Comments on "Temporal interactions within a phrase and sentence context" [J. Acoust. Soc. Am. 56, 1258–1265 (1974)]*

John J. Ohala

Department of Linguistics, University of California, Berkeley, California 94720

Bertil Lyberg

Department of Phonetics, Institute of Linguistics, Stockholm University, Fack, S-10405 Stockholm 50, Sweden
and Telefonaktiebolaget LM Ericsson, Long Distance Division, S-12625 Stockholm, Sweden
(Received 9 December 1974; revised 15 December 1975)

The interpretation given in the subject article to the incidence of negative correlations between the durations of adjacent speech segments, viz., that they show evidence of temporal compensation, is criticized. It is shown that the negative correlations are most likely due to measurement error and to the author's normalization of the duration of his test utterances.

Subject Classification: [43] 70.40, [43] 70.70, [43] 70.20.

INTRODUCTION

What determines the timing of the units of speech: Is there some underlying temporal schedule of events the brain must follow in delivering motor commands to the speech articulators, or does the brain simply keep track of the order in which events occur, the timing of a given motor command being determined by when the brain receives sensory information that the preceding motor commands have been or are being successfully executed? These two possibilities have been called the "preplanning" and "chaining" models, respectively, of speech timing. (These models are relevant only to the question of the timing of the units of speech, not to their ordering or their selection. There is no dispute that these latter two processes have to be preplanned.)

The preplanning model would predict that deviations from the schedule would be made up or compensated for by appropriate adjustments in the timing of successive units within the same utterance. Thus Wright, in the article under discussion,[1] suggests

> ... if temporal compensation takes place within a linguistic unit, a degree of interaction will be revealed when correlations are made between its smaller adjacent segment durations. Significant negative correlations between adjacent segment durations reflect a high degree of temporal compensation, while zero or positive correlations indicate an absence of compensation. [p. 1258.]

There are complications, however, that Wright ignores.

I. THE EFFECT OF MEASUREMENT ERROR

Wright had his subjects repeat a phrase and a sentence several times and found significant negative correlations between the durations of certain adjacent stretches of speech within these utterances. It was on this basis that he claimed that there is evidence of temporal compensation on those intervals of speech spanning both stretches. In general, though, there are two possible sources of negative correlation between adjacent intervals: One is temporal compensation, i.e., a systematic effort on the part of the brain to compensate for temporal deviations in given speech intervals—to keep things "on schedule"—by appropriate adjustment of the durations of successive intervals within the same utterance. The other source is measurement error, i.e., since the adjacent intervals share a boundary, an erroneous placement of the boundary will increment the duration of one interval while decrementing the duration of the other interval by the same amount. This by itself will introduce negativity into the correlation between the measured intervals. This point can be demonstrated formally as follows.

Assume two adjacent speech intervals X and Y which upon N repetitions are produced with the "true" durations $x_1, \ldots, x_i, \ldots, x_N$ and $y_1, \ldots, y_i, \ldots, y_N$, respectively. If the boundary between X and Y is randomly mislocated, there will be a measurement error $W = w_1, \ldots, w_i, \ldots, w_N$, such that the measured durations of the two intervals will be $x_1 + w_1, \ldots, x_i + w_i, \ldots, x_N + w_N$ and $y_1 - w_1, \ldots, y_i - w_i, \ldots, y_N - w_N$, respectively (see Fig. 1).

Assume the measurement error W is nonsystematic,

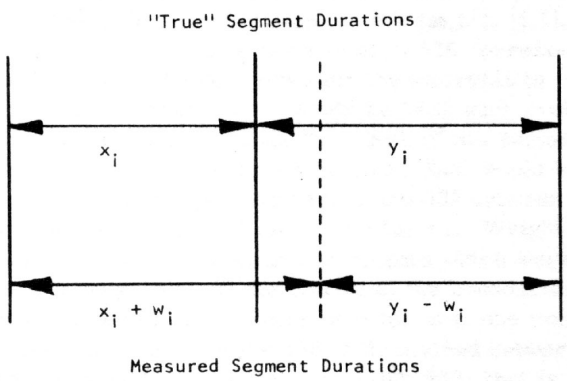

FIG. 1. Boundary mislocation of adjacent speech intervals X and Y.

