Journal of Experimental Psychology:
Human Perception and Performance
1988, Vol. 14, No. 1, 113-121

# The Role of Strong Syllables in Segmentation for Lexical Access

Anne Cutler and Dennis Norris
Medical Research Council Applied Psychology Unit, Cambridge, England

A model of speech segmentation in a stress language is proposed, according to which the occurrence of a strong syllable triggers segmentation of the speech signal, whereas occurrence of a weak syllable does not trigger segmentation. We report experiments in which listeners detected words embedded in nonsense bisyllables more slowly when the bisyllable had two strong syllables than when it had a strong and a weak syllable; *mint* was detected more slowly in *mintayve* than in *mintesh*. According to our proposed model, this result is an effect of segmentation: When the second syllable is strong, it is segmented from the first syllable, and successful detection of the embedded word therefore requires assembly of speech material across a segmentation position. Speech recognition models involving phonemic or syllabic recoding, or based on strictly left-to-right processes, do not predict this result. It is argued that segmentation at strong syllables in continuous speech recognition serves the purpose of detecting the most efficient locations at which to initiate lexical access.

Speech recognition is the process by which meaning is derived from the acoustic signal. A recognizer (be it a human or a machine) keeps in its memory a set of discrete meanings and locates in this memory the meanings that correspond to each input.

The number of potential utterances with which a recognizer might be presented is infinite. Therefore, the recognizer cannot store complete utterances in memory. Instead, it must store the discrete units from which utterances may be constructed. For simplicity, we will refer to these lexical units as words and beg the question of whether or not they actually correspond to orthographic words.

Speech signals are continuous. Before a recognizer can access the meaning of any word occurring in an input, it must decide where the word begins. This would be no problem if speakers provided reliable cues that marked such points in the signal. Speech researchers have so far failed to discover such cues, however, and much research effort in speech recognition has been devoted to the question of where to start lexical access in the absence of reliable information about where words begin.

One fairly crude solution, adopted by many machine recognition systems (see Holmes, 1984, for a review), is to match arbitrary stretches of the speech signal against stored acoustic templates. Because a potential match could be found starting at any point, however, this approach forces the recognizer to initiate a very large number of access attempts, by far the majority of which are futile.

An alternative solution, adopted by most psychological models of speech recognition (see Norris & Cutler, 1985, for a review), is to preprocess the signal and undertake some prelexical classification. For instance, if the speech could be analyzed into phonetic segments, then the phonetic sequence could be used as a basis for initiating lexical access: Stored representations would be in phonetic form, the recognizer would construct a prelexical representation of the signal as a sequence of specific phonetic segments, and a lexical access attempt could be begun at every segment. This procedure, too, would result in a majority of fruitless access attempts, but these could perhaps be reduced by considering, say, two-segment sequences and ruling out access attempts when the sequence postulated to begin the word was phonologically illegal in the language (e.g., [vn]).

A still more efficient classification would be in terms of syllables: Stored representations would be in syllabic form, the recognizer would construct a prelexical representation of the signal as a sequence of specific syllables, and lexical access could be attempted starting at every syllable. This procedure would result in a comparatively small proportion of wasted access attempts.

Indeed, there is direct evidence that human listeners do divide speech input into syllables: Detection of syllable-sized targets is significantly faster when the target matches the actual syllabification of the speech input, as shown by Mehler, Dommergues, Frauenfelder, and Segui (1981). But Mehler et al.'s experiment was run in French, and subsequent investigation showed that the syllabification effect, though highly reliable in French, did not hold in English (Cutler, Mehler, Norris, & Segui, 1983, 1986). Cutler et al. (1986) ascribed this difference to differences in the phonology of French and English. In French, syllable structure is relatively regular, and speakers' intuitions about syllable boundaries are clear. In English, however, which is a stress language, there is an enormous range of syllable structures (the words *a* and *scrounged* are both monosyllables), a large difference in per-

This research was supported by a grant from British Telecom Research Laboratories. Acknowledgement is made to the Director of Research of British Telecom for permission to publish the article.

We also thank John Williams and Bill Barry for recording the tapes; Steve Bartram and John Williams for testing the subjects and measuring the materials; Ian Nimmo-Smith for statistical advice; and Sally Butterfield, Kim Silverman, Mary Smith, and John Williams for helpful discussions.

Experiments 1 and 2 were presented to the 109th meeting of the Acoustical Society of America, Austin, Texas, in April 1985.

Correspondence concerning this article should be addressed to either author at the Medical Research Council, Applied Psychology Unit, 15 Chaucer Road, Cambridge, CB2 2EF, United Kingdom.

ceptibility between stressed and unstressed syllables, and speakers are frequently unclear about where to place boundaries between syllables. These factors combine to make the syllable per se an unsatisfactory segmentation unit for English.

Because English is a stress language, its most noticeable structural characteristic, in fact, is that it has two very different categories of syllable: strong and weak. Strong syllables contain full vowels; the words *eye*, *pill*, *crypt*, and *scrounge* are all strong monosyllables. Weak syllables contain "reduced" vowels. Usually this is the vowel schwa, as in the second syllable of *ion* or *scrounges*; but it may also be a very short form of another vowel, as in the second syllable of *pillow* or *cryptic*.

An alternative form of the syllable classification hypothesis, applicable to a stress language like English, holds that speech input is classified into *feet*. The foot is the rhythmic unit of stress languages; in English it consists of one strong syllable plus optionally one or more following weak syllables. Under such a classification system, stored representations would be in foot form, the recognizer would construct a prelexical representation of the signal as a sequence of specific feet, and lexical access could be attempted starting at every foot, that is, at the beginning of every *strong* syllable.

Investigations in our laboratory, however, have produced evidence counter to this proposal. Recall that the evidence for syllabic classification in French was that target detection was faster when the target corresponded exactly to a syllable than when the target was smaller or larger than a syllable. Target detection in English is *not* faster when the target corresponds exactly to a foot than when the target is smaller than a foot. For instance, the target *gar* is detected no faster in *gargoyle* (which consists of two strong syllables, i.e., two feet) than in *gargle* (which has a strong and a weak syllable and therefore is all one foot).

