THE PERCEIVED RHYTHM OF SPEECH

Andrew Donovan and C.J. Darwin, Laboratory of Experimental Psychology, and Centre for Research in Perception and Cognition, University of Sussex, Brighton, England

Introduction

Attempts to model the duration or rhythm of the segments of connected speech fall into two broad groups (see Fowler, 1977, for a review). On the one hand are those which allow the syntactic structure of the utterance to perturb a segment's ideal duration (e.g. Lindblom & Rapp, 1973; Klatt, 1975) but which recognise no overall rhythmic patterning; on the other hand are those which allow an overall rhythmic pattern to be perturbed by limits on the segmental durations which must be compressed or expanded into that pattern (Abercrombie, 1964; Witten, 1977). This latter family of models has taken for its underlying rhythm a sequence of isochronous beats occurring on adjacent stressed syllables marking out rhythmic units called feet. The choice of an isochronous foot is based partly on linguistic intuition ("English utterances may be considered as being divided by the isochronous beat of the stress pulse into feet of (approximately) equal length". Abercrombie, 1964), and partly on the observable phonetic fact that syllables tend to be shorter, the more there are in a foot (Huggins, 1975; Fowler, 1977). In choosing between these two approaches a crucial question is the status of the isochronous beat. It is clearly not a phonetic fact since a foot with many syllables tends to be longer than one with fewer (Halliday, 1967; Allen, 1975). Is isochrony then a significant linguistic insight, or merely a poetic fiction? Lehiste (1973; 1977) has argued that the discrepancy between the linguistic intuition and the phonetic data may be due to a perceptual illusion. Perhaps we hear speech as more isochronous than it actually is. Indeed such an illusion is precisely what we would expect if perception undid those perturbations required by segmental constraints on an underlying regular rhythm, presenting to the listener the underlying rhythm of the speaker. Evidence for the perceptual reality of isochrony would thus argue for its inclusion in models of speech production.

Experiments

Our experiments extend the earlier observations by Lehiste (1973) and Coleman (1974) on listeners' inability to perceive the rhythm of speech veridically. We have used two tasks, a rhythm matching task and a tapping task. The first two experiments used the rhythm matching task. Subjects adjusted the times between four noise bursts to match the overall rhythm of either a sentence or a control sequence of non-speech sounds. They could listen to the sound whose rhythm they were to match or the adjustable noise burst sequence by pressing one or other of two buttons; thus they could not hear the two stimuli simultaneously but were able to listen to each separately as many times as they liked while making the adjustments.

The sentence used in Experiment 1 was "A bird in the hand is worth two in the bush", synthesized on PAT from parameters derived from real speech. The noise burst sequence that subjects adjusted had four strong bursts corresponding to the four stressed syllables with appropriate intervening weaker bursts representing the unstressed syllables. By adjusting either of three knobs subjects could adjust the interval between adjacent stressed bursts, but the rhythm of the intervening weaker bursts was kept constant, scaled in tempo to the new inter-stress interval. Subjects matched the rhythm of two versions of the sentence; one had the natural pitch contour, the other a monotone. They also matched the rhythm of a control sequence of tones whose onsets were at the same time intervals as the stressed syllables of the sentence. Subjects performed seven matches (the first two of which were not analysed) to each of the three stimuli. The control was always done first followed by the two speech conditions in an order counter-balanced between subjects.

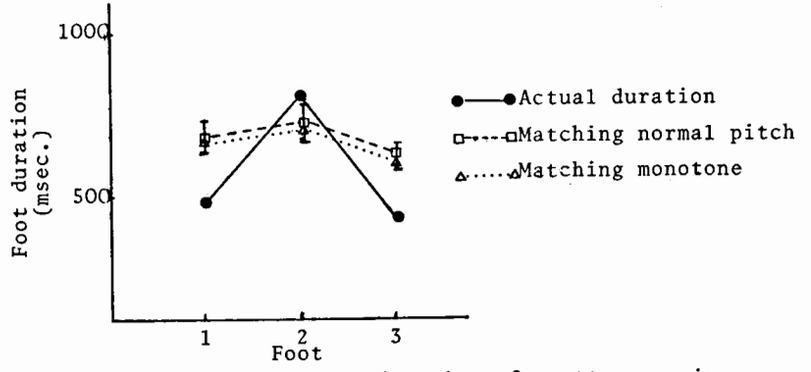

Fig.1. Actual and perceived foot durations for utterance in Experiment 1. (vertical bars represent ± 1S.E. of the mean)

The actual and mean matched durations of the first three feet (that is the intervals between the four stressed syllables) are shown in Figure 1. To test any tendency for the matched durations to be more isochronous than the original utterance the quantity:

$|(1-a_i/a_{i+1})| - |(1-p_i/p_{i+1})|$ where a_i = actual duration of i^{th} foot
p_i = matched " " " "

was calculated for i=1,2. A positive value for this quantity indicates that the perceived durations are more isochronous (over the two feet in question) than the actual durations. Such a tendency towards perceptual isochrony was reliable ($p<.001$) for both foot-ratios when subjects matched the two sentences, but was not found when they matched the control, non-speech rhythm. The natural and the monotone speech are both perceived as more isochronous than they really are, but the non-speech tonal pattern is not.

