Journal of Speech and Hearing Research, Volume 32,912-920, December 1989

*

THE PERCEPTION OF RHYTHM AND WORD BOUNDARIES IN NOISE-MASKED SPEECH

MARY R. SMITH ANNE CUTLER SALLY BUTTERFIELD IAN NIMMO-SMITH
MRC Applied Psychology Unit, Cambridge, U.K.

The present experiment tested the suggestion that human listeners may exploit durational information in speech to parse continuous utterances into words. Listeners were presented with six-syllable unpredictable utterances under noise-masking, and were required to judge between alternative word strings as to which best matched the rhythm of the masked utterances. For each utterance there were four alternative strings: (a) an exact rhythmic and word boundary match, (b) a rhythmic mismatch, and (c) two utterances with the same rhythm as the masked utterance, but different word boundary locations. Listeners were clearly able to perceive the rhythm of the masked utterances: The rhythmic mismatch was chosen significantly less often than any other alternative. Within the three rhythmically matched alternatives, the exact match was chosen significantly more often than either word boundary mismatch. Thus, listeners both perceived speech rhythm and used durational cues effectively to locate the position of word boundaries.

KEY WORDS: speech perception, prosody, rhythm, word boundaries, noise-masking

A six-syllable sequence in English could be made up of any number of words from one *(identification)* to six (7 *don't have a case on).* If the speech is heard simultaneously with noise that is intense enough to mask the segmental structure, can listeners tell one of these sequences from the other?

The evidence to date on this question is mixed. Certainly, listeners do not always perceive word boundaries accurately. Psycholinguists have for many years collected and analysed the "slips of the ear" that occur in conversation; many of these contain word boundary misplacements, such as "won't bother me" —> "lobotomy," or "disguise" —> "the skies" (Bond & Games, 1980; Browman, 1978).

Experimental studies of word boundary perception, however, have not been numerous. Most have focussed on the use of segmental variations. At this level, it is known that the presence of a word boundary can produce such effects as an intervocalic glottal stop or aspiration of word-initial stop consonants (Bolinger & Gerstman, 1957; Garding, 1967; Lehiste, 1960, 1964). Christie (1974) established in an experiment using synthesized English stimuli that prevocalic stop aspiration is an effective cue to a word or syllable onset. Zwanenburg, Ouweneel, and Levelt (1977), using French material, showed that listeners can distinguish minimal juncture pairs (two identical sequences of phonemes, which differ in the location of a word boundary) when they have been spliced out of context. Shimizu and Dantsuji (1980) likewise found that Japanese listeners could distinguish minimal pairs in Japanese. Nakatani and Dukes (1977) also found this to be true with English pairs, and found, furthermore, that the effective cues to the presence of a juncture were located in word beginnings rather than word endings; some further effects in word-medial positions provided evidence for the *absence* of juncture. A series of studies in Dutch by Quene (1985, 1987a, 1987b) found that speakers frequently differed considerably in how they signalled word boundaries, but that listeners could make use of whatever cues speakers provided. The most accurately located word boundaries were those falling after a consonant and before a vowel; in this context, the consonant duration was a stronger cue than the duration of the vowel rise time. Sonorant consonants provided better prejunctural cues than fricatives or plosives.

It is known that segmental duration varies with position in a word. Consonants, for example, tend to be longest in word-initial position, somewhat shorter in word-final position, and shortest in word-medial position (Klatt, 1974, 1976; Oiler, 1973; Umeda, 1977). Vowel duration is primarily determined by stress rather than position in the word (Umeda, 1975). Syllable duration also varies with position in the word; if stress and syllable weight are controlled, word-final syllables are somewhat longer than nonfinal syllables (Klatt, 1975; Oiler, 1973), though some investigators have failed to find such effects (Harris & Umeda, 1974). In fact, such durational effects are so small and so variable that Klatt (1976, p. 1213) has speculated that durational cues to word boundary location in English are unlikely to be of much perceptual use.

Durational variation in speech, however, occurs at several levels, of which segmental duration is only one. At the highest level there is syntactically conditioned durational variation, and this has been shown to be perceptually useful in locating syntactic boundaries (Bouwhuis & de Rooij, 1977; Lehiste, Olive, & Streeter, 1976; Scott, 1982; Streeter, 1978). In stress languages like English, there is stress rhythm—the pattern of strong and weak syllables, where strong syllables are those containing a full vowel (stressed syllables and secondarily stressed syllables), and weak syllables are those containing a central or reduced vowel (unstressed syllables). Stress rhythm is also known to be easily perceptible. Lieberman (1965) showed that whether a syllable was strong or weak could be accurately identified in electronically filtered English speech, in which no segmental

0022-4685/89/3204-0912$01.00/0

information was present. Carlson, Granstrom, Lindblom, and Rapp (1972) found that trained Swedish listeners' reproduction of the pattern of strong and weak syllables from reiterant and synthesized reiterant speech was very accurate. (In reiterant speech a natural utterance is mimicked by a series of repetitions of a single syllable such as "ma.") Svensson (1971, 1974) similarly found that Swedish listeners presented with hummed speech produced candidate transcriptions that accurately preserved the pattern of strong and weak syllables.

