Phonetica
Editor: K. Kohler, Kiel

Separatum
Publisher: S. Karger AG, Basel
Printed in Switzerland

Phonetica 1988; 45: 175–197

Motor Programs and Hierarchical Organization in the Control of Rapid Speech

Saul Sternberg[a], *Ronald L. Knoll*[b], *Stephen Monsell*[c], *Charles E. Wright*[d]

[a] Department of Psychology, University of Pennsylvania, Philadelphia, PA., USA;
[b] AT&T Bell Laboratories, Murray Hill, N.J., USA;
[c] Department of Experimental Psychology, University of Cambridge, Cambridge, U.K.;
[d] Department of Psychology, Columbia University, New York, N.Y., USA

Abstract. We provide a summary of our recent research on the control of rapid action sequences in speech production, emphasizing findings about the advance planning and hierarchical organization of such utterances. The effects of number of elements in the utterance (its 'length') and other factors on maximum production rates of short utterances lead us to infer that a 'motor program' for the whole utterance, prepared in advance, controls the execution of each of its 'units'. Findings from studies of typewriting as well as speech production have led us to a model in which the performance of each unit is controlled by two processes arranged in sequence: one *(subprogram selection)* whose duration increases linearly with sequence length, and the other *(command)* whose duration depends on type of unit. Quantitative aspects of the production of utterances composed of different types of element suggest that the action unit in speech is the stress group or metrical foot. The virtual identity of the timing of word and nonword utterances implies that the utterance program is sufficiently detailed so it can be executed without reference to learned routines for words stored elsewhere in memory. We review our search for properties of performance that are suggested by the model: First, the time from a reaction signal to the first unit (the latency) increases linearly with utterance length. Second, the maximum length utterance controlled by one program depends on unit size. Third, the effect of utterance length on production timing is localized (intermittent), rather than affecting all parts of the articulatory stream. And fourth, the effect of utterance length on production timing appears in just one epoch per unit.

1. Introduction

1.1 Investigation of the Control of Action Sequences

The work described in the present paper is part of a program of research aimed at understanding how people organize and control rapid sequences of actions. We have been studying performance by skilled subjects in two domains, speech production and typewriting, and have performed similar experiments in the two domains, with similar

outcomes. In typewriting we have studied the temporal aspects of short, rapidly produced sequences of keystrokes. In our work on speech production, some of which we review here, we have investigated rapid utterances composed of up to twelve words or pseudowords. Performances in these domains are remarkable feats, in which people produce complex and precisely coordinated movements at very high rates – as fast as nine keystrokes or spoken syllables per second – rates that appear too high to permit much processing of information or planning between one action and the next.

One salient difference between typing and speech is that whereas the rapid performance we study is a normal aspect of typing skill, it is probably less 'natural' for our speakers. Similarity of outcomes in the two domains, however, shows that our principal effects are not sensitive to this difference, and raises the question, touched on below, whether the same mechanisms underlie rapid and natural speech.

1.2 Advance Planning of Whole Sequences: Motor Programs

Since Lashley's [1951] time, one of the ideas often discussed in relation to skilled action is the idea of the advance planning of entire sequences. It has been suggested that in fluent rapid action there is too little time for the specification of each element to be generated after the previous element has been executed, as in an associative-chain mechanism [e. g., Keele and Summers, 1976]. Instead it has been argued that there must exist a *motor program* (for speech, an *utterance program)* for a whole sequence of actions – a detailed and integrated specification that is established before the sequence begins and that controls its execution.

The length of an action sequence controlled by one program probably has a maximum (see section 4.2). Long sequences might then consist of a series of short subsequences, each controlled by a separate motor program. How these programs would be concatenated or otherwise integrated is unknown; the implications of such a structure of multiple programs have been little discussed [see, however, Butterworth, 1980; Fujimura, 1987, section 4.4].

Some investigators [e. g., Keele, 1968] have taken the view that advance planning generates only command sequences that can be executed without any feedback ('open loop'), rather than, for example, programs that include instructions for sensing and responding to feedback, programs that can themselves be altered in response to feedback, or even programs that consist of ordered sets of 'response images' to which feedback from the movement sequence is compared. We believe that it is inappropriate to restrict the idea of 'program' to cases of sequence control without feedback. Questions about the existence and extent of advance planning are separable from questions about precisely what roles are played by feedback.

Suppose there *was* such a mechanism: How could we study it and determine its properties? To find evidence for the existence of such motor programs has been one of the principal objectives of our work. We shall present data that make sense in terms of a simple notion of how a motor program might be used, and what it is.

Given this objective, we have chosen experimental procedures that appear to maximize the opportunity and likelihood of the advance planning of entire sequences. First, we generally use *short* sequences, which contain up to five or six spoken words and have durations of no more than about 1 s. We hope this is short enough so the entire sequence might be controlled by a single motor program. (The existence of a substantial 'eye-hand span' in copy typing and the 'eye-voice span' in oral reading [Butsch, 1932; Levin, 1979] makes it plausible that

planned sequences with a duration of about 1 s may be used in such tasks.) Second, we specify the sequence well in advance, to provide an opportunity for advance planning. This also permits us to study the organization of output uncontaminated by the processing of external stimuli. Third, we require rapid execution, because we believe it is in rapid, fluent performance where advance planning is most likely to be required, or at least where it is most useful. Equally important, we believe that by forcing performance to its limiting speed we are more likely to discover its fundamental constraints.

1.3 Hierarchical Structure of Sequences: Units and Subunits

A second idea often discussed in relation to skilled performance – one that dates back to the turn of the century – is that fluently performed action sequences tend to be structured hierarchically [Book, 1908; Gallistel, 1980; Johnson, 1972; Miller et al., 1960]. Thus, for purposes of their execution, disjoint subsequences of actions are thought to be organized into higher-level action units, larger than what is defined as a 'single action', but smaller than the complete sequence controlled by a program.

The existence of intermediate-length subsequence units defines the simplest (most shallow) sequence structure that can be described as hierarchical, a structure with only one level between sequence and 'single action'. In a deeper hierarchy the subsequences would be further partitioned into smaller disjoint subsequences, also larger than single actions. Note that the existence of intermediate-length subsequence units is only one of several meanings of the term 'hierarchy' in relation to motor control. The term has also been used to refer to the existence of representations of movement at different 'levels' of specificity, where more detailed aspects of control are allocated to 'lower' levels, closer to execution [Greene, 1972; Saltzman, 1979; Szentagothai and Arbib, 1975]. For the skilled production of learned action sequences there is a surprisingly large disparity between the enthusiasm for the hypothesis of hierarchical organization [Gallistel, 1980, chapter 12; Johnson, 1972; Keele, 1987; MacKay, 1982; Miller et al., 1960] and the persuasiveness of the evidence cited for this hypothesis. For example, Bryan and Harter's [1899] claims were for hierarchical structure in the receipt, not the transmitting, of Morse code, and other claimed evidence does not exclude a perceptual locus [Leonard and Newman, 1964]. The often-cited conclusions of Book [1908] about multiple-stroke units in typewriting depend exclusively on introspective evidence.

