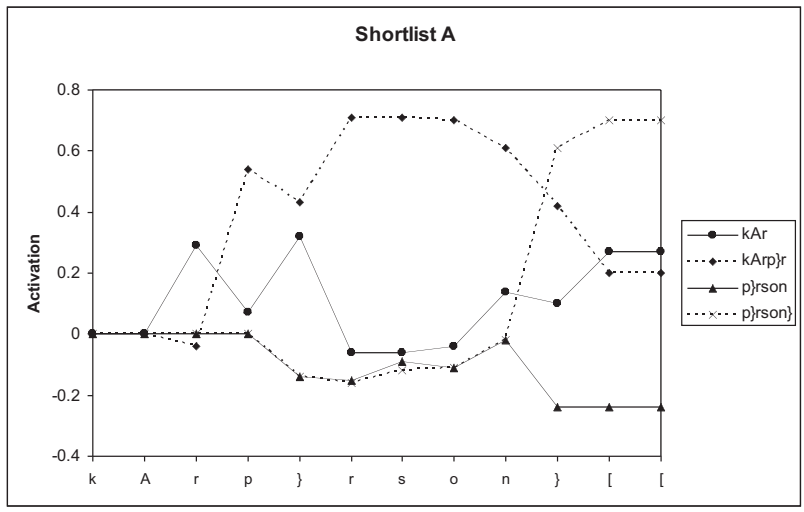
Word Frequency

Spoken word recognition is strongly influenced by the frequency with which a word occurs in the language (Connine, Mullennix, Shernoff, & Yelen, 1990; Dahan et al., 2001; Howes, 1957; Luce, 1986; Marslen-Wilson, 1987; Pollack, Rubenstein, & Decker, 1959; Savin, 1963; Taft & Hambly, 1986). However, neither TRACE nor Shortlist A gives any account of how frequency influences spoken-word recognition. Even though Dahan et al. (2001) have investigated ways of incorporating frequency into TRACE, there is no principled reason for preferring one of those methods over another.

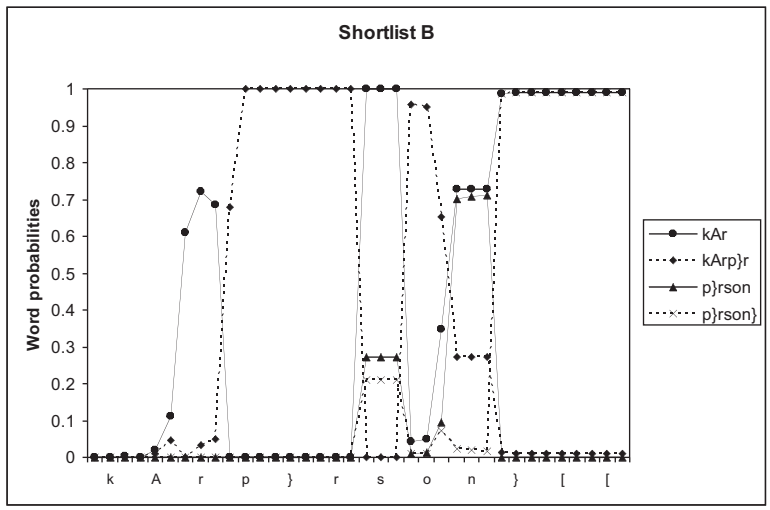
The Bayesian approach, in contrast, forces a specific account of word frequency effects. The essence of Bayesian statistics is to use evidence to revise prior beliefs. The expected frequency with which a word occurs in the language generally provides our best

[4] The source code for running Shortlist B simulations (and Merge B simulations, see below) is available at http: http://www.mrc-cbu.cam.ac.uk/~dennis/ShortlistB

A



B



C

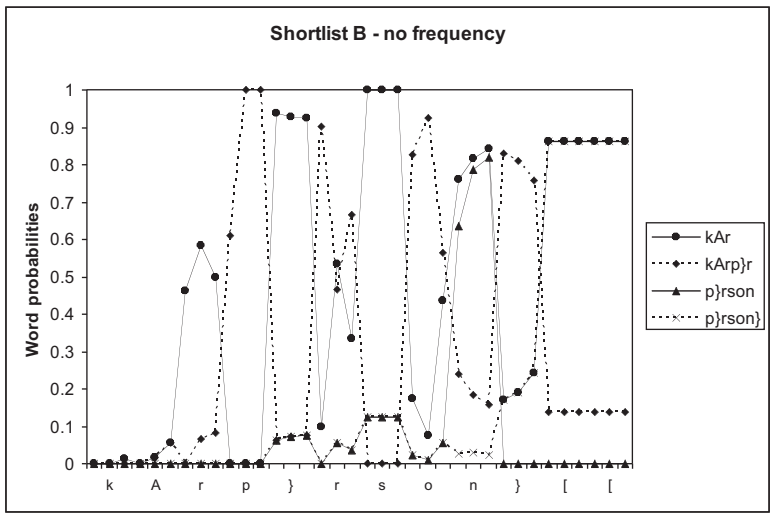


Figure 3. Patterns of lexical activation in Shortlist A (3A) compared with patterns of word probabilities in Shortlist B (3B and 3C) given the input *kar personen* ("kArp}rson}", ending with two silent segments "[[[">). 3C shows probabilities from Shortlist B with sensitivity to word frequency disabled (by setting all word priors to the same value). Note that there is one input time slice per segment in Shortlist A, but three per segment in Shortlist B.

initial estimate of the prior probability of encountering a particular word. In the present simulations we simply assume that P(Word) is given by the frequency of the word in the CELEX database. But many factors other than frequency can alter the probability of encountering a word in a particular context. The most obvious of these are the local semantic and syntactic context. In ASR systems, these contextual prior probabilities are usually incorporated into what is known as a language model. Frequency itself is a unigram language model. In the Bayesian approach, frequency and context are therefore given a unified explanation.

Norris (2006) provides an in-depth discussion of the advantages of a Bayesian interpretation of word frequency effects along with a number of simulations of word-frequency effects in visual word recognition. For example, he shows that the commonly observed logarithmic relationship between identification time and frequency (e.g., Whaley, 1978) falls directly out of a Bayesian model, even though the model is driven by standard linear probabilities (as it must be for Bayes's theorem to be valid).

An important feature of the way frequency influences P(Word|Input) in Bayes's rule is how frequency trades off against perceptual information. The better the perceptual information, the less the effect of frequency (Luce & Pisoni, 1998). As the perceptual evidence for a word increases, the denominator in Equation 2 tends to become dominated by the frequency-weighted evidence for that particular word. The overall probability of identifying the word therefore asymptotes to 1.0, regardless of the frequency of the word. That is, word frequency has a potentially large effect when perceptual evidence is poor, but this decreases as the perceptual evidence improves. This is a highly desirable state of affairs. In the absence of reliable perceptual evidence, it makes sense to be influenced by prior knowledge of the probability of the word. In the limiting case, when there is no perceptual evidence, frequency is the only available basis for responding. However, once the perceptual evidence becomes completely unambiguous, frequency should never override it. Because frequency and context are treated identically in the Bayesian approach, in that they both alter prior probabilities, the influence of context will also be dependent on the reliability of perceptual evidence. Although context will influence recognition when the perceptual evidence is poor, context will never be able to override reliable perceptual evidence. This provides the ideal way of taking full advantage of contextual information without running the risk of hallucinating.

This interplay between perceptual and contextual information has been studied extensively by Massaro and colleagues (Massaro, 1979; Massaro & Cohen, 1983a, 1983b; Massaro & Friedman, 1990; Massaro & Oden, 1995). The trade-off in Shortlist B between perceptual evidence on the one hand and frequency/contextual evidence on the other has a direct parallel with a similar trade-off in the FLMP model. Indeed, the "American football" pattern (e.g., Massaro, 1987), where context plays a greater role in the ambiguous region in the middle of a phonetic continuum than at the unambiguous endpoints, has been taken as a kind of trademark for the FLMP.

The Bayesian procedure for combining perceptual and frequency information has significant advantages over other ways of incorporating a frequency bias in models of word recognition. In the logogen model (Morton, 1969), frequency is represented as a constant additive bias on resting levels or thresholds. The effect of frequency is to reduce the amount of perceptual evidence required for recognition, regardless of the absolute amount of perceptual evidence for either that word, or for any other word. This way of incorporating a frequency bias into a model runs the risk that, if the frequency bias is too strong, low-frequency words might never be recognized. Forster (1976) pointed out that if a low-frequency word has a high frequency neighbor, then the higher frequency word could always have more activation than the low-frequency word. This might be a particular problem in speech recognition where many words do not become unique until their final phoneme. If the frequency bias were sufficient to make a high-frequency word be recognized early, then low-frequency words in the same cohort would be likely to be misidentified as the high-frequency word. Equally badly, a very high-frequency word, if it were embedded in a sequence of very low-frequency words, might dominate its competitors and lead to a complete misanalysis of the input. The Bayesian approach to word frequency in Shortlist B avoids these problems.

Although this approach is an important feature of the new model, there are no data on frequency using Dutch materials that we can simulate directly. Instead, we illustrate the behavior of the model by constructing sets of Dutch stimuli modeled as closely as possible on the English stimuli used in experiments by Luce and Pisoni (1998) and by Dahan et al. (2001). All subsequent Shortlist B simulations will be of experiments carried out in Dutch. These simulations therefore use the exact stimuli used in those experiments.

Frequency and Neighborhood Effects

Luce and Pisoni (1998) reported a lexical decision experiment (Experiment 2) where they orthogonally manipulated word-frequency, neighborhood density, and neighborhood frequency of CVC words. We selected eight sets of 34 Dutch CVC words that mirrored Luce and Pisoni's original English stimuli. The characteristics of the English and Dutch stimuli are given in Table 2 and Luce and Pisoni's lexical decision data are shown in Table 3. The frequency difference between high- and low-frequency words is much smaller for our stimuli than for those of Luce and Pisoni. This is because of the constraints imposed by matching our stimuli closely across conditions. We thought it more important to produce matched sets of stimuli than to equate our frequencies with those of Luce and Pisoni.

The main points to note about their data are that there were significant effects of both word frequency and neighborhood frequency in both RTs and errors, but the effect of neighborhood density was less consistent. For low-frequency words, the effect of neighborhood density was inhibitory in RTs but facilitatory in errors. The results of simulations using the Dutch stimuli are plotted in Figure 4.

Shortlist B simulates the human data well. The probability of high-frequency words is higher than that of low-frequency words. In addition, there are smaller effects of both neighborhood frequency (word probability is higher for words in low-frequency neighborhoods) and neighborhood density (word probability is higher for words in low-density neighborhoods). We confirmed that these patterns were robust across items by performing an analysis of variance on item probabilities for all input slices. The effects of frequency ($F2(1, 264) = 25.01$, $p < .0001$), neighbor-

Table 2
Properties of the English Materials from Luce & Pisoni (1998) and the Matched Dutch Materials Used in Shortlist B Simulations

	English			Dutch		
Property	Mean word frequency (per million)	Mean density	Mean neighbor frequency (per million)	Mean word frequency (per million)	Mean density	Mean neighbor frequency (per million)
High word frequency						
High density						
High neighborhood frequency	254	22	370	47	24	320
Low neighborhood frequency	254	22	46	48	22	73
Low density						
High neighborhood frequency	254	11	370	47	16	447
Low neighborhood frequency	254	11	46	47	12	43
Low word frequency						
High density						
High neighborhood frequency	5	22	370	3	21	304
Low neighborhood frequency	5	22	46	3	21	121
Low density						
High neighborhood frequency	5	11	370	2	15	443
Low neighborhood frequency	5	11	46	2	13	161

Note. Luce and Pisoni (1998) provide only the means for each pair of conditions, not the individual cell means; here we are assuming that there was no variability among cells within conditions.

hood frequency ($F2(1, 264) = 7.37$, $p < .01$) and neighborhood density ($F2(1, 264) = 4.00$, $p < .05$) were all significant.

Luce and Pisoni (1998) did not in fact present any NAM simulations of their lexical decision data. Simulations using NAM would produce only a single probability value for each word. Because the confusion data collected by Luce and Pisoni come from identification in noise, phonemes are never identified with a probability near 1.0. In NAM, factors such as lexical neighborhood size and word frequency therefore have their effect on the overall probability of correct identification. However, with some exceptions, listeners can identify the input very reliably by the end of a diphone. Consequently, by their end, most words in Shortlist B are identified almost perfectly, and competing words become completely inconsistent with the input. This means that almost all of the interitem variability in Shortlist B occurs before the end of the word, very much as it does in TRACE and Shortlist A. As can be seen in Figure 4, the effects of frequency, neighborhood frequency, and neighborhood density mainly influence how quickly the probabilities rise over time and not their asymptotic value. A more general point is that in longer stretches of input, most alternative paths or interpretations of the input die out quite quickly, and there are only multiple paths covering the last few phonemes on the input. Informal observation suggests that multiple paths rarely extend more than two words back. However, if the input is degraded in any way, multiple paths will become far more prevalent. The effects of stimulus quality on identification can be seen in the next simulation.

Table 3
Mean Reaction Time (RT; in ms from Word Offset) and Mean Error Rate (%) in Auditory Lexical Decision, From Luce & Pisoni (1998)

Property	Mean RT	Mean error
High word frequency		
High neighborhood density		
High neighborhood frequency	409	7
Low neighborhood frequency	392	5
Low neighborhood density		
High neighborhood frequency	382	7
Low neighborhood frequency	377	6
Low word frequency		
High neighborhood density		
High neighborhood frequency	451	11
Low neighborhood frequency	445	10
Low neighborhood density		
High neighborhood frequency	463	18
Low neighborhood frequency	421	16

Word Frequency and Stimulus Quality

The Dutch words selected to match the stimuli used by Luce and Pisoni (1998) can also be used to illustrate the way that the influence of frequency in the Bayesian framework varies as a function of the reliability of perceptual information. We can do this by modifying the phoneme confusion probabilities. That is, we can calculate the confusion probabilities that we would expect to obtain if the listening conditions were to permit listeners to make more accurate responses. The details of the procedure are given in Appendix B. Figure 5 shows simulations averaged over all high- and all low-frequency words. The upper panel shows the separate probabilities for high- and low-frequency words with both the empirically determined confusion probabilities and the modified probabilities that make the phonemes less confusable. The lower panel shows the difference between high- and low-frequency items under the two conditions.