Directly relevant to the present research are two studies offering tantalizing evidence that word boundary cues might be provided by durational variation at a level somewhere between stress rhythm and segmental duration. Both studies used reiterant speech, thereby factoring out most segmental information. The first study is that by Carlson et al. (1972), mentioned above. Carlson et al. found that, in general, listeners' ability to locate word boundaries in reiterant and synthesized reiterant speech was not significantly different from chance. However, Carlson et al. also found that their materials fell into two groups—with some utterances, listeners performed significantly better than chance with both natural and synthetic reiterant speech, whereas with others they performed below chance. Carlson et al. concluded that "the timing information . . . does in fact contribute to the word boundary identification" (p. 18); however, they also speculated that the utterances in which listeners could more successfully locate word boundaries might be those in which word boundaries occurred in positions that were, in Swedish, highly probable. (Unfortunately, Carlson et al. did not test this suggestion directly.)

The second study, by Nakatani and Schaffer (1978), found that listeners performed significantly better than chance at distinguishing reiterant speech versions of adjective-noun sequences such as "noisy dog" and "bold design" (i.e., sequences matched for stress rhythm), and that the most effective cue was the relative syllable duration of the phrases. This study embodies the most direct claim that durational variation contains cues to word boundary location.

The present study provides a new test of such durational cues. We presented listeners with natural speech, masked by noise to an extent that removed all but the grosser segmental information. We assessed listeners' perception both of stress rhythm (i.e., the pattern of strong and weak syllables) and of word boundary location. Two further variables were included. The first was listeners' sensitivity to prosodic probabilities in English. Carlson et al. (1972) suggested that the relationship between stress rhythm and word boundaries in Swedish is not unconstrained; likewise, there are significant regularities in English—about three-quarters of the words in the vocabulary begin with strong syllables (Cutler & Carter, 1987). Therefore, we investigated whether listeners' responses would be sensitive to this property of their native language, by varying the degree to which the words within utterances began with strong or with weak syllables. The second additional factor was the nature of the vowels in the strong syllables we presented. This was included as a control factor. In the presence of noise, it may be that vowel quality, intrinsic acoustic attributes of vowels, and the frequency of occurrence of different vowels in the language interact in a complex way; for instance, the perceived quality of different vowels becomes differentially affected by the presence of masking noise (Pickett, 1957). If listeners are sensitive to the distribution in the language of vowels in strong syllables, then they may report rhythmic similarity with less success if they perceive the noise-masked vowels as those less likely to occur in strong syllables. Alternatively, the masking noise may differentially distort information about intrinsic duration, so that listeners would be less successful in the task when the signal contained phonetically short stressed vowels than when it contained phonetically long stressed vowels. Or again, utterances with phonetically short stressed vowels may simply have shorter total durations and, thereby, provide listeners with less persistent information for the task. To assess whether such factors might be affecting our results, we systematically varied the distribution of vowels in the stressed (strong) syllables.

METHOD

Materials

Forty-eight unpredictable utterances of six syllables ("rings amused the sultan"; "conduct ascents uphill") were constructed. Each utterance had an alternating stress rhythm of strong (S) and weak (W) syllables. In half the cases the rhythm was SWSWSW ("rings amused the sultan"); in the other half it was WSWSWS ("conduct ascents uphill"). Note that each of these rhythmic structures allows very many different possible divisions into words, and each is a very common pattern in English; thus we did not offer our subjects the opportunity to exploit such factors as the maximum permissible number of weak syllables in sequence.

Two further factors were varied systematically in the materials. One was where word boundaries occurred with respect to the rhythm. One third of the utterances had only weak word-initial syllables ("conduct ascents uphill"; "mean baboons detained him"—note that although in the latter example the very first syllable is strong, the first syllable is to a certain extent irrelevant, because subjects have no choice about whether or not it is word-initial). A further one third had only strong word-initial syllables ("dusty senseless drilling"; "an eager rooster played"); and the remaining third had a mixture of strong and weak word-initial syllables ("rings amused the sultan"; "achieve her ways instead").

The remaining factor was the nature of the vowel in the strong syllables. These were chosen from a set of three phonetically short vowels (/E/, /I/, /A/) and a set of three phonetically long vowels (/e/, /i/, /u/). One quarter of the utterances contained all long vowels in the strong syllables ("mean baboons detained him"); one quarter con-

tained all short vowels ("conduct ascents uphill"); and the remaining half contained a mixture of long and short vowels ("rings amused the sultan").

These 48 sentences were the base utterances that were presented auditorily *to* the subjects. For each of these base sentences four further sentences were constructed, to serve as the response set *to* be presented to the subjects. These sentences were also each six syllables in length, with three strong and three weak syllables. Each of the four items in a given response set had the same three vowels in the strong syllables as the base sentence had. However, the rhythm and word boundary placement in the response set was varied, in the following manner: (a) one member of the response set was exactly matched to the base in rhythm and word boundary placement; (b) one member was mismatched on rhythm; (c) the remaining two members were matched to the base on rhythm but mismatched on where word boundaries occurred. These latter two differed only in that in every case one contained more strong word-initial syllables than the other.