This suggests that English listeners do not classify speech input into feet. But as pointed out by Norris and Cutler (1985), the process of *classification* is logically distinct from the process of *segmentation*. Segmentation means making a division at some point in the signal. Classification means identifying units occurring in the signal. In order to *classify* speech into any sequence of units (phonemes, syllables, or feet) the recognizer must indeed segment the speech signal at the boundaries of these units. But the reverse is not true: It is possible to segment speech without classifying it. That is, the recognizer could segment the signal by choosing points at which to begin lexical access attempts, without necessarily constructing any prelexical representation of the signal as a sequence of specific phonetic segments, syllables, or feet.

Thus for a language like English, one might still hypothesize that the recognizer segments speech by starting a lexical access attempt at every strong syllable, despite the evidence that there is no classification into feet.

The success rate of such a procedure, in English at least, would be high. Statistical studies of the English vocabulary show that the number of lexical words (i.e., content words, excluding functors) beginning with strong syllables is approximately three times as large as the number beginning with weak syllables; moreover, those beginning with strong syllables occur, on average, twice as frequently as those beginning with weak syllables (Cutler & Carter, in press). This implies that, on average, we hear six times as many lexical items beginning with strong syllables as with weak syllables. This in turn implies that a recognizer that started lexical access at strong syllables would actually miss very few word beginnings. The false alarm rate would also be low in comparison with a lexical segmentation procedure that considered each phoneme or syllable to be a potential word onset location.

The proposal that lexical access starts with strong syllables is not a new one. It has been repeatedly suggested that weak syllables may be disregarded in computing a first-pass lexical access code (Bradley, 1980; Cutler, 1976; Grosjean & Gee, 1987; Taft, 1984). There is also evidence that listeners assume that weak syllables are not word-initial (do not start words). Taft (1984) presented listeners with ambiguous strings of a strong plus a weak syllable, such as [lɛtəs], and found that one-word readings (*lettuce*) were chosen far more often than two-word readings (*let us*); in contrast, the proportion of two-word choices was far higher when the second syllable was strong, as in [ˈnvɛsts] (*invests*; *in vests*).

How could one put to a test, in a way that directly measures speech recognition processes, the hypothesis that segmentation for lexical access occurs at strong syllables? Some researchers (e.g., Taft, 1984) claim that the hypothesis predicts that words beginning with strong syllables should be recognized more rapidly than words beginning with weak syllables. Thus *petrol* should be recognized faster than *patrol*, for instance. This tends to be true of auditory lexical decision responses, but it is probably accounted for by effects of word length (Cutler & Clifton, 1984). Other researchers, however (e.g., Grosjean & Gee, 1987), claim that words beginning with weak syllables can be accessed via their strong syllables (so that *patrol* would be accessed via *trol*) and that this mode of access should be just as efficient as access via a strong syllable that happens also to be a first syllable. Thus simple word recognition times offer no easy test of the hypothesis.

Suppose, however, that we were to construct a situation in which the occurrence of strong syllables led to segmentations that were inappropriate in that they did not actually correspond to any lexical item. For instance, suppose that real words were embedded in nonsense: Nonsense syllables that were strong should trigger inappropriate segmentations and competing lexical access attempts, whereas weak nonsense syllables should have no such effect.

In Experiment 1 we required listeners to detect real words in nonsense strings. (This task can be said to have a certain ecological validity, in that identifying real words in acoustic input is the task of speech recognition). For example, listeners were presented with *mint* embedded either in *mintayve* or *mintesh*. In *mintayve*, the second syllable, *tayve*, is strong. According to the strong syllable segmentation hypothesis, the string will be segmented and a lexical access attempt initiated at *tayve*. Detection of the word *mint*, which belongs partly to both syllables, will be interfered with by this inappropriate intersyllabic segmentation because successful detection will require assembly of material across a point at which the signal has been segmented. On the other hand, when the second syllable is weak (as in *mintesh*), the hypothesis predicts no segmentation and hence no interference. That is, *mint* should be detected faster in *mintesh* than in *mintayve*.

## Experiment 1

### *Method*

*Materials.* Thirty-two monosyllabic words were chosen in 16 pairs. All words had short vowels and ended in a two-consonant cluster. The members of each pair rhymed. Examples are *mint, hint, act, fact.* None of the words could be turned into another word by removing the last consonant (as, for instance, *tint* would make *tin*, or *pact* would make *pack*).

Two alternative vowel and consonant endings were added to the words to turn them into bisyllabic nonwords. One ending had a full vowel; the other ending had the weak vowel [ə]. Thus each word occurred in the context of two strong syllables (SS) and in the context of a strong first and a weak second syllable (SW). In the SS contexts, the first syllable was always more highly stressed than the second. The same two endings were used for the two members of any pair. For *mint* and *hint*, for example, the endings were *-ayve* and *-esh*, producing *mintayve, hintayve, mintesh*, and *hintesh*; for *fact* and *act*, the endings were *-uve* and *-em*, giving *factuve, actuve, factem*, and *actem*. The full set of items is listed in the Appendix. Seventy further bisyllabic nonwords were constructed, which did not begin with words. Examples are *bozzen, grivelom, scrornive*, and *crenthish*.

Because we did not know in advance how difficult the task of detecting a word in the initial portion of a bisyllabic nonword would prove for our subjects, we constructed 20 further items that were intended to extend the range of difficulty of the task. All of these "test" items began with real words; 10 of the items were intended to be considerably easier than our experimental items (e.g., *bookving*; *stretchib*), whereas the other 10 were intended to be considerably harder (e.g., *redgeling* [rɛdʒlɪŋ], which begins with *red*, and *panksim* [pæŋksɪm], which begins with *pang*).

