# Temporal Organization of Spoken Language*

Ilse Lehiste

*This study was supported by PHS Research Grant No. 1 R03 MH-18122-01 from the National Institute of Mental Health.

# Temporal Organization of Spoken Language

Ilse Lehiste

This paper reports the results of a pilot study dealing with the temporal organization of spoken language. In particular, it deals with the temporal structure of monosyllabic and disyllabic words in English.

It is assumed in this study that the production and perception of spoken language takes place in terms of phonological units. These units may be of various sizes, ranging hierarchically from a single speech sound through syllables and phonological words to phonological phrases. Evidence for the existence of such units comes from various sources, for example from studies of coarticulation (Öhman, 1967; MacNeilage and DeClerk, 1969). Another source of evidence is the study of suprasegmental patterns (Lehiste, 1970). All suprasegmental patterns are patterns in time; any contrastive arrangement of fundamental frequency or intensity is crucially dependent on the time dimension. The arrangement of articulatory events along the time dimension may likewise have suprasegmental function, and may serve to establish higher-level phonological units.

One way in which a phonological unit may be specified is with reference to its temporal organization. Several recent studies (Kozhevnikov and Chistovich, 1965; Slis, 1968) have shown that when a

speaker repeats the same utterance many times, at the same rate of articulation, the durations of adjacent phonemes are quite strongly negatively correlated. Thus, if an error is made in the duration of one phoneme, the error is largely compensated for in the following phoneme, which finishes at the originally planned time, despite the fact that it started late. This negative correlation suggests that articulatory events are programmed, at some (here unspecified) higher level, not in terms of single phonemes, but in terms of higher-level articulatory units. One way to determine the extent of these higher-level units would be to establish the domain over which such temporal compensation takes place, since it seems reasonable to assume that the sequences of sounds which are subject to temporal compensation constitute a single articulatory program.

The question might now be asked whether such articulatory units (defined as the domain of a single articulatory program) are universal or language-specific. Different researchers, working with CVC-sequences in different languages, have found a closer correlation between either the initial CV sequence (Russian, Kozhevnikov and Chistovich, 1965), or between the VC sequence (Dutch, Slis, 1968). The observations regarding English which are reported in this paper support the view that in English, there is a closer connection between a vowel and a following consonant than between an initial consonant and a following vowel.

While this question is of intrinsic importance, it would be still more interesting to know whether two phonemically identical, but morphologically different linguistic items have identical time programs.

An example might be provided by the word pair weighed and wade, the first being the past tense of the verb to weigh, the second the infinitive of the verb wade. The past tense form contains two morphemes: the verbal stem weigh and the past tense marker -d. The word wade is monomorphemic. If the two words weighed and wade are produced with identical timing patterns, one may assume that the morphological process of assigning the past tense marker to weigh has taken place at a level which precedes the programming of motor commands for the realization of the phonemic sequence, which, according to traditional descriptions, is common for both weighed and wade. On the other hand, a difference in the temporal organization of the two sequences might indicate a difference in the level at which the utterance, about to be generated, is converted into a sequence of motor commands.

The specific aim of this pilot study was to test the temporal compensation hypothesis for English, and to establish the domain over which temporal compensation takes place.

I selected a set of ten words: steed, staid, stayed, stead, skid, skit, stay, steady, skiddy, and skitty. The words were chosen to provide an opportunity to study several different aspects of the problem, and also for the sake of relative ease of processing. I intended to analyze the tapes by means of a pitch meter and an intensity meter, and display the curves on a Mingograph. The initial clusters /st/ and /sk/ were selected because it is relatively easy to measure the duration of an initial /s/ from intensity curves with high-frequency pre-emphasis. The plosive following an initial /s/ is unaspirated in

English, which makes it possible to establish the duration of the plosive and the onset of the vowel with considerable precision. The set of words contains the pair staid ('stodgy') and stayed (past tense of the verb stay, which was likewise included), providing a chance to compare two words with identical phonological structure, but different morphological structure. The three disyllabic words steady, skiddy, and skitty are derived from the monosyllabic words stead, skid and skit through another morphological process--the suffixation of -y. I was interested in this particular word type, because in the Midwestern dialect of American English these words would normally be pronounced with a so-called 'voiced /t/'--a flapped allophone of the sounds that are realized initially as /t/ and /d/. The flapped /t/ occurs only intervocalically; its occurrence signals that another vowel has to follow, and I interpret this to mean that the articulatory program must obligatorily encompass the whole CVC sequence.