Thus for "rings amused the sultan" the exact match was "things confused the culprit"; the rhythmic mismatch was "a drink boosted results"; and the word boundary mismatches were "clinging movements rustle" (all word-initial syllables strong) and "pink balloons disgust them" (most word-initial syllables weak). For "conduct ascents uphill" the exact match was "adjust attempts until"; the rhythmic mismatch was "combustion prevents spills"; and the word boundary mismatches were "a rustic settled hill" and "robust as seven mills." It proved impossible to achieve exact matching between the four alternatives within each set with respect to the degree to which they shared phonetic segments with *the* base utterance (note that they always shared all vowels; only the degree of consonantal sharing varied). However, many of the shared consonants occurred in the weak syllables, and hence, were assumed to be less perceptible under noise masking. Furthermore, a shared consonant in one response alternative was often paralleled in the other response alternatives by a consonant in the same position differing from the base utterance's consonant by only one feature; under noise masking, two such minimally different consonants would be readily confused. On average, the base utterances contained 11.5 consonants, and the mean number of consonants shared identically with the base was 4.7 for the exact matches, 3.2 for the rhythmic mismatches, and 2.8 and 2.6 for the two types of word boundary mismatch. The complete set of 48 base sentences with their respective response sets is listed in the Appendix.

The 48 base sentences, along with a short set of practice sentences, were recorded in a sound-dampened room by a native speaker of Standard Southern British English. The recording was made on a Revox B77 reel-to-reel tape recorder using a Shure 565SD unidirectional dynamic microphone. The 48 base sentences were then copied, in a different sequence, so that the final recording contained the initial practice set followed by 96 experimental trials, with each of the 48 base sentences occurring twice in the sequence of 96. Each sentence was repeated each time it occurred; the interval between trials (where a trial was a single sentence uttered twice) was approximately 5 s. The speaker's voice gave, prior to each trial, the number (from 1 to 96) of the trial and, prior to each repetition, the word "again." To this recording, noise was added coincident with each sentence, using a white noise generator. The spectrum of the noise was essentially flat across the frequency range of speech. The sentence numbers and the word "again," which signalled each repetition, were not masked. The output of the noise generator was varied with the amplitude of the 48 base sentences; the signal-to-noise ratio, averaged across the utterances, was approximately -10 dB. To derive this ratio, the speech and noise waveforms were digitized at 20 kHz and the ratio was calculated of the rms amplitude of the nonsilent portions of the speech to the rms amplitude of the white noise within the bandwidth 0-8 kHz. It is reported that at this signal-to-noise ratio listeners can accurately detect the presence of speech but cannot recognize it—that is, they can perceive only the grossest segmental structure (Erber, 1971).

Three response sheets were constructed. Because subjects were to be presented on each trial with two response alternatives, and because the complete response set for each trial contained four items, there were six possible pairings of items from the response set: AB, AC, AD, BC, BD, CD. For each base sentence, the AB and CD pairs occurred on one response sheet, the AC and BD pairs on another, and the AD and BC pairs on the third. The pairings were counterbalanced across response sheets so that each response sheet contained an equal number of exact matches, rhythmic mismatches, and word boundary mismatches in equal combinations. Order of alternatives within pairs was also counterbalanced across the set as a whole. The response alternatives for each trial were numbered as the trials were numbered on the tape. The response sheets required subjects both to make a choice within each pair and to indicate the confidence (high, medium, or low) with which they had made that choice.

Two control studies were conducted. First, to assess the inherent plausibility of the response alternatives, the response sheets as used in the main experiment were presented to 18 subjects without the accompaniment of any auditory stimuli. The subjects were asked to choose the most plausible member of each pair of alternatives, and to indicate the confidence with which they had made that choice. Overall, the proportions of choices fell between .20 and .30 for each of the four response alternative types; there was no sign of a strong bias towards any alternative. We concluded that subjects' responses would not be determined by inherent plausibility of the response alternatives available to them.

Second, to assess the effectiveness of the noise masking, the experimental tape was played to 3 listeners with some phonetic sophistication, who were asked to transcribe as much of the masked speech as they could. Overall, 16.25% of the phonetic segments in the masked speech were correctly transcribed. However, only 13.3% of consonants were transcribed correctly; the listeners

were more successful at identifying vowels (21.9% correct). Because all the response alternatives contain the same vowels as the base utterances, identifiability of vowels cannot affect the outcome of the experiment; what is important is that the consonants proved difficult to identify. They were not, however, *impossible* to identify, a fact that we took into account in the data analysis. In order to check whether any results we obtained over our materials set as a whole might have been due, in part, to differing similarity between the response alternatives and the base, we identified a subset of our materials that were almost precisely matched on the degree to which the response alternatives shared consonants with the base. The subset comprised 20 items (numbers 3, 7, 8, 9, 13,14, 18, 19, 21, 27, 28, 31, 32, 33, 38, 40, 43, 44, 46, and 48 in the Appendix). Across this subset, the exact match averaged 3.5 consonants shared with the base, the rhythmic mismatch 3.45, and the two word boundary mismatches 3.45 and 3.4. Moreover, the set was balanced not just in the mean number of shared consonants, but in the number of times each response alternative type had the most shared consonants, which was six times for all four alternatives (a total of 24 because there were four ties). All statistical analyses carried out on our materials set as a whole were repeated on this subset.