With respect to the evidence from speech errors cited for hierarchical structure [Dell, 1984; Fromkin, 1971, 1981; Shattuck-Hufnagel, 1983], the distinction has often not been made between the idea that *some* representation during the evolution of an utterance contains intermediate-size subsequence units, and the idea that this is true of the ('final') representation that is actually used in producing the utterance. Thus, it is important to distinguish between *units of planning* and *units of execution* [Monsell and Sternberg, 1981; Rosenbaum et al., 1983]. Errors may arise during either or both planning (which might best be understood in terms of a linguistic hierarchy, composed of entities such as phonemes, words, and syntactic units), and execution (which might best be understood in terms of a production hierarchy composed of entities such as onsets, rhymes, syllables, feet, and prosodic units). Several investigators have recently expressed skepticism about hierarchical sequence control, in both learned skills [Marteniuk and Romanow, 1983; Namikas, 1983] and locomotion [Wetzel and Howell, 1981]. However, some evidence consistent with hierarchical structure of the final representation of action sequences has recently begun to emerge [Collard and Povel, 1982; Povel and Collard, 1982; Gordon and Meyer, 1987; Knoll and Sternberg, 1987; Rosenbaum et al., 1983, but also Klein, 1983, and Rosenbaum, 1983; Vorberg and Hambuch, 1984].

What would be the consequences if two or more action segments were part of the same higher-level unit? *Are* there multiple-action units in rapid sequences, and if so,

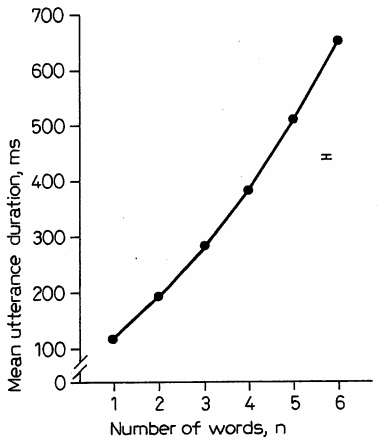

Fig. 1. Mean utterance duration for sequences of random digit or letter names, estimate of ± SE, and fitted quadratic function. Results are averaged over 4 subjects, 2 days, and two vocabularies; about 240 observations per point. The fitted function is $61 + 57n + 8n(n-1)$ ms.

what are they? A second objective of our work has been to seek evidence for such hierarchical structure and to identify units of action, as well as to define what 'unit' might mean.

1.4 Experimental Procedure

The typical events on a trial are as follows: a list of letters, digits, words, or pseudowords is displayed, often sequentially. There is an interval of 3 or 4 s between the display and the reaction signal. In most experiments the reaction signal is preceded by two countdown signals, to maximize the subject's certainty about when it might occur; in some experiments the foreperiod is deliberately varied. (Main features of the results are the same.) To discourage anticipations we omit the reaction signal on some trials. The subject's task is to recite the sequence so as to complete his or her response as fast as possible after the reaction signal, while maintaining adequate intelligibility; we use feedback and payoffs to encourage such performance. Subjects are informed about any errors after each trial, and about the average time from signal to response completion after each block of about 20 trials; they are encouraged by scores and cash bonuses to respond as rapidly as possible consistent with high accuracy. We measure the latency (interval between signal and speech onset) and the duration of the response and in some experiments also use a record of the acoustic signal, supplemented by sound spectographs, to measure the timing of words, syllables, and smaller segments. To define speech onset we have typically used an energy threshold that must be exceeded by the acoustic signal for no less than a brief critical duration; to define offset we have required that the energy drop below threshold and remain there for no less than a (longer) critical duration. Words are generally either randomized or balanced over utterance length and serial position.

In the sections below we first describe several findings about production rates in speech. Then we introduce a model that explains these findings, and discuss additional experiments that permit us to elaborate it. Finally, we review our search for four properties of performance suggested by the model.

2. Some Production-Rate Phenomena

2.1 Duration of Spoken Strings of Random Letters or Digits

Figure 1 shows mean duration of an utterance as a function of the number of words it contains. The utterances contained from one to six randomly ordered digit names or letter names, such as *3, 8, 9, 7, 2, 5,* or *C, F, J.* If each additional word added the same mean duration increment, we would expect a linear increase in duration with n. Instead, the curve is concave up, and indicates that longer utterances are produced at slower rates. For example, whereas the increase in duration from one to two words is about 70 ms, the increase from five to six words is about 140 ms – twice as great. The quadratic function shown fits well; the error bar shows 2 SE, calculated so as to be appropriate for evaluation of goodness of fit of the function. The qua-

dratic coefficient, here 8 ms, reflects the amount of curvature. Another way to describe the slowing of rate is that the average duration per word is greater in a longer utterance. In section 2.4 we show that a quadratic duration function is produced by a linearly increasing mean duration per word (element duration) of each of a linearly increasing number of words, plus possible end effects that are independent of utterance length.

2.2 Duration of Spoken Strings of Monosyllables versus Disyllables

In a second experiment we used strings of randomly arranged nouns, with each string containing either all monosyllables or all stress-initial disyllables. [Each disyllabic word contained (an approximation of) one of the monosyllabic words as its first syllable; two examples of strings are thus *bay, rum, limb* and *baby, rumble, limit*.] One purpose of this experiment in addition to testing the generality of the effect of length on rate was to determine how 'utterance length' should be defined in the present context: For example, is it the number of words an utterance contains, or the sum of their 'intrinsic durations' (greater for our two-syllable than our one-syllable words), or the number of sublexical constituents (syllables, phonemes,...), or some combination, that influences speech rate? Results are shown in figure 2a. Again quadratic functions describe the data well. Note the error bar, indicating the high precision of the data. Clearly the two-syllable words take longer. The linear coefficient in each function can be taken as an estimate of the mean intrinsic duration of the elements: it differs greatly between one-syllable and two-syllable words. The constant coefficient may reflect phrase-final lengthening: the last syllable of an utterance is prolonged, especially when it is a stressed syllable [Klatt, 1976; Nakatani et al., 1981]. Note, however, that the amounts of curvature for the two kinds of list, indexed by the quadratic coefficients 12 and 14 ms/word^2, are about the same, even though the durations differ greatly. This observation will turn out to be important.