The critical feature of these simulations is that the frequency effect decreases as the perceptual evidence improves. This is particularly apparent in the asymptotic levels of performance. Using the empirically determined probabilities, there is residual ambiguity at the end of the word that allows room for an influence of word frequency. In contrast, when using the modified probabilities, the ambiguity is effectively removed, so frequency plays a

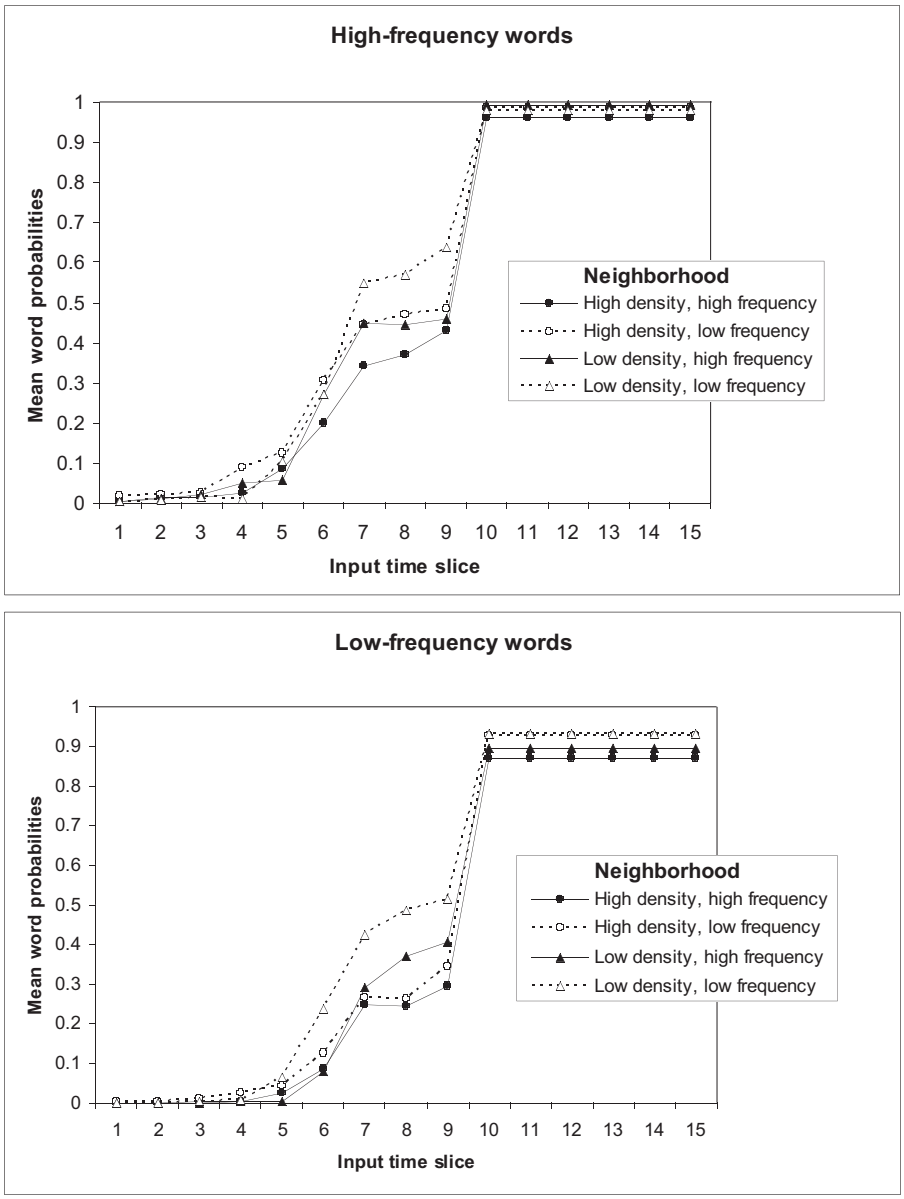


Figure 4. Mean word probabilities in Shortlist B for Dutch materials based on the Luce and Pisoni (1998) study. The upper panel shows the average results for four sets of 34 high-frequency CVC words: those with high density and high frequency neighborhoods, those with high density but low frequency neighborhoods, those in low density but high frequency neighborhoods, and those in low density and low frequency neighborhoods. The lower panel shows the average results for four sets of 34 low-frequency CVC words in the same four conditions.

smaller role in recognition. In terms of Equation 1, *P(Evidence|Hypothesis)* tends towards zero for all competitor words, therefore the numerator and denominator both reduce towards *P(Evidence|target word)* × *P(target word)*, so that *P(target word|Evidence)* tends towards 1.0.

Shortlist B can therefore simulate word frequency effects and how these effects interact with lexical neighborhood characteristics and with stimulus quality. We now turn to a detailed examination of how the influence of word frequency changes over time.

The Time-Course of Word-Frequency Effects

Dahan et al. (2001) studied the effect of word frequency of a target word and its neighbors in an eye-tracking study. On each trial participants saw a display containing pictures of four objects. In the critical conditions, the names of three of the objects overlapped phonologically, and the name of the fourth object was phonologically unrelated. For example, on one trial, participants saw pictures of a *bench*, a *bed*, a *bell* and a *lobster*. Participants were instructed to "Pick up the bench." The target words had a

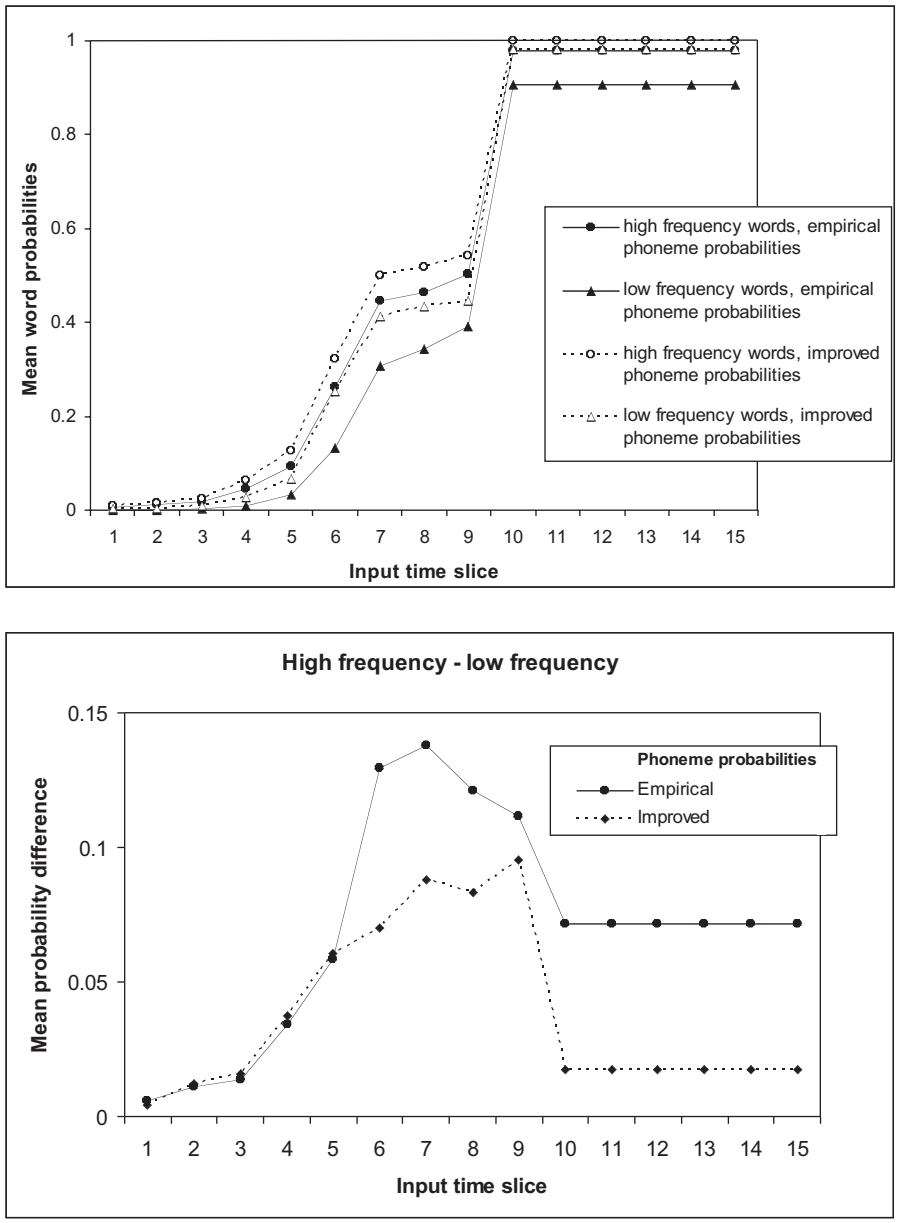


Figure 5. Mean word probabilities in Shortlist B for Dutch high- and low-frequency materials based on the Luce and Pisoni (1998) study, averaged over the neighborhood density and frequency conditions. The upper panel shows the mean word probabilities for these two sets of words using the model's default parameters (as in Figure 4), that is, using the confusion probabilities derived from Smits et al. (2003), with no changes to phoneme likelihoods (empirical probabilities), and also where the estimated variance of the probability density functions for the empirically determined phoneme likelihoods was halved and the likelihoods recomputed (improved probabilities), simulating the effect of perceptually clearer input (see Appendix B). The lower panel shows the mean probability difference between the two sets of words for the same two simulation runs.

mean frequency of 14.5 per million (Francis & Kucera, 1982). One of the nontarget competitors was low frequency (10 per million) and one was high frequency (138 per million).

Dahan et al.'s (2001) data (from their Set A items) are reproduced in Figure 6 along with simulations from Shortlist B. Once again, it was necessary to generate a set of Dutch stimuli that were matched to the English materials in the original study. Twenty triplets consisting of three words that shared their first two or three phonemes were selected from CELEX. One of the words in each triplet was the analogue of the target word (mean frequency 9 per million), a second was the low-frequency competitor (mean frequency 9 per million), and the third was the high-frequency competitor (mean frequency 139 per million). Because these materials were not intended for use in an actual eye-tracking study, it was

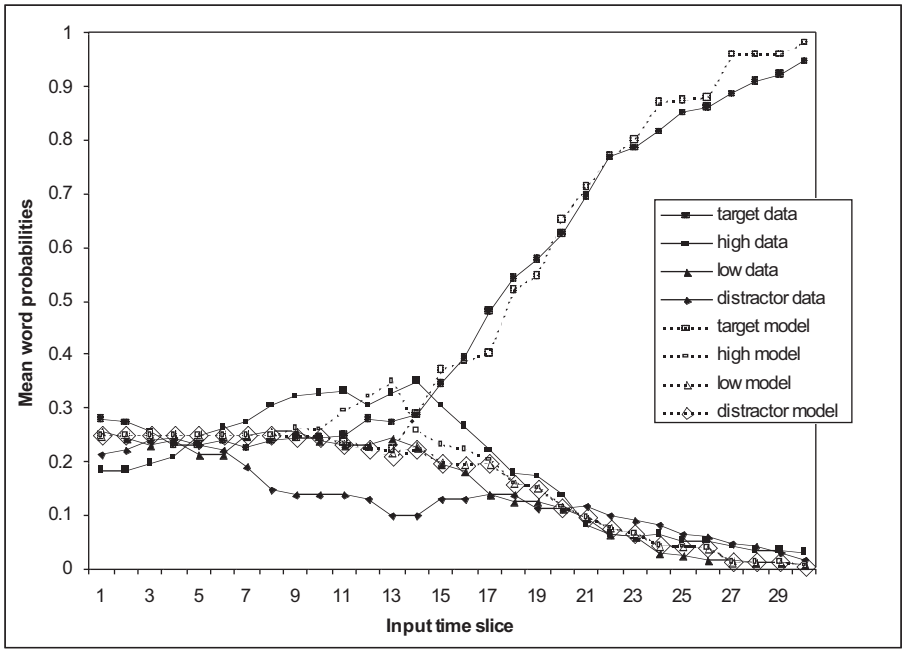


Figure 6. Fixation probabilities for the English materials in Dahan et al. (2001; data), aligned with those from Shortlist B on analogous Dutch materials (model). Fixation probabilities are shown for target words (e.g., *bench;* target data/model), high-frequency phonological competitors (e.g., *bed;* high data/model), low-frequency phonological competitors (e.g., *bell;* low data/model), and unrelated objects (e.g., *lobster;* distractor data/model).

not necessary that all of them refer to picturable objects. The unrelated distractor in each triplet was a target from one of the other triplets that began with a different phoneme. Mean probabilities derived from Shortlist B for these 20 sets of items are shown in Figure 6.

In presenting their TRACE simulations, Dahan et al. (2001) corrected for the fact that, because there are only 4 visual targets, the baseline fixation probability for each visual target is 0.25. In the Shortlist B simulations, the probabilities shown are therefore calculated as follows: Corrected fixation probability = ((1.0—sum of raw word probabilities for all 4 targets)/4) + raw word probability. That is, we assume that any probability not taken up by one of the four target words is distributed equally among all of the targets.

Note also that, in common with Dahan et al. (2001) and with other simulations of eye-movement data (e.g., Allopenna et al., 1998), we have adjusted the position of the simulated results on the time axis to allow for lag in initiation of eye movements. As noted earlier, because of the nature of the diphone database, changes in probability in the model are likely to occur earlier than probabilities computed by listeners. Listeners in the Smits et al. (2003) study were not required to identify the diphones quickly, so the database tells us how much information listeners can extract from the input at a given point and not how long it takes listeners to actually extract that information. The simulated results are shifted back in time by eight slices. The time scale is mapped onto the data using the same procedure as in Dahan et al. Our targets were, on average, 5.3 phonemes long; their targets were, on average, 5.4 phonemes long. This corresponds to about 16 time-slices in Shortlist B. Because Dahan et al.'s Set A items were 498-ms long, on average, one time-slice in Shortlist B is thus equivalent to about 31 ms. Each slice in the model therefore maps almost exactly onto one of the 33-ms time-slices in the Dahan et al. data.

The simulations in Figure 6 closely parallel the Dahan et al. data. Fixations to the high-frequency competitors are most probable (in the model and the data) until the effect of the phonological divergence of the target from the competitors starts to emerge in the eye-movement record. Furthermore, there is little difference in fixation probabilities between targets and low-frequency competitors in either the data or the model until the target starts to dominate the fixation pattern. The only difference between the data and the model is that the model tends to overestimate the proportion of fixations to the unrelated distractor. With respect to the three phonologically related conditions, Shortlist B thus captures the time-course of fixation probabilities very accurately.

Dahan et al. presented simulations of their data using TRACE and contrasted three different methods of implementing word frequency in TRACE. They compared implementing frequency in terms of resting level, connection weights, or in a postactivation decision stage. In terms of ability to simulate the overall pattern of data, all three methods were roughly equivalent and little different from simulations with no word frequency mechanism. The main difference between the simulations was in terms of their ability to account for the difference between high- and low-frequency competitors. Here the best fit was obtained when frequency was implemented by varying the connection weights to words. In the data, fixation probabilities only become greater for high- than low-frequency competitors after about 100 ms. In fact, before that, there is a slight reversal. However, the resting level and postactivation decision simulations both show a high-frequency advantage

from the outset. Only in the connection weight simulation does the frequency difference build up over time. Shortlist B shows the same pattern.

Figure 7 shows the Shortlist B simulations of the difference between high- and low-frequency competitors along with the equivalent data from Dahan et al. (2001) and their connection-weight simulation of these data in TRACE. The Shortlist B simulation is very similar to the data and close to the TRACE simulations. The main difference is that the peak is narrower in the Shortlist B simulation. As noted earlier, a more complete model would include internal noise in the simulations. One effect of including noise would be to smooth this peak.

The TRACE simulations use 20 parameters, and the connection-weight mechanism is just one of a number of possible procedures that might be added to TRACE to make it sensitive to frequency. Significantly, there is no theoretical basis to prefer the connection-weight procedure to any other. In contrast, the Shortlist B simulation depends only on a single free parameter, representing the time-alignment between the simulation and the data. This parameter, which corresponds to 264 ms (eight 33-ms slices), is roughly what would be expected on the basis of eye-movement research. Even the simplest eye movements have a latency of 150–175 ms (Rayner, Slowiaczek, Clifton, & Bertera, 1983). In the context of a more complex linguistic task, such as reading, the time to program and execute saccades is generally estimated to take about 100 ms more. For example, Version 7 of the E-Z Reader model (Reichle, Rayner, & Pollatsek, 2003) uses an estimate of 245 ms as the time required to program a saccade. Furthermore, a common assumption in visual-world studies (including that of Dahan et al., 2001) is that programming and launching a saccade introduces a delay of at least 200 ms between auditory stimulation and a resulting eye movement. Other than this fixed time-alignment parameter, there are no free parameters in the Shortlist B simulation. For example, the way the model simulates frequency effects is independent of the number of candidates or the number of paths. The behavior of the model follows entirely from the underlying assumption of optimality. There is really nothing we could do to make the model behave differently.