Two lists were constructed, each containing all 70 nonword items, all 20 test items, and one version of each experimental item. Type of context (SS versus SW) was counterbalanced across pairs and lists. Thus *mintayve* and *hintesh* occurred in one list, *mintesh* and *hintayve* in the other. For each pair in each list, one member of the pair occurred in the first half of the list, and the other member occurred in the second half of the list. The lists were recorded by a male speaker of British English. The interval between items was 3 s. Only one recording of the whole set of materials was made; this recording contained both versions of each experimental item. Tape 1 was then made by copying all but the List 2 experimental items, and Tape 2 was made by copying all but the List 1 experimental items. Thus all of the items common to both tapes were acoustically identical on both tapes. A short set of practice items was also recorded.

*Subjects.* Thirty members of the Applied Psychology Unit subject panel took part in the experiment and were paid for participating. The responses of 4 of these subjects were not analyzed because they missed one third or more of the experimental items. Responses from a further 2 subjects were lost due to equipment failure. Twelve of the remaining subjects heard each tape.

*Procedure.* Subjects were tested individually. They were instructed that they would hear nonsense words and that they should press the response key whenever they heard a nonsense word beginning with a real word. They should then say aloud the word they had detected.

Reaction times for the experimental items were measured from a signal aligned with the burst of the final consonant of the embedded word. (For the test items, which did not always contain a stop consonant, we aligned the signal with the word onset.) Timing and data collection were under the control of a PDP 11/23 computer.

Subjects' spoken responses were recorded and checked. When a subject spoke any word other than the intended word, that response was discarded from the reaction time analysis.

The experimental items were digitized and measured. By adjusting the reaction times for these measurements we were therefore able to analyze responses from word onset as well as from the embedded word's final consonant.

### *Results*

The responses to our test items were inspected first. The mean detection latency for the "easy" words was 963 ms, and the miss rate was 4%. The "hard" words, on the other hand, were missed 24% of the time (and 6 of the 10 were missed by 50% of more of the subjects); when they were detected, the mean detection latency was 1,135 ms. This difference is significant, $t(23) = 5.09, p < .001$. The grand mean of response times to our experimental items, measured (as were the responses to the test items) from word onset, was 1,022 ms, that is, in between the easy and hard test items. We concluded that the overall difficulty of the experimental task was within a satisfactory range.

Some of our experimental items, however, apparently fell into the hard category. Specifically, four items were missed by 50% or more of the subjects. Three of these were words with low frequency of occurrence: *frond, vend*, and *apt*. The fourth was *fence*, which was an ill-chosen item, because it in fact contained the embedded word *fen*. Accordingly, we discarded these items, and, in order to maintain the balanced

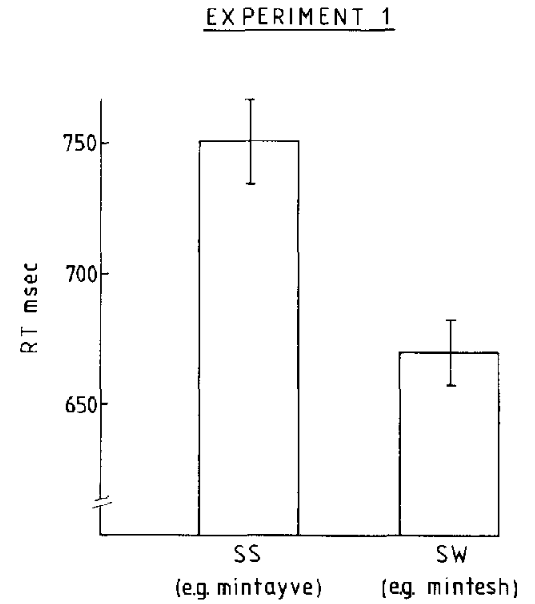

*Figure 1.* Mean word detection response times (milliseconds) for SS (two strong syllables) and SW (strong first, weak second syllable) items, Experiment 1. (The vertical lines give standard deviation values.)

structure of the materials sets, discarded their matched pairs (*blond, spend, crypt*, and *sense*) as well. Thus for each subject there were 12 items in each condition.

Separate analyses of variance were conducted with subjects and items as random factors. Mean response times for SS (two strong syllables) and SW (strong + weak) contexts are shown in Figure 1. It can be seen that detection latency was considerably slower in SS contexts (751 ms) than in SW contexts (669 ms). This difference is significant: $F(1, 22) = 11.54, p < .005$; $F_2(1, 12) = 22.29, p < .001$. The response times in Figure 1 are measured from the burst of the final consonant of the embedded word. Measuring from word onset, the detection latencies are 1,061 ms for SS contexts and 983 ms for SW contexts. This difference is also significant: $F_1(1, 22) = 10.13, p < .005$; $F_2(1, 12) = 18.85, p = .001$.

### *Discussion*

The results of this experiment are precisely as predicted by our model of segmentation at strong syllables: Words like *mint* were detected significantly faster in SW contexts (e.g., *mintesh*) than in SS contexts (e.g., *mintayve*). We argue that occurrence of the second strong vowel in SS contexts triggers segmentation, and segmentation of the bisyllable means that detection of the embedded word requires assembly of speech material across a point at which the signal has been segmented; this delays the detection process in comparison with that in SW contexts, where no segmentation occurs.

One potential alternative account can be easily dismissed. It might be suggested that the identity of the following vowel affected detection of the final consonant of the embedded word. Diehl, Kluender, Foss, Gernsbacher, and Parker (1985) have shown that detection latency for syllable-initial phonemes varies directly with the length of the following vowel. Because the vowels in the strong second syllables were longer than the vowels in the weak second syllables, perhaps longer response times to *mint* in *mintayve* than in *mintesh* simply reflect varying detection time for the [t]. However, Diehl et al.'s (1985) study investigated only full vowels. Our argument is that differences with full vowels are irrelevant, because our result reflects a difference in *kind* between strong and weak syllables, that is, between full vowels and reduced vowels. Moreover, other phoneme-detection studies show that Diehl et al.'s (1985) finding does not generalize to the case of schwa. Schwa is a very short vowel, yet phoneme-detection studies have shown that detection times for syllable-initial phonemes are *longer* if the syllable is unstressed, that is, contains a schwa. In the study by Cutler and Foss (1977), for example, phonemes followed by full vowels were detected on average 89 ms faster than phonemes followed by schwa. Therefore any potential confounding is in the wrong direction, because the present study showed that *mint* followed by schwa was detected relatively rapidly. If detection of the final phoneme of *mint* was indeed sensitive to whether the following vowel was full or schwa, this effect was swamped by the predicted inhibition of detection in the SS case.