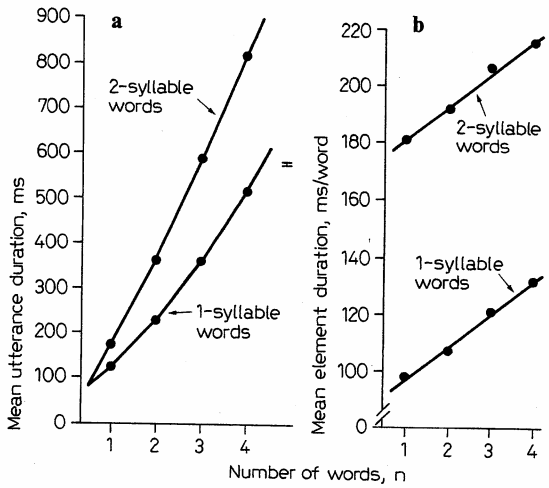

Fig. 2. Experiment comparing sequences of words of either one or two syllables: The results are averaged over 4 subjects and 8 days; about 400 observations per point. **a** Mean utterance durations, estimate of ± SE, and fitted quadratic functions: 44+78n+14n(n−1) ms (monosyllabic words), and 5+168n+12n(n−1) ms (disyllabic words). **b** Mean element-duration functions, based on word elements. The fitted linear functions, constrained to have the same slope, are 85+12n (monosyllabic words), and 169+12n (disyllabic words).

2.3 Robustness of the Effect of Utterance Length on Production Rate

The effect of utterance length on the rate of speech production is remarkably robust; the same effect appears, occasionally with

quantitative variations, in words produced in familiar and unfamiliar sequences, in utterances composed of distinct words and of repetitions of the same word [Sternberg et al., 1978, 1980], in sequences of unfamiliar pseudowords (section 3.3), and in utterances produced by both unpracticed and highly practiced subjects [Monsell, 1986]. Also the latency and the duration of the initial stress group in a sentence exhibit identical effects of utterance length as in a phonologically similar list utterance [Monsell, 1986].

(Examples of sentences are 'Barbara tricks a dean', and 'Barbara tricks a rather rueful dean'. Corresponding list utterances are 'Barbara, Trixie, Dean', and 'Barbara, Trixie, Arthur, Reuben, Dean'.)

2.4 The Element-Duration Function: Definition and Linearity

Another useful way to consider rates of production is in terms of the duration of the average element as a function of utterance length, rather than the duration of the whole utterance. Because these are *mean* element durations we need start with only the durations of whole utterances of one, two, or more words, or other elements:

$D_1, D_2, \ldots, D_n, \ldots$

Duration of an element here includes the durations of all events that occur from its beginning to the beginning of its successor or (for the last element) to the end of the utterance. Utterance duration is thus the sum of the durations of its elements. Furthermore, the 'element' in terms of which we describe utterances is arbitrary, and depends largely on how an experiment is designed. The same set of data can be described in terms of different element definitions. (Examples are stress groups, words, and syllables, in speech.) One of our aims is to discover that definition that brings the element into correspondence with a theoretically relevant action 'unit'. As will be seen below, one clue may be found in that definition of 'element' that leads to a 'simple' description of the data.

The empirical law relating D_n to n is expressed by a quadratic function; to enhance parameter interpretation in terms of the model we develop in section 3, we have chosen the parameterization

$$D_n = \alpha + \beta n + \gamma n(n-1). \tag{1}$$

According to one quantitative realization of our model, and assuming that elements correspond to action units, each element is associated with a selection process (mean duration γn) and a command process (mean duration β). However, whereas the utterance duration incorporates all of the n command processes (mean total duration βn) the selection process associated with the first unit is part of the latency, so the duration incorporates only n–1 such processes (mean total duration $\gamma n\,(n-1)$). This explains the form of the duration function. It also explains the role of the added $\hat{\gamma} = \hat{\gamma}n/n$ in the definition of the mean element duration; this corrects the mean for the selection process that is missing from the duration.

The fitted *utterance duration* is then

$$\hat{D}_n = \hat{\alpha} + \hat{\beta} n + \hat{\gamma} n(n-1). \tag{2}$$

To obtain the *mean element duration* for an utterance of length n we adjust the duration for end effects by subtracting the fitted constant, $\hat{\alpha}$, divide by the number of elements, and add $\hat{\gamma}$:

$$d_{\cdot n} = \frac{D_n - \hat{\alpha}}{n} + \hat{\gamma}. \tag{3}$$

The fitted element duration function is then a linear function, with parameters $\hat{\beta}$ and $\hat{\gamma}$:

$$\hat{d}_{\cdot n} = \frac{\hat{D}_n - \hat{\alpha}}{n} + \hat{\gamma} = \hat{\beta} + \hat{\gamma} n. \tag{4}$$

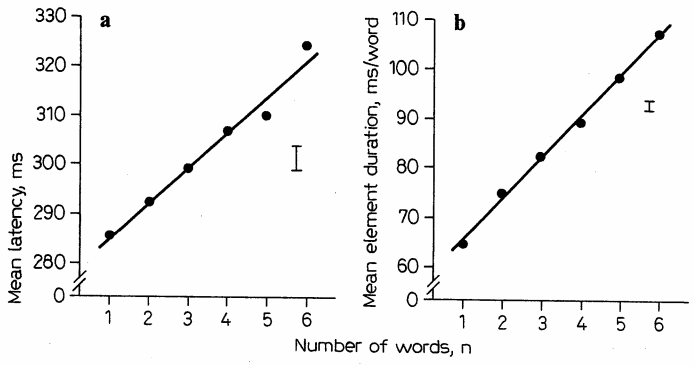

Fig. 3. Mean latency (**a**) and element duration (**b**) for sequences of random digit or letter names, with estimates of ± SE and fitted linear functions: 278+7n ms (**a**) and 57+8n ms (**b**).

Thus, insofar as *utterance* duration is describable as a *quadratic* function of utterance length, the mean *element* duration increases *linearly* with utterance length. Furthermore, if the mean durations of two classes of utterance differ by a constant, independent of utterance length, this difference will be reflected in α, not in β or γ, and hence will have no effect on the element-duration function. Figure 3b shows element durations from the experiment with strings of random letter or digit names. The linear function fits well.

2.5 Implications of Parametric Parsimony for the Conjectured Unit in Rapid Speech

Let us again consider the data from the experiment comparing one- and two-syllable words. The *utterance* durations already shown (fig. 2a) were plotted as functions of the number of *word* elements. Let us consider the *element* duration functions under two different definitions of the element. When the element is taken to be the *syllable,* we find that the average time per syllable increases linearly with utterance length in syllables, but the *rates* of increase differ by a ratio of approximately 4 (one-syllable words) to 1 (two-syllable words). Thus, the effect of utterance length in *syllables* on mean *syllable*-element duration depends on *element type,* which, for a syllable element, is determined by the number of syllables of the word that contains it. In fig. 2b are shown element durations for the same data, but here the element is the *word.* The parallel lines fit well. Thus an alternative plausible definition of the unit leads to parallel functions and hence additivity: the effect of utterance length in *words* on mean *word*-element duration is *independent* of element type, which, for a word element, is determined by the number of syllables the word contains. Because the description in the latter case is simpler, we prefer it on grounds of parsimony. It leads to the conjecture that if the utterances in this experiment can be said to contain action units they are more likely to be words than syllables. In Section 4.4 we present other evidence that favors this conjecture.