Speech Segmentation: The Possible Word Constraint

Although segmentation can be achieved using only a lexical competition mechanism (either the interactive-activation type in Shortlist A or TRACE or the path-based search in Shortlist B, as the *kar personen* simulation showed), there is more to segmentation than this. Human listeners are able to make use of a range of cues to help them segment the input into words. Shortlist A has been extended to simulate the effect of segmentation cues, such as those provided by metrical information (Cutler & Norris, 1988; McQueen et al., 1994; Vroomen & de Gelder, 1995) and phonotactic information (McQueen, 1998). A unified account of segmentation effects in a competition-based model was presented by Norris et al. (1997). In these Shortlist A simulations, segmentation cues affected word recognition by modulating lexical activation. Specifically, words that are misaligned with cued lexical boundaries had their activation levels reduced according to the operation of what Norris et al. termed the Possible Word Constraint (PWC).

In experiments using the word-spotting task, Norris et al. (1997) showed that listeners found it far harder to spot words (e.g., *sea*) embedded in nonsense words, such as "seash," than in nonsense words such as "seashub." In the former case, the nonsense word has to be parsed into the word "sea" plus the single phoneme residue /ʃ/. In the latter case, the residue is a syllable. Norris et al. therefore proposed that segmentation is driven by the PWC: The preferred segmentation of the input is always in terms of units that are possible words. The single consonant /ʃ/ is not a possible word

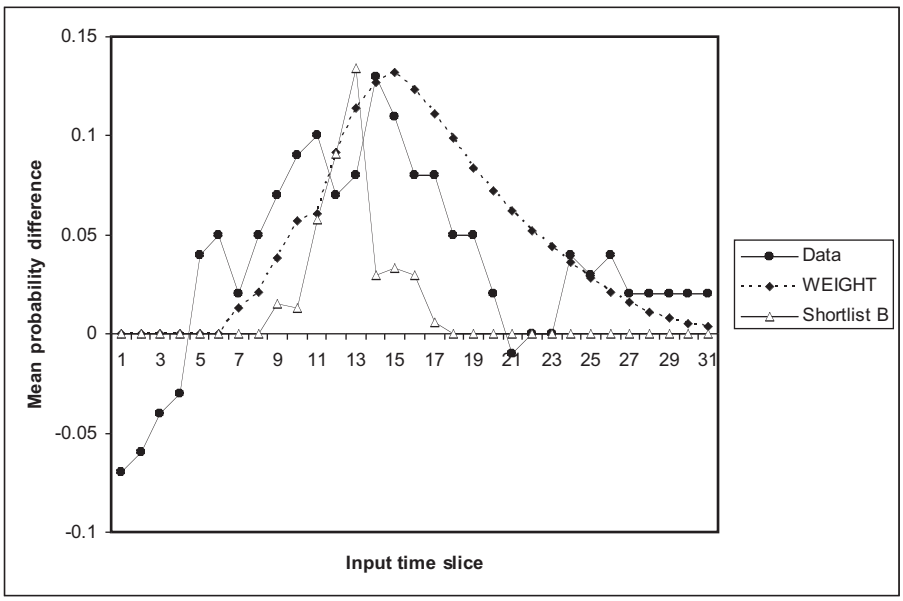


Figure 7. Differences in fixation probabilities between high- and low-frequency competitors: the Dahan et al. (2001) data (Data), Shortlist B simulations (Shortlist B), and connection-weight simulations in TRACE (WEIGHT).

in English, whereas the syllable /ʃʌb/ is. The word "sea" in "seash" is therefore penalized because it is not part of a segmentation consisting of possible words.

In the experiment just described, the PWC applies to the residue between the end of the candidate word and the silence at the end of the nonsense word. More generally, however, the PWC operates between the ends of candidates (their onsets or their offsets) and the nearest likely word boundary. Such a boundary can be indicated by any one of a number of different segmentation cues (e.g., a phonotactically signaled boundary, McQueen, 1998, or the onset of a strong syllable, Cutler & Norris, 1988). The extent to which a candidate that violates the PWC is inhibited is assumed to depend on the reliability of the segmentation cue (cf. Mattys et al., 2005).

One of the central motivations for postulating the PWC was that it would help deal with unknown words. To some extent the competition process in Shortlist A will break down if parts of the input do not correspond to known words. The PWC ensures that the input is always parsed in terms of units that could possibly be words. Norris, McQueen, Cutler, Butterfield and Kearns (2001) and Cutler, Demuth, & McQueen (2002) have shown that the PWC appears to be a language-universal constraint. That is, the residue does not need to be a possible word in the listener's language; it need only be a syllable. Johnson, Jusczyk, Cutler, and Norris (2003) have also shown that 12-month-old infants behave as if they are observing the PWC. Infants should indeed benefit from the PWC as, in the course of acquiring the vocabulary of their native language, they continually encounter words that are unknown to them.

How can the PWC be implemented in Shortlist B? The central requirement of the PWC is to penalize paths containing words whose boundaries are misaligned with known syllable boundaries in the input. That is, even if a path consists of a series of contiguous words that fully span the input, that path should be penalized if one or more of those words violates the PWC. This can be achieved straightforwardly by reducing the probability of words that violate the PWC. This, in turn, will reduce the probability of the paths those words lie on. This procedure applies both to real words and to dummy words. In practice, the path probability is multiplied by a very small number (the *PWC parameter*, 10^{-9}). This has exactly the same effect as inserting a very low-frequency word into the path. The influence of the PWC has parallels with the effect of dummy words: The PWC parameter penalizes paths that violate the constraint; but, if the only possible path is one that violates the PWC, that path will still be assigned a high probability.

In the example in Figure 1, the connections shown with dashed arrows are parts of paths which contain dummy words which fail the PWC (e.g., the single segment /l/ between *inner* and *eye* is a dummy word, but this path violates the PWC). The connections shown with dotted arrows, however, are parts of paths with dummy words that pass the PWC (e.g., the vowel-consonant sequence /ɒg/ between *cattle* and *in* is a dummy word with a vowel, on a path that is consistent with the PWC). Consequently, paths leaving a syllable as a residue have a far larger probability than paths leaving a consonant as a residue.

Earlier we saw that, in order to deal with unknown words, the model's beliefs had to be modified to include a dummy word. The model's beliefs also have to be modified to allow it to carry out experimental tasks where the input has different statistical properties from normal speech. One example of this is the word-spotting task. The word-spotting task is unusual in that the input never corresponds to a complete word. All word targets are embedded in nonword carriers and there is never a path consisting of a sequence of words that fully spans the input.

In the study by Norris et al. (1997), the target "sea" in "seash" can only be recognized by parsing the input in terms of "sea" + "sh," where "sea" violates the PWC. In contrast, the dummy word "seash" matches the input (by definition) and does not violate the PWC. Given that the probability of "sea" will thus be less than that of "seash" (because "sea" violates the PWC), the embedded word will never be recognized. Outside of the context of a word-spotting experiment, this seems to be exactly the right behavior. Try randomly interjecting "seash" into conversations and count the number of times someone says 'Oh, you mean "sea," but you've added "sh" onto the end.'

Participants in word-spotting experiments, however, need to revise their prior beliefs about where words might be located in the input. In the following simulations, we do this by reducing the probability of dummy words that fully span the input, so as to increase the probabilities of embedded words. That is, we reduce the probability that the entire input (e.g., "seash" or "seashub") will be a dummy word. We do so using a fully-spanning-path parameter: any path that consists of a single dummy word has its probability multiplied by a very small number (10^{-10}). This parameter would not apply in normal listening situations, nor indeed in an auditory lexical decision task, where the listener is told that some inputs will not be real words.

A third parameter reflects the probability that the input will contain sequences that will not correspond to known words (irrespective of whether those sequences span some or all of a complete input utterance). This dummy-word parameter in the following simulations is again a very small number (10^{-12}): All dummy words have this prior probability assigned to them, just as if they were words with an extremely low frequency of occurrence (the parameter thus acts in the model in exactly the same way as word frequency does for real words). As with the fully-spanning-path parameter, the dummy-word parameter will have to change depending on the nature of the linguistic input. In the lexical-decision simulations reported below, for example, the probability of a dummy-word interpretation of the input becomes much higher because half of the stimuli are nonwords. Listeners need to take this kind of prior knowledge into account when computing response probabilities in different experimental tasks. This flexibility is required not only for listeners to perform optimally in different psycholinguistic experiments, but also in different real-world listening conditions. In a commentary on an international football match, for example, the names of many of the players may be unknown words. It would be easier to parse this kind of input if the probability of dummy words were increased to reflect this increased frequency of novel words. The Bayesian approach in Shortlist B thus provides a straightforward and theoretically motivated account of the way in which listeners respond to different probabilities of encountering unknown words in different listening situations.

The PWC, fully-spanning-path and dummy-word parameters thus control the segmentation behavior of Shortlist B. The first parameter is essential for the model to be able to capture effects of

the viability of sequences of the input as possible words, but simulations show that the model's behavior was stable when this parameter was varied over a range of numerical values. The other two parameters are less critical to the model's segmentation performance per se, but are required for the model to be able to give an adequate account of performance across a range of tasks. Once again, model behavior does not depend on precise numerical values for these parameters, other than that the fully-spanning-path parameter must be nonzero for the model to simulate word-spotting data. Note that the same values for all three parameters are used in all Shortlist B simulations reported in this article. The value of these parameters has a negligible effect in the simulations where there is a fully spanning path consisting entirely of words.

We now turn to simulations of segmentation experiments. McQueen and Cutler (1998) examined the operation of the PWC in Dutch. Their materials can therefore be used in Shortlist B simulations. As in Norris et al. (1997), listeners were asked to spot real words embedded in either single consonant or syllabic nonsense contexts. Target words were 24 bisyllabic verbs and 24 bisyllabic nouns, with preceding consonant, bisyllabic, strong-syllable or weak-syllable contexts, as shown with examples in Table 4. A further 24 targets were included in contexts that did not test the PWC, and there were 144 fillers that did not contain embedded words. As in the Norris et al. study, therefore, the probability that a trial would include a real word target was .33. As shown in Table 4, both types of target word were harder to spot in single-consonant contexts than in any type of syllabic context.

Simulations were carried out using the same 48 target words, each in three nonsense contexts. The results of these simulations are shown in Figure 8. As in the human data, the model performs much better on verbs and nouns in syllabic contexts than on these same words in single-consonant contexts. These simulations thus show that Shortlist B, with its implementation of the PWC, can segment continuous speech into words. Path-based probability computations provide the basic means by which continuous speech can be segmented, even when there are no cues to the location of word boundaries that could help. These cues are nonetheless used, when available, via the PWC, to reduce the probability of paths that contain impossible words.

Vroomen and de Gelder (1995) also examined segmentation of continuous Dutch. They used a cross-modal identity-priming technique to examine the joint influence of metrical structure and lexical competition on segmentation. Participants heard bisyllabic spoken sequences containing a Dutch CVCC word followed by a VC sequence with either a strong or a weak vowel. The strong second syllables were either consistent with many other Dutch words or with few Dutch words (see Table 5 for examples), and the weak syllables were consistent with no Dutch words. Participants saw visual letter strings 250 ms after the end of the embedded CVCC words and made lexical decisions to those visual stimuli. Relative to an unrelated control condition, lexical decisions were faster when the visual targets matched the spoken prime words, but the amount of priming showed a stepwise pattern. Priming was largest when the second syllables of the spoken sequences contained weak vowels, smaller when they contained strong vowels with few lexical competitors, and smallest when they contained strong vowels with many competitors. We have previously interpreted these results as being consistent with the operation of the PWC in Shortlist A (Norris et al., 1997). Words are poorer hypotheses (and thus generate weaker priming) in strong-strong than in strong-weak sequences because the onset of the second strong syllable is a likely word boundary but that of the weak syllable is not (Cutler & Norris, 1988). The CVCC words are thus misaligned with those boundaries and have the PWC penalty applied to them. For example, *melk*, milk, in *melkaam* is misaligned with the likely word boundary before the /k/ because the /k/ is not a vowel and thus not a possible word. In addition, the number of words beginning at that segmentation point influence recognition: The more words there are beginning at the second syllable, the stronger the competition between them and the target CVCC words.

Table 4
Design, Example Stimuli, and Data (Mean Reaction Times in ms and Mean Percentage Error Rates in Parentheses) from McQueen and Cutler (1998)

	Context			
Item type	Consonant	Bisyllable	Strong syllable	Weak syllable
Verb targets	dwonen	dukewonen		kewonen
(*wonen* = to live)	739 (16%)	432 (8%)		413 (5%)
Noun targets	blepel		kulepel	selepel
(lepel = spoon)	667 (9%)		435 (5%)	380 (4%)

The results of the Shortlist B simulations using the Vroomen and de Gelder (1995) stimuli are shown in Figure 9. The model captures the pattern in the human data. Thus, although the simulations of the McQueen & Cutler (1998) study show how the operation of the PWC in Shortlist B influences how the model segments continuous speech, the present simulations show in addition how this segmentation process is modulated by lexical competition. Prime probabilities in both strong–strong conditions are lower than in the strong–weak condition because of the application of the PWC penalty. In addition, the more paths there are with different words beginning at the onset of the second strong syllable, the lower the probability of the prime.

Note that the data being simulated here come from a cross-modal identity priming task. That is, responses cannot have been driven directly by the probabilities of the spoken words. Instead we assume that the probabilities of the spoken prime words can modulate the prior probabilities of the visually presented target words, and it is this change in priors that produces priming. This is exactly the account of priming in visual word recognition proposed in the Bayesian Reader model (Norris, 2006). As we explained earlier, the critical factor influencing recognition in the Bayesian framework is the prior probability of each of the lexical hypotheses. Although our main emphasis here has been on priors determined by frequency of occurrence, they will also be altered by the context in which the word appears. If information in the speech signal makes a particular word more probable, then this change in priors will speed recognition of that word when it appears visually. The changes in word probabilities observed in the Shortlist B simulations should therefore lead to parallel changes in the speed of recognition of the visual target words given the three types of speech context tested by Vroomen & de Gelder (1995).

