The Temporal Regulation of Speech

John J. Ohala

Phonology Laboratory, Department of Linguistics
University of California, Berkeley, Calif.
U.S.A.

The questions this paper is concerned with are the following: (a) what factors determine the timing of speech utterances? and (b) what experimental techniques will reveal these factors? More concretely, given an utterance such as 'Joe took father's shoebench out', what determines the length of the time interval between the initial [dʒ] of 'Joe' and the [b] of 'shoebench' as well as all other intervals between the segments in the utterance? To start, I think three simple hypotheses can be entertained:

1. Some units of speech, perhaps syllables, stresses, or morae, are uttered in time to some underlying regular rhythm, e.g. the [b] of 'shoebench' will be uttered after the [dʒ] of 'Joe' an interval which is an integral multiple of the period of this underlying rhythm.

2. The units of speech are executed according to some underlying pre-programmed time schedule although there may be no isochrony in this schedule.

3. There is no underlying time program or rhythm; a given speech gesture is simply executed after the preceding gestures have been successfully completed, that is, one unit is simply strung after the other.

Hypothesis (1) is assumed by some linguists and language teachers to be true of English, Japanese, and a few other languages, specifically, they claim that there tends to be equal intervals between stresses in English and that all morae in Japanese tend to have equal duration (an impression no doubt derived in large part from Japanese poetic conventions and the near-syllabic *kana* orthographic system). But it has been difficult to verify these claims. Lenneberg (1967) who posits an underlying rhythm of 6 Hz for speech, outlines a method for testing this point. He suggests sampling running speech and measuring the intervals between several thousand successive releases of voiceless stops, or, presumably, any other easily detected speech event associated with syllable onset. The sampling technique also must necessarily be one that will miss some syllable onsets, e.g., one that detected syllables beginning with voiceless stops would miss all syllables beginning with other than voiceless stops. If there is some kind of periodicity underlying speech these intervals should coincide with the basic period of this rhythm. An interval histogram formed from these measured intervals ought to show a multimodal distribution, the distance between the peaks of the histogram being equal to the period of the underlying rhythm.

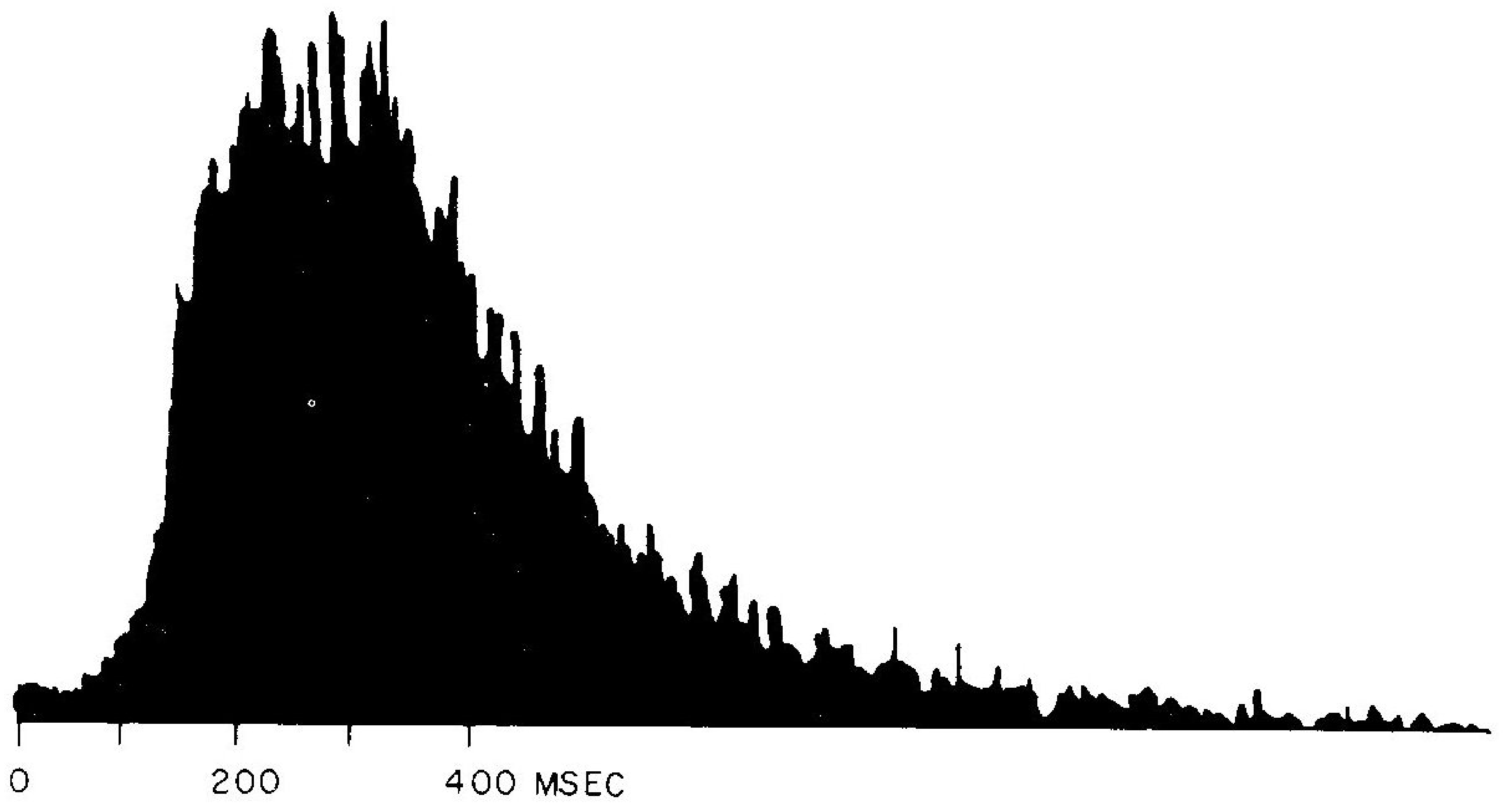

Figure 1 Histogram of the intervals between some 10 000 successive jaw openings in running speech (reading).

Fig.1 shows one such histogram (from Ohala, 1970 and 1972). This represents about 10,000 intervals between successive jaw openings, i.e., local maxima in the jaw displacement function. The subject (the author) read technical prose for about 1 1/2 hours; jaw movement was tracked optically (Ohala *et al.*, 1968); the intervals were measured automatically by a small computer. The histogram shows some high frequency noise between 0 and 100 ms, perhaps an artefact of the jaw tracking system. It also was not possible to exclude pauses, but these are no doubt limited to the larger intervals. In general the histogram seems quite smooth and reveals no obvious multiple peaks. There is a large single peak around 250 ms, which may be the modal syllable rate or the preferred frequency of the mandible. This data, then, gives no support to the claim that there is any isochronic principle underlying speech, at least, the speech of this particular English speaker.