One possible empirical criterion that a class of units could be asked to satisfy is a *quantal-effect* criterion: Incrementing utterance length by a member of the class should have an effect on some measure (such as time per unit) that is invariant across interesting differences within the class. Given the present findings, the class of *words* appears to satisfy the criterion; the interesting difference within the class is in *word length,* measured in syllables (or in correlated variables, such as duration). A second possible

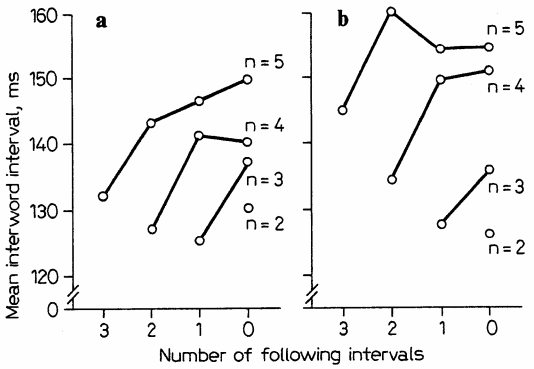

Fig. 4. Mean interword interval for each serial position in lists of monosyllabic words of lengths n=2, 3, 4, and 5. Results are averaged over 4 subjects; about 50 observations per point. **a** Homogeneous lists. **b** Heterogeneous lists. Intervals early in a list have more intervals following them and appear toward the left of the plots.

criterion is *process discreteness:* A unit might be defined as that segment of performance that results from a single rendition of a particular process. If both quantal-effect and process-discreteness criteria are met, as they will be for the action units discussed in the present paper, we are led to the idea that (planning of) a single rendition of the process (the subprogram-selection process discussed in section 3.1) has a quantal effect on the duration of the action sequence, i.e., an effect that is independent of unit size. Other empirical criteria for units are also of interest, of course. For example, the parts of a unit might be expected to cohere particularly well in the sense of influencing each other preferentially [Fowler, 1981], and in the sense of not participating independently in errors of omission or misordering [Fromkin, 1981]. Also the time intervals between features within a unit might be expected to be less variable than the intervals between features in different units [Vorberg and Hambuch, 1984].

2.6 Distribution over Elements of the Effect of Utterance Length on Mean Element Duration

The effect of utterance length on the duration of the *average* element is distributed over *all* the elements, as shown in fig. 4.

Data in figure 4 are from an experiment that is described in greater detail in sections V–VII of Sternberg et al. [1980]. Elements were drawn from a vocabulary of five monosyllabic words: *bee, cow, day, pie,* and *toe.* We chose these words (all CVs starting with stop consonants) in the hope that utterances would have amplitude envelopes containing n peaks separated by n–1 troughs, so that word boundaries would be clearly indicated by minima in the amplitude envelope; the primary aim of the experiment was to permit us to measure the interword time interval (from the 'beginning' of one word to the 'beginning' of the next) as a function of serial position. We used two types of list: *homogeneous* lists, containing repetitions of the same word, and *heterogeneous* lists, containing n distinct words. We determined interword intervals as follows: We digitized subjects' speech at 10 kHz and determined the mean absolute sample value (an amplitude measure) in adjacent 10-ms intervals. Linear interpolation between successive means then defined an amplitude envelope that usually showed a single peak for each word in the utterance and a single trough between each word and the next. For 86% of the correct utterances we were able to choose a criterion amplitude value that intersected the ascending flanks of exactly n amplitude peaks. The time differences between successive intersections then defined a series of n–1 intervals between the 'beginnings' of successive words. (We confirmed that the ascending flank approximates the consonant-vowel transition by measurements of both oscillograms and spectrograms.) We chose an algorithm that produced mean amplitude criteria that depended relatively little on n, and also established that the resulting interword intervals were relatively independent of the value of the amplitude criterion – an important check, given the crudeness of the segmentation procedure.

Figure 4 shows interword interval as a function of position within the string, for strings of different lengths, lined up at the final interword interval. (Lining up the curves at the final interval rather than the initial interval makes the relations among the curves more obvious.) Mean interword intervals for lists containing repetitions of the same word are shown in figure 4a, and

for lists containing n distinct words in figure 4b. Consider the last interword interval. Even though all the other words have been produced, the time from the next-to-last to the last word is prolonged when the utterance contains more words. Now consider the first interword interval. The time from the first to the second word increases as the number of words to follow increases. We already know that despite the complexity of these serial-position functions, when we *average* over serial positions to get the mean element duration the result is simple: a linear function of length. (The simplifying effect of averaging is another fact that requires explanation.) The implication of these data that we wish to emphasize here, however, is that the execution of each element is influenced by a characteristic of the whole utterance – its length. This suggests the existence of a representation of the whole utterance – a program – that is prepared in advance and controls the execution of each element.

2.7 Additivity of the Effects on Mean Element Duration of Element Type and Number of Elements

Figure 2b shows that parallel lines provide a good description of the data. Such parallel functions reveal *additivity* of the effects of two factors on mean element duration. One factor is the *type of element* (words of one versus two syllables). The other factor is the *number of elements*. By *additivity* we mean that the effect of each factor is the same for all levels of the other factor. Thus the effect of utterance length on mean element duration is the same for two types of speech element. This observation will have an important impact on the formulation of a model.

For general discussion of the use of additivity of factor effects on reaction time in making inferences about the structure of mental processes, see, e.g., McClelland [1979; but also Ashby, 1982], Roberts [1987], Schweickert [1985], Sternberg [1969, 1984] and Townsend [1984]. In the present context the arguments are easily extended from reaction time to mean element duration.

2.8 Summary of Principal Phenomena

We ask subjects to produce short utterances, with both the opportunity and incentive for advance planning. We find that utterance duration increases as a quadratic function of the number of elements. This corresponds to a linear increase in the duration of the average element with the number of elements, an increase that is distributed (albeit not with perfect uniformity) over all the elements in a string. Thus a characteristic of the whole utterance influences the execution of each of its elements. This finding suggests that a representation of the whole utterance – an utterance program – is used in the execution of each element. The program must therefore exist before production of the utterance begins. If we measure utterance length in terms of the number of spoken words, then the magnitude of the length effect is the same for words of different durations: the element-duration functions are parallel and vertically displaced.

