On the Dynamic Use of Prosody in Speech Perception*  

C. J. Darwin†

ABSTRACT

Two roles that prosodic variables might play in the perception of speech are reviewed and two experiments described. Prosody is seen on the one hand as helping to direct attention to a particular speaker and to the potentially most informative parts of his speech. It is also seen as modifying the hypotheses a listener might entertain about a sentence as he listens to it. The use of models of perception taken from natural language understanding programs is suggested as a way to begin modeling this second role of prosody.

There is little doubt that prosodic variables make a significant contribution to the intelligibility of speech. We lack, however, a suitable framework for modeling, in perception, the relation between prosodic variables and segmental information. Part of the reason for this is that many very different types of information can be conveyed by prosodic variables, and it would be a brave soul who denied any of them a role in intelligibility. The segmental distinctions carried by normally prosodic dimensions, such as change in pitch as a weak cue to voicing (Fujimura, 1961; Haggard, Ambler, and Callow, 1970) and lexical distinctions carried by stress (e.g., see Fry, 1970), are not particularly difficult to integrate into a scheme for the perception of lexical items from segmental cues. But these aspects of prosody exclude perhaps the bulk of the contribution that prosody makes to the intelligibility of speech. Sentences made up of concatenated words spoken in isolation are less intelligible than those spoken fluently (Stowe and Hampton, 1961; Abrams and Bever, 1969), a finding that is almost paradoxical when one considers (Huggins, 1972) that words excised from fluent speech are less intelligible than the same words spoken in isolation (Lieberman, 1963). The resolution of this paradox can be achieved in part by

---

*This paper was prepared for the Symposium on Dynamic Aspects of Speech Perception, held at the Instituut for Perceptie Onderzoek, Eindhoven, 4-6 August 1975. It is to be published in Structure and Process in Speech Perception, ed. by A. Cohen and S. G. Nooteboom. (Berlin: Springer-Verlag).

†Also University of Connecticut, Storrs; on leave of absence from University of Sussex, Brighton, England.

Acknowledgment: The first experiment reported here was carried out at the University of Sussex; the second, at the University of Connecticut. The work was supported in part by a grant from the British S.R.C.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]
knowing that coarticulation is no respecter of word boundaries, but it is likely, if not compelling, that the prosodic information present in fluent speech helps perception of the fluent utterance. When coarticulatory artifacts are made less likely, prosodic perturbations less drastic than word splicing still influence intelligibility. When sentences sharing a common five-word portion, but having major syntactic boundaries at different places within this common string, are cross-spliced so that an intonation contour inappropriate to the sentence's syntactic structure is heard, the intelligibility of these mosaic sentences is lower than those with normal intonation (Wingfield and Klein, 1971; Wingfield, 1975). This technique of cross-splicing introduces two perturbations into the prosody; first it produces an abrupt change in both pitch and rhythm at the splice points, and second it gives an inappropriate placement of the intonational cues to the syntactic boundary. While both of these changes may contribute to the reduced intelligibility of the mosaic sentences, the latter certainly has a specific effect since subjects' transcriptions of these sentences include ones with the syntactic boundary at the point suggested by the intonation contour. This result can be compared with the outcome of other experiments that have shown that subjects can extract considerable information about the stress pattern of speech even when there is virtually no segmental information present, either by virtue of its being hummed (Svensson, 1974) or spectrally rotated (Blesser, 1969). If indeed information that is a potentially useful indicator of syntactic structure can be extracted independently of segmental information, do we need to postulate any interaction between segmental and prosodic information while the sentence is being perceived? Perhaps Wingfield and Klein's (1971) subjects simply reinterpreted the sentences they heard in the light of an independently perceived intonation contour. Not very dynamic, but certainly easy to model!

Life would be more interesting and speech perhaps easier to perceive if prosody played a more dynamic role. This paper aims to review the evidence that bears on a potentially dynamic role of prosody in the perception of speech and to present two new fragments of data that contribute to this discussion. The review will be based on two putative roles of prosodic variables: the role of rhythm and pitch contour in both allowing temporal prediction and giving continuity to an attended channel, and the role of prosodic variables in delimiting higher syntactic structures.