However, one could argue that this study contained various features which would obscure an underlying rhythm if one did exist: first, the speech was not spontaneous and thus the speaker might not give free rein to the natural rhythm of speech; second, the particular event used to obtain the intervals, that of peaks in the jaw displacement function, is not reliably correlated with any underlying neurological speech event, that is, presumably the events of interest are those the speaker's brain uses and the attainment of peak jaw opening may not meet this criterion. To remedy these difficulties, another interval count was performed, this time with spontaneous speech (of about 1 hour's duration) and also measuring the intervals between successive drops in oral pressure, such as would occur upon the release of voiceless obstruents. (The author was the subject; oral air pressure was sampled via a short catheter introduced into the pharynx via the nose and connected to a strain-gauge pressure transducer; the measurements were performed automatically by a small computer; intervals of 40 ms or less were discarded.) The histogram of some 4000 intervals so obtained is shown in Fig.2. Again it appears there is some high frequency noise near the left end of the histogram, and again, there is a large peak around 200-250 ms. In this histogram, however, there are apparent sub-peaks, approximately 50 ms apart, although they are enveloped in noise and do not always seem to be spaced evenly. Only further such studies can reveal whether these sub-peaks are real or just noisy apparitions. For the present, then, this data still provides no obvious evidence for an underlying isochronic rhythm for speech but suggests that it would be useful to keep looking for one.

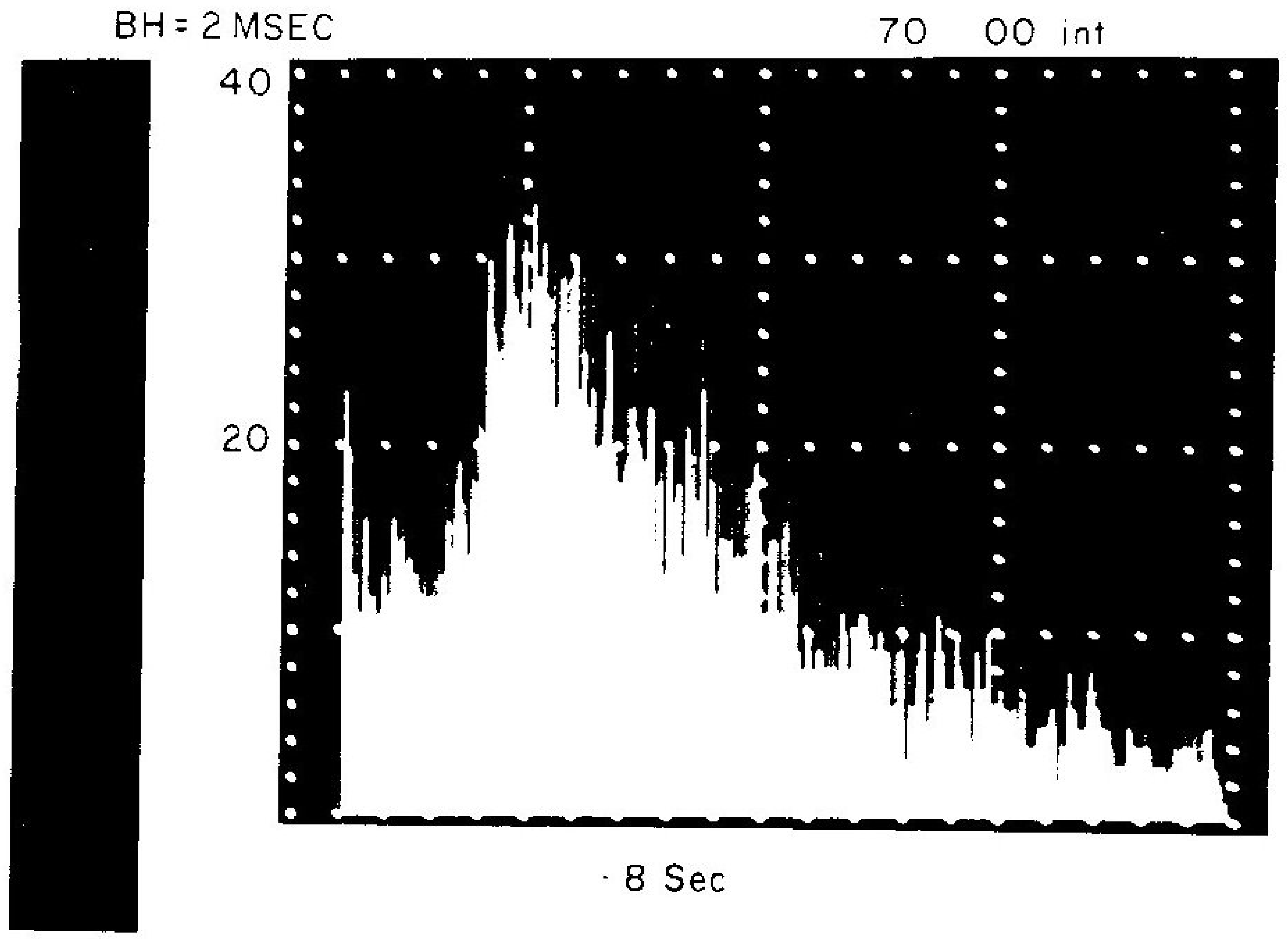

Figure 2 Histogram of the intervals between some 4000 successive drops in oral pressure accompanying the release of voiceless obstruents in spontaneous running speech.

Comb vs. chain model[1]

How can we test whether hypothesis (2) or hypothesis (3), above, applies to speech? Kozhevnikov and

[1] The discussion to follow supercedes that in Ohala (1970:145-152) which contains conceptual flaws.