Each of these ten words was recorded by two subjects, who repeated the word approximately 110 times at what was deemed a subjectively constant rate. The speakers were selected solely on the basis of their dialect: the Midwestern variety of General American, in which flapped allophones of /t/ and /d/ are the rule rather than exception. In other respects the two speakers differed a great deal. Speaker DS has a lowpitched (male) voice; he speaks slowly and steadily, with a clearly developed rhythm and fairly equal spacings between the productions of individual tokens of the test words. Subject JK, a high-pitched female speaker, speaks very fast and irregularly; she speeds up and slows down within a list of words, and is apparently unable to control

her rate of articulation very well, although it turned out that changes in tempo were mostly reflected in the spacings between the words rather than in the duration of the words themselves. Considering the great difference between the speakers with regard to the spacings between words, it was quite surprising that the results of the temporal compensation study turned out as similar for the two speakers as they did; however, I intend to control the rate of articulation much more rigorously in recording further subjects.

The recordings were processed through a pitch meter and intensity meter (designed by Børge Frøkjaer-Jensen, Copenhagen) and displayed on a Mingograph (Elema-Schönander, Stockholm). The output of a Mingograph is a set of time-correlated curves and an oscillogram, from which quite reliable time determinations can be made for each segment. Some decisions had to be arbitrary--for example, in the word stay I considered the peak of the last non-laryngealized vocal fold flap to constitute the end of the utterance. With a paper speed of 10 centimeters per second, one millimeter corresponds to 10 milliseconds. The precision of measurement depended ultimately on the width of a pencil line drawn to indicate segment boundaries; the final results are given in milliseconds, but the measurements are probably accurate within two or three milliseconds rather than half a millisecond which the numbers might imply. Tokens which for some reason were not easily measurable were not included in the calculations.

After making the measurements, I calculated the following for each set of test words: the average duration of each segment; the variance for each segment; the relative variance; and the standard

deviation. Relative variance, a concept recently introduced by George Allen (Allen, 1969), is simply variance divided by average duration. By taking into account differences in the average duration of segments, relative variance provides a good measure of articulatory variability.

The goal of the study was to test whether there was any temporal compensation within sequences of segments. By assumption, a negative correlation between the durations of two successive segments was taken to imply that the two are programmed as a unit at some higher level at which articulatory sequences are programmed. There is a negative correlation between the durations of two segments, if the variance of the duration of the sequence of two segments is less than the sum of the variances of the segments considered separately.

On the other hand, a positive correlation reflects the influence of changing tempo: if the rate of articulation increases, all segments are shortened, although not necessarily at the same rate, and conversely, if the rate of articulation decreases, all segments are lengthened. It is possible to eliminate or reduce tempo effects by a normalization procedure which I did not employ in this pilot study, but intend to use during later stages of the project of which this article constitutes the first report.

I calculated the variances of all individual segments and of all successive pairs of sounds. In addition, I treated the initial cluster as a unit and calculated the variance of the sequence consisting of the initial cluster and the following vowel. I also calculated the variance for the whole word, and compared it with the sum of variances for the individual segments. To compensate for the differences

between average durations, I calculated the relative variances by dividing variances by average durations. Table I summarizes the results for the seven monosyllabic words for speaker DS.

Table I

Difference between the relative variances of successive segments taken individually and considered as a co-articulated sequence, calculated on the basis of monosyllabic words produced by DS.

| Word | $C_1C_2$ | $C_2V$ | $C_1C_2V$ | $VC_3$ | $C_1C_2VC_3$ |
|---|---|---|---|---|---|
| steed | -0.73 | -0.38 | -0.58 | -0.66 | -0.79 |
| staid | +0.07 | +0.26 | +0.13 | -0.48 | -0.05 |
| stayed | +0.48 | +0.14 | +0.26 | -1.07 | -0.79 |
| stead | -0.40 | -1.29 | -0.36 | -4.17 | -2.80 |
| skid | -0.06 | -0.33 | -0.32 | -0.27 | -1.14 |
| skit | +0.01 | -0.20 | -0.20 | -1.26 | -0.53 |
| stay | -0.55 | -0.12 | -0.08 | | |