3. A Subprogram-Selection Model and Its Elaboration

3.1 A Model of the Production of Rapid Action Sequences

Despite its simplicity, the flow-chart model shown in fig. 5 not only captures important aspects of performance, but has

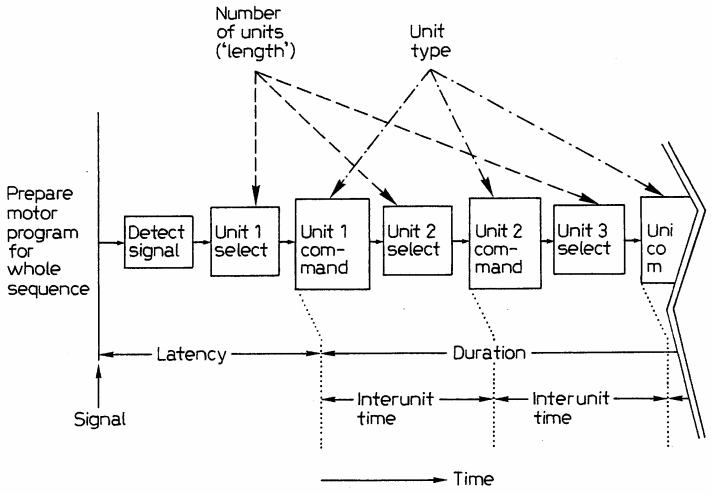

Fig. 5. The subprogram-selection model.

also suggested several fruitful lines of experimentation, as we discuss below. The model, developed for action sequences in general, describes one way an utterance program might be used. The program is prepared before the signal and stored in a motor-program buffer, distinct from the short-term phonological store that limits performance in memory span and similar verbal short-term memory tasks. (We present some of the evidence for this dissociation in section 4.1.) We suppose that the program consists of a set of *subprograms,* one for each *unit* in the sequence. The program is operated upon by a series of selection and command processes.

Before production of a unit can be initiated, the relevant subprogram must be *selected* from the buffer.

By 'selection' of a subprogram we mean nothing more than gaining access to it or passing control to it. The term 'selection' has been chosen somewhat arbitrarily for this process, which might also be termed 'activation' or 'retrieval'. 'Selection' has the disadvantage of suggesting, erroneously, that what is being selected is the action unit controlled by the subprogram (presumably already selected), rather than the subprogram itself. 'Retrieval' has the disadvantage of suggesting the transfer of information in the subprogram to another location, which we do not wish to postulate. 'Activation' tends to be thought of as having a continuum of levels, rather than just two, of changing levels slowly, and possibly of 'spreading' to other units. The term 'buffer' should also not be overinterpreted: We mean no more than a temporary representation of speech units and their ordering; no implication is intended as to how this representation is organized.

The *command* process then causes it to be 'executed'. Utterance production is thus controlled by an alternating sequence of selection and command processes. We assume that the duration of the *selection* process is influenced by the *number* of subprograms, but not by the length of the subprogram selected or the size or type of unit. Although not a central part of the model, the mechanism by which selection is accomplished is worth considering. The linearity of selection time with utterance length can follow from various simple processes of serial search through a set of directory entries or subprogram addresses, one for each of the units contained in the utterance. Al-

ternatively, mechanisms of parallel search or direct access can be coaxed to produce similar effects. Also, some search processes can easily reconcile the simplicity of the element-duration function with the complexity of the serial-position functions.

If each selection process during the production of an utterance involves a self-terminating sequential search in a fixed order from the same starting point, for example, and units in different positions in the utterance have different mean buffer locations over trials, then number of units searched will in general differ from one serial position to another, but the *mean* number of units searched per serial position, averaged over positions (on which the utterance duration and mean element duration depend) will be linear in n. Because the first selection process is incorporated in the latency rather than the duration, this argument applies only approximately to the duration, but should apply exactly to the total time: latency plus duration. Note that the authors of the present report disagree about which type of search, if any, is most plausible.

To justify any search process, however, we would have to explain either why units are not arranged in a buffer in order of execution, or, if they are, why, for each unit after the first, the search cannot continue at the position of its predecessor (which would eliminate the effect of number of units on time per unit). (One possible explanation is that the 'place-keeping' process that would be required for the latter requires resources also needed for the command process.) Note that to describe any search fully one must specify the search target; possibilities here include, for example, a 'tag' associated with each unit that specifies the unit's serial position.

We assume that the duration of each *command* process is influenced by the size or type of the unit about to be executed, but not by the number of units in the utterance. The unit duration is then the sum of the durations of the selection and command processes. Furthermore, we have assumed that number of units and unit type influence these processes *selectively*. One consequence of the seriality of selection and command processes, together with selective influence of the two factors, is that their effects on unit duration are additive, as required [Sternberg, 1969]. Note that the model provides a theoretical definition of the action unit: The unit is that utterance segment that is produced on the basis of a single selection operation (and hence a single command process). We now turn to two elaborations of the model.

3.2 Elaboration of the Model: Identity of the Action Unit in Speech

In the experiment comparing utterances composed of one- versus two-syllable words we have seen that the effect of length on these two classes of utterance is the same when measured in words, but not syllables. This suggests that the unit is the word. That is, there is one selection process per word, regardless of the number of syllables. Another possibility, however, is the 'stress group' or 'metrical foot' – composed of a stressed (strong) syllable (with a full, unreduced vowel) and any associated following unstressed (weak) syllables (with reduced vowels, usually schwa). In the speech experiments described above, the number of stress groups in an utterance equaled the number of words. Fowler [1981] recently provided evidence for the articulatory cohesiveness of the stress group, consistent with its functioning as an action unit: She reported relatively strong backward durational effects of weak syllables within the stress group on the strong syllable, corre-

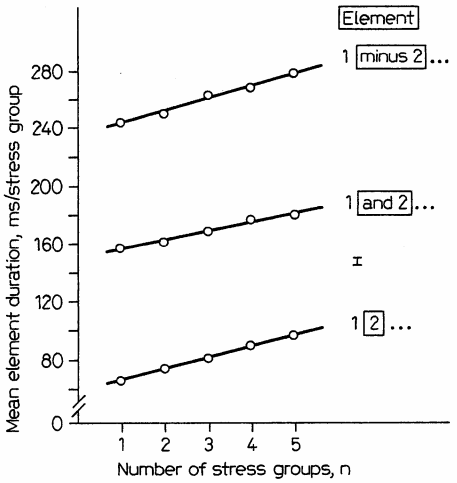

Fig. 6. Mean element-duration functions based on stress-group elements for control utterances and for utterances with interpolated unstressed words ('minus', 'and'), with estimate of ± SE and fitted linear functions: 58+8n (control); 151+6n ('and'); 235+9n ('minus'). Results are averaged over 4 subjects and 3 days; about 140 (70) observations per point for control (other) conditions.