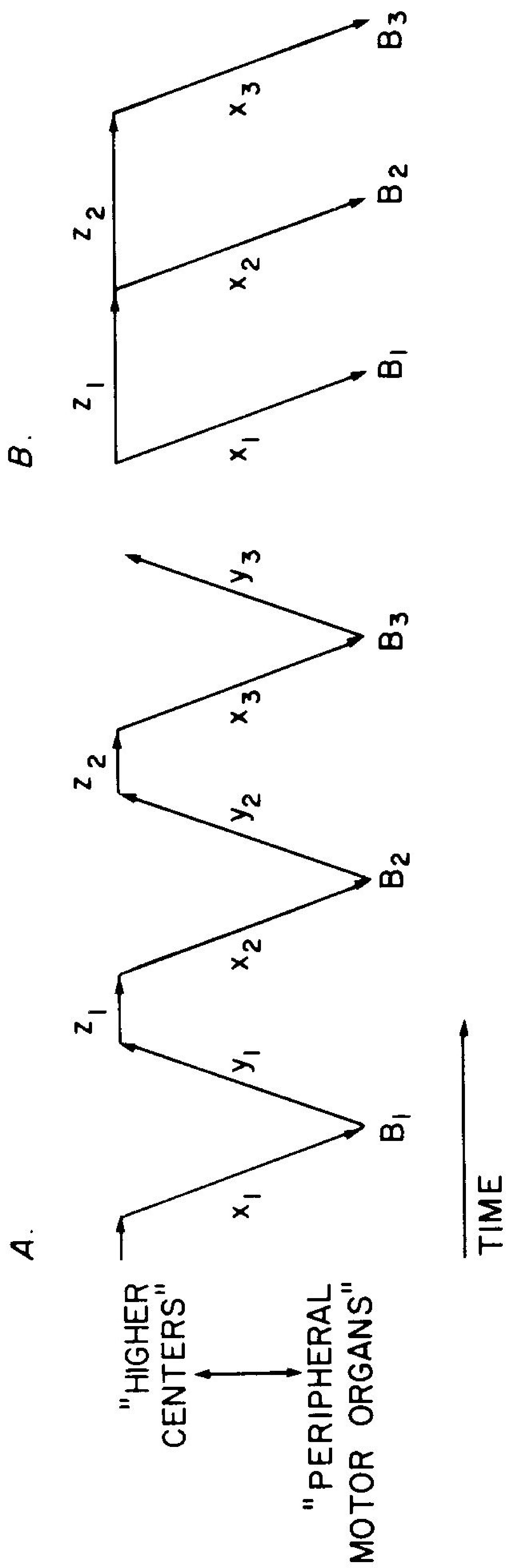

Figure 3 A: The 'chain' model. B: The 'comb' model.

Chistovich, in their pioneering 1965 monograph, *Speech: Articulation and Perception*, proposed that this can be tested by performing a special statistical analysis of the durations of segments in a given utterance repeated many times by a subject. They represented the two hypotheses by the two graphical structures in Fig.3. Fig.3b represents the structure corresponding to hypothesis 2; Fig.3a that appropriate for hypothesis 3. Following the terminology of Bernstein (1967) we can refer to Fig.3b as representing the 'comb' model and Fig.3a as representing the 'chain' model. In the figures the segments marked x represent the transmission of a 'command' from the brain to the peripheral motor organs (tongue, lips, velum, larynx, etc.), y segments (not part of the comb model) represent the transmission of information from the periphery back to the brain regarding the progress of the execution of the commands; in other words y represents sensory feedback, and z segments represent the transmission of impulses in the brain which give rise to the next command. All of this hypothetical neurological structure is inaccessible, of course. We can only observe and note the moments of occurrence of the speech events resulting from the execution of the delivered commands. These events are marked B_1, B_2, etc.

Restated in terms of these figures, the problem is that we can readily measure the inter-event interval, B_1B_2, B_2B_3, B_1B_3, etc., but how can we discover which of the two neurological structures underly these intervals? Kozhevnikov and Chistovich noted that if a speaker repeats the utterance containing intervals B_1B_2, etc., many times there will be some variability in the magnitude of these intervals because there will necessarily be some variability in the steps that are part of these intervals. In the case of the chain model there will be some variability in the underlying steps y_1, z_1, x_2, or, in the case of the comb model, in the underlying steps x_1, z_1, x_2. However, they argue, these two models should yield a different relation between the variance of any large interval, say B_1B_3, and the sum of the variances of its component intervals, B_1B_2 and B_2B_3. Specifically, in the case of the chain model it should be found that

(1) $$V(B_1B_3) = V(B_1B_2) + V(B_2B_3)$$

(where V represents 'variance of'), whereas in the case of the comb model it should be found that

(2) $$V(B_1B_3) < V(B_1B_2) + V(B_2B_3)$$

These relations follow from three points:

1. The classical statistical relation:

(3) $$V(X + Y) = V(X) + V(Y) + 2COV(X,Y)$$

(where COV represents 'covariance of'), or, more generally,

(4) $$V(\Sigma X_i) = \Sigma V(X_i) + 2 \sum_{i<j} COV(X_i,X_j)$$

2. The assumption that the errors (i.e., the individual deviations from the mean) encountered in the various transmission paths in the chain model will be uncorrelated, i.e., that the last term in (4) will be zero.

3. The observation that in the comb model the error in any x segment that is shared by two adjacent intervals (e.g., in Fig.3b, x_2 is shared by intervals B_1B_2 and B_2B_3) will contribute to the error of both intervals in equal magnitude but opposite sign, thus maing adjacent intervals negatively covarying, i.e., making the last term in (4) negative.

Kozhevnikov and Chistovich noted that in the event that there are variations in the overall rate at which each sentence is uttered, for the chain model the assumption in point (2) will not be valid. Instead there will be a positive correlation between the z segments and thus a positive correlation between the separate measured intervals $B_{i-1}B_i$, B_iB_{i+1}, themselves, that is, the last term in (4) will be positive and thus relation (5), below, will also be characteristic of the chain model.

(5) $$V(B_1B_3) > V(B_1B_2) + V(B_2B_3)$$

Kozhevnikov and Chistovich found relation (2) to hold in the speech material they studied. Allen (1969) and Lehiste (1971 and 1972) both found negative correlations existing between adjacent intervals of speech material they measured. It should be clear from the above that these are equivalent findings. These authors concluded that this was evidence for the existence of some sort of time program or schedule at least as long as the word and perhaps as long as the whole sentence, i.e., in the terms of this paper, they found conditions answering best to the comb model. However, data presented by Ohala (1970), if analyzed in this way, as in Fig.4, would show that relation (5) holds, which points to the chain model.[2]