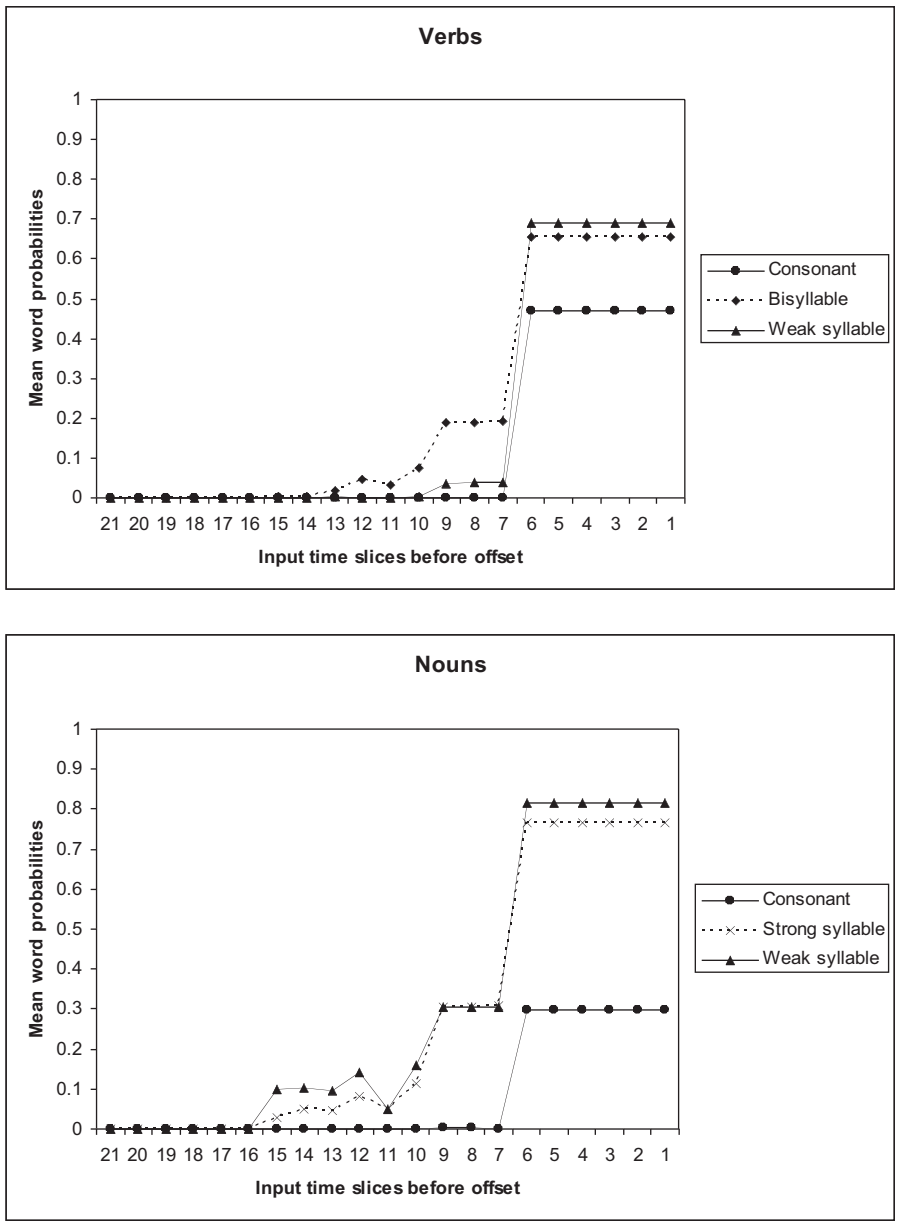


Figure 8. Mean word probabilities in Shortlist B for the materials from McQueen and Cutler (1998). The upper panel shows the average results for 24 verbs in consonant, bisyllable and weak-syllable contexts. The lower panel shows the average results for 24 nouns in consonant, strong-syllable and weak-syllable contexts.

Shortlist B can thus simulate how segmentation is influenced by competition among spoken words and can, in principle, account for priming effects.

Perceptual Match and Goodness of Fit

Word recognition necessarily involves a comparison of the evidence in the current acoustic input with stored knowledge about the phonological form of words. Models of speech recognition generally assume that the degree of match between the signal and a lexical representation is determined by a similarity metric reflecting the perceptual distance between the input and the lexical representation. Experiments examining the effect of mispronunciations on lexical access, for example, have shown that the degree of disruption caused by the mispronunciation depends on the phonetic similarity of the mispronounced segment to the correct segment. The substitution of a phonetically unrelated segment can completely block lexical access (Marslen-Wilson & Zwitserlood, 1989), whereas substitution of phonetically similar segments does not necessarily do so (Connine, Blasko, & Titone, 1993). There appears to be more support for lexical hypotheses as the phonetic similarity of the mispronounced segment to that in the intended word increases (Connine, Titone, Deelman, & Blasko, 1997;

Table 5
Design, Example Stimuli, and Data, Mean Reaction Times (RTs) in ms, From Vroomen and de Gelder (1995)

Condition	Spoken prime	Visual target	RT
Control	lastem	MELK	621
SS-many	melkaam	MELK	602
SS-few	melkeum	MELK	589
SW	melkem	MELK	578

Note. Conditions are defined in the text; *melk* means "milk," and *last* means "load."

Marslen-Wilson, Moss, & van Halen, 1996), and distortions of the phonetic properties of a segment that do not change the segment's identity influence lexical processing as a function of the size of those distortions (Andruski, Blumstein, & Burton, 1994; McMurray, Tanenhaus, & Aslin, 2002; but see van Alphen & McQueen, 2006).

Models of spoken-word recognition account for these findings using similarity metrics based on perceptual distances. In TRACE, for example, overlap in terms of features and phonemes between the signal and the lexicon determines degree of lexical activation. In contrast, in a Bayesian approach, posterior probabilities are driven by likelihoods and not by any simple measure of perceptual or physical similarity. In Figure 2, whether or not the input I_x provides more support for phoneme A or phoneme B does not depend on the distance between I_x and the mean of the likelihood functions for A and B but only on the likelihoods (the height of the pdf) at I_x. That is, the critical measure is how likely we are to observe a particular acoustic-phonetic signal, given that the signal was generated by that phoneme, and not by how similar the representations of the phoneme and the signal are (though of course similarity and likelihood measures may often be closely related). This means that the pattern of variation in the realization of a particular word or phoneme is more important than any measure of absolute distance in some physical or perceptual space (cf. Newman, Clouse, & Burnham, 2001). In fact, if decisions are to be Bayesian optimal, posterior probabilities must be driven by the likelihood functions and not by perceptual distance or similarity.

As an illustration of this, consider the case of a phoneme that is always realized in almost exactly the same way. As shown in Figure 2, there is little variability between tokens of Phoneme A. In contrast, for another phoneme there might be wide variation in how it is realized (Phoneme B in Figure 2). In the former case, even a small deviation from the modal representation of that phoneme will greatly reduce *f(Evidence|Phoneme)*. That is, even a small deviation will make it very unlikely that the input signal originated from that phoneme. But in the latter case, a similar deviation may do little to alter *f(Evidence|Phoneme)* because prior experience has shown that that phoneme can be produced in a larger variety of different ways. If follows that an input that falls exactly half way between the peak values of A and B will not give equally strong support for A and B. *P(Phoneme B|Input)* will be greater than *P(Phoneme A|Input)* because at that point the likelihood of B is greater than A. Note that although there is no guarantee that the likelihood functions that listeners learn will take the simple Gaussian form shown in Figure 2, this argument does not depend on these distributions being strictly Gaussian.

The critical difference between computation of goodness of fit based on perceptual similarity versus likelihoods becomes clear

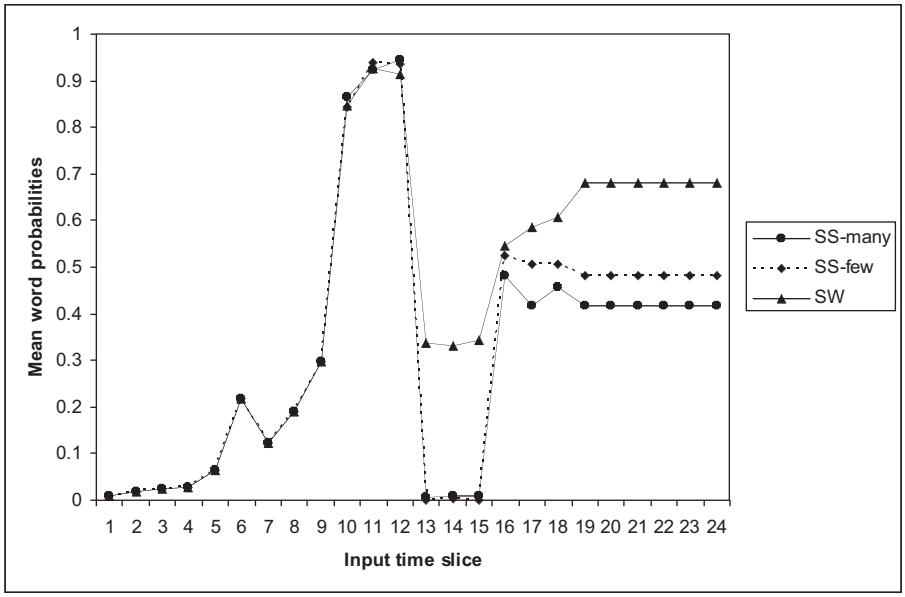


Figure 9. Mean word probabilities in Shortlist B for the materials from Vroomen and de Gelder (1995). Average results are shown for 44 CVCC prime words in three conditions: where the strong-syllable prime was followed by a strong syllable consistent with many lexical candidates (SS-many), where it was followed by a strong syllable consistent with few lexical candidates (SS-few), and where it was followed by a weak syllable (SW).

when one considers asymmetries in perceptual confusions. For instance, one phoneme may be more often misperceived as another phoneme than the reverse (as indeed occurred, e.g., in the diphone gating study, Smits et al., 2003; Warner et al., 2005). This cannot be explained in the simplest version of a model in which goodness of fit is based only on perceptual distance. On any similarity metric (e.g., the number of shared acoustic-phonetic features or the difference in a value of a continuous phonetic variable), one phoneme will be as different from another as the reverse. Any explanation of asymmetries in perceptual confusions about these phonemes would thus require some additional mechanism in a model based on perceptual distance, such as a decision bias. In a model in which perception is based on phoneme likelihoods, however, asymmetric patterns of confusion will arise naturally whenever the relevant likelihoods are asymmetric. In Figure 2, because of the difference in the width of the two functions (and not simply the distance between the peaks of the distributions), there is a wider range of values on the perceptual dimension for which Phoneme B is likely to be misrecognized as Phoneme A than the reverse. Tenenbaum and Griffiths (2001) explain how distributional properties of exemplars can give rise to different patterns of generalization in a Bayesian category-learning model. This explanation for asymmetries in phonetic perception is an important motivation for the assumption in Shortlist B that phoneme recognition is based on likelihood functions.

As we have already suggested, the knowledge necessary to compute likelihoods is probably initially acquired in infancy, as a result of exposure to the distribution of phonetic variability of phonological categories in the language the infant hears (Maye et al., 2002). But because this knowledge reflects the cumulative effect of prior experience with speech sounds, it should continue to change over the listener's lifetime. Importantly, in order to maintain optimal performance, the listener's estimate of the distributions of speech sounds should be continuously updated. Recent results indeed suggest that speech perception can be altered in response to the current input. That is, perceptual learning about speech sounds can occur in adulthood. For example, adult listeners appear to be able to adjust their phonetic categories as a result of the combination of prior lexical knowledge and very limited exposure to a talker speaking in an unusual way (Norris, McQueen, & Cutler, 2003). After a group of listeners were exposed to an ambiguous sound, midway between /f/ and /s/, in lexical contexts that indicated that the sound should be interpreted as /f/, those listeners interpreted more sounds on an /f/–s/ continuum as /f/ than another group of listeners who had been exposed to the same ambiguous sound but in /s/-biased lexical contexts. In terms of the density functions sketched in Figure 2, the distributions of phonemic categories were altered given lexical knowledge and only brief exposure to an idiosyncratic talker. These listeners had thus learned to adjust the estimate of the density functions that we argue are used to compute *f(Evidence|Phoneme)*.

Clayards, Aslin, Tanenhaus, and Jacobs (2007) have recently shown that adult listeners are sensitive to the distribution of phonetic cues. Listeners who were exposed to more strongly peaked bimodal distributions of voiced and voiceless stops on a VOT continuum produced a sharper category boundary in their identification responses to that continuum than listeners exposed to broader distributions. As Shortlist B predicts, as the variability of two phonemes on a phonetic dimension decreases, those phonemes' likelihood functions will become steeper; and, thus, for values on the dimension spanned by either of those functions, *f(Evidence|Phoneme)* will increase, leading in turn to a sharper category boundary between those phonemes. Clayards et al., thus, show that adult listeners are tracking the distribution of the phonetic realization of phonemes exactly as is required in a model in which word recognition is based on calculation of likelihoods, such as Shortlist B. Feldman and Griffiths (2007) illustrate the value of a Bayesian approach in understanding categorical perception and, more specifically, the perceptual magnet effect (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). They assume that in phonetic discrimination tasks, in addition to simply identifying the appropriate phonetic category of a speech sound, listeners attempt to extract phonetic detail and to recover the speaker's original target production. In order to perform optimally, the listener's inferences must be influenced by their prior knowledge of the structure of their phonetic categories. Feldman and Griffiths present simulations of data from Iverson and Kuhl (1995) to show that this leads to the perceptual warping characteristic of the perceptual magnet effect, whereby speech sounds near the center of a category are treated as being closer together in perceptual space whereas sounds near category boundaries are perceived as being further apart. Once again, it is worth pointing out that the Bayesian approach is not restricted to speech perception. Huttenlocher, Hedges, and Vevea (2000) have presented a similar Bayesian analysis of category effects in the judgment of visual stimuli.

The suggestion that *f(Evidence|Phoneme)*, or *f(Evidence|Word)*, should play a role in speech recognition has important implications for how we should explain psycholinguistic data on the consequences of a mismatch between input and the canonical form of words. Using a cross-modal priming task in Dutch, Marslen-Wilson and Zwitserlood (1989) showed that words such as *honing* (*honey*) were not reliably accessed when nonwords with a different initial sound (e.g., *foning*) were presented. That is, the mismatching phoneme appeared effectively to block lexical access. At least subjectively, however, it seems quite easy to appreciate that 'shigarette' is an instance of *cigarette*. Whether this is true or not is ultimately an empirical issue, but there is an important difference between the two cases. Listeners are unlikely to ever have heard *honing* pronounced as /fonIN/. That is, *P(Evidence is like* /f/ | *intended phoneme is* /h/) is likely to be close to zero, and so *P(Evidence is* /fonIN/ | *Word is honing*) is also likely to be near zero. In contrast, given suitable experience of listening to drunks, *P(Evidence is like* /S/ | *stimulus is* /s/) might be nonzero. If this probability is nonzero, and the probability of alternative words is zero, /SIg{rEt/ should be recognized as *cigarette*. In this specific example it happens to be the case that there is phonetic similarity between /s/ and /S/, caused by drunks' poor control over their articulators. But the same effect (recognizing *cigarette* given /SIg{rEt/) would hold even if /s/ and /S/ were highly distinctive. That is, there need be no correlation between mismatch and similarity. Furthermore, this is an example of the asymmetries that can arise in perceptual confusability that we discussed earlier. Speakers sometimes produce tokens of /s/ as /S/, but rarely produce /S/ as /s/. /S/ should therefore be more confusable with /s/ than /S/ is with /s/.