TEMPORAL PATTERNING, PREDICTION, AND ATTENTION

Reaction-time techniques provide a useful way of studying the ongoing process of sentence perception. The most extensively used technique, the phoneme-monitoring task, was introduced by Foss and Lynch (1969). Here, subjects, while listening to a sentence that they subsequently must recall, have to press a key whenever they hear a word beginning with a particular phoneme (usually /b/). Changes in this reaction time have been found to depend on a number of syntactic and semantic variables. Reaction time is longer when the target word occurs in a self-embedded than in a right-branching sentence (Foss and Lynch, 1969), in sentences with deleted rather than intact relative pronouns (Hakes and Cairns, 1970; Hakes and Foss, 1970), in sentences with reduced rather than intact complements (Hakes, 1972), and after high rather than low frequency words (Foss, 1969; at least for adjectives of different frequencies, Cairns and Foss, 1971). These differences have been interpreted as reflecting changes in the processing load imposed by the sentence (cf. Aaronson, 1968) and should not be thought of as reflecting directly the extraction of segmental information. The conscious
access to the phonemic level that this task requires is made after (or at least is influenced by) decisions about what the word containing the target is. Reaction times are shorter if the target starts a word rather than a nonword (Rubin, Turvey, and Van Gelder, in press; see also Foss and Swinney, 1973; Treisman and Tuxworth, 1974, for discussion of related points).

The phoneme monitoring task has been used in two contexts that relate to the use of prosodic information. An experiment by Cutler and Foss (1973) provides a convenient starting point. Following an earlier finding that reaction time to an initial stop consonant was faster when it appeared at the beginning of a content than of a function word, they showed that this difference was attributable to differences in the stress with which function and content words are normally spoken. When stress and lexical category were varied independently, the faster reaction time accompanied the stressed word. Although unstressed function words gave slower reaction times than unstressed content words, they felt this could be owing to different degrees of stress.

This experiment alone can support a number of interpretations: (1) stressed words may be articulated more precisely so that there is more segmental information available, (2) stressed words may be processed more efficiently on account of their intrinsic stress, (3) stressed words may be processed more efficiently through the preceding stress pattern, suggesting that they will be stressed. This last hypothesis is compatible with recent comments on the use of rhythm by Martin (1972).

The first two interpretations attribute the faster reaction time for stressed words to factors intrinsic to the word itself, and can now be dismissed with some confidence. Shields, McHugh, and Martin (1974) measured reaction time to nonsense disyllables beginning with a target phoneme. The disyllable was pronounced with the stress either on the first or the second syllable. When the word occurred as part of a fluent sentence, subjects were faster to a target in a stressed syllable than in an unstressed one, provided that the target did not occur too close to the end of the sentence. However, when the same target words were spliced out and presented in isolated list form, the differences between stressed and unstressed syllables disappeared. The curious lack of difference when the targets occur at the end of the sentence may be simply a floor effect, since reaction times decreased throughout the sentence for both types of syllables. This study is complemented by a recent experiment by Cutler (cited in Cutler, 1975) in which target words were spliced into a sentence context, whose prosody suggested that the target word would or would not be stressed. She found that reaction time was faster when subjects expected a stressed word than when they did not. The stress difference thus seems to be independent of the intrinsic stress of a word and depends rather on the preceding prosodic pattern, which allows the subject to anticipate the forthcoming stressed word.