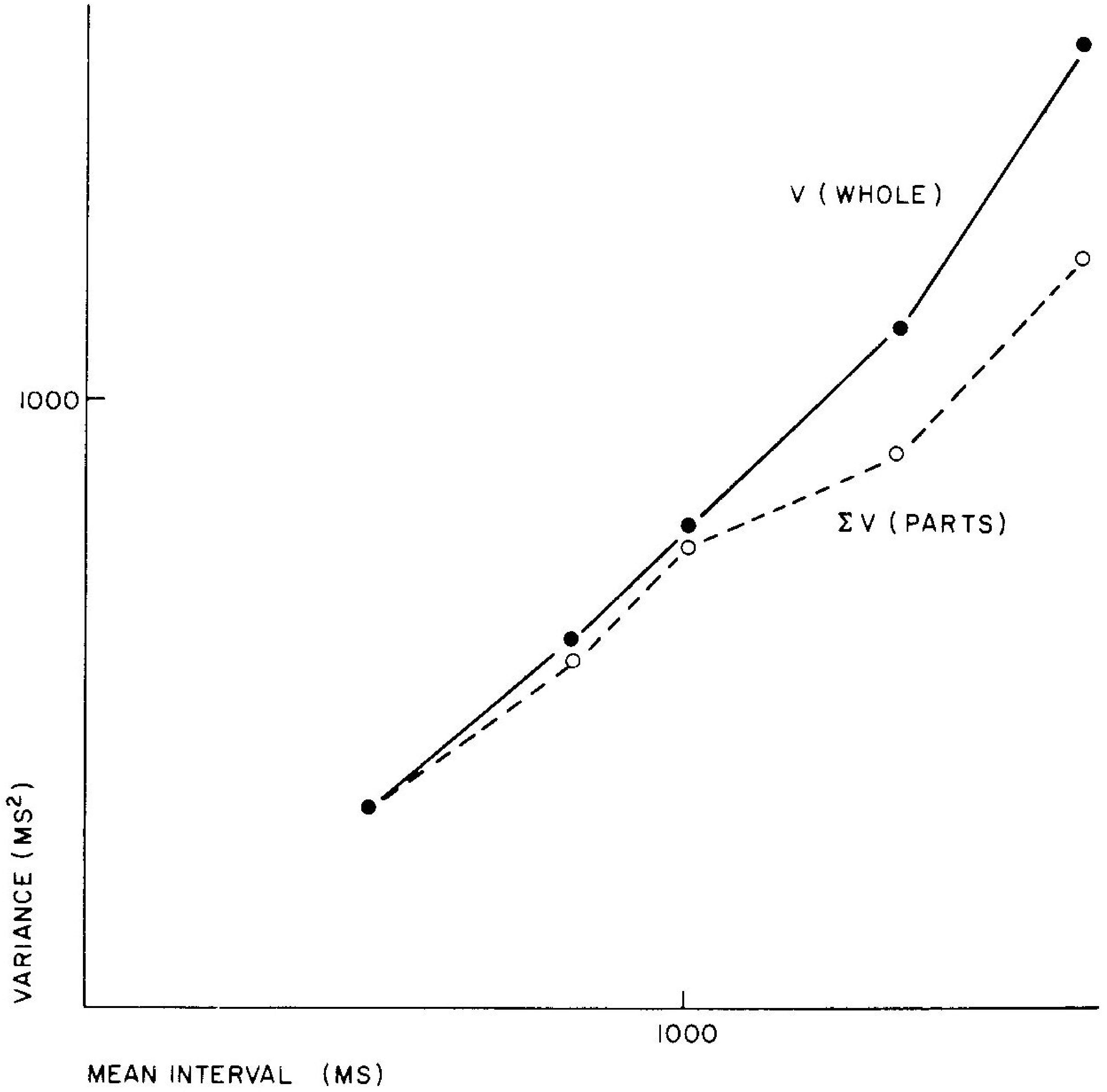

Figure 4 Variance vs. mean interval of speech from the data of Ohala (1970). Solid line: variance of given whole interval; broken line: sum of the variances of the component intervals of given whole interval.

In fact, this statistical metric cannot be used by itself to determine whether the chain or the comb model better applies to speech.[3] This metric might work if

we could be sure to eliminate measurement error and changes in the rate at which the subject speaks the test sentence from one repetition to the next. But we can be fairly sure that any data we obtain *will* be contaminated by these factors. Measurement error can be of two sorts: one which amounts to sloppiness, i.e., making a mistake in the segmentation of the speech material or mis-measuring an interval, and another which amounts to not knowing how to segment our speech material because one does not know what the brain of the speaker considers to be an 'event'. It may be possible to estimate the magnitude of or perhaps to control somewhat the first kind of measurement error. However there is nothing that can be done about the second kind. As was noted by Kozhevnikov and Chistovich, the effect of measurement error is to contribute to the variability of adjacent intervals in equal magnitude but in opposite sign, i.e., to make adjacent intervals negatively co-varying. This is the same effect as that due to the sharing of the error in the x segments in the case of the comb model. But measurement error will be present no matter which model may apply. Thus the last term in (4) would tend to be negative and may yield relation (2) no matter which model holds.

Further, variations in the over-all rate at which the subject speaks each repetition of the test sentence may lead to further problems. In terms of the models in Fig.3, speaking some sentences consistently a bit fast and others consistently a bit slow would mean that the z segments in either model would show some positive covariance. As noted above, this would tend to make the last term in (4) positive which would tend to yield relation (5). Kozhevnikov and Chistovich suggest that this effect of variation of rate would

[2] The variances presented in Ohala (1970) were erroneous; Fig.4 provides the corrected variances.

[3] This point grew out of discussion with Paul Tukey.

intervals in terms of relative error since it gives the false impression that there is more fluctuation on small intervals than on large intervals, therefore suggesting that some of these fluctuations are cancelled out or compensated for over longer time intervals and thus pointing to the comb model. But it is obvious from the data all investigators have obtained that the temporal fluctuations or variability of speech *increases* monotonically with the mean interval size.

Speech timing and feedback

In order to discover how the timing of the gestures of speech is regulated, it seems necessary to have some sort of direct intervention in the speech generation process. Viewed quite simply, the difference between the comb model and the chain model is that in the former no sensory feedback is used to determine when a given gesture will be executed, whereas in the latter sensory feedback *is* used for this purpose. It seems likely that we can find out if sensory feedback is necessary or important for maintaining the precision in timing in speech by seeing if the speaker's temporal precision suffers any if we reduce the amount of information getting to the brain by blocking (even partially) one or two of the feedback channels normally used in speech. Past studies of speech produced under sensory deprivation have indicated that there may be an adverse effect on the precision of the articulation of speech, or even the ordering of the speech gestures, but there have been no studies which looked for any change in the *temporal* precision of speech.

A preliminary version of this type of test was run as follows: two subjects, both young adult males, one a speaker of American English and the other a speaker of Japanese, participated in the study. Each spoke a corpus of sentences in their respective languages, under three experimental conditions (and in the following order): (1) control, in which there was no sensory impairment, (2) masking noise, in which broadband noise of sufficient intensity to mask speech was fed to the subjects' ears over earphones, and (3) anesthesia in which the surface tactile sensation of the subject's tongue and palate was reduced slightly by an oral application of Xylocaine Viscous - a mild surface anesthetic designed to relieve the pain of sore throats. The corpus of utterances which the subjects spoke in each condition consisted of 150 sentences representing a randomized ordering of three sentence types, each sentence type occurring about 50 times in the corpus. The sentences were designed to permit easy segmentation of the acoustic signal. The acoustic signal was picked up by a microphone sensitive to low frequencies, recorded, and later written out on paper. Noise bursts due to the release of stops were used for segmentation; measurement of the intervals between these points was done by hand. Of the three sentence types uttered by the speakers, the one which yielded the largest number of separate measurable intervals was chosen for complete statistical analysis. The English sentence (with the segmentation points indicated by arrows) was

$\overset{\downarrow}{\text{C}}$athy $\overset{\downarrow}{\text{t}}$ook some $\overset{\downarrow}{\text{c}}o\overset{\downarrow}{\text{c}}$oa in $\overset{\downarrow}{\text{P}}ec\overset{\downarrow}{\text{k}}$ 'n Peck again.

