Perception of the Rhythm of English and of Nonspeech Analogues*

ALAN BELL

CAROL FOWLER
DARTMOUTH COLLEGE AND
HASKINS LABORATORIES

1. Introduction. The results of Lehiste 1973, 1977 and of Donovan and Darwin 1979 suggest that English speech rhythm is perceived as isochronous, and that this percept is special to speech. In the experiments of Lehiste 1977, for example, sentences containing unequal intervals between stresses (which we will refer to as feet or inter-stress intervals, or ISI's for short) were presented to listeners. They were asked to report whether a given ISI in a sentence was longer, shorter, or the same as another ISI. Since her listeners were unable to distinguish longer from shorter ISI's in the test sentences, Lehiste concluded that the ISI's were perceptually of the same duration. On the other hand, when listeners were asked to make similar judgments about the intervals between clicks in sequences whose timing matched that of the test sentences, they performed accurately, suggesting that the perceptual isochrony of ISI's in sentences were not the result of a general inability to distinguish temporal durations of certain magnitudes and organizations, but represented some special property of speech.

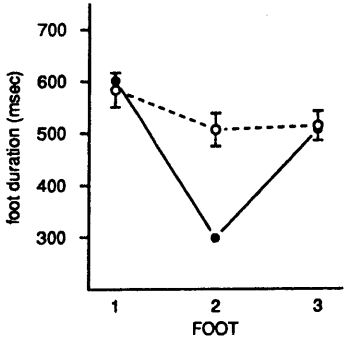

Figure 1. Comparison of stimulus rhythm and tapped rhythm for the Donovan and Darwin sentence *He 'turned up by 'ten 'talking of 'terrorism.* The actual foot durations of the stimulus sentence are represented by the filled circles.

Suggestive as the results were, one might be permitted to doubt that inability to judge the relative durations of pairs of feet is the same thing as perceiving the rhythmic structure of an unequal sequence of stresses as isochronous. Donovan and Darwin largely overcame this objection in their experiments. They asked listeners to respond directly to rhythmic structures by producing a matching rhythm, for example by tapping it out immediately after hearing the sentence. The results were very similar to Lehiste's: the responses to sentences were approximately isochronous, even though the intervals between the stresses of the sentences varied by as much as two to one; whereas the responses to sequences of tone bursts with the same timing as that of the stresses in the sentences matched the timing of the tone bursts very closely. In Figure 1, the actual rhythm (displayed in terms of the foot durations) is compared to the average of the response rhythms for one of Donovan and Darwin's sentences. Even though the number and the variety of test sentences used by Donovan and Darwin was not great, their results appeared to substantiate the hypotheses that English utterances are perceived as composed of approximately equal feet or ISI's and that this percept is specific to speech. Moreover, the paradigm of rhythm-matching appeared to be a powerful tool for further investigation of rhythmic perception in speech.

2. Experiment 1. To insure that we could obtain the same results, our first experiment was a replication of the Donovan and Darwin procedure. We chose two sentences as stimuli, both containing four stresses (and hence three feet). One, with a short middle foot, was the same as one of Donovan and Darwin's:

He 'turned up by 'ten 'talking of 'terrorism

The other sentence had a long middle foot:

Be'tween the 'carp swam 'tiny tadpoles

For each of these sentences we constructed matching nonspeech controls, consisting of sequences of tone bursts with the same timing as the onsets of the stresses in the sentences. Sixteen subjects tapped to the rhythm of the two sentences, and sixteen different subjects tapped to the rhythm of the tone bursts. Overall our results were similar to those of Donovan and Darwin. The averaged intervals between sentence response taps were much closer to isochrony than the intervals between stresses. But the response taps to the tone bursts matched the stimuli rhythms quite closely.