If some form of anticipation is the cause of the stressed-syllable advantage, then we might expect two further effects; some time should be necessary for subjects to get the rhythm of the utterance on which prosodic predictions might be based, and local disturbance of the rhythm should upset the stressed-syllable advantage. There is evidence for both. Aaronson (1968) asked her subjects to monitor a list of digits, spoken at a constant rate, for the occurrence of a target digit. There was a decrease of about 100-msec reaction time to the target over the first three serial positions. This occurred whether subjects subsequently had to recall the list of digits or not. Reaction times remained
steady in subsequent serial positions for subjects who had only to monitor the lists, but they increased after about the third item until the end of the list for subjects who had to recall the lists as well as monitor. We will return to this latter finding below; for the present it is enough to note that the initial decrease in reaction time may well reflect subjects' accommodating to the rhythm of the list. Cutler (1975) showed that local temporal disturbances influence phoneme-monitoring reaction times. She found that inserting a quarter-second period of silence before a pair of monosyllabic words embedded in a fluently spoken sentence influenced the reaction time to a phoneme target at the beginning of either of the words. The reaction time was slowed to the second word, which was initially spoken with stress, and quickened to the unstressed first word. Cutler failed to find any advantage for the stressed word over the unstressed when the silent interval was absent; this may possibly be attributable to the target's position in the sentence (cf. Shields et al., 1974).

The picture emerging from these studies, then, is of subjects' attention being allocated preferentially toward the stressed syllables in a sentence (Shields et al., 1974). This is made easier by the timing of speech (at least in a stress-timed language) being determined by stressed syllables. Allen (1972), for example, showed that subjects can tap with less variability to stressed syllables than to unstressed (even when the sentence in which they were embedded was heard repeatedly), and Huggins (1972) finds subjects to be more tolerant of timing distortions that preserve stressed-syllable separations. As Cutler and Foss (1973) point out, allowing processing to be directed toward the stressed parts of the sentence allows the focus of the speaker's sentence to control the listener's perception. However, we might note in passing that it is not yet clear how much the results of these experiments depend on using a temporal response measure. The task of targeting for a phoneme might cause an undue advantage to those portions of the speech stream that are temporally predictable. It would strengthen the case if other measures, such as perhaps the detectability of feature substitutions, showed a similar increase in stressed syllables. Some support comes from an experiment by Dooling (1974). He found that subjects' immediate recall of sentences heard in noise was improved if they had previously been exposed to sentences with a similar stress pattern. His results also showed that repetition of stress pattern was much more effective at improving performance than was repetition of syntactic structure, when these two variables were independently manipulated.

Although Martin's (1972) theorizing emphasizes the predictive nature of rhythm, others have shown that pitch contour can play a similar role. The phenomenon of "primary auditory stream segregation" (Bregman and Campbell, 1971) illustrates this, and again, like Martin's work, embraces both speech and music. A random sequence of six notes (three high, three low), when played rapidly, will perceptually segment into a high and a low tune, despite lack of any greater rhythmic cohesion within than between tones. The analogy with speech perhaps lies both in the use of frequency continuity of formants to help in their tracking (cf. Dorman, Cutting, and Raphael, 1975) and in the use of continuity of pitch to help in attending to one voice against competing sounds. This last point has been pursued in an experiment carried out in 1973 by Darwin and Davina Simmonds. While this experiment was motivated by the Bregman and Campbell (1971) findings, it does not in fact distinguish between rhythm and pitch, but looks at the influence of prosodic variables in general on the ability to attend to a particular speech source.
EXPERIMENT I

The technique of shadowing, although complex for the subject, provides a valuable way of studying the process of selective attention to a particular speaker. Treisman (1960) asked subjects to shadow a passage of continuous speech led to one ear and to ignore a similar passage led to the other. At some point during the passage the two channels were switched so that the passage that had been shadowed continued in the other ear. She found that subjects occasionally gave, as part of their shadowing response, words that had in fact occurred immediately after the switch on the ear they had been instructed to ignore, and that these intrusions were more common the more redundant the passages. Inasmuch as this effect has any sensory basis, we can ask whether the tendency for subjects to shadow words from the unattended ear is being determined by some semantic/syntactic priming, as Treisman maintained, or rather whether the continuity in rhythm and pitch across the ears at the switch point is sufficient to cause a momentary change in the ear from which the attended auditory input is drawn.