The second experiment differed from the first as follows:

1) Four sentences of natural (female) speech were used which contained different numbers of syllables in each foot.

2) The stressed syllables in each utterance all began with a stop consonant (/t/) and there were no other occurrences of this sound in the utterance. This made it easier to specify to the subjects where the major stresses fell as, instead of saying match the rhythm of the 'syllable beats' or the 'tapping points', they could be told to hit the /t/'s.

3) The noise-burst sequence was made up of five bursts only; an initial low amplitude burst corresponding to the first, unstressed syllable in each utterance, and four 'stressed' bursts corresponding to the four stressed syllables.

4) Subjects were explicitly encouraged to use a strategy that we had observed in the first experiment, namely repeating the sentence to oneself while listening to the noise bursts. In case subjects' own articulations were more isochronous than the original, recordings were made of each subject speaking each sentence. In fact we found they were not more isochronous and the following results still hold if matched durations are compared with subjects' own productions.

For the four sentences as a whole, four of the eight foot-ratios gave a significant ($p<.01$) tendency towards perceived isochrony, three gave no difference (partly because their foot ratios were actually quite close to unity already) and one gave a significant tendency away from perceptual ischrony. The results from

this deviant sentence and from one of the others are shown in Figure 2. Notice first in the right-hand panel that although subjects' judgements are quite reliable they are massively inaccurate at judging the duration of the middle foot. It is not clear though whether this huge overestimation of the middle foot should be attributed to perceptual isochrony. If it were, then we would not expect the similar, though more variable overestimation of the middle foot found in the left-hand panel for a sentence whose middle foot is already relatively long. Alternative explanations, which could account for the data from all four of the sentences, are that subjects overestimate the length of a foot containing (a) a major syntactic boundary or (b) a tone group boundary. Our third experiment looks at this question.

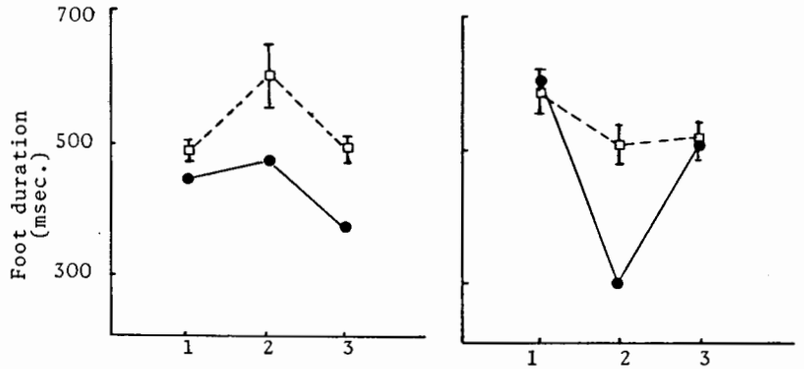

Fig.2. Actual and perceived foot durations for two of the utterances in Experiment 2.

To investigate the possible contribution of intonation to perceived rhythm a change of experimental technique was required. Subjects' imitations of sentences were, as we have seen, no more isochronous than the originals, but they did differ markedly in intonation. To ensure that subjects matched a sentence with the original intonation we asked them to tap in time to a sentence.

This new task differs from that used by Allen (1970) in that subjects tapped to <u>every</u> stressed syllable rather than just to a selected one on each trial. Here we are interested in subjects' perception of the entire rhythmic pattern. Subjects were not explicitly told to tap on the stressed syllables so the fact that they did provides an objective verification of the notion of the

rhythmic foot. The three utterances, which differed in number of tone groups and in syntactic structure, but which had identical foot durations were:

1) //1 Tim's in / Tuscany's / Training / Troops //
2) //1 Tim's in / Tuscany / Training / Troops //
3) //1 Tim's in / Tuscany //1 Training / Troops //

Here, following Halliday (1967), we bound tone groups by double slashes and indicated the type of tone group by a number. Both 2) and 3) contain a major syntactic boundary in the middle foot but in utterance 2) this was not marked by a tone group boundary. 1) and 2) were acoustically identical except that the /s/ of "Tuscany's" was spliced out for sentence 2) and replaced by four additional pitch periods of /ɪ/ and an appropriate amount of silence to maintain the same foot length.

Fifteen subjects were divided into three groups, each group receiving a different order of presentation of the three utterances. Subjects heard each utterance 15 times and were told to start tapping after the third token. Each utterance was preceded by a warning tone 750 msec. from the onset of the utterance. Only the last 10 trials in each condition were analysed. Before each block of 15 trials, subjects heard the utterance three times and were given a context in which the utterances could occur. For example 3) might be the response to the question "Where's Tim and what's he doing?", while the same utterance with one tone group (sentence 2) might be the response to the question "What's Tim doing with the troops in Tuscany?" This was done to ensure that the subjects had a good idea of the syntax and tone group structure of the sentences they were listening to.