The question, then, is whether Shortlist B will be able to recognize a word with an initial mispronunciation (like *cigarette* given *shigarette*) or not (like *honing* given *foning*). Simulations of

the recognition of two word-initial mispronunciations are shown in Figure 10. As can be seen from the solid probability functions, Shortlist B successfully recognizes *chianti*, chianti, when presented with /pijAnti/, but does not recognize *sigaret*, cigarette, when presented with /SixarEt/. This might appear surprising, given that the phonetic differences between the correct and mispronounced are well matched (both changes involve alteration of only one phonetic feature, that of place of articulation). The reason for this difference is not that Shortlist B is privy to knowledge that Dutch drunks are more likely to mispronounce their alcohol than their tobacco (should that even be true). The radically different behavior of the model on the two mispronunciations is due instead to differences in probabilistic knowledge. It happens to be the case that, on the last gate of the /Si/ diphone, the listeners in the Smits et al. (2003) study made no /s/ responses, while on the last gate of the /pi/ diphone there was at least one /k/ response. Because P(response /s/ | /Si/) = 0 when all

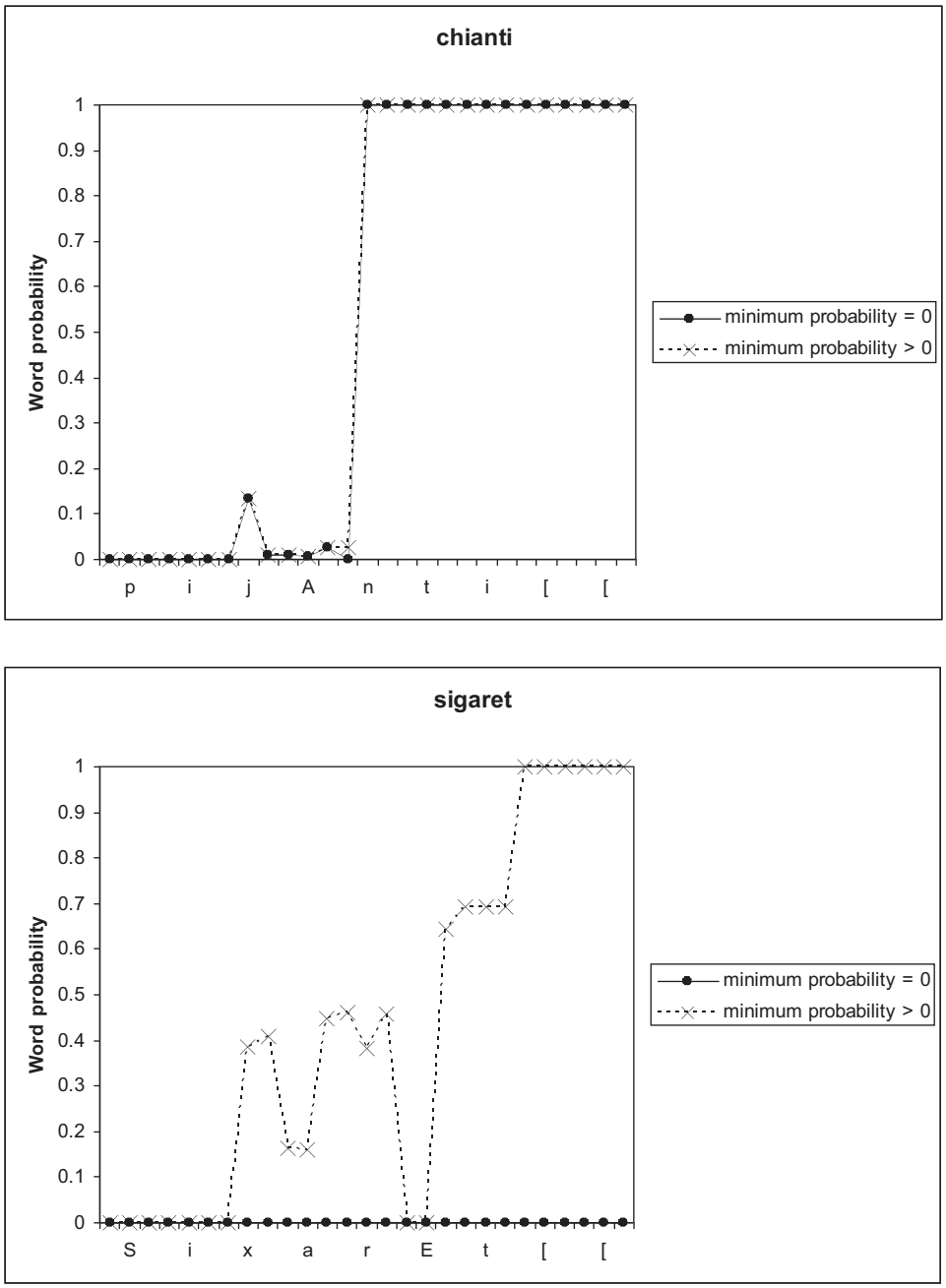


Figure 10. Patterns of word probabilities in Shortlist B given the mispronunciations *pianti* (pijAnti[[; upper panel) and *shigaret* (SixarEt[[; lower panel), with the minimum probability parameter set to zero or to a nonzero value (10^{-18}). The word *chianti*, chianti, is recognized in both cases, but *sigaret*, cigarette, is recognized only when minimum phoneme probability is greater than zero.

of the diphone has been heard, then, following Equation 7, the probability of *sigaret* will also be zero. Even though P(response/k/ | /pi/) is very small (0.028), it is still enough to keep *chianti* in the running; ultimately, because there are no other plausible paths, *chianti* dominates the probability landscape.

This difference between examples is clearly a consequence of a rather arbitrary difference in the diphone database. It could easily have been the case that the response probabilities in the two contexts were reversed. But the difference nevertheless serves to exemplify the point that word recognition in Shortlist B is determined by likelihood, that is, *P(Evidence|Word)*, and not simply by perceptual similarity. It is of course the case that the model is strongly influenced by the patterns of perceptual confusions in the diphone database, such that *P(Evidence|Word)* is modulated by phonetic similarity; but, as we have just shown, this is not the only modulating factor. It is reasonable to assume that the participants' responses in the diphone experiment were driven in part by prior probabilistic knowledge. For example, when listeners heard a sound that was consistent with two alternatives, they may well have chosen one based on a probabilistic bias (e.g., that one of those two sounds is more often confusable; though note that Warner et al., 2005, showed that simple transition probabilities did not have a strong influence on listener behavior in the diphone study). These biases are part of the diphone database and, thus, of the operation of the model. Thus, although the model does not have specific knowledge that some mispronunciations may be more likely than others nor that they may be more likely in some situations than others (e.g., listening to a sober vs. a drunk person), we can see how the model would work if the model were enriched in that way. Critically, the Bayesian approach offers a principled account of how recovery from mispronunciations in word recognition can vary as a function of the mispronunciation involved and as a function of different listening situations. When there is a change in the likelihood that the source of an input is a given word, the probability of recognizing that word changes.

It is clear from the *sigaret* example, however, that the probabilities in the diphone database are inappropriate for the modeling of experiments on the recognition of mispronounced words. Specifically, the forced-choice nature of the task in the diphone gating task resulted in many situations (particularly at later gates) where the probabilities of many responses are zero. It is plausible to assume that, although the probabilities of some responses for a given input may be very small, they should not be zero. That is, there is some nonzero probability that any apparent phoneme in the input could in fact be any other phoneme. A minimum-probability parameter was therefore added to the model. All phonemes for which P(response|input) is zero are assigned that probability value (10^{-18}). As can be seen from the dashed probability functions in Figure 10, this small adjustment has no effect on the recognition of *chianti* given /pijAnti/, but now allows *sigaret* to be recognized given the input /SixarEt/. That is, as soon as the probability that the initial segment is /s/ is not zero, the word *sigaret* easily becomes the most likely interpretation of this input. The minimum-probability parameter was switched off in all other simulations presented here, but, with the limitation that the parameter value must be very small, model behavior in those simulations does not change substantially across parameter values. Note that it is appropriate for the parameter to be switched off when modeling experiments (other than those on mispronunciations) where high-quality laboratory speech was used.

The key insight offered by Shortlist B into how listeners deal with mispronunciations, therefore, is that recognition of a mispronounced word is determined ultimately not by perceptual similarity but by the listener's estimate of the probability that that type of mispronunciation might occur. Thus, although perceptual similarity can of course influence such likelihoods, it can also be the case that the likelihood of a mispronounced word can change across contexts where perceptual similarity is the same. Just as there may be differences in likelihoods under different listening conditions in the everyday world (e.g., listening to drunks rather than sober people) and adjustments in these probabilities due to perceptual learning (Norris et al., 2003), there can also be differences across experiments as a function, for example, of changes in experiment-internal probabilities of particular events (Clayards et al., 2007). This analysis suggests that an important issue for future research will be to establish the range of tolerance that listeners have for perceptual mismatches.

Lexical Influences on Phoneme Identification

One of the central theoretical motivations driving Shortlist A was to demonstrate the viability of a completely bottom-up model of spoken word recognition. This argument was developed further by Norris, McQueen, and Cutler (2000a, 2000b), who argued that many phenomena that appeared to be attributable to a top-down effect of lexical information on prelexical processing were actually fully consistent with a completely bottom-up feed-forward architecture. In support of this argument, they developed the Merge model. Merge is an elaboration of Shortlist A designed to simulate the effects of lexical knowledge in tasks such as phoneme identification and categorization. In Merge, there are phoneme decision units that integrate information from lexical and prelexical levels (see Figure 11). Responses in tasks requiring phoneme identification are determined by these decision units and not by the prelexical phoneme units themselves. Prelexical processing is therefore completely independent of lexical processing: There is no feedback from lexical to prelexical processing.

In this section we show that Merge is compatible with the Bayesian approach taken in Shortlist B. It is in fact more than compatible: the Bayesian approach necessarily forces us to adopt the same feedforward in Shortlist B as in Merge. Furthermore, the Bayesian approach provides a more principled motivation for the Merge architecture and leads to a model (Merge B) that is computationally simpler than the original network implementation (Merge A).

The central argument that Norris et al. (2000a) presented against the use of top-down lexical feedback in prelexical processing was that it can be of no benefit. The best that any recognition system can do is to match its input against the representations in memory and to select the closest match. Feedback cannot improve this process. Note, however, that our discussion of Bayesian decision making should make it clear that there is an important qualification to this statement: The decision process should also take prior probability into account.

If feedback cannot improve the process of matching perceptual input onto lexical representations, why should there be lexical effects on phoneme identification at all? The answer is that, under

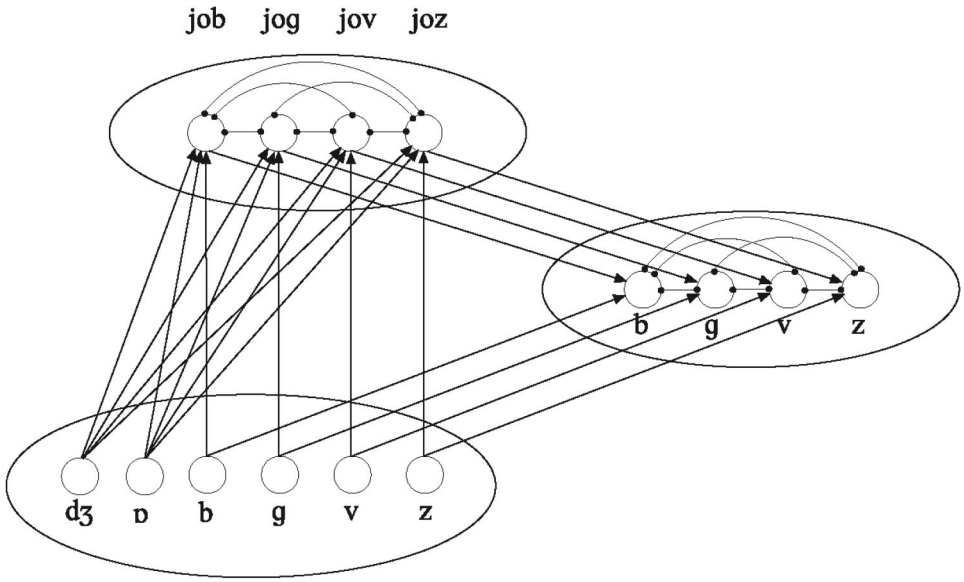


Figure 11. The architecture of the Merge model (Norris et al., 2000b). Information is fed forward (lines with arrows) from input nodes to lexical and phoneme decision nodes and from the lexical nodes to the decision nodes. Inhibitory competition (lines with closed circles) operates at the lexical and decision levels but not at the input level.

some circumstances, lexical information can improve phoneme identification. This possibility is a key feature of Merge B. An essential restriction on this Bayesian model, however, is that any lexical influences on phoneme identification must not form part of a feedback loop. Lexical information should not feed back to alter any prelexical processes involved in word recognition itself. That is, word recognition must remain a feed-forward process.

Consider what should happen if a listener is required to identify the final phoneme in a word like /fɪʃ/. If the listener believes that the input is a word, there are two sources of information that could be used to identify the phoneme. The first is the bottom-up perceptual evidence for /ʃ/. The second is lexical information. If the first two phonemes have been identified as /fI/, then this places constraints on the identity of the final phoneme. The two sources of information are quite independent (/f/ɪ/ and /ʃ/) and can therefore be combined. The standard way of viewing this is in terms of multiplying likelihood ratios. Any number of independent sources of information can be combined by multiplying the corresponding likelihood ratios. Another way to think about the process is that one source of evidence can revise the prior probabilities, and the second source of evidence can then revise the priors once again. Lexical information can update the prior probability of the occurrence of the phonemes (e.g., making /ʃ/ more probable than /s/), and these probabilities can then be revised again in the light of the perceptual evidence. In other words, the optimum way to identify phonemes is to combine the independent sources of evidence from lexical and prelexical processing, exactly as in Merge A and Merge B. This is also the basic principle underlying the account of context effects in phoneme identification given by the FLMP (Massaro, 1989b; Massaro & Oden, 1995). This is to be expected as the FLMP has the same basic form as Bayes's theorem (Massaro, 1987; Massaro & Friedman, 1990; for further discussion, see the section *FLMP and Shortlist B*).

The idea that phoneme identification might be a Bayesian decision process has also been suggested by Mirman, McClelland, and Holt (2005), but they also advocate feedback from lexical to prelexical processes. So, what would happen if the lexical information were allowed to feed back to modify the prior probabilities on the prelexical representations driving word recognition? If lexical information revised a prelexical prior, this would increase the posterior probability of the phoneme for a given input. This in turn would increase the posterior probability for the word. If this feedback were allowed to continue, the input would always be identified as the most frequent word and as containing the phoneme most consistent with that word. Far from improving recognition, any feedback at all will therefore make word recognition suboptimal. This analysis reinforces the claim made by Norris et al. (2000b) that feedback can never help word recognition unless prelexical processing is somehow suboptimal. Optimal recognition is achieved by pooling relevant sources of information without feedback. An important implication of this analysis is that any claim that word recognition does involve on-line feedback implies that the word recognition system is suboptimal.

To illustrate the importance of preventing lexical information from feeding back to modify prelexical prior probabilities (as opposed to feeding forward to influence decision priors), let us consider the case where a listener must determine whether the final phoneme of an ambiguous stimulus such as /fɪ?/ is /ʃ/ or /s/. First of all consider the case of word identification where the task is to discriminate between the words /fɪʃ/ and /fɪn/, and where *fish* is higher in frequency than *fin*. Other things being equal, there will be a bias to identify the input as the higher frequency word. If that information were used to modulate the prelexical priors, then it would induce a further bias towards *fish*. That is, the lexical-level bias would be exaggerated by the feedback. If the lexical information were then fed back through the system again (as in an

interactive-activation model) there would be a positive feedback loop, and presenting only the two phonemes /fɪ/ would always activate the word *fish* as much as all three phonemes in /fɪʃ/. This would make it harder to recognize *fin*, and its final /n/. Feedback can therefore make both word recognition and phoneme identification worse.

This analysis shows that the optimal way to combine evidence from word and phoneme levels is to process the two sources of evidence independently, as can be done either using Bayes's theorem, using fuzzy logic—as in the FLMP—or using an interactive-activation network such as that in Merge A. As we will now show, however, the Bayesian approach makes it is possible to implement Merge in a far simpler and more principled way. To illustrate the operation of Merge B, we will show how it can simulate the data on subcategorical mismatch that was critical for the evaluation of Merge A. We will not review all the literature on the modularity debate here; it is presented in detail in Norris et al. (2000b) and McQueen, Norris, and Cutler (2006). As McQueen et al. argue, first, no data have yet been found which show convincingly that there is feedback from the lexicon to the prelexical level, and, second, the data of Pitt and McQueen (1998) directly challenge the notion of lexical-prelexical feedback. Thus, although the debate about the data on feedback is still ongoing (see, e.g., McClelland, Mirman, & Holt, 2006), there are, in addition to the theoretical arguments just discussed, also empirical grounds to reject lexical–prelexical feedback.

Please note also that the current version of Shortlist B does not lend itself directly to the fine-grained simulations of RT and error rate that are required in this analysis. That is why we developed Merge B instead.