Our experiment was basically a repetition of Treisman's but with independent manipulation of prosodic and semantic factors. Pairs of passages of about 50 words each were selected from short stories by H. E. Bates. From each pair of passages, four recordings were made by the same female speaker. Two were of the original passages, and two were made by reading the first part of one passage followed smoothly by the second part of the other. The switch point between the passages was later than halfway through each passage and was always prior to a word beginning with a stop consonant (to facilitate subsequent splicing) but was otherwise placed at random (although never at a major clause or sentence boundary). From these four original recordings, four different dichotic conditions were made: a normal condition in which the two original passages were paired together (aligned to give simultaneity of the stop closures); a semantic change condition, made by pairing together the other two original recordings; an intonation change condition, made by switching the latter pair of passages after the stop closure; and a condition in which both semantics and intonation changed, made by switching the two original passages after the stop closure. The first and last of these conditions repeat Treisman's (1960) conditions, the other two vary independently semantic and intonational continuity on the attended ear. All four conditions were made up with the same number of splicings and rerecordings to prevent the preparation of the tapes selectively introducing artifacts into the experiment.

Each of 14 subjects was instructed to shadow the passage on one ear with as little lag as possible between hearing the speech and saying it. They were told to try not to chunk the speech into phrases before speaking it, and were given five practice trials on similar dichotic passages that did not have any switches. They then took eight different dichotic passages with each of the four experimental conditions appearing twice in a counterbalanced order.

Two types of errors are distinguished in the results. A particular trial is classed as having an omission error if the subject misses at least two words over the break point, and it is classed as an intrusion error if the subject shadowed any words from the unattended ear over the break point. Table 1 shows the distribution of errors over the four different types of break.
TABLE 1: Percentage of trials on which omission or intrusion errors occurred for different types of discontinuity and for subjects who either shadowed continuously or "chunked" their responses.

<table>
<thead>
<tr>
<th>Type of discontinuity on shadowed ear</th>
<th>No break</th>
<th>Intonation</th>
<th>Semantic</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chunked (5 Ss)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrusions</td>
<td>0</td>
<td>20.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Omissions</td>
<td>10.0</td>
<td>40.0</td>
<td>90.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Continuous (9 Ss)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrusions</td>
<td>0</td>
<td>77.8</td>
<td>5.6</td>
<td>55.6</td>
</tr>
<tr>
<td>Omissions</td>
<td>0</td>
<td>11.1</td>
<td>55.6</td>
<td>22.2</td>
</tr>
<tr>
<td>TOTAL (14 Ss)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrusions</td>
<td>0</td>
<td>57.1</td>
<td>3.6</td>
<td>35.7</td>
</tr>
<tr>
<td>Omissions</td>
<td>3.6</td>
<td>21.4</td>
<td>67.8</td>
<td>32.1</td>
</tr>
</tbody>
</table>

In the results for all the subjects combined, there is a significant change in the number of intrusion and omission errors across the three experimental conditions (p < .01 and < .05, respectively). This variation is due to there being significantly more intrusion errors in both the conditions with intonation changed than when only the semantics changed (p < .05), and through there being significantly more omission errors in the condition with the semantics alone changed than in the other two experimental conditions (p < .01). A closer look at the subjects responsible for the various types of errors showed, in addition, that very few intrusion errors were made by the five subjects, who despite their instructions, chunked their shadowing responses. Omission errors, by contrast, were distributed more evenly between the subjects, but tended to be made more by the subjects who chunked their responses. In other words, for subjects who shadow continuously, an abrupt change of the intonation contour between the ears causes intrusion errors to occur, while an abrupt change in the semantic content of the message, but without any switch in the intonation contour, gives omission errors. Subjects who chunk their shadowing responses likewise produce a large number of omission errors on semantic-switch trials, but show a different pattern from the subjects who shadow continuously on the intonation-switch trials, giving more omission than intrusion errors.