The results of this experiment (Figure 3) showed that while the number of tone groups has a distinct effect on perceived rhythm, the syntactic structure does not. In particular we found no tendency towards perceived isochrony in sentence 3, which contained two tone groups, but we did find a significant ($p<.01$) tendency towards perceived isochrony for both foot-ratios in sentence 1 and 2. Sentences 1 and 2 did not differ from each other significantly in this respect, but both differed significantly from sentence 3 ($p<.01$).

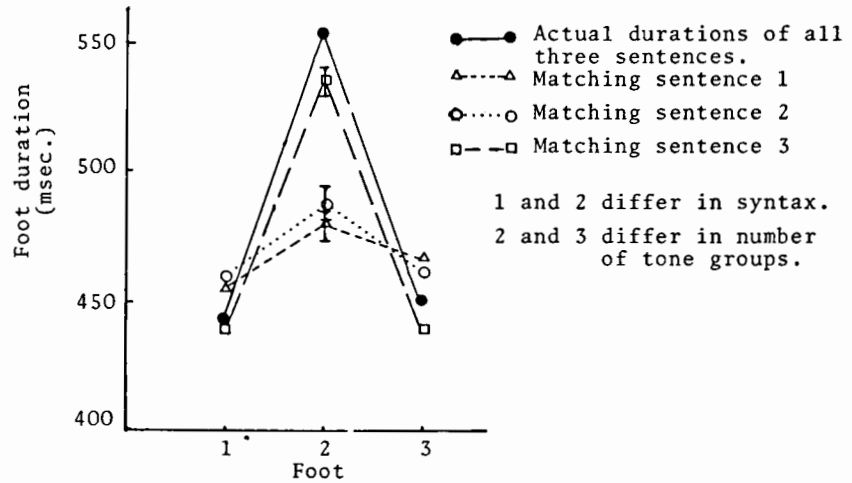

Fig. 3. Actual and perceived foot durations for the three utterances in Experiment 3 (see text for details).

It is apparent from these results that subjects are responding differently to the one and two tone-group utterances irrespective of the syntactic structure and despite the fact that the foot durations are the same in all three cases. Rees (1975), building on Halliday's (1967) work, has proposed that the tone group is a unit of rhythm as well as a unit of intonation so that isochrony need be maintained within but not between tone groups; it may be that this puts constraints on the limits of perceptual isochrony as well as on the tendency towards isochrony in production. It is clear from the experiments reported here that people are consistently inaccurate when judging speech rhythms and, furthermore, that they tend to hear these rhythms as more regular than they really are, at least when the utterance is bounded by a single tone group. Within the tone group, long feet tend to be underestimated, even when they contain a major syntactic boundary, while short feet tend to be overestimated.

Conclusions

Our results have broadly confirmed Lehiste's proposal that isochrony is partly a perceptual phenomenon. But we would make two points in addition. First, it is a perceptual phenomenon which is not independent of intonation. Second, we feel that it is a perceptual phenomenon, confined to language, reflecting underlying processes in speech production. Our results strengthen the case for models of the timing of English that incorporate an underlying

rhythmic organisation within tone groups. Conversely, they question the value of seeking direct links between syntax and segmental durations rather than indirect ones via an overall rhythmic structure which is also determined by the pragmatic and semantic context of a sentence (cf. Cutler & Isard, in press).

References

Abercrombie, D. (1964): "Syllable quantity and enclitics in English", in In Honour of Daniel Jones, D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, J.L.M. Trim (eds.) London: Longmans, 216-222.

Allen, G. (1970): "The location of rhythmic stress-beats in English: An experimental study", UCLA Working Papers 14, 80-132.

Allen, G. (1975): "Speech rhythm: its relation to performance universals and articulatory timing", JPh 3, 75-86.

Coleman, C. (1974): A study of acoustical and perceptual attributes of isochrony, Ph.D. thesis, Univ. Washington.

Cutler, A. and S.D. Isard (in press): "The production of prosody", in Language Production, B. Butterworth (ed.) New York: Academic Press.

Fowler, C.A. (1977): Timing Control in Speech Production, Ph.D. thesis, Univ. Connecticut, Connecticut, Ind: Indiana Univ. Linguistics Club.

Halliday, M.A.K. (1967): Intonation and Grammar in British English, The Hague: Mouton.

Huggins, A.W.F. (1975): "On isochrony and syntax", in Auditory Analysis and Perception of Speech, G. Fant and M.A.A. Tatham (eds.), London: Academic Press, 455-464.

Klatt, D. (1975): "Vowel lengthening is syntactically determined in a connected discourse", JPh 3, 129-140.

Lehiste, I. (1973): "Rhythmic units and syntactic units in production and perception", JASA 54, 1228-1234.

Lehiste, I. (1977): "Isochrony reconsidered", JPh 5, 253-263.

Lindblom, B. and K. Rapp (1973): "Some temporal regularities of spoken Swedish", Papers from the Institute of Linguistics, University of Stockholm.

Rees, M. (1975): "The Domain of Isochrony", Edinburgh Univ. Dept. of Linguistics, Work in Progress, 8, 14-28.

Witten, I.H. (1977): "A flexible scheme for assigning timing and pitch to synthetic speech", L & S 20, 240-260.