Table I. Expected and observed slopes (γ) of element-duration functions

	Expected slope			Observed slope
	syllable unit	word unit	stress-group unit	
1 2...	8	8	8	(8.0)
1 and 2...	32	32	8	6.2
1 minus 2...	72	32	8	8.7

lated with relatively strong forward coarticulatory effects of the strong syllable on weak-syllable second formants. Furthermore, if the unit of action corresponds to the unit of initial perceptual segmentation for lexical access, and Cutler and Norris [1988] are correct in suggesting that the latter unit is the stress group, then the possibility of a stress-group action unit becomes even more credible.

To test this possibility we interpolated unstressed words in a sequence of stressed words, as indicated in fig. 6. In the control condition subjects produced counting sequences of various lengths, such as *1, 2, 3.* In experimental conditions they produced utterances like *1 and 2 and 3* and *1 minus 2 minus 3* in such a way that the connective terms were unstressed. (A fourth condition using utterances such as *1 by 2 by 3* is mentioned in section 5.) Such interpolation of unstressed words leaves the number of stressed syllables (hence stress groups) invariant, but it does change the number of words. The analysis shown in fig. 6 is based on defining the *element* to be the *stress group*. The three functions are close to being parallel, despite large differences in mean element duration (reflected in the vertical separation of the three functions). Based on a slope of 8 ms fitted to the control data, table I gives the slopes expected in this analysis from three different definitions of the *unit* – syllable, word, and stress group. Note the remarkably clean discrimination among alternatives provided by this test. Observed slope values, shown in the last column, clearly favor the stress group, which is our best current definition of the action unit in rapid speech under our conditions.

Boundaries of the elements indicated in fig. 6 differ from stress-group boundaries, although adding one element to the utterance has the desired effect of adding one stress group, and although our findings indicate that adding one element adds one action unit. In this connection we note that although boundaries of the action unit in the present context are plausibly the same as those of the stress

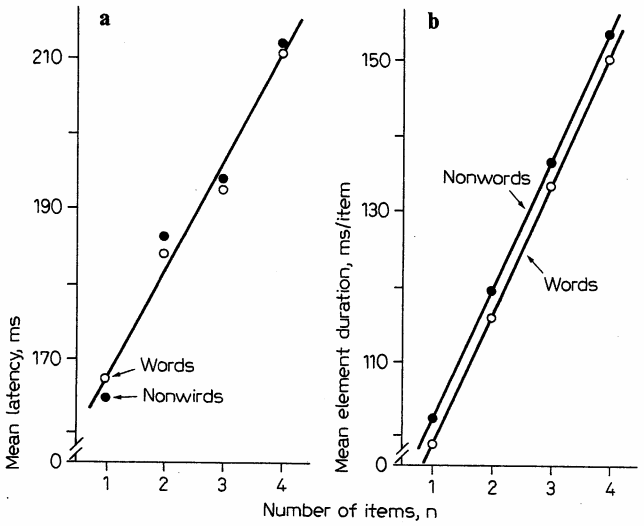

Fig. 7. Mean latency (**a**) and element duration (**b**) for utterances of words and phonologically matched nonwords, with fitted linear functions: 153+15n (**a**); 82+17n (**b**, words); 85+17n (**b**, nonwords). Results are averaged over 7 subjects; about 140 observations per point.

group (which starts with a stressed syllable) boundary location is in fact not dictated by our findings. Given additional assumptions, the results discussed in section 4.4 suggest that boundaries of action unit and stress group may indeed differ. We note also that for utterances in the experimental conditions the initial 'element' (always *one* with no connective) is unrepresentative of the other elements, which generates an unusual end effect that will be reflected in the constant term, α, in the fitted quadratic duration function (section 2.4).

3.3 Elaboration of the Model: Level of Specification in the Program

A second aspect of the model that we now consider is the level of detail in the utterance program. We shall see that the associated experiment also bears on the nature of the production unit. One possibility is that the program contains all the information necessary to control articulation. For familiar words, however, or for stress-group units composed of familiar words, there is an alternative. It is possible that articulatory routines for whole words are learned and stored in long-term memory. The program might then contain only the *addresses* of relevant routines, rather than containing articulatory information explicitly, and would 'call' these routines while it was being executed. To attack this question we examined the production of utterances composed of *nonwords*. If normal performance depends on learned routines for words, the mechanism should function quite differently with nonwords (for which such routines would not have been learned) and we would expect qualitative differences in performance. We used monosyllabic words and nonwords, carefully matched phonologically, and subjects were asked to pronounce the two kinds of utterance with the same stress pattern.

The existence of coarticulation effects implies that not only should the sets of constituent phonemes be the same for matched sets of words and nonwords, but also at least the set of pairs of adjacent phonemes, both within items and across item boundaries. An example of a pair of matched sets that satisfies these criteria is {*vote, hone, vain, hate*} and {*vate, hane, vone, hote*}. We used four such pairs of sets in the experiment.

Element-duration functions together with fitted parallel lines are shown in

fig. 7 b. Linear functions fitted both sets of data remarkably well. Performance is qualitatively the same, and quantitatively very similar: Although the duration difference is reliable, it amounts to only about 1.6 %. We conclude that stored routines for familiar words are *not* called for during execution of the program. Instead each subprogram explicitly specifies the appropriate series of articulatory gestures. Note that this result also adds to the evidence that the unit is defined in terms of stress pattern, rather than lexically.

An alternative possibility not excluded by our findings is that the program calls on stored articulatory routines for phonemes (or other sublexical constituents) whose populations are the same for the words and nonwords we used. Coarticulation would then be produced in the course of program execution, rather than being specified in the program. Neither of these possible accounts is complete, however, since neither explains the small but reliable performance difference between words and nonwords. (Our conjecture is that the difference results from an inherent limitation in any matching procedure for utterances composed of words containing only three phonemes: One can match phonemes and phoneme pairs, as we did, but cannot match longer subsequences and still have nonwords; yet coarticulation may extend beyond the adjacent phoneme.)

4. Four Properties of Performance Suggested by the Elaborated Subprogram-Selection Model

4.1 Effect of the First Selection Process: Increase of Latency with Utterance Length (First Property)

Two Variants of the Model. Given the model, selection of the first subprogram must precede the start of the response. One possibility is that this selection process occurs *before* the signal. There is then no reason to expect an effect of utterance length on response latency (i. e., on the time from the signal to the start of the response). A second possibility is that the first selection process *awaits* the signal, as shown in figure 5. We then expect that latency should increase with utterance length. Furthermore, since it incorporates the duration of a single selection process, the mean latency should grow with utterance length in the same way as does the mean element duration (i. e., linearly, and with the same slope).