Another potential confounding is less easy to dismiss. It could be that the embedded words as they were spoken in SS contexts sounded less like their canonical lexical templates than they did when they were spoken in SW contexts. If this was the case, a simple template-matching account might be able to explain the result. The lexicon cannot contain exact acoustic templates for every word because it is quite rare that the acoustic form in which we first hear a word is exactly reproduced on subsequent encounters. If there were whole-word lexical templates, they would have to be normalized and abstracted from acoustic representations. It is possible that *mint* in *mintayve* is spoken in such a way that it approximates less well to such an ideal lexical template for *mint* than does *mint* in *mintesh*. A template-matching account might claim, then, that the response time difference in SS versus SW contexts was due to acoustic factors, not to segmentation.

To rule out this alternative explanation, we conducted a second experiment, in which the same recordings of the same words were presented but without their nonsense endings. If *mint* in *mintayve* was indeed a less satisfactory exemplar of *mint* than *mint* in *mintesh*, then it should still be less satisfactory when *-ayve* has been edited off. Therefore the alternative explanation based on template-matching would predict that Experiment 2 should show the same result as Experiment 1: The *mint* from which *-ayve* has been removed should be detected significantly more slowly than the *mint* from which *-esh* has been removed. If, on the other hand, the claim of the segmentation model is correct, and the difference between SS and SW contexts in Experiment 1 is entirely due to the nature of the second syllable, then there should be no detection time difference when there are no second syllables. That is, the segmentation model predicts that detection time for *mint* from *mintayve* and for *mint* from *mintesh* should be equal.

## Experiment 2

### *Method*

*Materials.* The nonexperimental items from Experiment 1 were also digitized. All items were then edited down to monosyllables by using a waveform editor. For the experimental items, the final vowel-consonant (VC) sequence was removed, so that *mintayve, mintesh, factuve*, and *factem* became *mint, mint, fact*, and *fact*, respectively. The division was made so as to preserve as much as possible of the original item without including any of the second syllable's vowel. When there was a [t] burst in the original, for instance, this was included. Similar manipulations were performed on the other items; thus *bozzen, grivelom, scrornive*, and *crenthish* became *boz, grive, scrorn*, and *crenth*, respectively, whereas *bookving, stretchib, redgeling*, and *panksim* became *book, stretch, redge*, and *pank*. (By making the hard "test" items from Experiment 1 into nonwords, we preserved the effective word-nonword ratio of the earlier experiment.)

Two experimental tapes were constructed, which mimicked the orders of the previous experiment. Thus where *mintayve* had occurred on Tape 1 of Experiment 1, Tape 1 of Experiment 2 contained the *mint* from which *-ayve* had been removed, and Tape 2 contained the *mint* from which *-esh* had been removed.

*Subjects.* Twenty-four members of the Applied Psychology Unit subject panel took part and were paid for participating. Twelve subjects heard each tape.

*Procedure.* We attempted to keep the procedure as close as possible to that of Experiment 1. Subjects were instructed to press the button as soon as they heard a real word and then to say the word aloud. Subjects were tested individually, and the data were collected as in

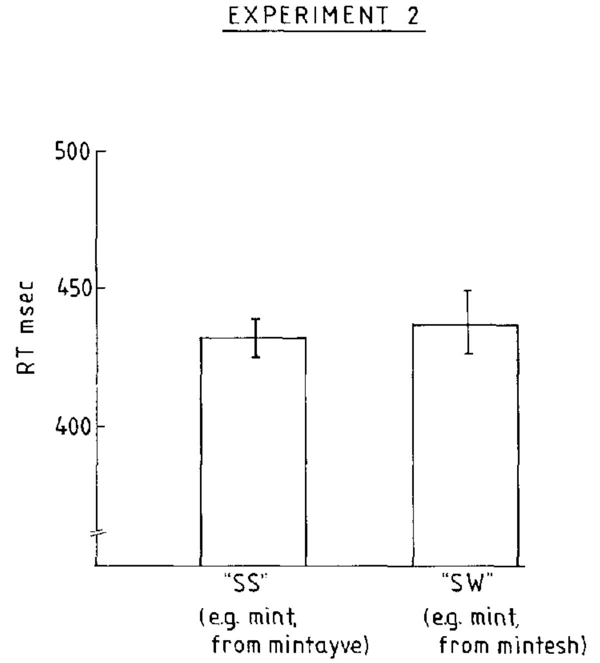

*Figure 2.* Mean word detection response times (milliseconds) in Experiment 2, for items that were SS (two strong syllables) versus SW (strong first, weak second syllable) in Experiment 1. (The vertical lines give standard deviation values.)

Experiment 1. Subjects' spoken responses were recorded and checked in the same way as in Experiment 1, and the durational measurements of the items again allowed us to compute responses from word onset or offset.

## *Results*

In order to maintain complete comparability with Experiment 1, we discarded the data from those items that had been discarded in the previous study. (In fact, the same items caused problems for subjects in this experiment. Although the overall miss rate was 15%, this was chiefly due to the items that were rejected from the previous study because they were missed by half the subjects—these had a mean miss rate of 47%. Otherwise, no item in this experiment had a high miss rate. It appears that at least some of our subjects were not acquainted with these low-frequency words.)