So far so good. A closer examination of the data, however, raised two troubling questions. Although the average data show a clear tendency toward isochrony, not all subjects responded as if they perceived an isochronous rhythm. In fact, a slight majority (9 of 16) tapped out a veridical rhythm which matched the sentence stimulus rhythm. Figure 2 compares the responses of a typical "veridical" subject and a typical "isochronous" subject for the CARP sentence. There is of course an explanation for this that is consistent with the hypothesis of a speech-specific isochronous percept for English. It is that the isochronous subjects were processing the sentences in a "speech mode" and that their responses reflect the speech-specific nature of the isochronous percept, whereas the veridical subjects were not processing the sentences as speech in this experimental setting, and consequently matched the temporal structure of the stresses, much as the subjects did who heard the tone burst controls. Upon a bit of reflection it becomes apparent that the subjects' behavior can also be explained in a way that is not consistent with the perceptual isochrony hypothesis. Perhaps the veridical subjects are the "good" subjects, and the isochronous ones the ones that require an explanation. The subjects were after all instructed to match the rhythm of the stresses in the sentences.

Under this interpretation, the veridical subjects are the ones who are capable of perceiving the rhythmic pattern of the stresses. The isochronous subjects, on the other hand, were not able to perceive the rhythm clearly, and consequently produced a more-or-less even pattern of taps. Under this interpretation, obviously, there is no such thing as an isochronous percept, and the tendency toward isochrony that appears in the responses averaged over subjects is an artifact of the experimental task.

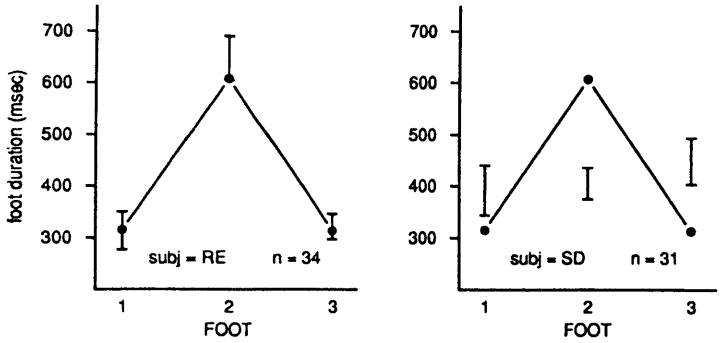

Figure 2. Responses of two subjects to the sentence CARP in Experiment 1. The vertical bars indicate the range of duration between taps for each foot.

The second question concerned what constitutes appropriate nonspeech controls. A series of clicks or tone bursts that match the temporal structure of the stress onsets in sentences is a natural choice, for it is indisputably nonspeech, and it is generally assumed that the temporal structure of stresses is the most important component of the stress-based rhythm of a sentence. On further reflection it is clear that stripping away all segmental and prosodic structure except the temporal structure leaves a stimulus whose rhythm may differ enough that it no longer qualifies as a control. The principal problem is that it is the nontemporal prosodic structure of a sentence that in large part assures the rhythmic coherence of the sentences, i.e., that it consists of a single rhythmic unit, which for sentences of the type under consideration here amounts to a single tone group. An example will make this clear. The tone group counterpart of a sentence with a long middle foot, such as the CARP sentence above, would have a temporal structure that can be represented schematically as X X X X. For the foot durations in question, between 300 msec and 600 msec, most listeners will perceive such sequences of identical auditory events as two rhythmic groups, not one, probably each with an initial beat: DA da, DA da. Since it is well established that durations between rhythmic groups are perceived more accurately than durations within groups, it is thus to be expected that such stimuli will be matched more accurately by tapped rhythms than the corresponding sentences. Such controls are thus not properly regarded as nonspeech controls, but rather as response controls that assure that subjects are able to match at least some rhythms with sequences of taps.

