COMPONENTS OF PROSODIC EFFECTS IN SPEECH RECOGNITION

ANNE CUTLER

MRC Applied Psychology Unit
15 Chaucer Rd.
Cambridge CB2 2FF, U.K.

ABSTRACT

Previous research has shown that listeners use the prosodic structure of utterances in a predictive fashion in sentence comprehension, to direct attention to accented words. Acoustically identical words spliced into sentence contexts arc responded to differently if the prosodic structure of the context is varied; when the preceding prosody indicates that the word will he accented, responses are faster than when the preceding prosody is inconsistent with accent occurring on that word. In the present series of experiments speech hybridisation techniques were first used to interchange the timing patterns within pairs of prosodic variants of utterances, independently of the pitch and intensity contours. The time-adjusted utterances could then serve as a basis lor the orthogonal manipulation of the three prosodic dimensions of pilch, intensity and rhythm. The overall pattern of results showed that when listeners use prosody to predict accent location, they do not simply rely on a single prosodic dimension, but exploit the interaction between pitch, intensity and rhythm.

Speakers place accent on the most important words in an utterance. Thus by finding accented words, listeners can efficiently locate the most central parts of a speaker's message. Previous studies have shown that listeners do indeed actively use sentence prosody to tell them where accented words arc going to occur. Cutler [2] produced pairs of sentences varying in prosodic contour. An example is (I):

(I) (a) The couple had quarrelled over a BOOK they had read.
(b) The couple had quarrelled over a book they hadn't even READ.

Upper case represents sentence accent. In (Ia) the main sentence accent falls on *book,* in (Ib) on *read.* These sentences were used as materials in a phoneme-monitoring experiment, in which listeners are asked to respond as quickly as possible to the presence of a specified word-initial phoneme. In (I), the target phoneme is /b/, so the target-bearing word is *book. T*argets on accented words are responded to faster than targets on unaccented words in this task. In Cutler's experiment, the target-bearing word itself was actually spliced out of both sentence contexts and replaced in each by identical copies of a neutral rendition of the same word. The result of this manipulation was a pair of sentences with acoustical!} identical target-bearing words, which were preceded by identical sequences of words; the only difference between the members of each pair was the prosody applied to the words preceding the target. In one case the prosodic contour in which the target-bearing word occurred was consistent with accent falling upon that word; in the other, it was consistent with the target-bearing word being unaccented. Under these conditions, the 'accented' targets still elicited faster responses than the 'unaccented' targets, and since the only relevant differences between the two sentences in each pair lay in the prosody, Culler concluded that listeners must have used cues in the prosody to direct their attention to the location where sentence accent would fall.

Prosody, however, is not a unitary phenomenon. The separate dimensions of rhythm, pitch and intensity all contribute to the prosodic structure of an utterance. Culler's experiment did not examine *how* listeners were exploiting prosody to predict accent, or whether any one prosodic dimension was more informative than others.

Culler and Darwin [3] subsequently found that removing pilch information - i.e. monotonising the sentences - did not remove the accent effect; in monotonised spliced sentences like (I) the 'accented' targets are still responded to significantly faster than the 'unaccented' targets.

From this, Culler and Darwin concluded that pilch information could not be a necessary component of the accent prediction effect. They speculated that no prosodic dimension might prove necessary for listeners to predict upcoming accents, but variation in any prosodic dimension might prove sufficient.

In the present studies, the three prosodic dimensions of pilch, rhythm and intensity are separately manipulated in an attempt to analyse the accent effect in further detail. Unlike the study by Cutler and Darwin, which simply removed the dimension of pitch by setting it to a single value across each utterance, the present studies investigate the effects of the separate prosodic dimensions when they arc *interchanged* between the two members of a sentence pair. To begin with, using dynamic time-warping techniques in a system developed by Jeffrey Bloom at the Polytechnic of Central London [I], we exchanged the rhythmic patterns within each pair of sentences (for examples like [1], where naturally different contours were produced by having a slight variation in the text at the end of the sentences, the rhythmic patterns were exchanged up to the point at which the two members of the pair diverged). Thus (Ia), for example, was given the rhythm of (Ib) but retained its original pitch and intensity contours; (Ib) had the rhythm of (Ia) but its own pitch and intensity patterns.

In Experiment I, phoneme-monitoring response limes were measured in these rhythmically manipulated sentences, and in the same sentences with intact prosody. The intact sentences were LPC-analysed and resynthesised to control for acoustic effects of resynthesis. The words bearing (he target were acoustically identical in all four sentences belonging to a set such as (I).