The Japanese sentence was

$\overset{\downarrow}{\text{T}}a\overset{\downarrow}{\text{k}}$e o $\overset{\downarrow}{\text{t}}a\overset{\downarrow}{\text{k}}$i no $\overset{\downarrow}{\text{k}}$ami ni $\overset{\downarrow}{\text{t}}a\overset{\downarrow}{\text{t}}e\overset{\downarrow}{\text{k}}a\overset{\downarrow}{\text{k}}$e nasai.

yield relation (5) only in the case of the chain model; but this effect could be present in the comb model, too.

Thus it is evident that there will always be various effects present which will tend to make the last term in (4) positive, zero, and negative, that is, variations due to rate, 'neuromuscular noise', and measurement error, respectively. Depending on the relative magnitude of these effects it is quite possible that we might obtain relation (1), (2) or (5) and whichever relation is obtained will not indicate whether the chain or comb model better accounts for the timing of speech gestures.

It is possible to show the validity of this analysis by using it to account for certain aspects of the data already obtained on temporal variation in speech. We can, for example, characterize roughly *how* the variance of a given interval will vary as a function of the interval size.

Variations due to rate, V_r, will vary proportionately with the square of the mean interval size. That is, if we assume a given large interval consists of n units or sub-intervals, and if the standard deviation is σ for each unit, then the standard deviation of n units will be $n\sigma$, since rate variation is essentially a multiplication of the duration of intervals by the same constant. The variance of n unit intervals will be the square of the standard deviation or $n^2\sigma^2$. σ^2 we can assume will be constant and since n will vary directly with the interval size we arrive at variance due to rate varying directly with the square of the mean interval, i.e.,

$$V_r = k_1\bar{I}^2 \tag{6}$$

The errors due to neuromuscular noise or randomness in the transmission of neural impulses we assume must be uncorrelated. The variance in a measured interval due to this source will simply be the sum of the individual fluctuations or errors in the underlying units which the large interval consists of. Thus variation due to noise, V_n, will vary proportionately with the mean interval, i.e.,

$$V_n = k_2\bar{I} \tag{7}$$

The third source of variation is measurement error which ought to be the same no matter what size interval is measured - thus it will be constant, i.e.,

$$V_m = k_3 \tag{8}$$

Therefore the total variance for a given interval will be the sum of these three separate variances, i.e.,

$$V(\bar{I}) = V_r + V_n + V_m = k_1\bar{I}^2 + k_2\bar{I} + k_3 \tag{9}$$

This, of course, is a simple quadratic equation - the function describes a parabola which intercepts the y-axis above zero. That this is a correct determination of the way variance varies can be shown by examination of existing data on variance. Fig.5 is data from Kozhevnikov and Chistovich and is similar to data obtained by me, Allen, Lehiste, and others. Here the variance of speech intervals is plotted as a function of the magnitude of the interval. As can be seen a parabolic function fits the data points rather well.

As mentioned above, Kozhevnikov and Chistovich proposed that the comb model would be accepted if the

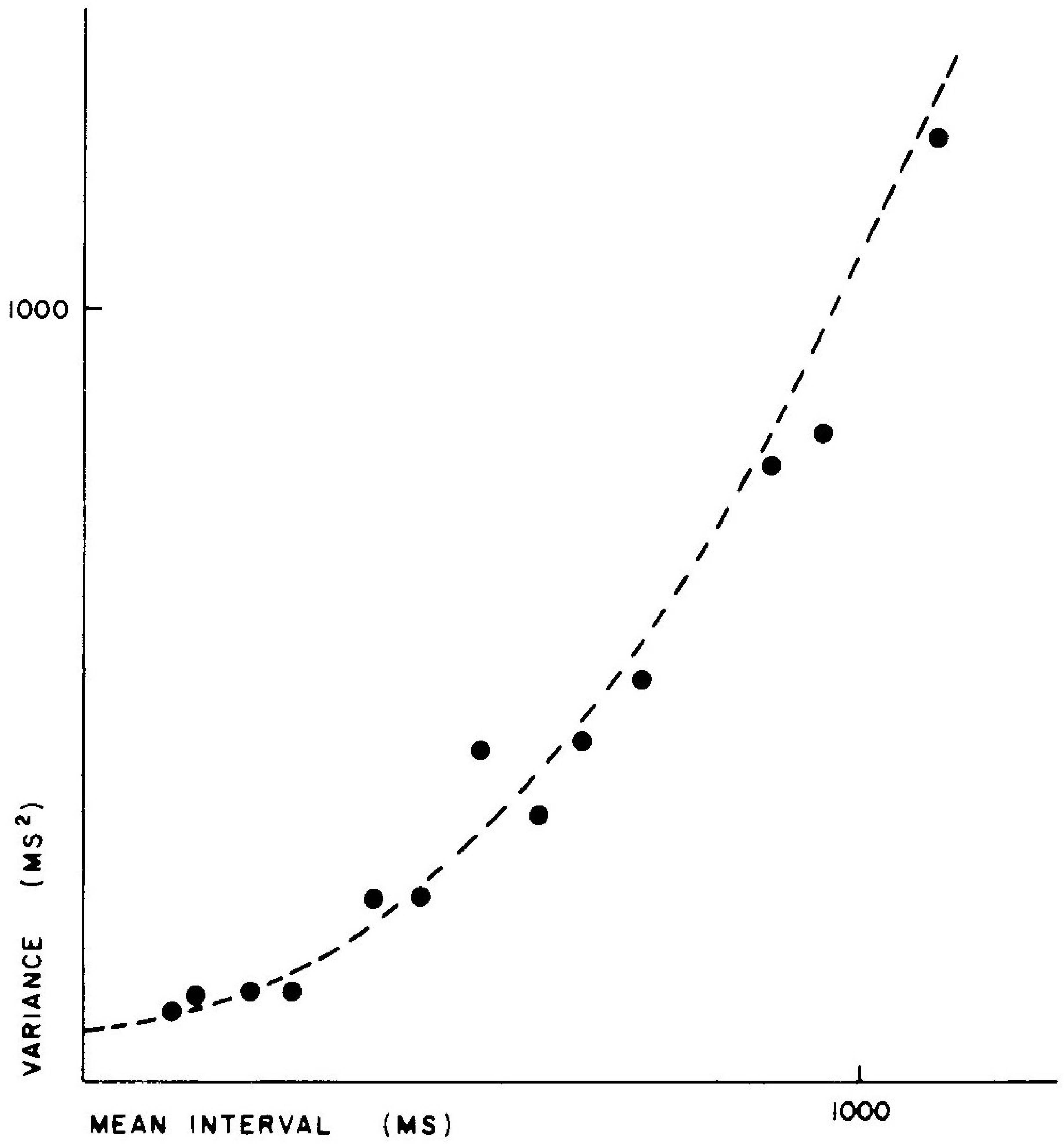

Figure 5 Variance vs. mean interval of speech from the data of Kozhevnikov and Chistovich (1965).