3. Experiment 2. Experiment 2 was designed primarily to answer whether the results of the tapping task come from an isochronous percept of speech or from an artifact of the task, and secondarily to explore alternative nonspeech controls for rhythmic perception. Let us call the two hypotheses to be distinguished Hypothesis I and Hypothesis II:

	Hypothesis I	Hypothesis II
Isochronous responses	speech mode	poorly perceived rhythm
Veridical responses	nonspeech mode	clearly perceived rhythm

The stimuli for Experiment 2 consisted of five sentences, which included the two sentences used in Experiment 1. All sentences contained four stresses and hence three feet. Two of the sentences had a short middle foot:

He 'turned up by 'ten 'talking of 'terrorism

The 'pain from a de'cayed 'tooth can be 'terrible

Two of the sentences had a long middle foot:

Be'tween the 'carp swam 'tiny 'tadpoles

She 'tied the 'tarpaulin with a 'piece of 'cord

One sentence consisted of feet of about the same duration:

The 'desks'll be 'covered with 'paper and 'pencils

These sentences were recorded and then digitally processed. There were four experimental conditions, corresponding to four different forms of the stimulus. Sentences in the first condition were LPC resynthesized versions of the original sentence without further modification. For the second and third conditions, the stimuli were versions of the sentences produced by the technique of sinewave sythesis. Sinewave synthesized speech is produced by the summation of three pure tones whose pitches correspond to the three lower formants or frequency-domain intensity peaks at any given moment in the corresponding natural speech (Remez et al 1981). For many people, the resulting signal is not perceived or recognized as speech, but more like a warbling whistle. If one is told that it is speech, however, or for some people after practice listening to it, it may become perceived as speech. This effect can be accentuated by either cueing the stimuli with more or less speech-like practice materials in practice sessions before the stimuli are presented. Thus conditions two and three used the identical stimuli, but in condition two the sinewave synthesized sentences were cued as speech and in condition three they were cued as nonspeech. The stimuli in condition four were produced by bandwidth broadening, a modification of the LPC representation of the original sentence which roughly consists of synthesizing the sentences from LPC parameters recalculated after setting the formant bandwidths to 1000 Hz. The result is a signal that is very much speech-like, but is quite unintelligible, rather like what one might imagine hearing in a very poor quality public announcement system in a noisy bus terminal.

There were a total of 48 subjects, 12 for each condition. Each subject tapped to all five sentences. The order of sentences was counterbalanced across subjects.

Our expectations were that the two sinewave synthesis conditions would enable us to distinguish between Hypothesis I and Hypothesis II, and that in addition the bandwidth broadened condition would provide a comparison of perceived rhythm in a speech-like signal in which the rhythmic structure was moderately degraded.

Colorado Research in Linguistics, Vol. 9 [1986]

https://scholar.colorado.edu/cril/vol9/iss1/1
DOI: https://doi.org/10.25810/cjqv-9y28

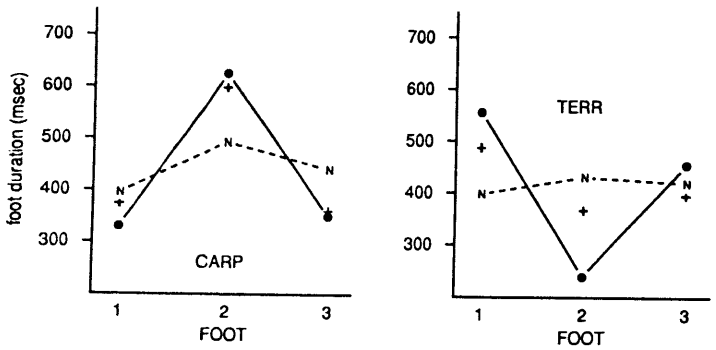

Figure 3. Responses to the CARP and TERRORISM sentences in Experiment 2. The stimulus foot structure is indicated by the dots joined by solid lines. The responses to the LPC synthesis condition are indicated by crosses, and the responses to the nonspeech sinewave synthesis condition is indicated by the N's. Responses to the other two conditions, speech sinewave synthesis and bandwidth broadening, were close to those for LPC synthesis, and are omitted here for clarity.