There were 20 such sentence sets. Forty listeners, in four groups of ten, took part in the experiment. Each group heard only one sentence from each set, and the two variables of 'accented' versus 'unaccented' targets, and intact versus rhythmically manipulated prosody, were counterbalanced across subject groups.

Subjects were tested individually. Response times, measured from a click (inaudible to the subjects) aligned with target onset, were collected by a microcomputer using programs developed hy Norris [4]. After the experiment subjects were given a short recognition test, and their response times were analysed only if they scored at least two-thirds correct on this test.

The results of this experiment are shown in Fig. I. The intact sentences, in which rhythm, pitch and intensity contours are preserved from the original utterance, show the advantage of 'accented' over 'unaccented' largets which was found in the earlier experiments. This indicates that the resynthesis alone is not interfering with listeners' ability to use prosodic contours to predict the location of acccnl. The diflerence in this condition is significant (Fl(1,3.6) = 21.36, p < .001). In the rhythmically manipulated sentences, however, the advantage of originally accented over originally unaccented largets is less than half as large as the diflerence in the prosodically intact sentences, and n is not statistically significant (F1(1,36) = 3.55, p >.05).

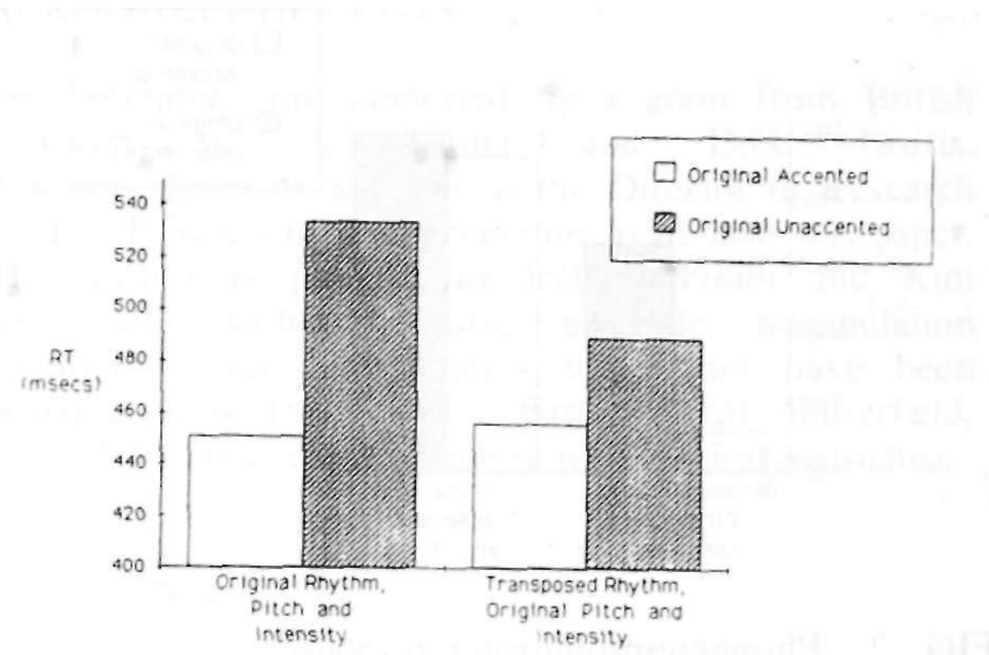

FIG. 1. Phoneme-monitoring response time (msecs.), Experiment I.

This experiment shows that the rhythmic manipulations have severely affected the accent effect. Kach of ihe utterances which had undergone this rhythmic manipulation had an unnatural, indeed a conflicting, prosodic structure - pilch and intensity contours signalled one prosodic pattern while the rhythm signalled another. It is clear that listeners did not base their prosodic processing on one aspect of the prosodic contour alone.

One possible interpretation of this result is that listeners are simultaneously processing all three prosodic dimensions, and that the separate contributions of each prosodic dimension to the predicted accent effect are simply additive. The attenuated, but still positive, effect in the rhythmically manipulated sentences would, on this simple story, be attributable to the combination of positive effects contributed by the pilch anJ intensity contours, set against a negative effect contributed by the rhythmic contour.

This interpretation was tested in Experiment 2. This experiment investigated prosodic manipulations which were the reverse of those in Experiment I. The pitch and intensity contours were transposed between originally accented-target and unaccented-target members of a sentence pair, leaving the rhythmic contour, alone, intact.