Given this second possibility together with the model, the latency reflects the selection process for one of the serial positions in the list (the first). Our expectation about the latency then depends on the assumption that the effect of utterance length is independent of serial position. Without knowing the correspondence between positions in lists of different lengths, this assumption is difficult to test.

Latency of Spoken Strings of Letter or Digit Names. Let us consider the latency data from two of the experiments whose element duration functions were discussed above. Consider first the experiment with random digit or letter names (section 2.1). Figure 3 b shows the element-duration function already discussed for this experiment. The latency data in figure 3 a are not as clean, but are nonetheless reasonably well fitted by a linear function with approximately the expected slope. This result supports the idea that a selection process occurs after the signal and contributes its duration to the response latency.

Latency of Word versus Nonword Strings. Next, consider the experiment comparing words and nonwords. We have already discussed the element-duration functions in figure 7 b. Mean latencies, in figure 7 a, are again approximately linear and with ap-

proximately the correct slope. Incidentally, the closeness of the two sets of latencies further confirms the similarity of performance for strings composed of words versus nonwords.

Whereas the element durations for nonwords were a reliable 1.6% greater than for words, there is no corresponding mean latency difference. According to our model as applied to speech, both latency and element durations incorporate the duration of a selection process, but only the element duration incorporates the duration of a command process. Our finding is therefore consistent with the idea that the small word-nonword effect is located in the command process. Furthermore, this locus of the effect is consistent with the conjecture that it originates in coarticulation of nonadjacent phonemes.

Is the Effect of Utterance Length on Latency due to Ordinary Memory Load? We describe a test of one of several alternative explanations of the effect of utterance length on latency. (We have considered other alternatives as well, and rejected them; for discussions see Sternberg et al. [1978, 1980]). In our experiments, lists must not only be produced fast, but must also be retained in short-term memory to permit such production. The latency could increase with length because of the increasing load on short-term memory, rather than because of a subprogram-selection process. (One possibility is that the longer the list, the more of a limited general-purpose processing capacity is required for maintaining it in memory, and the less is then available for processing the reaction signal and initiating the response.)

We have tested this idea in two experiments, with similar results. We outline one of these experiments here: On each trial subjects were simultaneously shown two lists, a *fast list* (composed of digits) and a *slow list* (composed of digits or letters; results were the same). They had to prepare and produce the fast list under our usual conditions, then recall the slow list without time pressure. Thus, during preparation and production of the fast list, both lists contributed to the memory load. To guarantee accuracy we arranged that the sum of the lengths of the two lists never exceeded five; within this constraint the lengths of the two lists were varied independently. We measured the latency of the fast list. If the effect depends on short-term memory load, we expect the lengths of both lists to influence the latency. But if the effect is due to a selection process associated with the fast list, we expect only *its* length to have an effect. The results were as follows: The mean effect of each item in the fast list on the latency was 11.1 ± 2.9 ms, a typical value. (The second quantity in $a \pm b$ is the standard error based on between-subject variability.) In contrast, the mean effect of each item in the slow (load) list was a negligible 0.4 ± 1.3 ms. The latency therefore depends only on the length of the fast list. This finding argues against any explanations in terms of load on ordinary short-term memory. Similar results were obtained in a second experiment [Monsell, 1986] where the fast list consisted of a familiar sequence of up to five names of days of the week and the slow list contained up to five random digits.

4.2 Maximum Length of a Programmed Utterance Depends on Unit Size (Second Property)

If we continue to increase the length of the utterance we ask the subject to prepare, a limit will presumably be reached at which the effect of utterance length upon latency will break down. The utterance length at which such a limit was reached could pro-

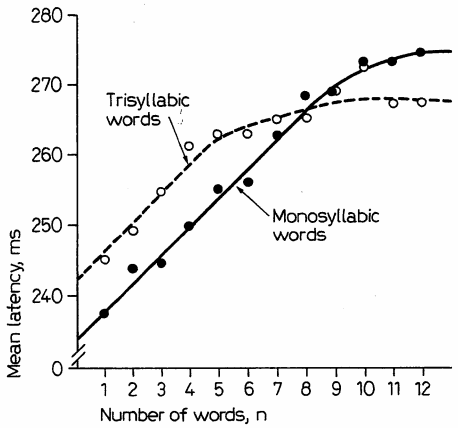

Fig. 8. Mean latency for overlearned sequences of monosyllabic or of trisyllabic words as a function of utterance length. Fitted functions are the averages of bilinear functions. Results are averaged over 6 subjects and 6 days; about 200 observations per point.

vide a measure of the maximum capacity of the utterance program. However, there is a well-known and logically distinct limit, measured by memory span, in our ability to retain an unfamiliar and arbitrary sequence of words. In the hope of measuring the program limit unconstrained by this memory limit, we have examined the production of long utterances derived from highly overlearned sequences.

In our first experiment of this kind, we had 4 practiced subjects prepare and produce utterances containing from 2 to 16 digits or letter names. These were cyclic counting sequences beginning with *one* or *five* (e.g., *five six seven eight nine ten one two)* or subsequences of the alphabet beginning with *A* or *H*. For each subject the latency data were reasonably well described by fitting a bilinear function: an initial segment, linearly increasing up to a critical number of words, and a constant (i.e., flat) final segment. (The mean of this critical utterance length is estimated by the mean abscissa value of the intersection point of the two segments, which was 7.8 ± 0.5 words.) This result suggests that the utterance program does have a maximum size, and that for utterances that exceed this maximum, subjects construct a program for as much of the utterance, starting at its beginning, as a maximum-size program can represent.

Given that the utterance program has a maximum size, together with the idea that the program is not merely a list of subroutine addresses, but incorporates information that explicitly specifies the appropriate sequence of articulatory gestures (section 3.3), if we add the assumption that a subprogram that specifies more articulatory gestures requires more 'storage space', we would expect that the maximum program size would be expressed in terms of unit size as well as number of units. That is, the more articulatory gestures are specified for each (stress-group) unit, the fewer such units could be represented in an utterance program of maximum size.

To test this idea we conducted an experiment in which 6 subjects prepared and produced sequences of from 1 to 12 mono- or trisyllabic words derived from well-learned word cycles [Monsell and Nelson, 1983]. The latency data displayed in figure 8 show that the number of units that can be incorporated in a single program depends upon the size of the utterance fragment that each unit represents. The smooth fitted functions, which describe the data well, are obtained by averaging the two bilinear functions (increasing then flat, with a common slope fitted for the initial segments, and with unconstrained intersection points that provide estimates of the number of words

in the program of maximum size) fitted to each subjects' data. (The mean of several bilinear functions with intersection points distributed in some domain can yield a smooth curve over that domain.) Mean breakpoints were 10.3±0.6 monosyllabic words and 6.4±0.9 trisyllabic words.