The entries in the table represent the difference between the relative variances of successive segments taken individually (for example, the first consonant and the second consonant) and considered as a coarticulated sequence (for example, the initial cluster). A consideration of some entries in the first row will illustrate the procedure. The first number, -0.73, is the difference between the relative variances of the two consonants /s/ and /t/ taken separately and /st/ considered as a coarticulated cluster. The sum of variances

for /s/ and /t/ was 1,136.58; the variance of the /st/ cluster was 954.10. The average duration, of course, was the same in both cases, and amounted to 251 milliseconds. The relative variance for the sum was 1,136.58 divided by 251, which is 4.53; the relative variance for the cluster was 954.10 divided by 251, which is 3.80. The difference between 4.53 and 3.80 is 0.73; the minus sign indicates that the relative variance for the cluster was smaller than the relative variance for the sum of segments, which means that temporal compensation was present and there was a negative correlation between the durations of /s/ and /t/. In the 106 measurable productions of this word, there was obviously a certain amount of temporal compensation between each successive pair of segments, as well as within the whole word, as shown by the negative entries in all columns.

Now the results obtained for this first word would not solve the question whether there is a closer correlation between an initial consonant and a following vowel, or between a vowel and a following consonant. Temporal compensation was present between all successive pairs of sounds; unless we had a way of evaluating the significance of degrees of correlation, it would be impossible to conclude which of the sequences constitutes a more closely coarticulated unit. I have in fact calculated Pearson correlations for many of the pairs, some of which will be presented below; but I am not sure they are very meaningful, and for the following reason. It so happens that there may be a statistically significant negative correlation between /s/ and /t/ in the word steed; but there is a positive correlation, likewise significant, between /s/ and /t/ in the word stayed, recorded

during the same session. Steed and stayed have exactly the same amount of temporal compensation within the whole word (-0.79 in the last column of Table I for both steed and stayed). It seems to me that one should compare not only the correlations within each word, but also the patterns produced within one recording session; in other words, not only the entries within a row, but also the analogous entries within each column. What seems significant to me is the fact that we find both positive and negative correlations in all columns except the two last ones. Within this recording session, there was always a negative correlation present between the vowel and the following consonant, and within the whole monosyllabic word.

Table II presents the same data for the second speaker.

Table II

Difference between the relative variances of successive segments taken individually and considered as a co-articulated sequence, calculated on the basis of monosyllabic words produced by JK.

| Word | $C_1C_2$ | $C_2V$ | $C_1C_2V$ | $VC_3$ | $C_1C_2VC_3$ |
|---|---|---|---|---|---|
| steed | +0.20 | -0.08 | +0.35 | -0.19 | +0.22 |
| staid | -1.49 | +0.29 | -0.28 | -0.36 | -0.50 |
| stayed | +0.09 | +0.31 | +0.49 | -0.09 | +0.36 |
| stead | -0.13 | +0.45 | +0.35 | -0.58 | -0.25 |
| skid | +0.17 | -0.17 | -0.09 | -0.22 | -0.01 |
| skit | -0.19 | -0.33 | -0.13 | -1.22 | -0.94 |
| stay | +0.23 | +0.45 | +0.47 | | |

As was mentioned above, the second subject was a highly irregular speaker, who varied her speaking tempo to a much greater extent than the first. One might thus expect a greater amount of positive correlation, or perhaps a lesser degree of negative correlation, reflecting the influence of changing tempo. And indeed, the number of instances of positive correlation was doubled for this speaker. These were not simply additional cases; a comparison of the matrices for the two speakers shows that the pluses and minuses do not necessarily occur in the same slots. The one thing that is regular is the negative correlations in the next but last column, showing temporal compensation between a vowel and the following consonant. The tendency for negative correlation here was evidently strong enough to resist the influence of changes in tempo.

Table III presents similar data for the disyllabic words of speaker DS.