variance of the whole interval was found to be less than the sum of the variances of the component intervals. They found this relation to be true of their data. However my data (Fig.4) yields the opposite relation, that is, variance of the whole greater than the sum of the variances of the parts. This seems to be contradictory, but in fact, as is explained below, these findings are compatible - but still do not reveal whether the comb or chain model applies to speech.

If $\bar{I}$ is the mean whole interval and $\bar{I}/n$ is the

component interval duration, then, using the equation (9), we see that

$$
\begin{aligned}
(10)\quad V(\text{whole}) - \Sigma V(\text{parts}) &= V(\bar{I}) - n(V(\frac{\bar{I}}{n}))\\
&= (k_1\bar{I}^2+k_2\bar{I}+k_3) - n(k_1(\frac{\bar{I}}{n})^2+k_2\frac{\bar{I}}{n}+k_3)\\
&= k_1\bar{I}^2(1-\frac{1}{n})+k_3(1-n)
\end{aligned}
$$

This equation, giving the difference between the variance of the whole and the sum of the variances of the component parts, as a function of n, the number of parts the interval has been divided up into, is shown graphically in Fig.6.

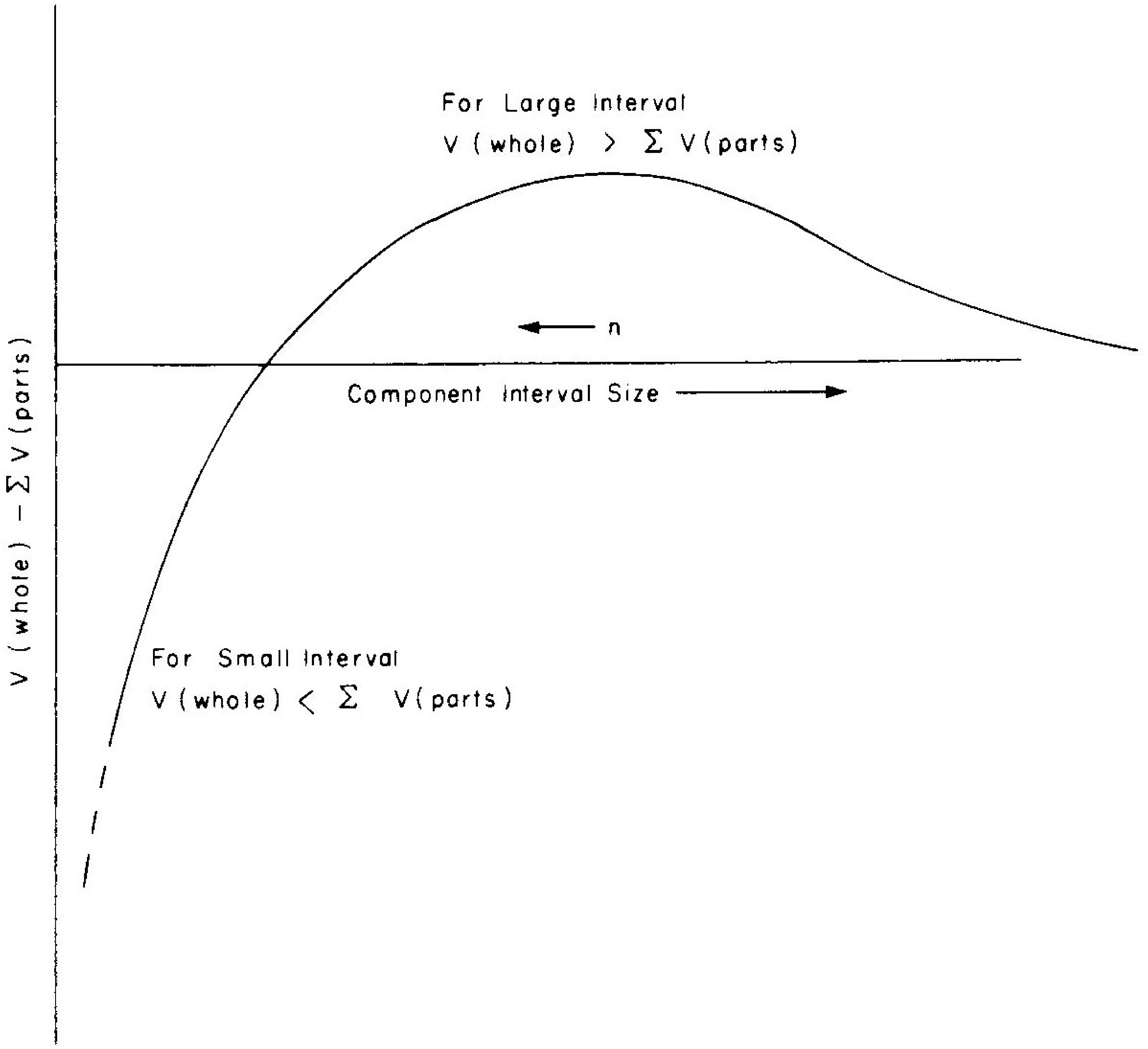

Figure 6 Graphical plotting of equation (10) in the text.

When a small sub-interval is chosen one would find as Kozhevnikov and Chistovich did, the V(whole) < ΣV(parts), but when large sub-intervals are chosed, as I did in my study, then the V(whole) > ΣV(parts).

Other statistical metrics

Two other statistical metrics require comment. The first is that of various authors' attempt to eliminate rate variations by 'normalizing' the durations of the whole utterance: Ohala (1970), Lehiste (1972) did this by limiting their statistical analysis to utterances which had durations closest to the mean; Allan did this by multiplying the total durations of his utterances and their component intervals by a normalizing factor which would give them the same total duration. Nothing useful is accomplished by these techniques. Insofar as they serve to artifically reduce the first term in equation (4) they necessarily cause the last term, that which expresses the covariance between intervals, to tend to be negative. Therefore it is no surprise that significant negative correlations between intervals are found in the data after applying these normalizations and consequently they give no evidence whatsoever of the comb model applying to speech.