A general idea of the results can be obtained from Figure 3, which compares the temporal structure of the stimulus feet with the structure of the response taps for the CARP and TERRORISM sentences. Of the four sentences with uneven foot structures, the CARP sentence yielded the most nearly veridical responses in the unmodified (LPC) speech condition, and the TERRORISM sentence yielded the least veridical (more nearly isochronous) responses. The other sentences were in between. Of the other conditions, the nonspeech sinewave synthesis condition was the most deviant, consistently being more nearly isochronous than the responses under the other conditions.

In order to analyze the results more systematically, the foot structure of the stimuli and the responses was transformed in the following way. Since we are concerned with the overall rhythmic structure, it is the relative durations of the three feet that is of interest. Accordingly, the three foot durations were normalized by dividing them by the sum of the durations. The rhythmic structure can then be represented by two independent parameters, X and Y, which are defined as follows:

$$X = \text{FOOT}_1 - \text{FOOT}_3$$

$$Y = \text{FOOT}_2 - \text{Average}\,(\text{FOOT}_1, \text{FOOT}_3)$$

X is the difference between the first and the last foot; Y is the difference between the middle foot and the mean of the other two feet.

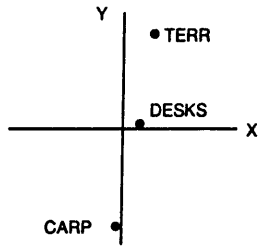

Figure 4. Rhythmic structure of three sentences in X-Y space.

In this space, the rhythmic structure of the stimulus sentences varies most in the Y dimension, as can be seen from Figure 4, in which the locations of the rhythmic structures of the CARP, DESKS, and TERRORISM sentences are plotted. There are two main statistics defined in terms of X and Y that are of interest to us. One, which I call DEV, is a two-dimensional measure of the deviation of the response from the stimulus rhythmic structure. DEV is obtained by a linear transformation of X and Y, first by shifting the origin to the location of the stimulus rhythm for each sentence, i.e. to the coordinates

$$X', Y' = X - X_{resp}, Y - Y_{resp},$$

and then rotating the axes so that the Y axis points toward the origin. See Figure 5. DEV is thus the difference between the stimulus and response rhythms in the transformed space, and consists of two components: the deviation of the response from the stimulus rhythm in the direction toward isochrony (the origin of the original X-Y space), and an orthogonal component which represents a shift away from the stimulus rhythmic structure.

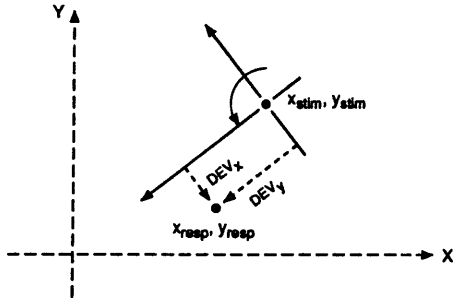

Figure 5. Transformation of the rhythmic structure variable (X, Y) to the vector DEV.

The other main statistic of interest, R, is a measure of deviation of the response from isochrony and is independent of the stimulus rhythm. It is simply the Cartesian distance

$$R = (x^2 + y^2)^{1/2}.$$

Some of the results of Experiment 2 are presented in Table 1.

Condition	DEV_x	DEV_y
LPC synthesis	-17	83
	7	100

Table 1. Averages of DEV_x andF DEV_y for the four stimulus conditions over all subjects and sentences. The normalized values of DEV_x and DEV_y are dimensionless. The conditions are significantly different ($p < .001$, MANOVA).

What can we conclude from these results? Let us consider the issue of isochrony of perceived rhythm first. One hypothesis might be that $R = 0$ for English sentences, i.e., that the perceived rhythm is indistinguishable from a sequence of even feet. The data for Experiment 2 do not support this, for R is significantly greater than 0 for all sentences, even the rather evenly stressed DESKS. The opposite hypothesis of no systematic deviation from the stimulus rhythms, i.e., that $DEV = 0$, is also unsupported, for DEV is significantly greater than 0 for all sentences and all conditions. Since DEV_y is consistently and substantially positive, the results indicate a response preference for more nearly even rhythms. (The results are of course similar to Experiment 1 and those of Donovan and Darwin in this regard.) Incidentally, DEV is significantly different for different sentences.