Again, we conducted separate analyses of variance with subjects and items as random factors. The mean detection times measured from the final consonant burst are shown in Figure 2. The mean detection time for originally SS items was 431 ms, and for originally SW items it was 437 ms, an insignificant difference (both $F_1$ and $F_2 < 1$). Measured from word onset, the mean detection times were 740 ms for originally SS items and 751 ms for originally SW items, again an insignificant difference (both $F_1$ and $F_2 < 1$).

## *Discussion*

The lack of difference between the two conditions of this experiment argues strongly against the suggestion that the detection latency difference of Experiment 1 reflected a difference in the way the words were spoken. Out of their following contexts, *mint* that once was followed by *-ayve* and *mint* that once was followed by *-esh* were equally quickly recognized as *mint*. Of course, the lack of a reaction time difference does not of itself demonstrate the lack of an articulatory difference; it is possible that words like *mint* are indeed spoken systematically differently when followed by strong versus weak syllables, but the difference does not render one significantly more *mint*-like than the other.

One further investigation of whether there is such an articulatory difference was suggested by the informal observation that neither of the authors was able to tell whether an individual item in the present experiment had previously had a strong or weak second syllable. The Cambridge linguistic community includes a number of highly trained phoneticians. Accordingly, we made a tape consisting of all of the experimental items from Experiment 2, plus a few of the other real-word items. The items occurred in random order on the tape. Seven experienced phoneticians agreed to listen to the tape. They were provided with a transcript that gave, in phonetic notation, the two possible bisyllables of Experiment 1 for each item and were asked to choose the bisyllable from which they thought each item had been extracted. For *mint* on the tape, for instance, they chose between [mɪnteɪv] and [mɪntəʃ]. The phoneticians listened to the tape at their own pace.

We reasoned that if there were differences in the way a given word was spoken in SS and SW contexts, phonetically trained listeners should be able to detect and correctly interpret the differences, thereby scoring significantly better than chance on this task. In particular, we predicted that if such differences existed, they should lead to a rather higher correctness score for originally SS items. In Experiment 1, words in SS bisyllables were detected significantly more slowly. If this was in any way due to the SS context's producing an utterance less like the isolation form than the utterance produced in the SW context, then the words extracted from SS contexts should offer more phonetic cues to their previous context than should the words extracted from SW contexts.

In fact, the phoneticians showed a bias toward choosing weak contexts ($z = 2.9$, $p < .005$). There was thus no trace of an advantage for words extracted from SS contexts. For the 48 words analyzed in Experiment 1, the mean percentages correct were as follows: for originally SS items, 45.8%; for originally SW items, 65.5%. For all 64 items the corresponding means were 46.9% and 66.5%. No subject achieved a higher score on originally SS words than on originally SW words. Overall, the mean percentages correct were as follows: for the 48 words analyzed in Experiment 1, 55.6%; for all experimental words, 56.7%; and for the filler items we included, 55.1%.

One must therefore conclude that there simply were no systematic differences in how our experimental words were spoken in SS and SW contexts. In particular, our subjects' bias toward choosing SW contexts suggests that the words all sounded as if they had minimal following context. Thus the results of this listening test, and the results of Experiment 2, allow us to conclude that the response time difference in Experiment 1 is highly unlikely to have resulted from *mint* in *mintayve*, for example, sounding less *mint*like than *mint* in *mintesh*. Our explanation of the Experiment 1 results was that they reflected segmentation effects. No segmentation occurred with the monosyllabic items of Experiment 2. Therefore there were no differences in detection response time.

Further potential confounding factors in Experiment 1 are addressed in our final experiment. Measurements showed that strong second syllables were longer than weak second syllables in Experiment 1. A possible alternative explanation of the detection time difference for words in SS and SW contexts could therefore be that subjects simply waited until the end of the item before responding. SS items were longer, and hence detection responses were delayed more.

Also, strong syllables have greater intensity than weak syllables. This allows yet another objection to be raised: Perhaps second syllables mask first syllables, and the greater the intensity, the greater the masking. On this account, words in SS contexts would simply be more difficult to detect because they were more effectively masked by following context.

It is not possible to counter these objections by removing the confounds in question. Short duration and lower intensity are among the defining characteristics that make a syllable weak. If a weak syllable is made as long as or longer than a strong syllable, or as loud as or louder than a strong syllable, it becomes strong. Therefore, strong second syllables will always result in greater overall item duration and greater second-syllable intensity.

However, it is possible to provide an indirect counterargument. Our segmentation-based explanation of the Experiment 1 results holds that the segmentation of SS sequences disrupts the detection of words that belong to both strong syllables. Consonant-vowel-consonant-consonant (CVCC) words like *mint* are hard to detect in bisyllables like *mintayve* because segmentation produces *min-tayve*, with part of *mint* belonging to each portion.

If the word to be detected belonged to only one portion, however, the segmentation model would hold that segmentation should have no effect on detection latency. Consider the set *mintayf*, *mintef*, *thintayf*, and *thintef*. Embedded in the first pair is the CVCC word *mint*, in the second pair the CVC word *thin*. Just as segmentation of *mintayf* would produce *min-tayf*, segmentation of *thintayf* would produce *thin-tayf*. However, because no part of *thin* belongs to the second portion *tayf*, the segmentation process should not in any way interfere with the detection of *thin*. The detection time difference for SS versus SW contexts that we found with embedded CVCC words should not be found with embedded CVC words.

Note that the fact that we predict segmentation of *thintayf* does not imply facilitation of the detection of *thin* in *thintayf*. Our model proposes (foot-based) segmentation without foot-based classification. Segmentation for lexical access occurs at strong syllables, that is, at the beginning of each foot; but because there is no classification of the signal into a sequence of specific feet, there is no need to determine what the foot is or even where it ends. Put another way, it is useful to the recognizer to know that lexical unit $X$ begins at point $t$, because this is precisely what the recognizer needs to know to initiate lexical access. But our model claims that the further information that $X$ ends at point $t + n$ is only of value in that it suggests that lexical unit $X + 1$ *begins* at $t + n$; the endpoint information may be useful for the processing of $X + 1$, but it is irrelevant to the processing of $X$. The segmentation of *thintayf* will therefore only affect the processing of the second syllable, *tayf*. Because our subjects are not making any response based on *tayf*, our model predicts that detection time for *thin* will not differ in *thintayf* and *thintef*.