Table III

Difference between the relative variances of successive segments taken individually and considered as a co-articulated sequence, calculated on the basis of disyllabic words produced by DS.

| Word | $C_1C_2$ | $C_2V_1$ | $C_1C_2V_1$ | $V_1C_3$ | $C_3V_2$ | $V_1V_2$ | $C_1C_2V_1C_3V_2$ |
|---|---|---|---|---|---|---|---|
| steady | -0.55 | +0.03 | -0.31 | -0.13 | +0.18 | -0.53 | -0.70 |
| skiddy | -0.04 | -0.09 | -0.22 | -0.32 | -0.61 | -0.38 | -0.92 |
| skitty | +0.33 | -0.03 | +0.35 | +0.01 | -0.37 | -0.73 | -0.86 |

As may be seen from the table, the intervocalic flapped /t/ does not

seem to have any closer correlation with either the preceding or the following vowel; the values in the fourth and fifth column show both positive and negative correlation, and no obvious pattern emerges. The last column shows a considerable degree of interaction within the whole disyllabic word, as had been the case for this speaker also with monosyllabic words. The next but last column shows that there was also a temporal compensation (i.e. negative correlation) between the durations of the two vowels. If this can be substantiated by further research, it seems that in such disyllabic words, the duration of the second vowel is adjusted to the duration of the first, and the sequence of two vowels constitutes a unit of programming at some higher level. Unfortunately the second speaker's results are very confusing, and the conclusion is therefore even more tentative than the other conclusions drawn on the basis of this exploratory study.

Table IV presents Pearson correlations between the syllable nucleus and the final consonant in the monosyllabic test words produced by the two speakers.

Table IV

Pearson correlations* between the syllable nucleus and the final consonant in monosyllabic words produced by speakers DS and JK.

| Word | Speaker DS | Speaker JK |
|---|---|---|
| steed | -0.35 | -0.18 |
| staid | -0.37 | -0.33 |
| stayed | -0.27 | -0.25 |
| stead | -0.76 | -0.47 |
| skid | -0.10 | -0.57 |
| skit | -0.38 | -0.61 |

* $$r = \frac{\sum_{i=1}^{n} \frac{x_i y_i}{n} - M_x M_y}{s_x \; y_x}$$

As may be remembered, both speakers had negative correlations in all test words between this pair of sounds. These data are presented for the sake of possible comparison with the relative variances; I hesitate to draw any conclusions from the difference in degree of negative correlation on the basis of this material alone, without consideration of the relationships between other segments within the word. Other factors have to be included in the consideration; for example, speaker DS always had a much larger standard deviation for the duration of the final consonant than for the duration of the syllable nucleus, while speaker JK's standard deviations showed no

such pattern. Clearly for speaker DS final position influenced the variability of the duration of a segment in such a way as to make the two standard deviations non-comparable.

The results of the study thus indicate rather strongly that in English, there is a close interaction between the durations of vowels and following consonants in monosyllabic words, and between the durations of all the sounds within a monosyllabic or disyllabic utterance. This seems to provide some independent phonetic evidence for the existence of phonological words, which I would like to define as the domain over which such temporal compensation takes place. There is further evidence for the existence of such phonological units in the average durations of segments within a word during one recording session. A comparison of these average durations shows very interesting compensatory effects.

Table V shows the average duration of segments and words in the four monosyllabic words steed, staid, stayed, and stay, produced by speaker DS.

Table V

Average durations of segments (in milliseconds) in four monosyllabic words produced by speaker DS. N = number of tokens.

| Word | N | $C_1$ | $C_2$ | V | $C_3$ | Total |
|---|---|---|---|---|---|---|
| steed | 106 | 130 | 121 | 301 | 168 | 720 |
| staid | 110 | 119 | 96 | 330 | 167 | 712 |
| stayed | 111 | 125 | 96 | 330 | 151 | 702 |
| stead | 110 | 133 | 123 | 307 | 149 | 712 |

It is obvious that for this speaker, the word constituted a unit of timing. Compare, for example, the relative arrangement of the durations of the segments in steed and staid. There is a difference in the intrinsic durations of /i/ and /e$^{I}$/; all other factors being kept constant, /e$^{I}$/ is longer than /i/. However, the greater length of /e$^{I}$/ was clearly compensated for in the shorter duration of the initial cluster; the difference in the durations of the words is very much smaller than the difference in the durations of the vocalic syllable nuclei. On the other hand, the absence of a final /d/ in stay was accompanied by lengthening of both members of the initial cluster.