Subjects

Subjects were 40 members of the Applied Psychology Unit subject panel. They were paid for their participation in the experiment, which was administered as the first experiment in a 1-hr session with four other short experiments. The data of 1 subject, who was rejected from the series of five experiments for failure to comply with instructions on one, were discarded. Each of the three alternative forms of the response sheet was received by 13 of the remaining 39 subjects.

Procedure

Subjects were tested in groups of from 2 to 6. No more than 2 subjects being tested in any one session received the same form of the response sheet. The experimental tape was presented from the Revox B77 over Sennheiser HD 420SL stereo headphones at approximately 64 dB SPL. Subjects recorded their own responses on their response sheets. They were instructed to listen to each noise-masked sentence and to choose from the two response alternatives given on the response sheet the one that more closely matched the rhythm of the auditory presentation. The instructions stressed that the rhythm was the crucial criterion, and illustrated this with an example. Subjects were also told to indicate for each trial the confidence with which they made their choice, on the 3-point high/medium/low scale. They were explicitly told that neither of the response choices they were offered was actually what they were hearing; it was emphasized that they were to choose whichever was the better match. If neither seemed a good match they were to make a choice anyway and give the match a low confidence rating.

Results

The proportions of choices for each of the four response types, averaged across items, are shown in Tables 1 (as a function of rhythmic pattern and word onset type) and 2 (as a function of rhythmic pattern and vowel sequence). It can be seen that in every row, the exact match was chosen more often, and the rhythmic mismatch less often, dian any other option. Recall that our procedure of two alternative forced-choice presented subjects with 50% of trials in which no exact match was available; thus, the maximum possible proportion of choices for the exact match (or for any other option) was .50.

We conducted three separate analyses of variance on the frequency with which a given response was chosen against an alternative, and three parallel analyses on the

TABLE 1. Proportion of response alternative choices as a function of rhythmic pattern and word onset type.

Choices	*Exact match*	*Rhythmic mismatch*	*Wb mismatch: more strong onsets*	*Wb mismatch: more weak onsets*
SW pattern				
AH strong onsets	.353	.168	.266	.213
All weak onsets	.359	.061	.237	.343
Mixed S & W onsets	.401	.069	.248	.282
WS pattern				
All strong onsets	.381	.051	.309	.258
All weak onsets	.375	.099	.255	.271
Mixed S & W onsets	.314	.122	.284	.280

The table shows the proportion of times each of the four response alternatives was chosen, as a function of the rhythmic pattern (SWSWSW or WSWSWS) and word boundary placement (all strong, all weak or mixed strong and weak word-initial syllables) of the base utterance. The proportions sum across each row to 1.00, but because the response alternatives were presented as paired comparisons rather than all at once, the maximum proportion for each cell is .500.

TABLE 2. Proportion of response alternative choices as a function of rhythmic pattern and vowel type.

Choices	*Exact match*	*Rhythmic mismatch*	*Wb mismatch: more strong onsets*	*Wb mismatch: more weak onsets*
SW pattern				
All long vowels	.406	.122	.192	.280
All short vowels	.374	.085	.293	.248
Mixed vowels	.351	.095	.259	.295
WS pattern				
All long vowels	.340	.043	.293	.325
All short vowels	.353	.105	.254	.288
Mixed vowels	.368	.108	.292	.233

The table shows the proportion of times each of the four response alternatives was chosen, as a function of the rhythmic pattern (SWSWSW or WSWSWS) and vowel pattern (all long, all short, or mixed long and short vowels) of the base utterance. The proportions sum across each row to 1.00, but because the response alternatives were presented as paired comparisons rather than all at once, the maximum proportion for each cell is .500.

relative confidence levels with which the choices were made (coded 1, 2, and 3 for low, medium, and high respectively). Each analysis included the factors: (a) rhythmic pattern (i.e., whether the base utterance was SWSWSW or WSWSWS); (b) word onset type (i.e., whether the base had all strong word-initial syllables, all weak word-initial syllables, or a mixture of strong and weak word-initial syllables); and (c) vowel sequence (i.e., whether the strong syllables of the base had all long vowels, all short vowels, or a mixture of long and short vowels). The analyses differed only in which of the six possible pairings of response alternatives were considered. The analysis of the perception of stress rhythm considered the three pairings in which one member was the rhythmic mismatch. The first analysis of the perception of word boundaries considered the two pairings in which one member was an exact match and the other a word boundary mismatch; the second word boundary analysis considered the remaining pair, in which the alternatives were two word boundary mismatches.