This manipulation was possible because the time-warping applied to the sentences in Experiment I produced pitch and intensity contours which, although they preserved the contour shape from the utterance they had originally belonged lo, were aligned with the rhythmic pattern of that utterance's pair. Therefore these contours could simply be transposed onto that pair. These transpositions were realised using prosodic editing routines devised by Kim Silverman.

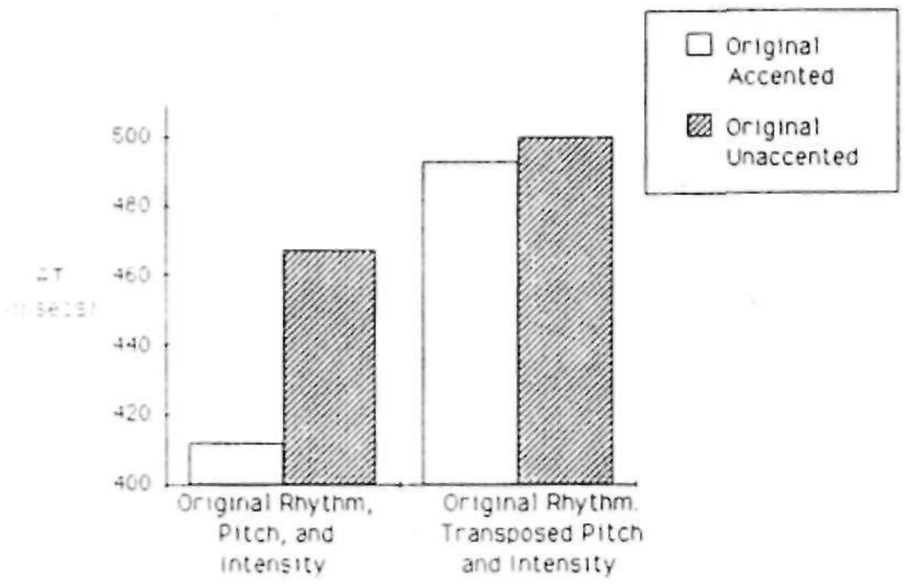

FIG. 2. Phoneme-monitoring response time (msecs.), Experiment 2.

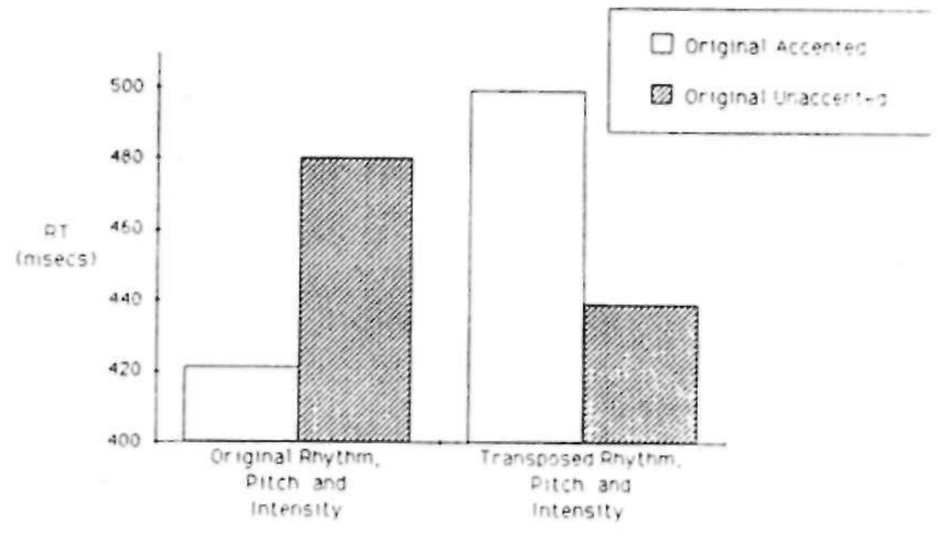

FIG. 3. Phoneme-monitoring response time (msecs.), Experiment 3.

Experiment 2, like Experiment I, included the resynthesised utteranccs with intact prosody; these were compared with the utterances in which of the original prosody only the rhythm was preserved intact, the pitch and intensity contours being transposed between members of a pair. Again, the target-bearing words were acoustically identical in all sentences from any set.