Alternative explanations based on length or intensity, on the other hand, would predict that the detection time difference should be the same for CVC words as it is for CVCC words. *Thintayf* should be just as much longer than *thintef* as *mintayf* is than *mintef*. Subjects should have to wait longer for the end of SS items than the end of SW items irrespective of whether the embedded word is CVCC or CVC. Similarly, the second syllable of *thintayf* should be just as much louder than the second syllable of *thintef* as the second syllable of *mintayf* is louder than the second syllable of *mintef*. If there is a masking difference in the *mint* pair, there should be a similar masking difference in the *thin* pair.

Therefore, an experiment similar to Experiment 1, but with embedded CVC rather than CVCC words, will resolve these remaining potential objections. If overall item length or relative second-syllable intensity determines response time, such items will show the same detection time differences as the items of Experiment 1. If, however, segmentation is the cause of the detection time difference in Experiment 1, the embedded CVC words will not show that difference.

In Experiment 3 we measured detection latency for CVC words embedded in SS and SW contexts. We also took the opportunity to replicate Experiment 1 by comparing the detection time for CVC words with detection time for matched CVCC words.

## Experiment 3

### Method

*Materials*. Thirty-two words were chosen, half of which ended in a consonant cluster and half in a single consonant. The words formed quadruples such as *mint*, *hint*, *thin*, and *sin*; that is, a rhyming pair of words ending in a cluster was matched with a rhyming pair that (a) had the same vowel, (b) ended in a consonant that was the first consonant in the other pair's cluster, and (c) could not be made into words by adding the second consonant of the other pair's cluster (that is, *thint* and *sint* are not English words).

Again, all of the words were made into bisyllabic nonwords by the addition of an extra syllable. Two alternative pairs of VC endings were constructed for each quadruple; as in Experiment 1, within each pair, one vowel was full and the other was schwa. In contrast to Experiment 1, the final consonant was constant within each pair. Thus for the example given above, the endings were *-ayf*/*-ef* and *-oog*/*-eg*, making *mintayf*, *mintef*, *thintayf*, *thintef*, *hintoog*, *hinteg*, *sintoog*, and *sinteg*. For the consonant-final words like *thin*, the VC endings were preceded by the final consonant from the matched words like *mint*. The complete set of experimental materials is listed in the Appendix.

As before, two tapes were constructed, each tape containing one version of each item plus 70 nonword bisyllables (most of which were the same as in Experiment 1) and 3 further bisyllables beginning with real words. Tape counterbalancing and recording was as in Experiment 1.

*Subjects.* Subjects were 42 undergraduate members of Churchill College, Cambridge, who were paid for participating. The verbal responses of 9 subjects were lost when a response tape was accidentally erased. Because it could not be ascertained whether these subjects had responded with the correct word, their response time data were also discarded. Two further subjects were rejected for missing too many items (according to the same criteria used in Experiment 1). Of the remaining subjects, 16 heard Tape 1 and 15 heard Tape 2.

*Procedure.* The testing procedure was as in Experiment 1 except that a portable microcomputer was used to control timing and data collection. Again, subjects' spoken responses were checked, and the materials were digitized and measured.

## *Results*

As in Experiment 1, it was necessary to discard some items. The word *numb* was missed by 22 of the 31 subjects. According to the criteria established in Experiment 1, we therefore discarded this item along with the remaining members of its matched quadruple, *gum, jump,* and *lump.* (In fact, the word *gum*—although it did not quite reach the rejection criterion of a 50% miss rate—received the second highest number of misses: 14 out of 31.) This left seven items in each condition for each subject. Separate analyses were conducted with subjects (an unequal $N$ analysis) and with items as random factors.

Mean response times for each word type in each context are displayed in Figure 3. These response times were again measured from the burst of the stop consonant within the item. The difference for the cluster-final items replicated the results of Experiment 1: Words were detected significantly more slowly in SS contexts (818 ms) than in SW contexts (726 ms), $F_1(1, 30) = 9.43$, $p < .005$, $F_2(1, 13) = 6.69$, $p < .025$. However, detection latency for the consonant-final items in SS contexts (705 ms) was not significantly different from that in SW contexts (697 ms); both $F_1$ and $F_2 < 1$). Response times from word onset for the cluster-final items were 1,234 ms in SS contexts and 1,135 ms in SW contexts: $F_1(1, 30) = 9.73$, $p < .005$, and $F_2(1, 13) = 7.06$, $p < .02$. Response times from word onset for consonant-final items were 1,102 ms in SS contexts, and 1,091 ms in SW contexts (both $F_1$ and $F_2 <$ 1).

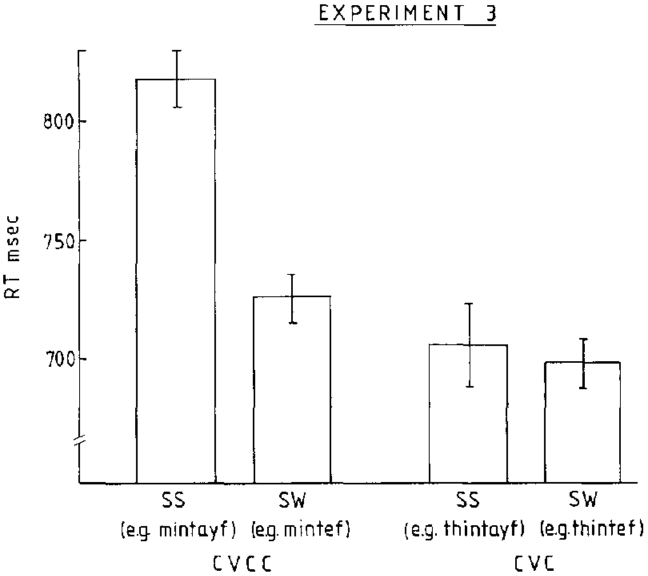

*Figure 3.* Mean word detection response times (milliseconds) for SS (two strong syllables) and SW (strong first, weak second syllable) items in Experiment 3 as a function of whether the embedded word ended in a cluster (CVCC) or a single consonant (CVC). (The vertical lines give standard deviation values.)