Perception of Stress Rhythm

Our initial analysis assessed the degree to which the rhythmic mismatch was chosen when it was presented as an alternative to any of the three rhythmically matching options. Thus, this analysis allows an evaluation of how well subjects perceived the stress rhythm of the masked utterances.

The rhythmic mismatch was strongly rejected. Overall, it was chosen only 18% of the time. Paired with the rhythmic mismatch, the exact match was chosen 88% of the time, the word boundary mismatch with more strong onsets was chosen 77.7% of the time, and the word boundary mismatch with more weak onsets was chosen 80.3% of the time. T tests of the significance of each mean as a deviation from the 50% value which would be hypothesized to occur by chance showed that even the smallest of these preferences (77.7%) is significantly different from chance [f(90) = 9.81, $p < .001$].

The difference between these three mean percentages of preference was significant [F(2, 90) = 3.61, $p < .04$]; post hoc analyses showed that the exact match was preferred to the rhythmic mismatch significantly more often that eitber of the word boundary mismatches was preferred to the rhythmic mismatch, but the two word boundary mismatches did not differ in how often they were preferred to the rhytbmic mismatch.

In this analysis, there was no significant main effect of any of the three other factors (rhythmic pattern, word onset type, and vowel sequence). There was only one significant interaction—between rhythmic pattern and word onset type *[F(2,* 90) = 10.17, $p < .001$]. This was due to the rhythmic mismatch being chosen more often with WSWSWS patterns than with SWSWSW patterns for base utterances with all strong word-initial syllables, but more often with SWSWSW than with WSWSWS patterns for base utterances of the other two onset types. We have no explanation for this interaction.

The analysis of the confidence ratings showed that listeners rejected the rhythmic mismatch with an average confidence rating of 2.29 and chose it with an average confidence rating of 1.72. Excluding those cases where no subject chose a particular alternative, there was a difference of 0.491 between the mean ratings [t(57) = 9.62, $p < .001$]. There was no significant difference as a function of rhythmic pattern, word onset type, or which rhythmic match was being offered in the degree to which the rhythmic mismatch was more confidently rejected than accepted. However, the vowel sequence did have an effect—the degree to which rejections were more confident than acceptances was greater for utterances with all or some long vowels than for utterances with all short vowels [F(2, 57) = 7.17, $p < .01$]. We suggested in the introduction that because utterances with all short vowels were shorter in overall duration, giving subjects less information on which to base their decision, performance might have been poorer on these items. This turned out not to be the ease, but subjects' confidence in making decisions was definitely lower for items with all short vowels.

The analysis of the materials subset matched for consonantal similarity showed the same pattern of results. The rhythmic mismatch was rejected 72.2% of the time [((79) = 8.62, p < .001], Listeners were also significantly more confident in rejecting the rhythmic mismatch than in choosing it [mean difference between the means 0.314; ((67) = 4.27, $p < .001$].

These results showed that our subjects accurately perceived the stress rhythm of these noise-masked utterances and were confident of the accuracy of their perception.

Perception of Word Boundaries

The following analyses excluded the rhythmic mismatch and assessed the choices made when only rhythmically matching options were available. First, we examined what happened when one of the available options was an exact match.

When this was the case, subjects showed a preference for the exact match, choosing it 67.2% of the time, which is significantly more often than chance [t(84) = 7.3, $p <$ .001]. There was no significant difference in this effect as a function of which word boundary mismatch was the other option, or of any of the other variables, either alone or in interaction.

The confidence ratings showed that subjects had much less confidence in their ability to discriminate between these alternatives than they had in choosing between different rhythms. Their average confidence in choosing an exact match was 2.17, in choosing a word boundary mismatch 1.95. Excluding again the cases where a particular alternative was never chosen gave a mean rating difference of 0.164, which was significant [((50) = 2.86, p < .01]. None of the other variables significantly influenced the size of the difference.

Second, we examined what happened when a choice was made between two word boundary mismatches. The word boundary mismatch with more strong word-initial syllables received 58.6% of choices, the mismatch with more weak word-initial syllables 41.4%; this difference was significant [((47) = 2.29, $p < .05$]. Again, there was no significant difference in this effect as a function of any of the other variables, either alone or in interaction.

The confidence ratings this time showed no significant difference in the confidence with which one word boundary mismatch was chosen versus the other. None of the other variables significantly influenced the size of the difference.

Again, the analysis of the materials subset matched for consonantal similarity showed a similar pattern. The exact match was chosen in preference to a word boundary mismatch 62.1% of the time [((39) = 3.1, $p < .005$]. Listeners' confidence in choosing versus rejecting the exact match was, however, not significantly different [mean difference between the means 0.132; ((36) = 1.4, p > .1]. The word boundary mismatch with more strong onsets was chosen 54.6% of the time in comparison to 45.4% choices for the word boundary mismatch with more weak onsets; in this small subset (only 20 pairs in this, the smallest, analysis) this effect did not reach significance [t(19) = .87, p > .1]. Nor was there a significant difference in the confidence with which either alternative was chosen.