Merge B

Subcategorical mismatch data. Norris et al. (2000b) used Merge A to simulate data from studies by Marslen-Wilson and Warren (1994) and McQueen, Norris, & Cutler (1999). These two studies examined the effects of subcategorical phonetic mismatch on phoneme categorization and lexical decision. The critical manipulation was to cross-splice stimuli such that the initial portion of a word or nonword provided misleading phonetic cues as to the identity of its final consonant. The details of the materials are shown in Table 6. All items consisted of the first part of one item, up to the end of the vowel spliced onto the final consonant from another item. Critically, the stimuli could form words or nonwords, and their initial portions could be derived from either a word or nonword. For example, the word *job* could be made by cross-splicing the initial portion of another word (e.g., *jo* from *jog*) onto a final /b/ release, creating what we label as a W2W1 item, or by cross-splicing the initial portion of a nonword (e.g., *jo* from *jod*) onto the same /b/ release (an N3W1 item). In addition to the cross-spliced items, there were identity-spliced items (W1W1 and N1N1) which were made by splicing together different tokens of the same word or nonword.

Table 6
Design and Example Stimuli From the Subcategorical Mismatch Studies (Marslen-Wilson & Warren, 1994; McQueen, Norris & Cutler, 1999)

Item type	Notation	Example
Word		job
1. Word 1 + Word 1	W1W1	job + job
2. Word 2 + Word 1	W2W1	jog + job
3. Nonword 3 + Word 1	N3W1	jod + job
Nonword		smob
1. Nonword 1 + Nonword 1	N1N1	smob + smob
2. Word 2 + Nonword 1	W2N1	smog + smob
3. Nonword 3 + Nonword 1	N3N1	smod + smob

Note. The underlined portions of the examples were spliced together.

The critical result was that the lexical status of the initial portion of the stimulus had an effect on phonetic categorization responses to nonwords but not to words. When the first part of a nonword stimulus was derived from a word (W2N1), responses were slower than when it was derived from another nonword (N3N1). There was no such difference for the two types of cross-spliced word (W2W1 and N3W1). Thus, although for both words and nonwords there was an inhibitory effect of the subcategorical mismatch in the cross-spliced items (relative to the identity-spliced items), there was in addition a lexical effect on the nonwords. This interaction can be attributed to lexical competition. When the entire string ends up sounding like a word, that word is the dominant lexical hypothesis; and other lexical hypotheses, including critically the word matching the initial portion of the stimulus (i.e., the W2 word), tend to lose the competition. There is thus an effect of cross-splicing (responses are slowed because of the misleading phonetic information) but no additional lexical effect. But when the entire string ends up being a nonword, W2 words in W2N1 sequences remain as viable lexical hypotheses, and these words thus exert an effect on phonetic categorization (i.e., over and above the bottom-up effect of the phonetic mismatch, the lexicon also indicates that the final sound is not what the postsplice information suggests). These data are important because they do more than show simply that lexical knowledge can influence phonetic decision making. They show further that there is a complex interaction between the effects of lexical knowledge and the effects of detailed phonetic information. They thus impose much stronger constraints on models; and their simulation is thus more valuable than simulation of a simple lexical effect of phonetic decision making, such as the Ganong effect in phonetic categorization (Ganong, 1980). These data thus serve as a key test of the adequacy of Merge B.

Model details. Relative to Shortlist B, Merge B is a simplified Bayesian model based on the sampling process used in the Bayesian Reader (Norris, 2006). In the Bayesian Reader, visual words are represented as points in perceptual space. Any letter string (a word or a nonword) can be represented as a point in that space. Input to the model consists of a series of samples generated by adding noise to the input. Both the input and the samples are vectors corresponding to points in perceptual space. The model calculates the standard error of the mean (σ_m) of the samples based on the distances between individual samples and the sample mean. (See Appendix A of Norris, 2006, for details and equations). Given the mean and σ_m at any time, the probability of each word, given the perceptual input, can be calculated on the basis of the likelihood and frequency of that word.

A critical difference between applying the principles of the Bayesian Reader to reading and applying them to speech concerns how information arrives over time. The Bayesian Reader assumes that all of the letters in a word are presented simultaneously. To

simulate the arrival of the speech signal over time in Merge B, each successive phoneme vector is presented every N steps. In the current simulations, new phonemes are presented every 100 steps. As with the Shortlist B simulations, each new phoneme remains present throughout the word. A phoneme at the beginning of a word can therefore carry on receiving further samples right through the word. In fact, in order to avoid making any assumptions about the duration of memory for perceptual input, the simulations allow samples to continue to be accumulated until a response is made. Other than this difference in timing, Merge B calculates phoneme probabilities in exactly the same way that the Bayesian Reader calculates word probabilities. Once phoneme probabilities have been calculated, the model can compute word probabilities in just the same way as in Shortlist B.

The task in these experiments is either to categorize the final phoneme (e.g., is it /b/ or /*g*/) or to perform lexical decision. To deal with the fact that half of the stimuli will be nonwords, the lexicon also contains a dummy word. As discussed above, dummy words match any input to some extent. If the input becomes inconsistent with any word, the dummy word will dominate the likelihood calculations and have a high probability. When the dummy word has a high probability, the input is likely to be an unknown word or nonword. In these simulations we use the simplest possible lexical decision procedure: respond 'Yes' when the probability of any word exceeds a 'Yes' threshold, and respond 'No' whenever the probability of the dummy word exceeds a 'No' threshold. Simulated RTs are given by the number of samples/time-steps required to reach threshold. Error rates are simply the proportion of trials on which the probability exceeds the wrong threshold. The critical parameters for performing lexical decision are therefore the 'Yes' and 'No' thresholds and a likelihood for the dummy word. The model also needs a parameter reflecting the standard deviation of the sampling noise. In the simulations presented here, this is always 0.5.

Simulations. Each phoneme is coded as a vector where one element is set to 1.0 to represent that phoneme and all other elements are set to 0. Words are simply a concatenation of phoneme vectors. At each time step in processing, the model receives a sample from the input vectors of all phonemes presented to that point. Each sample is constructed by adding zero-mean Gaussian noise to each element of the input phoneme. As sampling proceeds, the model calculates the mean location of the input samples in perceptual space, and the standard error of the mean of the samples. This is computed on the basis of the distances between each sample and the mean of the input samples. The next step is to calculate the distance between the mean of the input samples (i.e., the vector in which every element is the mean of the corresponding input elements) and each phoneme. This is then used to calculate the likelihood of each phoneme (determined by the height of the normal distribution with the calculated standard error of the mean at the given distance from the mean). These likelihoods are then used to calculate *P(Phoneme|Input)* values. *P(Word|Input)* is calculated from these phoneme probabilities, exactly as described for Shortlist B.

Cross-splicing is simulated by changing the vector representing the final phoneme at the splice point. For example, at the splice point, the vector representing /d/ might be replaced by the vector representing /*g*/. Consequently, the calculated mean location of the input vector will be a weighted sum of the samples from the /d/ vector and the /*g*/ vector. This will slow recognition of the final phoneme, because recognition will be delayed until samples from /*g*/ outweigh those from the /d/ in the presplice portion of the stimulus. Cross-splicing in this way will also make the standard error of the mean larger than it would be in the unspliced case, which will also delay recognition. In the simulations reported here, the splice point was 50 steps into the final phoneme.

Phoneme categorization is performed by combining the evidence from both the phoneme and lexical levels. In the experiments being simulated here, there are only two alternative phonemes on each trial, and the task is simply to decide which of these two phonemes has been presented. The probabilities of the two output phonemes are calculated in exactly the same way as for the prelexical phonemes, that is, from *P(Evidence | Phoneme)*, but using prior probabilities, that is, $P(Phoneme_i)$, derived from the lexicon and considering only the two critical phonemes.

The stimuli could be either words or nonwords. To reflect this, we need to reduce the lexical influence on phoneme identification. That is, the decision units should not assume that the input is a word. This was achieved by taking a proportion of the lexically determined prior probabilities for each phoneme and redistributing it among all phonemes. For example, in the simulations here this value was set to 0.2, so the probability of each phoneme was 0.2 multiplied by the original lexical prior plus 0.8 divided by the number of phonemes.. In other words, only 20% of the lexically determined priors were allowed to modulate the effective phoneme priors. Each phoneme prior was therefore at least 80% of what it would be in the absence of any lexical information. This parameter and a parameter representing the response threshold are the only two parameters specific to the phoneme categorization simulations.

Once suitable values for the response thresholds and the dummy word are selected so as to control the overall accuracy of the model's responses, the model reproduces the main features of both the lexical decision data and the phoneme categorization data reported by Marslen-Wilson and Warren (1994) and by McQueen et al. (1999). The results of the simulations of both phoneme categorization and lexical decision are shown in Table 7. These numbers are the means of 1,000 trials of the model in each condition. Each trial uses exactly the same input but a different random number seed.

As can be seen in the left part of Table 7, the model's behavior on the words shows only a cross-splicing effect. Phonetic categorization RTs estimated by Merge B to the identity-spliced W1W1 stimuli are faster than those to the cross-spliced W2W1 and N3W1 stimuli, and there is little difference between W2W1 and N3W1. This is the pattern found in the human data—in both studies. But in the nonwords, there is a lexical effect superimposed upon this cross-splicing effect. RTs estimated by Merge B to the identity-spliced N1N1 stimuli are faster than those to cross-spliced N3N1 stimuli, which in turn are faster than those to the W2N1 cross-spliced stimuli. Again, this mirrors what was found across the two studies in human listeners.

Merge B thus successfully captures this complex interaction of lexical status and subcategorical phonetic mismatch. When the sequence as a whole is consistent with a word, then that word dominates the probability landscape, and thus the lexical status of the first part of the cross-spliced stimuli does not influence phonetic categorization behavior. But when the sequence as a whole is

Table 7
Merge B Subcategorical Mismatch Simulation Latencies (in Samples) for Phonetic Categorization and Lexical Decision Compared to the Results (in ms) from Marslen-Wilson & Warren (1994, MWW) and McQueen, Norris & Cutler (1999, MNC)

	Phonetic categorization			Lexical decision		
Item type	Merge B	MWW Experiment 3	MNC Experiment 4	Merge B	MWW Experiment 1	MNC Experiment 3
Word						
W1W1	263	497	668	373	487	340
W2W1	366	610	804	447	609	478
N3W1	364	588	802	433	610	470
Nonword						
N1N1	338	521	706	389	537	425
W2N1	443	654	821	495	625	476
N3N1	403	590	794	442	553	451

Note. The model latencies have not been adjusted to account for any nondecision component of the human responses. The notation is explained in Table 6

a nonword, then the W2 word (i.e., the source of the initial portion of a W2N1 item) remains as a plausible lexical interpretation; and, thus, exerts its effect on behavior over and above the effect of the phonetic mismatch, making phonetic decisions to these items even slower than those to the N3N1 items. Merge B therefore simulates a lexical influence on phonetic decision making without feedback.

As shown in the right half of Table 7, Merge B also accurately simulates the same complex interaction of subcategorical mismatch and lexical status in lexical decision. The account of the model's behavior in these simulations is exactly parallel to that just given for phonetic categorization, except that it is based on the probability of words (or, for nonwords, the dummy word) rather than on the probability of phonemes.

The most important message from both of these simulations is that the behavior of Merge B follows directly from the underlying Bayesian principles. This contrasts with Merge A. Because Merge A began as an interactive-activation network, it could equally well have been set up to simulate a different pattern of data (see Pitt, Kim, Navarro, & Myung, 2006). There was no principled reason for Merge A to predict the detailed pattern of results actually observed. Several of the design decisions in Merge A were simply pragmatically determined to enable the model to simulate the data. For example, there was a difference in the architecture of the phoneme input units and that of the word and decision units. Whereas the latter two had between-unit inhibition, the former did not (see Figure 11). The reason for not having inhibition between the phoneme input units was to make sure that the early stages did not force categorical responses to ambiguous stimuli. However, this is just a problem with interactive activation models and their tendency towards winner-take-all behavior. It is not a general problem with the notion of relative evaluation of perceptual hypotheses. As has already been noted, there is relative evaluation in the Bayesian calculations. If the likelihood of one hypothesis increases, then the probability of others will decrease. This is equivalent, at a computational level, to inhibition in an interactive activation model. But when there is ambiguity in the input, the Bayesian calculations never behave in a winner-take-all fashion. For example, even if a prelexical phoneme identification stage assigned a phoneme a very low probability, this could still be overcome at the decision stage by a strong lexical bias. The same computational principles therefore apply to all components of Merge B.

Model complexity. Table 3 of Norris et al. (2000b) lists 12 Merge A parameters with nonzero values. The model also has an extra parameter corresponding to the activation level for producing a 'Yes' response in lexical decision. Two additional parameters would be required to control the threshold and deadline for a 'No' response, and a further parameter is required to determine the threshold for a phoneme categorization response. This gives a total of 16 parameters. As witnessed by the fact that most of the parameters are reported to three significant digits, the exact value of the parameters is quite critical (see Pitt et al., 2006, for a discussion of the sensitivity of the model). The Bayesian implementation is very much simpler, and its parameters are shown in Appendix C. The sampling noise is a scaling factor and mainly acts to speed or slow responses. The number of samples per segment is analogous to the number of cycles per input slice in Merge A. The response threshold parameters are not critical and were simply adjusted to produce approximately the correct levels of accuracy. The model therefore has a total of 7 parameters. Note that the parameter values were adjusted by hand and not optimized.

As well as allowing us to lose some of the parameters required by Merge A, the Bayesian implementation also enables us to eliminate some of the ad hoc assumptions in Merge A. For example, Merge A required a bottom-up priority rule (Carpenter & Grossberg, 1987) to ensure that decision units could never become activated purely on the basis of lexical input. This is a property that naturally follows from Bayes's theorem. If the likelihood of the phoneme is zero, then no amount of lexical evidence can raise its probability above zero.

So, where have all the parameters gone? Remember that our claim here is that people approximate optimal Bayesian recognizers and that this determines the functions that must be computed. However, an interactive activation model can compute a wide range of functions depending on the parameters. That is, models like TRACE, Shortlist A, and Merge A all have a large parameter space, but only part of that space comes close to reproducing the correct pattern of data (Pitt et al., 2006). Without additional constraints, these are all free parameters. However, imagine there was a network that could be configured to compute exactly the same

function as either Shortlist B or Merge B. The theoretical requirement to compute a specific function would effectively determine what values the network parameters must take. That is, the parameters would no longer be free parameters. Such a model would just have the same set of free parameters as the computational-level theory. What allows us to dispense with so many free parameters are the strong principles underlying the theory.

Lexical effects in Shortlist B without online feedback. We have shown that Merge B can explain lexical involvement in phonetic decision making without feedback from the lexical level to the prelexical level. Further, we have argued that the success of the model arises from its underlying Bayesian principles. Perhaps most importantly, Merge B's account of lexical effects is true to our initial assumption that speech recognition is optimal: Given this assumption, there should be no lexical–prelexical feedback.

Two questions remain. First, what is the relationship between Merge B and Shortlist B? The answer is that it is the same as that between Merge A and Shortlist A. Merge and Shortlist are really just implementations of different components of the same theory. All four of these implementations share key assumptions about levels of processing, about prelexical and lexical representations, and, of course, the assumption that there is no feedback from the lexical level to the prelexical level. In addition, Merge B and Shortlist B operate according to the same Bayesian principles. The differences between the models thus do not lie in any differences in theoretical assumptions; they instead lie in differences in the scope of the simulations they can perform. On the one hand, Shortlist B simulates word recognition in continuous speech, but has no means, in its current implementation, to make phonetic decisions. On the other hand, Merge B is a simplification of Shortlist B, with a much smaller lexicon and no ability to deal with continuous speech recognition, but it can make phonetic decisions. In short, the present Merge B simulations show that if that model's phonetic decision component were added to Shortlist B, then Shortlist B would be able to explain lexical effects on phonemic decision making without feedback. Ideally Merge B and Shortlist B would be combined into a single program, but the simulations are much more tractable if they are kept distinct.