Clearly, the subjects who chunk their responses are much more sensitive to the semantic constraints on the message they are shadowing, in that they show omission errors when there is a semantic discontinuity on the attended ear. Nevertheless they do not show any tendency for intrusions to occur from the other ear when only the semantic information switches ears. Although the subjects who shadow continuously are disrupted less by the semantic discontinuity than their chunking colleagues, they too show no tendency for intrusions to occur from the semantically appropriate unattended ear. Rather than assuming that intrusion errors are caused by some semantic priming of a contextually likely word, it seems, at least in this experiment, that they are occurring only when there is a continuity of intonation across the two ears. For a brief time after a switch in intonation has occurred, the continuity of the intonation contour overrides the ear of entry as a criterion for selecting the speech that needs to be attended. If subjects are shadowing continuously, this leads to intrusion errors from the ostensibly unattended ear. But if the subject is chunking his response, he has time between hearing the potentially intrusive words and making his response to omit the words from the inappropriate ear. This, however, causes
omission errors, particularly when there is no semantic continuity on the attended ear to help him retrieve the words that occurred there immediately after the break.

This finding, that prosody helps a listener to attend to a particular speaker, complements the work reviewed earlier on the predictive use of prosody in anticipating stress. Both types of experiment emphasize the role prosody plays in controlling which parts of the speech stream are attended to, whether the selection is made temporally or spatially.

PROSODY AND THE STRUCTURES OF SPEECH

As well as allowing anticipation of the speech stream, prosodic factors doubtless also play a role in segmenting speech into higher-order structures. One rather simple way in which prosody delimits complex structures in speech is through pauses. Goldman-Eisler (1968) claims that even at its most fluent, over two-thirds of speech consists of utterances of six words or less. Pauses virtually never appear within a word, and their length is determined by the type of syntactic juncture that they occur in. They tend to be longer at the end of sentences than at the beginning of subordinate clauses, and before a subordinate clause their length is determined by the type of clause (Goldman-Eisler, 1972). Pauses have been used as experimental variables in studying memory for lists of items that do not naturally fall into larger structures, and we might get some insight into their use in natural language, by looking for a moment at their effect in unnatural situations.

A brief period of silence after an item in the middle of a list of digits provides an opportunity for rehearsal of the preceding group of digits (Kahneman, Onuska, and Wolman, 1968; Ryan, 1969; Kahneman and Wright, 1971), as well as allowing the additional information from an auditory memory to be utilized (Crowder and Morton, 1969). Merely indicating by instructions or interposed tones in a regular list how that list should be grouped gives substantially poorer overall performance and a reduced scalloping of the serial recall curve (Ryan, 1969). This is perhaps because less capacity can be allocated to rehearsal when there are competing perceptual demands. As well as delimiting groups of items for rehearsal, pauses also serve to delimit larger structures for coding in long-term memory. The cumulative learning that occurs when supraspan strings of items are repeated occasionally throughout a short-term memory experiment (Hebb, 1961) is reduced when these strings are grouped differently by pauses (Bower and Winzenz, 1969), but this reduction only seems to occur when the items that compose the string are amenable to some higher-level recoding (Laughery and Spector, 1972).

Coming closer to real speech, reading a list of nonsense syllables in a natural intonation can increase the number of syllables remembered immediately after, provided the syllables contain additional bound morpheme clues to the intended syntactic structure (O'Connell, Turner, and Onuska, 1968). Neither intonation, nor the presence of bound morphemes alone is sufficient to increase the number of syllables recalled. Incidentally, both these additional cues also seem to be necessary to procure a right-ear advantage (Zurif and Sait, 1970; Zurif and Mendelsohn, 1972; Zurif, 1974).