These results show that our subjects were sensitive to the placement of word boundaries in the masked utterances. When given the option of choosing an exact match (i.e., a match both in stress rhythm and in word boundary placement) against a match in stress rhythm but not in word boundary placement, they preferred the exact match. Although it appeared from the pattern of choice responses that subjects might prefer mismatching alternatives that were more like the actual utterance over mismatching alternatives that were less like the actual utterance, these effects did not reach significance. There was a tendency to choose alternatives that conformed more closely to the prosodic probabilities of the English language. Of the two word boundary mismatches, the one with more strong word-initial syllables was chosen more often than the one with more weak word-initial syllables. Nonetheless, subjects' sensitivity was not matched by strong confidence in the judgments they were making; phenomenally, the rhythmically matching pairs were more like one another than any rhythmic match was with a rhythmic mismatch.

DISCUSSION

Previous research on the perception of word boundaries in the absence of segmental information has led to differing conclusions. Some researchers (e.g., Nakatani & Schaffer, 1978) have claimed that cues to word boundary location do exist outside segmental structure, particularly in durational variation; other researchers (e.g., Klatt, 1976) have doubted whether listeners could effectively exploit such cues. On the basis of the present study, one may conclude that durational cues to word boundary location do indeed exist, and that listeners are able to make at least some use of them.

We have found clear evidence, first, that listeners are highly sensitive to the stress rhythm of utterances. Using noise-masked speech, we have replicated the result of studies using reiterant speech (Carlson et al., 1972) and hummed speech (Svensson, 1974) by finding that listeners are capable of identifying an utterance's pattern of strong and weak syllables without relying on segmental information. Moreover, listeners know they can do this; not only did they reject rhythmically mismatched options, but they rejected them with high confidence. Stress rhythm is a salient property of spoken English. The greater the difference between strong and weak syllables (for instance, when strong syllables contain long vowels), the more confident listeners were of their rhythmic judgments. Reduction of this difference (strong syllables with only short vowels) led to a reduction in listeners' confidence in their rhythmic judgments.

Second, we have found that listeners can derive some word boundary information from the durational structure

of speech. Although listeners were not as accurate in deriving word boundary information as they were in making rhythmic judgments, and particularly, they had no great confidence in their word boundary choices, they could make correct decisions about the locations of word boundaries in a masked utterance significantly more often than they could by chance. Durational cues to word boundary location may not be salient in listeners' perceptual experience, but they are present and can be used.

Therefore, we have provided support for Nakatani and Schaffer's (1978) hypothesis that durational patterns can provide cues to word boundary location in speech recognition. We feel that this result has considerable implications for our understanding of human speech recognition performance. The problem of word boundary location in the recognition of speech is a very important one. Speech is continuous—there are no pauses in the speech stream to signal the boundaries between lexical units in the way white spaces signal such boundaries in most orthographies. But to understand speech, we have to match the incoming speech stream against stored representations in lexical memory, and because lexical storage capacity is not infinite, such stored representations must be discrete. In most cases the content of our lexicon will be words. Understanding spoken language, therefore, requires us to divide a continuous speech stream into words. Cues to word boundary location will obviously be very helpful in this task.

We suggest that listeners may exploit a variety of sources of information in performing the word boundary location task. Where segmental cues are available, listeners will use these (Quene, 1987a, 1987b). Duration of syllables will provide a further amount of information, as the present study has shown. Finally, listeners are also capable of drawing upon their knowledge of the distributional characteristics of their language in order to make best bets about where word boundaries are most likely to occur. Butterfield and Cutler (1988) presented the same 48 base utterances used in the present study to a new group of listeners; the utterances were presented minimally above each listener's individually estimated threshold of speech reception. The listeners were asked to write down what they thought each utterance had been. The word boundary location errors these listeners made showed a significant tendency towards the distributional skew characteristic of English: listeners inserted word boundaries that had not been there in the base utterance more often before strong than before weak syllables, and they deleted the base's word boundaries more often before weak than before strong syllables. In the present study, we also found a tendency for listeners choosing between two word boundary mismatches to choose the one with the greater number of strong initial syllables. Thus, listeners also seem able to call upon prosodic probabilities of their language in attempting to solve the word boundary location problem. Like many other aspects of speech recognition, locating word boundaries is a complex task, in the performance of which listeners exploit a number of different sources of information simultaneously.

ACKNOWLEDGMENTS

This research was supported by a grant (MMI 069) from the Alvey Directorate, U.K. to Cambridge University, the Medical Research Council and STC Technology Ltd. We thank Bill Barry, Paul Cosgrove, and John Holdsworth for technical assistance, and the reviewers of JSHR for helpful comments on the text. Mary R. Smith is now at Bellcore, Piscataway, NJ.

REFERENCES

BOLINGER, D. L., & GERSTMAN, L. J. (1957). Disjunctive as a cue to constructs. *Word, 13,* 246-255.

BOND, Z. S., & GARNES, S. (1980). Misperceptions of fluent speech. In R. Cole (Ed.), *Perception and production of fluent speech* (pp. 115-132). Hillsdale, NJ: Erlbaum.