Now the main question we began with in this experiment was how to interpret this preference. Is the preference largely a property of how English speech is perceived (Hypothesis I), or is it largely a property of the experimental task, perhaps related to the relative difficulty of perceiving the rhythmic structure of different kinds of stimuli (Hypothesis II)? Comparison of the two sinewave synthesis (sws-speech and sws-nonspeech) conditions obviously bears upon the question whether the preference is specific to speech or not. We posit the null hypothesis that the perceived rhythms under the two conditions show the same tendency toward isochrony, i.e.,

$$DEV_y(\text{sws-speech}) = DEV_y(\text{sws-nonspeech}).$$

It is opposed to the alternative that the percept under the sws-speech condition is more nearly isochronous,

$$DEV_y(\text{sws-speech}) > DEV_y(\text{sws-nonspeech}).$$

The results of course do not reject the null hypothesis, and hence provide no support for Hypothesis I.

Turning to the alternative explanation, if perceptual difficulty plays a major role in the tendency for some percepts to have a more nearly even rhythm than the stimulus, then we might expect that this tendency would be stronger under the sws-speech and bandwidth broadened (BWB) conditions than for the LPC condition, and if so, the difference would presumably not be

explainable by the more or less speech-like nature of the stimuli. Here the null hypothesis of no difference in tendency toward isochrony is

$$DEV_y(\text{LPC}) = DEV_y(\text{SWS-speech}) = DEV_y(\text{BWB}),$$

as opposed to

$$DEV_y(\text{LPC}) < DEV_y(\text{SWS-speech}) \text{ or}$$

$$DEV_y(\text{LPC}) < DEV_y(\text{BWB}).$$

Although Table 1 shows that DEV_y is larger for the sws-speech and BWB conditions, the difference is not significant ($p > .10$ for both pairs). Since again the results do not contradict the null hypothesis, this comparison affords no support for the perceptual difficulty explanation. If the lack of clarity of the perceptual structure of stimuli (or other source of difficulty) does shift tapped responses toward a more nearly even rhythm, then the difference among the stimuli in these three conditions was not large enough to have a detectable effect. The fact that subjects in all three conditions knew what the stimulus sentence was and were free to repeat it as they tapped out the rhythm may have contributed to this result.

Our general conclusion is that there is little support for the notion that English speech has the intrinsic and special property that some people perceive its rhythmic structure as more nearly isochronous than it is. Whether or not the observed tendency toward isochronous percepts can be explained in terms of clarity of rhythmic structure or other task-related source of difficulty remains an open question. (We note that much the same conclusions were reached by Isard, Scott, and deBoysson-Bardies in a report of somewhat different experiments to the May 1984 meeting of the Acoustical Society.)

**Paper presented to the 108th meeting of the Acoustical Society of America, Minneapolis, October 1984. Work supported in part by NSF, NICHD, and University of Colorado Council on Research and Creative Work.*

References

Donovan, A., and C. J. Darwin. 1979. The perceived rhythm of speech. *Proc. of the 9th International Congress of Phonetic Sciences*, Copenhagen.

Lehiste, Ilse. 1973. Rhythmic units and syntactic units in production and perception. *J. Acoustical Society of America* 54.1228-34.

_____. 1977. Isochrony reconsidered. *J. Phonetics* 5.253-63.

Isard, S. D.; D. R. Scott; and B. deBoysson-Bardies. 1984. The perceived rhythm of English and French as assessed by the tapping task. Paper presented to the 107th meeting of the Acoustical Society of America, Norfolk, May 1984.

Remez, R. E.; P. E. Rubin; D. B. Pisoni; and T. D. Carrell. 1981. Speech perception without traditional speech cues. *Science* 212.947-50.

Colorado Research in Linguistics, Vol. 9 [1986]