The second question concerns perceptual learning. As we have previously discussed, Norris et al. (2003) have shown that listeners can use lexical knowledge to retune their phonetic categories when they encounter a talker speaking in an unusual way. How is this possible in a model without feedback? The answer is that already given by Norris et al.: If one accepts that there is a distinction between on-line feedback and feedback for perceptual learning, the findings on lexical retuning of perception are completely consistent with Shortlist B (and Merge B). On-line feedback is what Shortlist and Merge do not have: lexical knowledge cannot modulate the prelexical analysis of a word as that word is being heard. We have just presented arguments why on-line feedback in speech recognition is unnecessary and indeed undesirable.

Feedback for learning is another matter entirely. As Norris et al. (2003) argued, perceptual learning can benefit speech processing, for example where adjustments in response to a given talker's idiosyncratic speech sounds can make it easier to understand what that talker is saying later. The input (Clayards et al., 2007) and the lexicon (Norris et al., 2003) are both sources of knowledge that listeners can use to make these adjustments, which, as we have already suggested, could take the form of changes to the likelihood functions of phonetic categories (cf. Figure 2). Such changes to *f(Evidence|Phoneme*) over time that are directed by lexical knowledge do not require there to be any effects of the lexicon on on-line processing. That is, feedback for learning and on-line feedback can involve distinct mechanisms and therefore do not entail one another. Thus, although a future version of Shortlist B might well include an implementation of lexical retuning of phonetic perception, that version could still have no on-line feedback.

As Norris et al. (2003) and McQueen, Norris, et al. (2006) argue, should convincing data that there is on-line feedback ever be forthcoming, it should probably best be taken as evidence for how feedback for learning is implemented. One possibility is that the mechanism for lexically guided learning would have subsidiary consequences for on-line processing (see Mirman, McClelland, & Holt, 2006, for one proposal). Lexical effects in on-line processing could thus potentially arise as an epiphenomenon of lexical involvement in perceptual learning. Such effects, should they ever be found, would therefore better be seen not actually as evidence for on-line interaction, which itself serves no useful function and may indeed be detrimental to recognition, but rather as further evidence for feedback for learning, which is beneficial for the listener.

FLMP and Shortlist B

In concluding this section on lexical influences on phoneme identification, it is important to compare the account offered by Shortlist B/Merge B with the FLMP (Massaro, 1987, 1989b; Oden & Massaro, 1978). As we have already noted, the FLMP combines lexical and prelexical evidence independently to perform phoneme identification and categorization, just as in Merge B. Although the equation underlying the FLMP has the same form as Bayes's theorem, the FLMP is assumed to generate a truth value rather than a posterior probability. As Massaro and Friedman (1990) note "Bayes's theorem and the FLMP are conceptually equivalent if the truth value can be interpreted as a conditional probability" (p. 232). However, in the FLMP, the truth values are interpreted as response probabilities and not posterior probabilities. Consequently, the FLMP is not strictly Bayesian and does not incorporate an optimal decision rule.

In the FLMP, responses are generated according to the R. D. Luce (1959) choice rule. The Luce choice rule, shown in Equation 11, gives the probability of generating a particular response as a function of the relative support for that response as a proportion of the total support for all responses.

$$P_r(a) = v(a) / \sum_{b \text{ in } R} v(b) \tag{11}$$

The Luce choice rule is often used to generate probabilistic choice behavior from deterministic systems. For example, it is used to translate the activation values from interactive activation networks into response probabilities (e.g., Dahan et al., 2001). As noted in our earlier discussion of the NAM, the Luce choice rule gives the probability of generating each response, not the posterior probability of the hypothesis given the evidence. That is, the choice rule is interpreted as giving the proportion of responses in each category and not the probability of the hypotheses. The Luce choice rule can be used to describe the average behavior of an optimal system, but it is not itself an optimal decision rule. The optimal

decision rule is to always select the response with the largest posterior probability. The Luce choice rule is a randomized decision rule that will always perform worse than the optimal Bayesian decision rule (Ferguson, 1967). Thus, although there are close formal similarities between the FLMP and Bayes's theorem (compare Equations 11 and 1), they are not identical. Most importantly, if the FLMP decision rule is applied at the level of the single trial, FLMP will not perform optimally.

In addition to these formal differences between the models, it is important to note that there are also significant differences between the explanations of lexical effects given by FLMP and Merge (Massaro, 2000; Norris et al. 2000a; Oden, 2000).

However, despite these differences, there is no doubt that the FLMP is close in spirit to the ideas we advance here. In particular, work within the FLMP framework has made some of the most important contributions to the feedback debate. Both Merge and FLMP incorporate the notion that perception involves combining the independent contribution of different sources of information. This is the hallmark of feedforward theories and the central contrast between them and interactive theories. Massaro (1989b) simulated the influence of lexical information on phoneme categorization in both the FLMP and TRACE and demonstrated that only the FLMP could simulate the data accurately. McClelland (1991) responded by developing the stochastic interactive activation model, which could account for the data by emulating the behavior of classical models, like the FLMP. In other words, the interactive model could only simulate the data to the extent that it behaved exactly like a noninteractive model. In fact, Massaro and Cohen (1991) argued that even this modified interactive model could not simulate the data as well as the FLMP.

From Activation to Bayes

We have shown that Shortlist B can account for key findings on segmentation, word frequency, and mispronunciations. The Merge B simulations show in addition that a Bayesian model with the same key assumptions as Shortlist B can explain data on lexical involvement in phonemic decision making. We now compare different versions of the model and, in particular, trace the development from the activation-based Shortlist A to the Bayesian Shortlist B.

Shortlist A and Shortlist B

The main aim of the Shortlist A article was to demonstrate the viability of a strictly bottom-up model of spoken-word recognition. As we have just seen, Merge B (and thus, by extension, Shortlist B) can explain lexical effects in a strictly bottom-up fashion. Norris (1994) also showed that, in contrast to TRACE, there was no need to have a copy of the entire lexical network associated with each segment in the input. In Shortlist A and in the new model, only a small subset of possible candidates need be considered at each segment. Furthermore, there is a clear distinction in both versions of the model between the process of lexical access, which generates candidates, and the process of competition among lexical candidates (Shortlist A) or paths (Shortlist B). This distinction corresponds to a contrast between representations of lexical types in the lexicon, where there is only a single representation of each word, and representations of candidate lexical tokens, where there may be many tokens of any given lexical type. This distinction between type and token representations is discussed extensively in Norris et al. (2006).

The two versions of the model thus share key assumptions about lexical representations. They are not identical in every way, however. One way to appreciate the relationship between the two versions of Shortlist is to separate out those properties of the original model that were fundamental theoretical claims (T) and those properties that were a consequence of pragmatic assumptions made simply to make it possible to construct a functioning computational model (M). Norris (2005) enumerated these two sets of assumptions as follows.

Core Theoretical Assumptions

T1. The flow of information from prelexical to lexical levels is bottom-up only. This was the central motivation for Shortlist A.

T2. Bottom-up selection of multiple lexical candidates is based on both matching and mismatching information (i.e., a claim about the procedure for computing a match between input and lexical entries).

T3. Matching lexical candidates (and only those candidates) enter into a competition process that optimizes the parsing of the input into words.

T4. There is no need for explicit lexical segmentation (i.e., the model does not need to be told where words begin and end in the input).

Assumptions Required for the Implementation of Shortlist A

M1. The input to the model is a string of phonemes.

M2. The input contains no phoneme deletions, insertions or substitutions (i.e., there are no errors in the perceptual analysis).

M3. The dictionary contains a single canonical representation of each word (i.e., no account of pronunciation variation).

M4. Lexical lookup is by means of a serial search through a dictionary.

M5. The match between the input and lexical entries is computed by counting +1 for each matching phoneme in the correct position, and −3 for each mismatching phoneme.

M6. Matches between input phonemes and the corresponding phonemes in a lexical entry are all-or-none (i.e., there is no account of phoneme similarity).

M7. The candidates are entered into the network just by wiring them in as required.

M8. Overlapping candidates are connected by inhibitory links.

M9. Competition is performed by an interactive-activation network.

M10. The model output is a pattern of lexical activations over time.

The only one of the core theoretical assumptions that has been revised in Shortlist B is the claim about the matching process being based on both matching and mismatching evidence (T2). The new model retains the spirit of this assumption, but, as we saw in the fourth set of simulations, match in the Bayesian formulation is determined by likelihood, rather than any simple similarity metric. Shortlist B incorporates one new theoretical claim: Word recognition is performed optimally. That is, word recognition is per-

formed by computing probabilities as determined by Bayes's theorem.

In contrast to the shared theoretical assumptions, most of the modeling assumptions have changed. The only similarities between the two versions of Shortlist are M2 and M3, and possibly M7 (but of course there is no network in the new model). Purely for reasons of computational efficiency, the new model no longer uses a strictly serial search of the lexicon (M4). The model is programmed such that there is a list of words associated with each diphone, so only words that really are potential candidates need be considered. It should be clear that these differences are nothing more than changes in the way the program is written. The precise algorithm chosen to implement the computer program makes no difference to the results of the simulations. Furthermore, these programming decisions most definitely do not reflect any theoretical claims about the nature of the lexical search process.

The most radical differences between Shortlist A and Shortlist B are the modeling assumptions M8, M9, and M10. Shortlist B no longer uses a connectionist network. It is important to emphasize that these are modeling assumptions and not theoretical assumptions. The interactive activation model used in Shortlist A was never anything more than a convenient algorithm for determining a near-optimal segmentation, and it was certainly never intended to be a claim about the neural implementation of the word recognition process. It might be possible to design a modified interactive activation model that could compute the correct posterior probabilities required by Shortlist B. However, an appropriately designed network would, by definition, compute exactly the same probabilities as the current computational implementation. A connectionist implementation of Shortlist B would therefore add nothing to the explanatory or predictive value of the theory. Worse still, there would be the possibility that a connectionist implementation might be a distraction from the critical insights provided by the Bayesian approach.

If a model is expressed as a connectionist network, the fact that the probability of one word is influenced by the probability of overlapping words is most readily implemented in terms of inhibitory links between the representations of competing lexical candidates. A computational-level theory need make no assumptions about the specific algorithms used to perform the computations or about the way those algorithms might be implemented.

The computational-level approach to model building we have adopted here is very different from connectionist models like Shortlist A and TRACE. But it is important to note that there is no deep philosophical incompatibility between the approaches. Rumelhart and McClelland (1985, 1986) suggested that, in Marr's terms, connectionist models could be considered to offer explanations at an algorithmic level. A complete account of human speech recognition would encompass both the computational and algorithmic levels. However, the only way to discover which algorithms might be used is to know what functions those algorithms need to compute. If theories are constrained to use the small set of existing connectionist architectures, there is no guarantee that the available architectures will be able to compute the necessary functions. For example, the interactive activation networks in Shortlist A and TRACE do not compute the functions required for Bayesian inference. These models are the wrong place to start. However, in principle, there might be infinitely many networks that could compute those functions. (For examples of neurally plausible Bayesian algorithms, see Rao, 2004, which contains suggestions as to how the cerebral cortex might implement Bayesian inference for an arbitrary hidden Markov model, and Bogacz and Gurney, 2007, for an illustration of how the basal ganglia and cortex might perform optimal decision making). So, why would one choose one kind of connectionist algorithm over another? Different algorithms might predict different behavior, and constraints from the implementational level might lead one to prefer some algorithms over others. But first we need to know whether algorithms that can compute Bayesian inference might have anything at all to contribute to the explanation of spoken word recognition.

We end this section by summarizing the advantages of Shortlist B over Shortlist A. First, the critical new theoretical claim in Shortlist B is that listeners approximate optimal Bayesian classifiers. As we have just argued, this computational-level claim is a more principled starting point for model building than the interactive activation algorithm on which Shortlist A (and TRACE) is founded. Second, the optimality assumption gives Shortlist B extra explanatory power, scope, and simplicity compared with Shortlist A. For example, the explanation of word frequency effects follows directly from this assumption, without the need to add any extra features or parameters. Of course, we might have been able to add extra features to Shortlist A to make it give an accurate simulation of frequency effects, in the same way that Dahan et al. (2001) did for TRACE. However, as with the TRACE simulations, that would still have fallen short of the achievement of Shortlist B, which is to explain why listeners behave as they do in response to differences in word frequency. Third, the input to Shortlist B, based on a very rich set of perceptual confusion data (Smits et al., 2003) allows the model to pass information continuously on to the lexical level of processing; this was not possible with the categorical phonemic input in Shortlist A. Finally, the new framework links the explanation of spoken-word recognition to a wider body of research on Bayesian models of perception and learning (e.g., Feldman & Griffiths, 2007; Huttenlocher et al., 2000; Tenenbaum & Griffiths, 2001; Tenenbaum, Griffiths & Kemp, 2006).

SpeM

The SpeM model of Scharenborg et al. (2005) is an implementation of Shortlist using techniques from ASR. It too shares the same theoretical assumptions as Shortlist A and Shortlist B, but has a very different implementation. It uses an automatic phone recognizer to generate a phoneme lattice. The phoneme lattice is used to generate a word lattice. A measure of word activation is then derived from a combination of the scores for individual words and the scores for the paths that they lie on. The word activation scores therefore reflect both the bottom-up perceptual support for the word and how well that word fits in with the best path through the lattice. In architectural terms, SpeM is very similar to the model being presented here, with the main difference being the form of the input: real speech versus confusion data. This means that SpeM can perform simulations by being fed with exactly the same speech stimuli used in a psychological experiment. For example, Scharenborg et al. demonstrated that SpeM can simulate data from Norris et al. (1997). However, because of some of the restrictions and complexities imposed by the need to recognize real speech using currently available ASR techniques, SpeM incorpo-

rates some necessary simplifications. In particular, the procedure it employs to compute word activations is not strictly Bayesian. The activations are not directly interpretable in terms of probabilities. Shortlist B, in contrast, dispenses with the notion of activation altogether, and the output of the model is a pattern of word probabilities that changes over time. This enables us to give a more thorough treatment of the implications of a Bayesian approach to speech recognition. A further limitation of SpeM is that the performance of the automatic phone recognizer is not perfect (see Scharenborg et al., 2005, for discussion). Consequently, some experimental stimuli are not correctly identified. Therefore, there is still a need for a model that can work with a phonemic transcription of the input.

Conclusions

The model presented here serves two purposes. First, it shows how data from a gating task can be used provide a psychologically plausible input to a model of continuous speech recognition. Second, and more importantly, it illustrates the implications that a Bayesian account of speech recognition has for a number of important theoretical issues. Although Bayesian techniques are at the heart of almost all statistical pattern recognition systems, they have not previously been used in psychological models of spoken-word recognition. The SpeM model of Scharenborg et al. was motivated by Bayesian principles but is not fully Bayesian.

The power of the Bayesian approach is that it offers a principled account of many phenomena that have previously been explained in an entirely ad hoc fashion. For example, the effect of word frequency can be simulated in an interactive activation model like TRACE in terms of changes in resting levels or weights (Dahan et al., 2001). However, there is no principled theoretical reason to prefer one account over the other. Moreover, as discussed by Norris (2006), most of the mechanisms proposed as explanations of the word frequency effect are actually detrimental to efficient recognition. In contrast, the use of frequency (or prior probability) in the present model provides the optimum way of combining perceptual evidence with knowledge of prior probabilities. Exactly the same argument must also apply to the explanation of the effects of context. Almost by definition, contextual constraints act to make particular words more or less probable. The effects of frequency and context must therefore both be modeled in terms of their influence on prior probabilities.