The second statistical practice of dubious value is that expressing the temporal variability of speech segments using *relative error* (E_r = (standard deviation/ mean interval) x 100). Kozhevnikov and Chistovich note that relative error is larger on small intervals (10-20%) than it is on large intervals such as the entire sentence (3%) and attach importance to this fact. Likewise, Allen (1968) noting their data and similar findings of his own, comments that

> In order for this reduction in variance to occur there must be timing information that extends over the whole phrase. (p.75)

But these values are directly derivable from quation (9) (with the appropriate constants) and the definition of relative error:

$$E_r = \frac{\sigma}{\bar{I}} 100 \qquad (11)$$

$$= \frac{(V)^{\frac{1}{2}}}{\bar{I}} 100$$

$$= \frac{(k_1\bar{I} + k_2\bar{I} + k_3)^{\frac{1}{2}}}{I} 100$$

Taking $k_1 = 7 \times 10^{-4}$, $k_2 = 0.25$, and $k_3 = 100$, for the data presented by the Leningrad group, the relative error can be calculated for various size intervals as in the table below.

Interval	Relative Error
50ms	21.4%
100	11.5%
1250	3.0%

These calculated values are in agreement with the values determined by Kozhevnikov and Chistovich, cited above. But since these figures are directly obtainable from the same basic data on the variance of speech intervals, they add no more information to the study and are subject to the same criticisms presented above, that is they provide no evidence regarding the applicability of the chain or comb model to speech. It is probably not advisable to express the fluctuations in speech

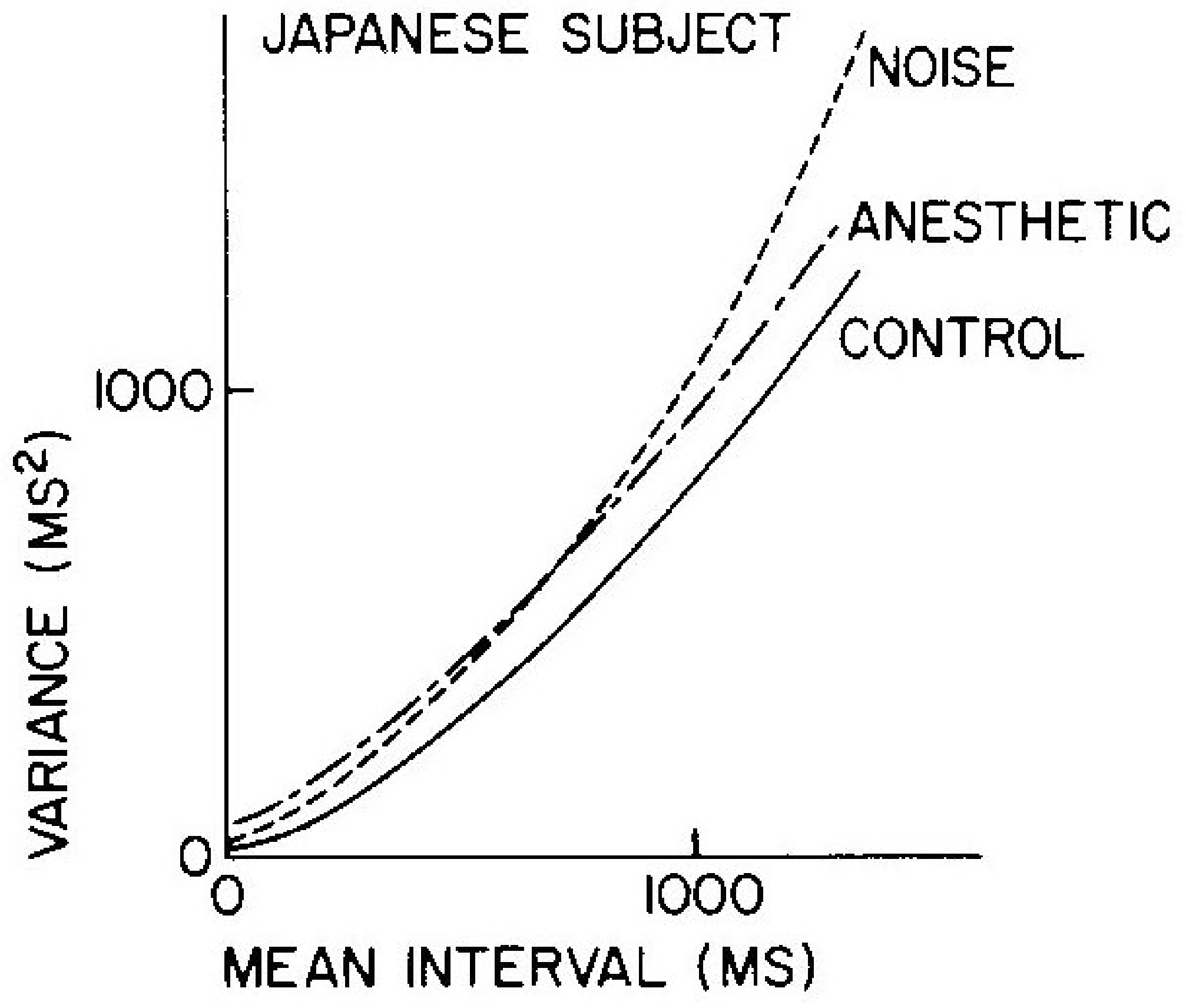

Figure 7 Variance vs. mean interval of speech for Japanese-speaking subject under three conditions: control, masking noise, and surface anesthetic. Explanation in text.