It is tempting to extrapolate from these experiments to a view of clause or sentence perception that delays syntactic processing of clausal unit until a
clause boundary has been reached, so that the end of the clause is, like the pause in a string of digits, a time of great mental effort (cf. Kahneman et al., 1968; Wright and Kahneman, 1971), but there is some evidence against this. As we remarked earlier, Aaronson (1968) found that reaction time to a target digit in a string that the subject had to remember increased toward the end of the list, reflecting the increased memory load. By contrast, there is very little evidence that a similar increase in reaction time occurs for a phoneme target in a sentence. Foss and Lynch (1969) did find such an increase, but all relevant subsequent studies have found a decrease in reaction time throughout the sentence (Foss, 1969; Hakes and Foss, 1970; Shields et al., 1974). Similarly, reaction time to a click played at the end of a clause decreases with increasing length of the clause, while reaction time to a click presented between clauses shows no consistent increase with length of the first constituent (Abrams and Bever, 1969). Nor is reaction time to a click consistently faster at the beginning of the second clause than in the clause break (Abrams and Bever, 1969; but see Bever and Hurtig, 1975). It is clear then that the constituent words of a clause are not being held in memory like so many digits. Something much more dynamic is happening. But what? Perhaps the best type of model to use here would be one based on Winograd's (1972) integrated language understanding system. In this program both syntactic and semantic information is used, as each successive word in a sentence is encountered, to construct procedures that can subsequently be used to take action on the sentence. The end of a clause does not in this scheme imply a particularly energetic syntactic activity on the part of the computer, provided the syntactic organization that it had presumed during perception of the preceding word string is comparable with the string's ending. Incompatibility requires backtracking, but the program's use of semantic constraints for eliminating alternative syntactic structures as it goes along reduces the likelihood of this being necessary. If these sorts of syntactic and semantic constraints are also being used dynamically in natural perception, then we might suppose that prosody is used in a similarly dynamic and interactive way to guide the search toward an appropriate syntactic organization of the sentence.

Some of the results of an experiment run recently by John Capitman in collaboration with myself and Susan Brady bear a little on these issues. Our experiment was designed primarily to be a replication of Wingfield and Klein's (1971) interesting finding that inappropriate intonation contours impair immediate recall of a sentence.

**EXPERIMENT II**

The sentences we used in this experiment were similar to those used by Wingfield and Klein (1971). Each sentence that the subjects heard was one member of a group of four sentences derived from pairs of sentences sharing a common string of words. Two of the sentences in the group of four were the originals with appropriate intonation, the other two were generated by cross-splicing the common string of words between these two sentences. This string always started with a stop consonant to facilitate silent splicing, and continued to the end of the sentence to reduce the rhythmic discontinuity that splicing can introduce. The sentences were between 10 and 14 words long and the common string was between 6 and 10 words long. Unlike the Wingfield and Klein material, the major syntactic boundary in each sentence was both preceded and followed by at least two words of shared material, this ensured that the intonationally suggested boundary in the mosaic sentences generated by cross-splicing was also
preceded and followed by at least two words of shared material. As in the Wingfield and Klein experiments the sentences were presented monaurally and the ear of presentation was switched within the sentence. This switch was always made during a stop closure to prevent extraneous clicks. The switch point always occurred within the shared material and was never at the intonationally suggested boundary, but was otherwise unconstrained. The recordings were made on the Haskins Laboratories pulse-code-modulation (PCM) facility.

Each of 24 subjects heard one sentence of the four generated from each of the 11 pairs in an order that counterbalanced conditions (normal versus cross-spliced). Subjects were instructed to write down each sentence as soon as they had heard it. In order to make the task more difficult, they were told to listen for the occurrence of stop consonants and to circle these in their answer. Their performance on this task was ignored.

On average, 55.5 percent of the normally intonated (unspliced) sentences were recalled correctly, while only 44.5 percent of the cross-spliced sentences were recalled correctly. In examining the errors that subjects made, four types of error are distinguished:

1. **Omission errors**: Each word omitted, whether or not this omission changed any other aspect of the sentence, was scored as an omission error.

2. **Lexical errors**: Each changed word in a string that did not affect the primary grammatical relations (the truth conditions of the sentence) was scored as a lexical error. Changes in tense, mood, and aspect were also counted as lexical errors.

3. **Syntactic errors**: Each change in any segment of a string that resulted in a change in the grammatical relations or the truth conditions of the sentence was scored as a syntactic error (except for errors of tense, mood, and aspect). No attempt was made to distinguish those changes that seemed to conform to the intonation pattern from those that did not. One other error type was defined that depended on the position of the error but not on its type:

4. **IB2 errors**: [This score applied only to the cross-spliced sentences.] Each error of any type occurring within one word on either side of the intonationally suggested boundary was scored as an IB2 error.