BOUWHUIS, D., & DE ROOIJ, J. J. (1977). Vowel length and the perception of prosodic boundaries. *IPO Annual Progress Report, 12,* 63-68.

BROWMAN, C. P. (1978). Tip of the tongue and slip of the ear: Implications for language processing. *UCLA Working Papers in Phonetics, 42.*

BUTTERFIELD, S., & CUTLER, A. (1988). Segmentation errors by human listeners: Evidence for a prosodic segmentation strategy. *Proceedings of SPEECH '88, Seventh Symposium of the Federation of Acoustic Societies of Europe, 3,* 827-833.

CARLSON, R., GRANSTROM, B., LINDBLOM, B.,& RAPP, K. (1972). Some timing and fundamental frequency characteristics of Swedish sentences: Data, rules and a perceptual evaluation. *Speech Transmission Laboratory (Stockholm): Quarterly Progress and Status Report, 4,* 11-19.

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

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

ERBER, N. P. (1971). Auditory detection of spondaic words in wideband noise by adults with normal hearing and by children with profound hearing loss. *Journal of Speech and Hearing Research, 14,* 372-381.

GRDING, E. (1967). *Internal juncture in Swedish.* (Travaux de rinstitut de Phonetique de Lund, 6.) Lund: Gleerup.

HARRIS, M. S., & UMEDA, N. (1974). Effect of speaking mode on temporal factors in speech .*Journal of the Acoustical Society of America, 56,* 1016-1018.

KLATT, D. H. (1974). The duration of [s] in English words. *Journal of Speech and Hearing Research, 17,* 51-63.

KLATT, D. H. (1975). Vowel lengthing is syntactically determined in a connected discourse. *Journal of Phonetics, 3,* 129-140.

KLATT, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America,* 59, 1208-1221.

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

LEHISTE, I. (1964). Juncture. *Proceedings of the Fifth International Congress of Phonetic Sciences,* 172-200.

LEHISTE, I., OLIVE, J. P., & STREETER, L. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America, 60,* 1199-1202.

LIEBERMAN, P. (1965). On the acoustic basis of the perception of intonation by linguists. *Word, 21,* 40-54.

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

NAKATANI, L. H., & SCHAFFER, J. A. (1978). Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America, 63,* 234-245.

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

PICKETT, J. M. (1957). Perception of vowels heard in noises of various spectra. *Journal of the Acoustical Society of America, 29,* 613-620.

QUENE, H. (1985). Word boundary perception in fluent speech: A listening experiment. *Progress Report, Institute of Phonetics, Utrecht, 10,* 69-85.

QTJENE, H. (1987a). Relative perceptual relevance of two word boundary markers. *Progress Report, Institute of Phonetics, Utrecht, 13,* 1-7.

QTJENE, H. (1987b). Perceptual relevance of acoustical word boundary markers. *Proceedings of the Eleventh International Congress of Phonetic Sciences, 6,* 79-82.

SCOTT, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America, 71,* 996-1007.

SHIMIZU, K., & DANTSUJI, M. (1980). A study on perception of internal juncture in Japanese. *Studia Phonologica, 14,* 1—15.

STREETER, L. A. (1978). Acoustic determinants of phrase boundary location. *Journal of the Acoustical Society of America, 64,* 1582-1592.

SVENSSON, S. G. (1971). A preliminary study of the role of prosodic parameters in speech perception. *Speech Transmission Laboratory (Stockholm): Quarterly Progress and Status Report, 2-3,* 24-42.

SVENSSON, S. G. (1974). *Prosody and grammar in speech perception.* (Monographs from the Institute of Linguistics, University of Stockholm, 2.) Stockholm: University of Stockholm, Institute of Linguistics.

UMEDA, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America, 58,* 434-445.

UMEDA, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America, 61,* 846-858.

ZWANENBURG, W., OUWENEEL, G. R. E., & LEVELT, W. J. M. (1977). La frontiere du mot en francais. *Studies in Language, 1,* 209-221.

Received November 14, 1988
Accepted March 4, 1989

Requests for reprints should be sent to Anne Cutler, MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF U.K.

APPENDIX

EXPERIMENTAL SENTENCES

The 48 utterances that were masked are listed, with their four response alternatives, in the following order: exact match (rhythm and word boundaries), rhythmic mismatch, word boundary mismatch with more strong word-initial syllables, word boundary mismatch with more weak word-initial syllables.