Similarly, there has been considerable debate in the literature as to the proper metric for computing the degree of perceptual match between the speech input and lexical representations. The standard approach has been to suggest that there is some perceptual similarity metric that can provide a measure of the perceptual distance between different segments, and that it is this distance that determines the degree of match. The Bayesian perspective shows that a simple metric based on perceptual form is inadequate. What counts is not perceptual distance itself but the likelihood that the input is an instance of the particular word or segment—$f(Evidence|Word)$ or $f(Evidence|Phoneme)$. Therefore a word or segment with a very variable pronunciation may be much more tolerant of mismatch than a word or segment that is always realized in the same way.

The effect of lexical competition in continuous speech recognition also follows inevitably from the assumption of optimality. Given a particular input, there is only one way to calculate the posterior probabilities of the words, and those probabilities must be influenced by the presence of other overlapping word candidates. More specifically, the effect of overlapping candidates must also be influenced by the viability of the path(s) that the word lies on. Overlapping candidates will only compete to the extent that they lie on paths with a high probability.

Finally, Bayesian principles provide a firmer theoretical underpinning for the case that lexical information should not influence prelexical processing during word recognition. In an optimally designed system, lexical knowledge should be able to influence decisions about the identity of phonemes in words, but that information should not feed back so as to influence the word recognition process itself.

It remains to be seen whether the Shortlist B account will stand up to future tests. Although the present analyses suggest that word recognition closely approximates optimal Bayesian decision making, new data may reveal that certain aspects of speech perception are not optimal. Pelli, Farell, and Moore (2003) have shown, for example, that visual word recognition is not as efficient as it could be. We therefore hope that Shortlist B not only provides important insights into speech recognition, but that it will also generate empirical tests of the optimality of the word recognition process.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52*(3), 163–187.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2*) [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics, 44*(5), 395–408.

Benki, J. R. (2003). Analysis of English nonsense syllable recognition in noise. *Phonetica, 60*(2), 129–157.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700–765.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation, 19*(2), 442–477.

Bouwman, G., Boves, L., & Koolwaaij, J. (2000). Weighting phone confidence measures for automatic speech recognition. In *Proceedings of the COST249 Workshop On Voice Operated Telecom Services* (pp. 59–62). Ghent, Belgium.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a selforganizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing, 37,* 54–115.

Church, K. W. (1987). Phonological parsing and lexical retrieval. *Cognition,* 25(1–2):53–69.

Clayards, M., Aslin, R. N., Tanenhaus, M. K., & Jacobs, R. A. (2007). Within category phonetic variability affects perceptual uncertainty, *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 701–703). Saarbrücken, Germany.

Cole, R. A., & Jakimik, J. (1978). Understanding speech: How words are

heard. In G. Underwood (Ed.), *Strategies of information processing* (pp. 67–166). London: Academic Press.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–163). Hillsdale, NJ: Erlbaum.

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language, 30*(2), 234–250.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language, 32*(2), 193–210.

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition, 16*(6), 1084–1096.

Connine, C. M., Titone, D., Deelman, T., & Blasko, D. G. (1997). Similarity mapping in spoken word recognition. *Journal of Memory and Language, 37*(4), 463–480.

Cutler, A., & Butterfield, S. (1992). Rythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31*(2), 218–236.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2,* 133–142.

Cutler, A., Demuth, K., & McQueen, J. M. (2002). Universality versus language-specificity in listening to running speech. *Psychological Science, 13*(3), 258–262.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*(1), 113–121.

Cutler, A., Norris, D., Mister, E., & Sebastian-Galles, N. (2004). La perception de la parole en espagnol: Un cas particulier. In L. Ferrand & J. Grainger (Eds.), *Psycholinguistique cognitive* (pp. 53–72). to Brussels, Belgium: De Boek.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*(4), 317–367.

Davis, M. H., Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Recognising embedded words in connected speech: Context and competition. In J. Bullinaria, G. Houghton & D. Glasspool (Eds.), Proceedings of the *Fourth Neural Computation and Psychology Workshop* (pp. 254–266). London: Springer-Verlag.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28*(1), 218–244.

Doyle, A. C. (1890). *The sign of four.* London: Spencer Blackett.

Elman, J. L. & McClelland, J. L. (1986). Exploiting the lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds*.), Invariance and variability in speech processes* (pp. 360–380). Hillsdale, NJ: Erlbaum.

Feldman, N. H., & Griffiths, T. L. (2007). A rational account of the perceptual magnet effect. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretical approach.* New York: Academic Press.

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. C. T. Walker (Eds.), *New approaches to language mechanisms* (pp. 257–287). Amsterdam: North Holland.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin Company.

Frauenfelder, U. H., & Floccia, C. (1998). The recognition of spoken words. In A. Friederici (Ed.), *Language comprehension: A biological perspective* (pp. 1–40). Berlin: Springer.

Ganong, W. F. (1980). Phonetic categorization in auditory perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110–125.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology, 45*(2), 220–266.

Gow, D. W., Jr., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21*(2), 344–359.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics, 38*(4), 299–310.

Grosjean, F. (1996). Gating. *Language and Cognitive Processes, 11*(6), 597–604.

Gussenhoven, C. (1992). Illustrations of the IPA: Dutch. *Journal of the International Phonetic Association, 22,* 45–47.

Healy, A. F., & Cutting, J. E. (1976). Units of speech perception: Phoneme and syllable. *Journal of Verbal Learning and Verbal Behavior, 15,* 73–83.

Howes, D. H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America, 29*(2), 296–305.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*(2), 220–241.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America, 97*(1), 553–562.

Johnson, E. K., Jusczyk, P. W., Cutler, A., & Norris, D. (2003). Lexical viability constraints on speech segmentation by infants. *Cognitive Psychology, 46*(1), 65–97.

Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. *Ohio State University Working Papers in Linguistics, 50,* 101–113.

Johnson, K. (1997b). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606–608.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics, 39*(3), 155–158.

Luce, P. A., Goldinger, S. D., Auer, E. T., Jr., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62*(3), 615–625.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*(1), 1–36.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman & Co.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition,* 25(1–2):71–102.

Marslen-Wilson, W., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*(6), 1376–1392.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101*(4), 653–675.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 576–585.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and

lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29–63.

Massaro, D. W. (1975). *Understanding language: An information processing analysis of speech perception, reading and psycholinguistics.* New York: Academic Press.

Massaro, D. W. (1978). A stage model of reading and listening. *Visible Language, XII*(1), 3–26.

Massaro, D. W. (1979). Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance, 5*(4), 595–609.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, N. J.: Lawrence Erlbaum Associates

Massaro, D. W. (1989a). *Experimental psychology: An information processing approach.* New York: Harcourt Brace Jovanovich.

Massaro, D. W. (1989b). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21*(3), 398–421.

Massaro, D. W. (2000). The horse race to language understanding: FLMP was first out of the gate, and has yet to be overtaken. *Behavioral and Brain Sciences, 23*(3), 338–339.

Massaro, D. W., & Cohen, M. M. (1983a). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*(5), 753–771.

Massaro, D. W., & Cohen, M. M. (1983b). Phonological context in speech perception. *Perception & Psychophysics, 34*(4), 338–348.

Massaro, D. W., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology, 23*(4), 558–614.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97,* 225–252.

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*(4), 1053–1064.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134*(4), 477–500.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–111.

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

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

McClelland, J. L., Mirman, D., and Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10,* 363–369.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*(5), 375–407.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*(2), B33–42.

McNeill, D., & Lindig, K. (1973). The perceptual reality of phonemes, syllabes, words and sentences. *Journal of Verbal Learning and Verbal Behavior, 12,* 419–430.

McQueen, J. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39*(1), 21–46.

McQueen, J. (2007). Eight questions about spoken word recognition. In G. M. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 37–53). Oxford: Oxford University Press.

McQueen, J., & Cutler, A. (1998). Spotting (different types of) words in (different types of) context. In *Proceedings of the Fifth International Conference on Spoken Language Processing* (Vol. 6). Sydney, Australia.

McQueen, J., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*(6), 1113–1126.

McQueen, J. M., Cutler, A., Briscoe, E., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes,* 10(3–4):309–331.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 621–638.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance, 25*(5), 1363–1389.

McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Science, 10*(12), 533.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27*(2), 338–352.

Mirman, D., McClelland, J. L., & Holt, L. L. (2005). Computational and behavioral investigations of lexically induced delays in phoneme recognition. *Journal of Memory and Language, 52*(3), 424–443.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13*(6), 958–965.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76*(2), 165–178.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America, 62*(3), 714–719.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America, 109*(3), 1181–1196.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*(3), 189–234.

Norris, D. (2005). How do computational models help us build better theories? In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 331–346). Mawah, NJ: Erlbaum.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*(2), 327–357.

Norris, D., Cutler, A., McQueen, J. M., & Butterfield, S. (2006). Phonological and conceptual activation in speech comprehension. *Cognitive Psychology, 53*(2), 146–193.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1209–1228.

Norris, D., McQueen, J. M., & Cutler, A. (2000a). Feedback on feedback on feedback: It's feedforward [Authors' response]. *Behavioral and Brain Sciences, 23*(3), 352–370.

Norris, D., McQueen, J. M., & Cutler, A. (2000b). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*(3), 299–370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238.

Norris, D., McQueen, J. M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34*(3), 191–243.

Norris, D., McQueen, J. M., Cutler, A., Butterfield, S., & Kearns, R. (2001). Language-universal constraints on speech segmentation. *Language and Cognitive Processes,* 16(5–6):637–660.

Oden, G. C. (2000). Implausibility versus misinterpretation of the FLMP. *Behavioral and Brain Sciences, 23*(3), 344.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85*(3), 172–191.

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature, 423*(6941), 752–756.

Pickett, J. M. (1957). Perception of vowels heard in various spectra. *Journal of the Acoustical Society of America, 29*(5), 613–620.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology* 7 (pp. 101–139). Berlin: Mouton de Gruyter.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113*(1), 57–83.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39*(3), 347–370.

Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31*(3), 273–279.

Rao, R. P. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation, 16*(1), 1–38.

Rayner, K., Slowiaczek, M. L., Clifton, C., Jr., & Bertera, J. H. (1983). Latency of sequential eye movements: Implications for reading. *Journal of Experimental Psychology: Human Perception and Performance, 9*(6), 912–922.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*(4), 445–476; discussion 477–526.

Rumelhart, D. E., & McClelland, J. L. (1985). Level's indeed! A response to Broadbent. *Journal of Experimental Psychology: General, 114*(193–197).

Rumelhart, D. E., & McClelland, J. L. (1986). PDP Models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing, Volume 1.* Cambridge, MA: M. I. T. Press, A Bradford Book.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*(1), 51–89.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America, 35*(2), 200–206.

Sawusch, J. R., & Jusczyk, P. (1981). Adaptation and contrast in the perception of voicing. *Journal of Experimental Psychology: Human Perception and Performance, 7*(2), 408–421.

Scharenborg, O., Norris, D., ten Bosch, L., & McQueen, J. (2005). How should a speech recognizer work? *Cognitive Science, 29,* 867–918.

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics, 68*(1), 1–16.

Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America, 113*(1), 563–574.

Tabossi, P., Burani, C., & Scott, D. (1995). Word identification in fluent speech. *Journal of Memory and Language, 34*(4), 440–467.

Tabossi, P., Collina, S., Mazzetti, M., & Zoppello, M. (2000). Syllables in the processing of spoken Italian. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 758–775.

Taft, M., & Hambly, G. (1986). Exploring the Cohort model of spoken word recognition. *Cognition, 22*(3), 259–282.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*(4), 629–640; discussion 652–791.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science, 10*(7), 309–318.

van Alphen, P. M., & McQueen, J. M. (2006). The effect of voice onset time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human Perception and Performance, 32*(1), 178–196.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory, 13*(2), 260–269.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9*(4), 325–329.

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

Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*(1), 98–108.

Wang, M. D., & Bilger, R. (1973). Consonant confusions in noise. *Journal of the Acoustical Society of America, 54*(5), 1248–1266.

Warner, N., Smits, R., McQueen, J. M., & Cutler, A. (2005). Phonological and frequency effects on timing of speech perception: A database of Dutch diphone perception. *Speech Communication, 46,* 53–72.

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology, 50,* 509–535.

Wessel, F., Schlueter, R., Macherey, K., & Ney, H. (2001). Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing, 9*(3), 288–298.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*(2), 143–154.

Young, S. J., Russell, N. H., & Thornton, J. H. S. (1989). Token passing: A simple conceptual model for connected speech recognition systems., *Technical Report CUED/F-INFENG/TR38.* Cambridge University Engineering Dept.

Appendix A

The Phoneme Inventory of Shortlist B in IPA Transcription and in the Machine-Readable Transcriptions Used by the Model

Consonants	IPA	b	d	ɡ	p	t	k	m	n	ŋ	l	r	ʋ	j	f	v	s	z	ʃ	ʒ	x,ɣ	h	ʤ
	Shortlist B	b	d	g	p	t	k	m	n	N	l	r	w	j	f	v	s	z	S	Z	x	h	_
Vowels	IPA	i	y	u	e	ɪ	ɛ	œ	o	ɔ	a	ɑ	ɛi		œy		ɑu		ʉ,ə				
	Shortlist B	i	y	u	e	I	E	|	o	O	a	A	K		L		M		}				

Appendix B

The diphone confusion matrix can be assumed to be generated from a noisy decision process operating on stimuli located in a multidimensional perceptual space. As indicated in Figure 2, the likelihood *f(input|phoneme)* will have a pdf. If we assume that the pdf is a Gaussian distribution with the same variance for all phonemes, then we can calculate the distances between pairs of phonemes that would be required to produce a likelihood that would result in the empirically determined *P(response|input)*. Given these distances, we can alter the variance of the pdf, and recompute a new set of likelihoods and probabilities. If we make the variance smaller, the new probabilities will correspond to what we might expect with perceptually clearer input. If we make the variance larger, the new probabilities will correspond to what we might expect with perceptually more ambiguous input. A *sharpen* variable was defined that controlled these variance adjustments. In the simulations summarized in Figure 5, the model was run in two ways. In one case, the model was run with its default parameters, that is, with no sharpening or broadening of the pdf variance, and thus with the empirically determined phoneme likelihoods (sharpen = 1). In the other case the variance was halved, and phoneme likelihoods were recomputed (improved probabilities; sharpen = 0.5).

To perform the simulations, these calculations are performed on the phoneme probabilities computed at each slice and not on the complete confusion matrix. This simplifies the computations, as the phonemes can then be treated as lying on a single perceptual dimension. The effectiveness of this procedure depends on whether the probabilities with which various phonemes are given in response to a particular target remain in the same ordinal relationship over changes in signal to noise ratio (variance). For example, if listeners consistently misidentified a particular phoneme in the diphone experiment, increasing the signal to noise ratio will exaggerate the error rather than make identification more accurate. However, these technical limitations should be of little concern here as the procedure is simply being used to illustrate the general relationship between frequency and the reliability of perceptual information.

Appendix C

Merge B Parameters

Parameter	Value
Sampling noise	0.5
Lexical decision ‘Yes’ threshold	0.8
Lexical decision ‘No’ threshold	0.01
Phonetic categorization threshold	0.999
Lexical decision dummy word probability (set to zero in phonetic categorization)	0.15
Lexical attenuation factor (not used in lexical decision)	0.8
Samples per segment	100

Received April 6, 2007
Revision received January 16, 2008
Accepted January 17, 2008 ■