## *Discussion*

As predicted by the segmentation model, strong second syllables slow the detection of embedded words only when the words actually belong partly to the second syllable. According to our model, *mint* is detected more slowly in *mintayf* than in *mintef* because *mintayf* is segmented into *min-tayf* so that *mint* has to be assembled from materials on either side of a segmentation point. This delays detection in comparison with detection of *mint* in *mintef,* where the second syllable is weak and hence does not trigger segmentation. *Thin,* however, is detected equally rapidly in *thintayf* and *thintef*—despite the fact that *thintayf* is segmented, whereas *thintef* is not—because segmenting *thintayf* into *thin-tayf* does not delay detection of *thin,* which belongs only to the first syllable.

Alternative explanations of the results of Experiment 1, suggesting that the detection-time delay in SS contexts is due to greater length of the second syllable or greater intensity of the second syllable, can therefore be rejected.

## General Discussion

The experimental evidence presented in this article strongly supports our model of segmentation based on strong syllables. We have shown that detection of a word is delayed when the word belongs to two strong syllables, but not when it belongs to a strong syllable followed by a weak syllable. We explained this result as an effect of segmentation triggered by the strong syllable; detection of the embedded word is delayed by the need to assemble speech information across a segmentation point.

Our findings call into question many basic models of speech recognition. Firstly, a simple phonetic classification model (e.g., Foss & Gernsbacher, 1983) has no way of predicting our result. If lexical access is based on a representation of the input in terms of phonetic segments, then lexical access attempts may be initiated at each segment; however, without elaboration of the model there is no basis for predicting that some potential segmentation points will be preferred and others disregarded.

Secondly and more seriously, our results are directly contrary to the predictions of syllabic classification models (e.g., Mehler, 1981; Segui, 1984). As we pointed out in our introduction, the syllabification of English is relatively ambiguous compared to that of some other languages. Thus it is perhaps not surprising to find that phonologists argue about whether a string such as *mintesh* should be syllabified *min-tesh*, *mint-esh*, or even *mint-tesh*. But all of these syllabifications are in conflict with our results. If *mintayve* is syllabified *min-tayve* (as phonologists seem to agree), and *mintesh* is syllabified *min-tesh*, then *mint* should be equally difficult to detect in each. Experiment 1 showed that this is not the case. If *mintesh* is syllabified *mint-esh* or *mint-tesh*, on the other hand, then *thintef* should be syllabified in the same way, and hence detection of *thin* in *thintayf* (*thin-tayf*) should be faster than detection of *thin* in *thintef* (*thint-ef* or *thint-tef*). Experiment 3 showed that this is not the case (as, of course, did the foot classification experiments mentioned in the introduction, in which *gar* was not detected faster in *gargoyle* than in *gargle*). Thus, syllabic classification, no matter where one draws the syllable boundaries, is refuted by our results.

Thirdly, our results are also directly opposed to the predictions of models of word recognition based on strictly left-to-right processes, such as the cohort model (e.g., Marslen-Wilson, 1980, 1987). According to the cohort model there should never be effects of following context on recognition of a word; yet effects of following context are precisely what we have demonstrated. In fact, this result is in accord with other recent demonstrations of following context effects in auditory word recognition. Taft and Hambly (1986) showed that the processing of a nonword string continues after the point at which there are no possible continuations that would make it a word. Grosjean (1985), using the gating paradigm, showed that words in a continuous speech context are frequently not recognized until after their acoustic offset. Our results provide a further argument that strictly left-to-right word recognition models are insufficient.

Our experiments suggest that speech recognition involves a process of segmentation that is triggered by the occurrence of a strong syllable. In the introduction we argued that such segmentation is motivated by the need to find the most efficient starting points for lexical access attempts. Into what more general framework could segmentation processes of this kind be incorporated?

We see two possibilities. Although there is evidence that English listeners do not classify speech input in terms of feet or syllables, we know of no experimental evidence against phonetic classification for English. Our postulated segmentation processes could be incorporated into a model involving phonetic classification in the following way. As the continuing classification process produces an output (a string of phonetic segments), the occurrence in this string of one of a small set of segments (the set of full vowels) could trigger initiation of a lexical access attempt, beginning from the vowel itself plus a syllabic onset (which could be, for instance, the maximal onset permitted by the given phonetic sequence, or possibly a specified number of phonetic segments; further research would be necessary to decide this issue). This procedure, in comparison with a policy of starting an access attempt at every phoneme, would have the great advantage of drastically reducing the number of lexical access attempts; moreover, it would concentrate access attempts at those points where they were most likely to be successful.

On the other hand, it is also clear that segmentation at strong syllables could be incorporated into a model involving no classification at all. Full vowels constitute quite reliably detectable portions of speech waveforms (they are highly resistant to casual misperception, for instance; Bond & Garnes, 1980). Suppose that a segmentation device simply monitored the incoming waveform for high-energy quasi-steady-state portions of a certain minimum duration (either absolute duration or duration relative to some standard obtained by monitoring the signal for, say, rate of occurrence of energy peaks). Upon encountering such a portion, the device could segment the signal at a point prior to the onset of the steady state (where once again the duration of the preceding waveform portion could be absolute or relative). A lexical access attempt could then be initiated from that point, in which the input to the lexicon could be a raw acoustic representation to be loosely matched against lexical templates. Our segmentation data as they stand are compatible with either of these general types of models.