Forty listeners, who had not taken part in Experiment I, were tested; design and procedure were as in Experiment I. The results are shown in Fig. 2.

It can be seen that once again the utterances with intact prosody showed a strong accent effect, i.e. response lime advantage for 'accented' over 'unaccented' targets. This difference was statistical!) significant (FI(I,36) = 6.85, p <.02). In the uiieranccs with transposed pitch and intensil) contours, there was virtually no response time difference between originally accented and originally unaccented targets (FI <1).

The results of this experiment rule out the very simple explanation of Experiment I offered above. Had listeners been simply evaluating all three dimensions of prosody in an additive fashion, we might have expected the reverse of the result found in Experiment I - that is, we might have expected an advantage of originally unaccented targets over originally accented targets of about half the magnitude of the difference in ihe opposite dircciion produced by the prosodically intact uiieranccs. However, the conflicting prosody in this case wiped out any difference in response times as a function of original accent location.

This result raises the possibility that transposition of prosodic contours might itself interfere with listeners' ability to predict accent location by extracting relevant information from the prosody. In order to rule out this possibility, a further experiment was conducted in which all three prosodic dimensions were transposed.

In Experiment 3, the resynthesised utteranccs with intact prosody were again tested, and compared in this case with utterances in which rhythm, pilch and intensity contours had all been transposed between members of a sentence pair. The manipulated utterances in this experiment therefore exhibited the maximum of transposition, in that every utterance had rhythm, pitch *and* intensity contours which had originally been applied to another utterance. However, they exhibited the minimum of prosodic conflict, since rhythm, pilch and intensity contours were always in accord.

As in the previous experiments, the target-bearing words were acoustically identical in all sentences from any set.

Forty listeners, none of whom had taken part in Experiments 1 and 2, were tested. Design and procedure were as in the preceding experiments. Ihe results are shown in Fig. 3.

Once again there was a significant advantage for 'accented' over 'unaccented' targets in the prosodicallv intact sentences (F1(1,36) = 10.38, p <.005). Moreover, there was a significant difference in the reverse dircciion, i.e. a response lime advantage of originally unaccented over originally accented targets, in the prosodically manipulated sentences (FI(I,36) = 6.83, p <.02). That is, when all three components of the prosodic contour signalled that accent would occur at the position where the target occurred, the target was responded lo faster; and this was true whether the consistent prosodv was applied to its original utterance or to its original utterance's pair.

This result allows us to dispose of the suggestion that prosodic transposition might interfere with listeners' prosodic processing. Instead, it is clear that what interfered most strongly with listeners' prosodic processing in the two preceding experiments was prosodic conllici. When one prosodic dimension was in conflict with the other two, listeners were unable to arrive at a consistent interpretation based on prosodic information. One effect of this was that significant accent effects disappeared.

However, the results from the prosodically manipulated conditions in Experiments I and 2, though they were both statistically insignificant, seem to differ. This might suggest that more sensitive experimentation could yet uncover differential contributions to the accent effect on the part of rhythm, pitch and intensity respectively. For the present, though, we may conclude with confidence that listeners' processing of prosody is not simply an additive evaluation of separate dimensions; the interaction between prosodic dimensions is of paramount importance. When the three dimensions rhythm, pitch and intensity agree, listeners exploit them effciently and consistently. When they conflict, this exploitation is significantly impaired.

ACKNOWLEDGEMENTS

This research wass supported by a grant from British Telecom to A. Cutler and D.G. Norris. Acknowledgement is made lo the Director of Research of British Telecom for permission to publish this paper. The author is grateful to Jeffrey Bloom and Kim Silverman, without whose prosodic manipulation techniques these experiments would not have been possible, as well as to Steve Bartram, Sally Butterfield, John Williams and John Culling for technical assistance.

REFERENCES

[I] Bloom, P.J. Use of dynamic programming for automatic synchronization of two similar speech signals. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* 1984, 2.6.1-2.6.4.

[2] Culler, A. Phoneme-monitoring reaction lime as a function of preceding intonation contour. *Perception & Psychophysics, 20,* 1976, 55-60.

[3] Culler, A. & Darwin, C.J. Phoneme-monitoring reaction lime and preceding prosody: Effects of slop closure duration and of fundamental frequency. *Perception & Ps\vhoph\sics, 29,* 1981, 217-224.

[4] Norris, D.G. A computer-based programmable tachistoscopc for non-programmers. *Behavior Research Methods, Instrumentation and Computers, 16,* 1984, 25-27.