i.e., that the expected values of X, Y, and W are

$$E\{X\}=\frac{1}{N}\sum_{i=1}^{N} x_i = m_X \ , \tag{1}$$

$$E\{Y\}=\frac{1}{N}\sum_{i=1}^{N} y_i = m_Y \ , \tag{2}$$

and

$$E\{W\}=\frac{1}{N}\sum_{i=1}^{N} w_i = 0 \ , \tag{3}$$

and therefore that

$$E\{X+W\}=E\{X\}=m_X \tag{4}$$

and

$$E\{Y-W\}=E\{Y\}=m_Y \ . \tag{5}$$

Then the covariance (COV) between the adjacent measured intervals $X+W$ and $Y-W$ would be

$$\mathrm{COV}(X+W, Y-W)=\mathrm{COV}(X, Y)+\mathrm{COV}(X,-W) + \mathrm{COV}(W, Y)+\mathrm{COV}(W,-W) \ . \tag{6}$$

If the measurement error W is independent of both X and Y, then

$$\mathrm{COV}(X,-W)=0 \tag{7}$$

and

$$\mathrm{COV}(W, Y)=0 \ , \tag{8}$$

and, of course,

$$\mathrm{COV}(W,-W)=-V(W) \ , \tag{9}$$

where V is the variance.

Therefore Eq. (6) reduces to Eq. (10):

$$\mathrm{COV}(X+W, Y-W)=\mathrm{COV}(X, Y)-V(W) \ . \tag{10}$$

The correlation coefficient between the two measured intervals is

$$r_{X+W,\,Y-W}=\frac{\mathrm{COV}(X+W, Y-W)}{[V(X+W)V(Y-W)]^{1/2}} = \frac{\mathrm{COV}(X, Y)-V(W)}{\{[V(X)+V(W)][V(Y)+V(W)]\}^{1/2}} \ ; \tag{11}$$

by Eq. (10) and the classical statistical relation (12):

$$V(A+B)=V(A)+V(B)+2\,\mathrm{COV}(A, B) \ , \tag{12}$$

and the assumption that the measurement error W is independent of both X and Y.

Now if the segments X and Y are independent of each other, then

$$\mathrm{COV}(X, Y)=0 \tag{13}$$

and Eq. (11) reduces to Eq. (14).

$$r_{X+W,\,Y-W}=-\frac{V(W)}{\{[V(X)+V(W)][V(X)+V(W)]\}^{1/2}} \le 0 \ . \tag{14}$$

For example, if $V(X)=V(Y)=V(W)$ and $\mathrm{COV}(X, Y)=0$, then

$$r_{X+W,\,Y-W}=-\frac{V(W)}{[2V(W)2V(W)]^{1/2}}=-\frac{V(W)}{2V(W)}=-0.5 \ .$$

The negative correlation of -0.5 is due entirely to the measurement error. Measurement error will always contribute some negativity to the correlation coefficient but the magnitude of this effect depends on the relative magnitudes of the variances of the intervals X and Y and the measurement error W. To consider another example, if $V(X)=V(Y)=100$ msec^2 and $V(W)=25$ msec^2 (i.e., $\sigma_X=10$ msec and $\sigma_W=5$ msec) then

$$r_{X+W,\,Y-W}=-0.2 \ .$$

It is clear that the variance of the measurement error has to be much less than the variance of the intervals studied if one shall be justified in neglecting the influence of the measurement error on the correlation coefficient. This strikes us as a virtual impossibility no matter how carefully one segments and measures the speech signal, especially when one considers that measurement error in this case is caused not only by slips of the pencil or calipers in segmenting the speech wave, but also by not knowing where the speaker's brain considers the boundaries between the speech units to be.[2] Wright assumes the boundaries are at discontinuities in the acoustic speech signal. But can anyone be sure that the brain does not use other, less obvious, boundaries, e.g., peaks in the tension of some speech muscles or perhaps in the velocity function of articulators, neither of which would necessarily coincide with any clearly definable acoustic boundary?