1. rust presents a nuisance
   crushed defence was useless
   the crust seldom improves
   rushing senseless humour
   tough expense misused them
2. collect enough adrift
   connect among assists
   contenders become fit
   a better budget shift
   attend a cousins film
3. eager rooster playing
   fevered stupor waning
   leafleting useful place
   leaks reduced the traces
   need secure campaigning
4. hay begins beneath it
   may impinge between it
   acclaim gives it belief
   Abel's finished cleaning
   stains resist a steaming
5. cadets are just unfit
   except it must submit
   professors instruct wits
   the clever stuff dismissed
   unrest among desists
6. achieve her ways instead
   appease his days ahead
   retreat made him upset
   the cheaper stays in bed
   belief constrained arrests
7. Lou's bereaved disgraced him
   Sue's relief amazed him
   choosing revealed a way
   music's even paces
   Bloom displeased a matron
8. depict a tool discussed
   equip to rule mistrust
   commit fools in disgust
   a pistol duelling thrust
   until confused entrust
9. a rustic settled hill
   a subtle special pill
   trouble as decks tilted
   distrust upsets the frills
   conduct ascents uphill
10. soon police were waiting
    June believed in saving
    losers agree at races
    ruling people waving
    tunes received pertaining
11. within reviewed results
    rescind revues repulsed
    contingent approved cuts
    a blinking lunar pulse
    the singers who entrust
12. readers playing lessons
    creatures making messes
    reach playful pensioners
    leaders' claims expect it
    Pete's dismay protects it
13 the newsmen seemed delayed
    the newer leans away
    a viewer received pay
    the news was seen for days
    assumed extreme in taste
14 jets adjust equipment
    let's construct within them
    sets are corrupt within
    better budget system
    never just convict them
15. making tinsel keyrings
    April's bringing clearings
    make crinkled vehicles
    angels pinned beneath it
    hay begins between it
16. rings amused the sultan
    things confused the culprit
    a drink boosted results
    clinging movements rustle
    pink balloons disgust them
17. the music's even pace
    the tulips' breeding place
    Lou was seemingly late
    reduce the steam today
    includes serene refrains
18. instruct the men confused
    among the ten pursued
    mistrust meant to resume
    a structured metal fuse
    misjudge revenge adduced
19. Tim approved results of
    hymns include among them
    extinct roots are among
    blinking lunar pulses
    rings amused the sultan
20. mean baboons detained him
    three cocoons contained it
    breeding resumed in May
    eager bugle playing
    teams removed the staining
21 angels pinned beneath it
    hazel tint intrigued them
    contagious in thin bees
    famous printing needles
    age within agreement

22. leaders' claims expect it
legions' aims direct him
allegiance in aims wrecked
lethal crates of pepper
reed remains effective
23. trusting tender viewers
rusting slender skewers
brush tended ruminants
rusted vents preclude it
rust unchecked removes it
24. they're making wrinkled jeans
and taking sprinkled teams
made in pink regency
display of ginger leaves
debate unskilled beliefs
25. machines create duress
ravines relate distress
completion awaits rest
and cleaning table sets
the leaders aimed ahead
26. dusty senseless drilling
custom seldom willing
customers mention frills
Doug's suspense was thrilling
sons expect enlistment
27. never just convict them
nectar judged unfitting
connected such big ones
better touch her mitten
red predicts revision
28. music's even paces
Susan's peevish faces
usages leaving space
soon police were waiting
Lou's bereaved disgraced him
29. the blinking lunar pulse
a thinking ruler sulks
links were soon multiple
begin the useless stuff
within reviewed results
30. sons expect enlistment
tuns protect enriched ones
muddles extend until
judges sentenced Richmond
Butler's sense eclipsed them
31. the eastern news remained
the teachers soon explained
an easy review game
to teach a student ways
repeat disputes in names
32. a better budget shift
the beggar's rubber skiff
getting the mud shifted
direct among the mist
collect enough adrift
33. distrust pretend balloons
disrupt intense disputes
instructing amends tunes
the trusting slender loons
the butler left bemused
34. leaks reduced the traces
leaps produced in places
leasing removed a stage
deepened prudent trading
peaks askew forgave it
35. blinking lunar pulses
thinking doing puzzles
trim gloomy governments
bigger views confront her
Tim approved results of
36. Butler's sense eclipsed them
buckled tents assist them
instructors sent lit ones
sculptured seven pigeons
rough condensed description
37. the hunters went fulfilled
the numbers meant unskilled
summer extends filming
the hunger sent a chill
results are best instilled
38. rust unchecked removes it
crushed cement improves it
mistrust wrecks a cartoon
trust a checkered schooner
rust presents a nuisance
39. debates are grim relief
estates for trim elite
disgrace brings in defeat
a bailiffs timbered eave
obey within regimes
40. includes serene refrains
produced foreseen restraints
school has repealed delays
the music's even pace
amused between the strains
41. and cleaning Mabel's pets
the leaning cable's best
she had made pedestals
receive a later rent
machines create distress
42. ornate distinct machines
arrange succinct decrees
orations depict scenes
they're making wrinkled jeans
a neighbour sings relieved
43. Pete's display corrects it
meets delayed collections
repeat eight with respect
readers playing lessons
please display the texture
44. the trusting slender loons
in plushly rendered rooms
sludge that went luminous
rebuff his censored views
distrust pretend buffoons
45. between secure campaigns
cuisine assured complaints
keyed with renewed disdain
the eager rooster played
at least withdrew today
46. better budget system
pepper buttered biscuit
benefits covered risks
bets corrupt her sister
jets adjust equipment
47. an eager rooster played
the recent suitor stayed
each a dutiful mate
proceed to move away
between secure campaigns
48. conduct ascents uphill
adjust attempts until
combustion prevents spills
a rustic settled hill
robust as seven mills