Summing together all the first three types of errors (omission, lexical, syntactic), Capitman found significantly more errors on the cross-spliced than on the normally intonated sentences. This difference is significant by a Wilcoxon test both across subjects (p<.01) and across sentences (p<.02). Moreover, this increase in errors is due to an increase in the omission (p<.02 by subject; p<.05 by sentence) and syntactic errors (p<.01 by both subject and sentence), but not to an increase in lexical errors. An exception is seen in three of the sentence pairs that show a decrease in syntactic errors in the crossed-intonation condition, though still an increase in omission errors. All of these sentences have an "if...then..." construction, which may impose greater
syntactic or semantic constraints on the subjects' responses, causing them to give up on their response rather than change the truth conditions of the sentence.

Although these general results confirm Wingfield and Klein's (1971) findings on intelligibility with a different set of sentences, they do not get us much nearer to the question of how the inappropriate intonation contour influences recall. Inspection of the TIB2 errors does bear on this question. When listening to the crossed-intonation sentences, subjects make more errors within one word on either side of the intonationally suggested boundary when this boundary precedes the major syntactic boundary than when it follows it (p<.025, by subject; p<.05 by sentence). This is not owing to subjects' making more errors earlier in the sentence, since the reverse tends to be the case. This asymmetry in the disruptive effects of an inappropriate intonational cue to a clause boundary is what one might expect on the hypothesis that intonational information is being used dynamically to restrict syntactic hypotheses as the subject listens to the sentence. If prosody were being used solely to restrict alternatives after initial perception of the sentence, it is difficult to see why the relative positions of the syntactic and intonational boundaries should matter. On a more dynamic view, it is possible that the inappropriate intonational boundary leads to backtracking throughout the previous incomplete clause; this clause will be much longer, and so the effects of backtracking will be more disruptive when the incomplete clause is the first one of the sentence. Hence, having the intonational boundary before the syntactic boundary will be more disruptive than having it after it. Looking at errors only in the region of the intonation boundary appears to be a more sensitive measure than looking at errors over the whole sentence, since there did not appear to be any significant difference in errors over the sentence as a whole as a function of these relative locations. This perhaps suggests that backtracking affects perception of the part of the sentence that is being actively perceived when it occurs. But we cannot entirely rule out the possibility that this effect is being caused in part by the rhythmic discontinuity that cross-splicing introduces immediately after the breakpoint. Although these conclusions are extremely speculative and the evidence on which they are based is not strong, we feel that the questions raised by this sort of approach are extremely interesting, and we hope to pursue them in subsequent experiments.

In summary, this paper has tried to draw attention to some of the possible dynamic roles that prosody might play in the perception of speech: the rhythmic and melodic aspects of speech may allow the listener to predict when potentially important speech material will arrive (and perhaps allow him to allocate his processing capacity accordingly); they may also allow him to attend selectively to one voice among many more readily. In addition, prosody undoubtedly plays a role in delimiting higher-order structures in speech, and the suggestion is made here that this is done by dynamically modifying the hypotheses the listener entertains while listening to a sentence. While we neither pretend that this exhausts the uses of prosody, nor believe that the study of isolated sentences is the most appropriate way to study a variable that depends so much on intersentence context (see Smith and Goodenough, 1971, for an interesting example of intonation interacting with context in a perceptual task), we do hope that the issues raised here will help to stimulate work on an unduly neglected area of speech perception.
REFERENCES


Bregman, A. S. and J. Campbell. (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J. Exp. Psychol. 89, 244-249.

Cairns, H. S. and D. J. Foss. (1971) Falsification of the hypothesis that word frequency is a unified variable in sentence processing. J. Verbal Learn. Verbal Behav. 10, 41-43.


Fujimura, O. (1961) Some synthesis experiments on stop consonants in the initial position. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 61, 153-162.


Brain Lang. 1, 391–404.
Zurif, E. B. and M. Mendelsohn. (1972) Hemispheric specialization for the per-
ception of speech sounds: The influence of intonation and structure. 
Neuropsychologia 8, 239–244.