In general there is no sure way to differentiate between negative correlations caused by temporal compensation and those caused by measurement error; however, we might expect that negative correlations due to measurement error would show up more between segments which share an indistinct boundary, since it would be more difficult to locate such a boundary reliably from one token of an utterance to the next. The test sentence Wright used, "To catch a fish is to hook it," would probably be spoken with stress on the words "catch," "fish," and "hook," the remaining words being unstressed. As unstressed syllables are spoken less distinctly than stressed ones, their boundaries should be harder to find and thus they ought to show higher negative correlations between segments on either side of the boundaries. Also, the boundary between [h] and surrounding vowels, which is notoriously difficult to find, especially if the [h] becomes partially voiced, would also be expected to be associated with higher negative correlations. The only clear boundaries, then, would probably be those between [kʰ, æ], [æ, tʃ], [f, I], [I, ʃ], and [U, k]. As there are a total of 816 correlation coefficients to be compared for the intervals in Wright's test sentence, there would be 1632 such coefficients to examine in the data from both of his subjects, of which 1100 would be between intervals that would be supposed to have unclear boundaries and 532 between intervals having relatively clear boundaries. Wright found 176 of these correlation coefficients which were significantly negative. If the clarity of the boundaries did not influence the correlation coefficients, one would expect the distribution of the 176 to be divided between the two boundary types in the ratio 1100 : 532, that is 118.6 : 57.4. The actual values reported by Wright are

153 significant negative correlation coefficients between intervals sharing what we might suppose are unclear boundaries and 23 between intervals sharing the clear boundaries. The difference between the observed and expected values is highly significant, $p \ll 0.001$.

Thus, the fact that Wright found more negative correlations in the latter half of his sentence does not show that "...temporal compensations...tended to be accomplished utterance finally...," only that the latter half of the sentence contained more fuzzy boundaries than the first half.

II. THE EFFECT OF NORMALIZATION

Furthermore, in the pilot experiment, some of the negative correlations Wright "discovered" he in fact put there himself. Wright started with 78 repetitions of the phrase he used and 68 repetitions of the sentence. However, for final statistical analysis he used only 54 tokens of the phrase which centered about the mean, discarding the rest. Similarly, the final analysis of the sentence involved only 48 tokens closest to the mean. The effect of this "normalization" procedure on the correlations between intervals can be gauged by considering the classical statistical relation (15), which is a more general form of Eq. (12) above.

For the variables $X_1, \ldots, X_i, \ldots, X_N$,

$$V\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} V(X_i) + 2\sum_{i<j} \mathrm{COV}(X_i, X_j) \,. \qquad (15)$$

The normalization procedure of discarding samples of sentence durations which were very far from the mean has the effect of limiting or reducing the variance of the utterance durations, i.e., of reducing the leftmost term in Eq. (15) without affecting, as much, the second term. Thus, if the equation is to balance, the rightmost term will be reduced as well. The covariances will exhibit more negativity than they would have before the normalization procedure.

For further discussion of the common pitfalls encountered in using these and similar statistical techniques to discover the underlying mechanism of speech timing, see Ohala.[3]

III. CONCLUSION

For the reasons mentioned, Wright's negative correlations reveal nothing about how speech timing is regulated. We must emphasize that we are not saying that Wright's claim that speech timing uses both "chaining" and "preplanning" is wrong; only that his data give no evidence on the point.

*This work was supported in part by the National Science Foundation.

[1]T. W. Wright, "Temporal interactions within a phrase and sentence context," J. Acoust. Soc. Am. **56**, 1258–1265 (1974).

[2]V. A. Kozhevnikov and L. A. Chistovich, "Speech: articulation and perception," U. S. Dept. of Commerce, JPRS 30, 543 (1965), pp. 101 ff.

[3]J. J. Ohala, "The temporal regulation of speech," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London, 1975) pp. 431–453.