They are also compatible with either of two different accounts of the precise source of the interference effect for word detection across a segmentation. We have suggested that when a division has been made in the speech signal, detection of a word that occurs partly on either side of this division may simply be rendered difficult by the necessity of reassembling speech material that has been divided. But we have also suggested that the primary motivation for postulating divisions in a continuous speech signal is the search for suitable points at which to initiate lexical access. Thus it is our contention not only that *tayf* is segmented off from *mintayf* but also that a lexical access attempt is initiated for *tayf*. We have in the present results no direct evidence for this lexical access attempt. But if we are right, then the interference with the detection of *mint* in *mintayf* may arise not merely from the difficulty of reassembling divided speech, but from competition of lexical hypotheses. That is, in the *mintayf* case one lexical access attempt would begin from the clear initial boundary, and another would begin from the boundary postulated in miditem. The first and second lexical hypotheses would then compete for the /t/, slowing the acceptance of the first hypothesis. In the *mintef* case, there would be no competing second hypothesis and hence no interference.

Further research will be necessary to distinguish between these accounts for the precise genesis of the segmentation effect that we have demonstrated. The argument to date, however, is that strong syllables trigger segmentation of continuous speech signals.

## References

Bond, Z. S., & Garnes, S. (1980). Misperceptions of fluent speech. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 115–132). Hillsdale, NJ: Erlbaum.

Bradley, D. C. (1980). Lexical representation of derivational relation. In M. Aronoff & M. L. Kean (Eds.), *Juncture* (pp. 37–55). Sarasota, CA: Anma Libri.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics, 20,* 55–60.

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

Cutler, A., & Clifton, C. E. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 183–196). Hillsdale, NJ: Erlbaum.

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech, 20,* 1–10.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1983). A language-specific comprehension strategy. *Nature, 304,* 159–160.

Cutler, A., Mehler, J., Norris, D. G., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25,* 385–400.

Diehl, R. L., Kluender, K. R., Foss, D. J., Gernsbacher, M. A. & Parker, E. M. (1985, April). *Consonant identification time depends on the length of the following vowel.* Paper presented to the Acoustical Society of America, Austin, Texas.

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

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

Grosjean, F., & Gee, J. (1987). Prosodic structure and spoken word recognition. *Cognition, 25,* 135–155.

Holmes, J. N. (1984). Speech technology in the next decades. In M. P. R. van den Broecke & A. Cohen, (Eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences* (Vol. IIA, pp. 59–62), Dordrecht, The Netherlands: Foris.

Marslen-Wilson, W. D. (1980). Speech understanding as a psychological process. In J. C. Simon (Ed.), *Spoken language generation and understanding* (pp. 39–67). Dordrecht, The Netherlands: Reidel.

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

Mehler, J. (1981). The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society, London, B295,* 333–352.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior, 20,* 298–305.

Norris, D. G., & Cutler, A. (1985). Juncture detection. *Linguistics, 23,* 689–705.

Segui, J. (1984). The syllable: A basic perceptual unit in speech processing? In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 165–181). Hillsdale, NJ: Erlbaum.

Taft, L. (1984). *Prosodic constraints and lexical parsing strategies.* Unpublished doctoral dissertation, University of Massachusetts.

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

## Appendix A
## Experimental Materials

| Experiment 1 | | Experiment 3 | |
|---|---|---|---|
| Stimulus set | Phonetics[a] | Stimulus set | Phonetics[a] |
| mintayve, mintesh, hintayve, hintesh | [eɪv], [əʃ] | mintayf, mintef, thintayf, thintef, | [eɪf], [əf] |
| meltive, meltesh, peltive, peltesh | [aɪv], [əʃ] | hintoog, hinteg, sintoog, sinteg | [ug], [əg] |
| factuve, factem, actuve, actem | [uv], [əm] | meltook, meltek, teltook, teltek | [uk], [ək] |
| riskime, riskel, whiskime, whiskel | [aɪm], [əl] | pelteesh, peltesh, yelteesh, yeltesh | [ɪʃ], [əʃ] |
| huskaze, husken, duskaze, dusken | [eɪz], [ən] | spendeek, spendek, glendeek, glendek | [ik], [ək] |
| softain, softej, loftain, loftej | [eɪn], [əʤ] | sendibe, sendeb, hendibe, hendeb | [aɪb], [ɔb] |
| stampoaj, stampent, stumpoaj, stumpent | [oʊʤ], [ɔnt] | flaskipe, flaskep, glaskipe, glaskep | [aɪp], [əp] |
| boltoach, boltra, joltoach, joltra | [oʊtʃ], [rə] | maskayth, masketh, paskayth, pasketh | [eiθ], [əθ] |
| lumpoid, lumpesh, jumpoid, jumpesh | [ɔid], [əʃ] | deskythe, desketh, meskythe, mesketh | [aɪθ], [əθ] |
| wristoin, wrister, fistoin, fister | [ɔin], [ə] | duskoov, duskev, fuskoov, fuskev | [uv], [əv] |
| liftude, liftel, giftude, giftel | [ud], [əl] | diskipe, diskep, miskipe, miskep | [aɪp], [əp] |
| nestume, nestes, westume, westes | [um], [əs] | riskeeb, riskeb, kiskeeb, kiskeb | [ib], [əb] |
| frondoiz, frondes, blondoiz, blondes | [ɔiz], [əs] | stampaig, stampeg, prampaig, prampeg | [eɪg], [əg] |
| vendite, vendel, spendite, spendel | [aɪt], [əl] | stumpeef, stumpef, drumpeef, drumpef | [if], [ɔf] |
| cryptove, cryptem, aptove, aptem | [oʊv], [əm] | jumpoov, jumpev, numpoov, numpev | [uv], [əv] |
| fensipe, fensej, sensipe, sensej | [aɪp], [əʤ] | lumpaysh, lumpesh, gumpaysh, gumpesh | [eɪʃ], [əʃ] |

[a] Phonetic transcriptions of the two second syllables of each set.

Received November 14, 1986
Revision received March 11, 1987
Accepted March 11, 1987 ■