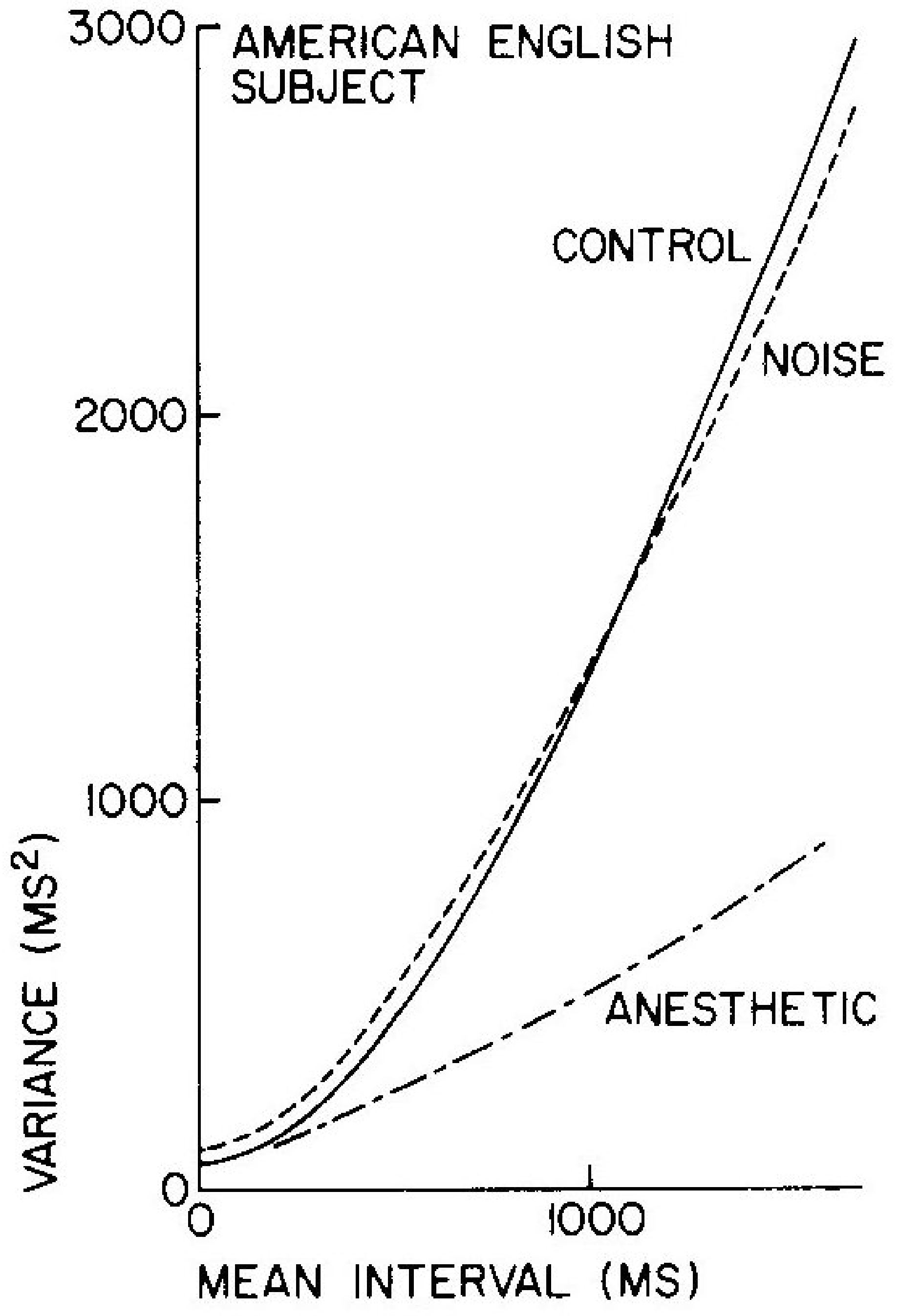

Figure 8 Variance vs. mean interval of speech for American English-speaking subject under three conditions: control, masking noise, and surface anesthetic. Explanation in text.

Figs.7 and 8 present the results for the Japanese and American English subjects, respectively. The trend of the variance as a function of the mean interval for the three conditions has been approximated by the three curves in each figure (fitted to the original data points by eye).

Unfortunately, the results are not conclusive. Although it is not clear how to obtain a measure of the overall temporal variability from figures such as these, it is clear that for the Japanese speaker the variability under all conditions was quite similar. This then points to the comb model. However, for the English speaker the variability was similar for the control and the masking noise condition, but variability was less for the anesthetic condition. This points to the chain model. Obviously, however, this preliminary experiment has too few controls: language-specific effects, order effects, practice effects, etc., need to be controlled in future studies. Furthermore, no doubt variability, like everything else, varies: probably several runs of each experimental condition for each subject is required to provide a reasonably accurate estimate of the variability of a subject in any given condition.

Timing and perception

Although further research is still called for on the question of how the timing of speech is controlled, there are results from perceptual and related studies that provide hints to the answer. We are asking how the timing of speech is regulated but we should first ask what the perceptual value of the time structure of speech is. There is abundant evidence in the literature that short-term variations in the timing of speech intervals have perceptual value (Huggins, 1972a and 1972b; Lehiste, 1970). But typically it is a vowel, a consonant, a vowel plus consonant sequence, or at most two adjacent syllables whose timing characteristics have perceptual import. Thus there is evidence that a speaker needs to maintain short-term temporal precision in his speech; there is yet no corresponding evidence that a speaker needs to maintain long-term temporal precision in his speech, that is, over a span of phrase or sentence length, in fact, some experimental results of Kozhevnikov and Chistovich (pp.114-5) suggest just the opposite. This being the case, it is likely that there is a pre-programmed time schedule the speaker must adhere to for short spans of speech, say over one or two syllables, but there is no such time schedule for longer stretches of speech. Thus a hybrid model is suggested: the chain model for long-term timing, the comb model for short-term timing.

Further research on the temporal variability of speech is clearly needed, but it is evident that this type of research, introduced by the Leningrad group, promises to provide us with insights into some of the neurological processes underlying speech.

I am grateful to G. Allen, B.E. Caspers, P. Denes, R. Harshman, and P. Ladefoged for help and advice on the research reported here. This research was supported in part by the United States National Science Foundation.

References

Allen, G.A. (1968). The place of rhythm in a theory of language. *Working Papers in Phonetics* 10, University of California, Los Angeles., 60-84

Allen, G.A. (1969). Structure of timing in speech production. Paper read at the meeting of the Acoustical Society of America, San Diego, 4 November 1969

Bernstein, N.A. (1967). The coordination and regulation of movements. Oxford: Pergamon Press

Huggins, A.W.F. (1972a). Just noticeable differences for segment duration in natural speech. *JASA* 51, 1270-1278

Huggins, A.W.F. (1972b). On the perception of temporal phenomena in speech. *JASA* 51, 1279-1290

Kozhevnikov, N.A. & Chistovich, L.A. (1965). *Speech: Articulation and Perception*. U.S. Dept. of Commerce translation, JPRS 30-543

Lehiste, I. (1970). *Suprasegmentals*. The M.I.T. Press

Lehiste, I. (1971). Temporal organization of spoken language. In *Form and Substance*, eds. L.L. Hammerich, R. Jakobson and E. Zwirner. Akademisk Forlag. 159-169

Lenneberg, E. (1967). *Biological Foundations of Language*. New York: Wiley

Ohala, J. (1970). Aspects of the control and production of speech. *Working Papers in Phonetics* 15. University of California, Los Angeles.

Ohala, J. (1972). The regulation of timing in speech. *1972 Conf. on Speech Communication and Processing*. IEEE. 144-147

Ohala, J., Hiki, S., Hubler, S. & Harshman, R. (1968). Photoelectric methods of transducing lip and jaw movements in speech. *Working Papers in Phonetics* 10, 135-144. University of California, Los Angeles