Coming back to the question of whether there is any difference between bimorphemic and monomorphemic words of the same phonemic structure, I must say that very little, if anything, can be concluded from a comparison of the words stayed and staid. Speaker DS had a

difference of relative variances of -0.48 between the syllable nucleus and the final consonant in staid and -1.07 in stayed, the Pearson correlations being -0.37 and -0.27 respectively. The two ways of expressing negative correlation provide contradictory evidence in this case. For speaker JK, the difference in relative variances was -0.36 for staid and -0.09 for stayed; the Pearson correlations were -0.33 and -0.25. This might be interpreted to mean that there was a higher degree of cohesiveness between the syllable nucleus and the final consonant in the monomorphemic word. However, these results should be compared with the difference in relative variances in the whole $C_1C_2VC_3$ sequence. For speaker DS, the word stayed considered as a whole had a much greater degree of temporal compensation than staid. For speaker JK, the situation was exactly opposite: stayed showed positive correlation, while staid showed negative correlation. Unless some further evidence is provided by later stages of the study, it must be concluded that the morphemic structure of a word does not have any influence on its temporal organization in English.

Table VI compares stead with steady, skid with skiddy, and skit with skitty, again for speaker DS.

Table VI

Comparison of average durations (in milliseconds) of segments in three monomorphemic and three bimorphemic words, produced by speaker DS.

| Word | $C_1$ | $C_2$ | $V_1$ | $C_3$ | $V_2$ | Total |
|---|---|---|---|---|---|---|
| stead | 133 | 123 | 307 | 149 | | 712 |
| steady | 94 | 98 | 133 | 20 | 173 | 518 |
| skid | 148 | 104 | 217 | 151 | | 620 |
| skiddy | 128 | 97 | 90 | 28 | 166 | 509 |
| skit | 156 | 104 | 185 | 115 | | 560 |
| skitty | 110 | 87 | 83 | 23 | 151 | 454 |

It is interesting to observe that in each case, the disyllabic word was shorter than the corresponding monosyllabic one, and that the shortening regularly involved the initial cluster. As was mentioned above, the two vowels of disyllabic words of this type are quite strongly negatively correlated. The observation might be added now that although skid and skiddy are longer than skit and skitty, the ratio between the durations of the two vowels in skiddy and skitty is practically identical: 0.54 for skiddy and 0.55 for skitty. The corresponding ratios for the other speaker were 0.89 and 0.84 respectively. Both speakers had a considerably different ratio between the two vowels in steady (although there was temporal compensation present between them); steady evidently constituted a different

disyllabic word type, although it too contained a flapped /t/.

Let us now return to the question regarding the relationship between morphological structure and phonological structure. Within morphology, steady, skiddy, and skitty are derived from the respective base forms by the addition of the derivative suffix -y, which produces adjectives from nouns. Within phonetically manifested phonology, we are not simply adding an [i] to the monosyllabic words stead, skid, and skit. For one thing, the bimorphemic words, which also contain a greater number of segments than the monomorphemic words, are consistently shorter, although one might expect them to be longer by something like the average duration of the final [i]. The bimorphemic words are realized as higher-level phonological units with some clearly definable phonetic properties of their own, such as the ratio between the vowels and temporal compensation between the two vowels rather than between the stem vowel and the following consonant. It is obvious that a simple distinctive features description, as might be given in a distinctive feature matrix constructed for the basic and the derived forms, would not reveal the essential differences in the temporal structure of the two word types.

This study of temporal compensation has thus produced evidence not only for the existence of temporal compensation between certain pairs of segments, but also within all the segments that constitute a word. I have tried earlier--in my studies of juncture--to define a phonological unit with reference to its boundaries; this is the first time I have found something to characterize a word as a whole, not by reference to its boundaries, but through the internal cohesiveness

of its component parts. And this appears to be a promising direction for future research.

## References

Allen, George D. 1969. Structure of timing in speech production. Paper presented at the San Diego meeting of the Acoustical Society of America, November 4.

Kozhevnikov, V.A., and L. A. Chistovich. 1965. Speech: Articulation and perception. Translated by J.P.R.S., Washington, D.C., No. JPRS 30,543. Moscow-Leningrad.

Lehiste, Ilse. 1970. Suprasegmentals. Cambridge; M.I.T. Press.

MacNeilage, Peter F., and Joseph DeClerk. 1969. On the motor control of coarticulation in CVC monosyllables. Journal of the Acoustical Society of America 45.1217-1233.

Öhman, S.E.G. 1967. Numerical model of coarticulation. Journal of the Acoustical Society of America 41.310-320.

Slis, I. H. 1968. Experiments on consonant duration related to the time structure of isolated words. IPO Annual Progress Report, No. 3.71-80. Institute for Perception Research, Eindhoven, Holland.