The breakpoint for monosyllables substantially exceeded the same subjects' memory span (7.4±0.6 words) for the same material, further indicating that the capacity under study is distinct from short-term memory. Moreover, the difference between capacity estimates in terms of number of words does not translate into equal measures of capacity in terms of spoken duration, as might be expected if program and short-term memory capacities were the same [Baddeley et al., 1975; Baddeley, 1986]: Utterance duration at their respective breakpoints is substantially and significantly shorter for monosyllable than for trisyllable utterances.

Also as expected from our model, mean latencies for utterances containing monosyllabic and trisyllabic words have equal slopes within their ranges of linear increase, consistent with the idea that the rate of increase of selection time (which determines the growth of latency within that range) depends on the number of units, but not on their size.

The difference (8.6±3.2 ms) in height of the functions within their linearly increasing ranges is somewhat larger than a corresponding difference we observed in the experiment comparing the timing of utterances composed of monosyllabic and disyllabic words (4.5±1.3 ms) [Sternberg et al., 1978]. Such height differences suggest that the use of stored information may be recursive, in the sense that after the subprogram for a *unit* is selected from among the *set of units* comprising the *utterance,* representations of each of a series of *subunits* must be selected from among the *set of subunits* comprising the *unit*. The model requires that the duration of this lower-level selection process [called 'unpacking' in Sternberg et al., 1978] be dependent at most on unit 'size' (e.g., number of subunits in the unit) but not on number of units.

In neither experiment did we observe marked pauses or perturbations in speech rate near the parts of longer utterances where the initially programmed section would be expected to terminate. Nor did mean element duration as a function of length reach its maximum at the end of that section, contrary to what we would expect if planning of units that follow that section could be accomplished without slowing execution. Our current hypothesis is that speakers can achieve fluency in a long utterance by programming later units while executing those already programmed, but that speech rate may be reduced when programming takes place concurrently with execution.

4.3 Intermittency of the Effect of Length on Production Rate (Third Property)

We now turn to the third property suggested by the model. The model (fig. 5) contains alternating selection and command processes such that there is one command process and one selection process between the beginnings of successive units. Only the selection process is assumed to depend on sequence length. The influence of sequence length is therefore *intermittent*. Consider what this suggests about the series of articulatory (or acoustic) events in speech. Suppose that the *command* process directly controls a subsequence of the series of articulatory gestures that constitutes a unit. Because this process is not influenced by utterance length, it is possible that the articulation rate within that subsequence might also be independent of utterance length.

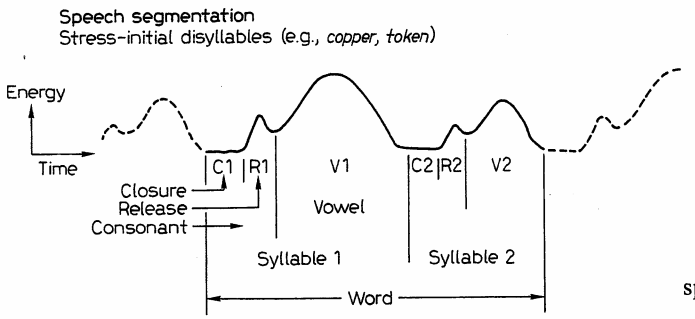

Fig. 9. Six segments of two-syllable spoken words.

The argument for this possibility depends on the idea of a sufficient degree of moment-to-moment coupling between articulation and the underlying processing operations. It would seem that the command process must have delayed effects as well, so as to permit movement control during the ensuing selection process. There are several control mechanisms (with distinct testable implications) by means of which the duration of the selection process could determine the rates of articulatory events. For example, in an *interruption mechanism* the time available for the articulators to continue moving toward a previously commanded vocal-tract target configuration (and thus the degree to which that configuration was approximated) could depend on the interval between the start of the selection process and the transmission of the first of the next unit's commands, which would interrupt the ongoing movement by defining a new target. [For discussions of articulatory control by target-sequence specification and the possibility of undershooting such targets at high rates, see, e.g., Lindblom, 1963; MacNeilage, 1970, and Perkell, 1980.] Insofar as such undershooting characterizes vowels more than consonants we would expect vowels to show larger length effects, which (as we shall see; fig. 10) they do.

Thus it is possible that within each unit, the extra time required when we lengthen the utterance is *localized,* perhaps toward the *end* of the unit, when the selection process for the *next* unit is taking place. An alternative explanation for such a finding might depend on some speech segments being susceptible to 'expansion' while others might not be. It would thus be important to demonstrate that the same segments whose insensitivity to utterance length demonstrates localization of the length effect also have durations that are systematically influenced by other factors (an 'expandability' criterion). The most interesting alternative to the length effect being intermittent is that it is *distributed* throughout the unit and slows all articulatory gestures: a global rather than a local effect. A test of this third property is combined with a test of the fourth in the following section.

4.4 Appearance of the Length Effect in One Epoch per Unit (Fourth Property)

We now consider the final property suggested by the model. Not only should the length effect be localized, but also it should have just *one* locus for each ostensible unit. That is, not only should we find segments that are *not* influenced by the number of units, but for each unit there should be only one segment or one contiguous set of segments that *are* so influenced. We have claimed that each *stress group* is a unit rather than each *syllable.* For testing the idea that the length effect is restricted to one epoch per unit, this claim suggests that we determine the loci of the length effect in multisyllabic words. Thus far, the argument

we have advanced for our claim about unit size in speech is based on nothing more than parametric parsimony: the simplicity of the additivity of effects of unit size and utterance length. If the claim is correct, it follows that in each two-syllable word the length effect should be localized in only one segment, or in one contiguous set of segments. (We might fail to observe such contiguity if there existed segments not susceptible to 'expansion', as discussed in section 4.3.)

Segmentation of Utterances Composed of Two-Syllable Words. We chose words such that segmentation based on the acoustic signal would be feasible. Each syllable started with an articulatory closure, and in all words the first syllable was stressed, as in *copper* and *token*. Figure 9 shows a schematic energy envelope of a stress-initial disyllable. We decomposed each word into six segments, corresponding roughly to the closure, release, and vowel of each syllable.(As suggested in the figure, the first syllable of stress-initial disyllables tends to have greater duration and amplitude than the second.)

The upper panel of figure 10 displays the element-duration function, which is similar to those found previously. The slope here is about 11 ms/word. The lower panel shows the segment-duration functions.

The segment-duration function for segment i is obtained by starting with the summed durations of all segments i in utterances of length n, for each n, and using the resultant set $\{D_{ni}\}$ to determine a set of mean segment durations $\{d._{ni}\}$, just as utterance durations $\{D_n\}$ are used to determine mean element durations $\{d._n\}$. It follows that the observed and fitted element-duration functions are sums of the observed and fitted segment-duration functions, respectively. (See section 2.4.)

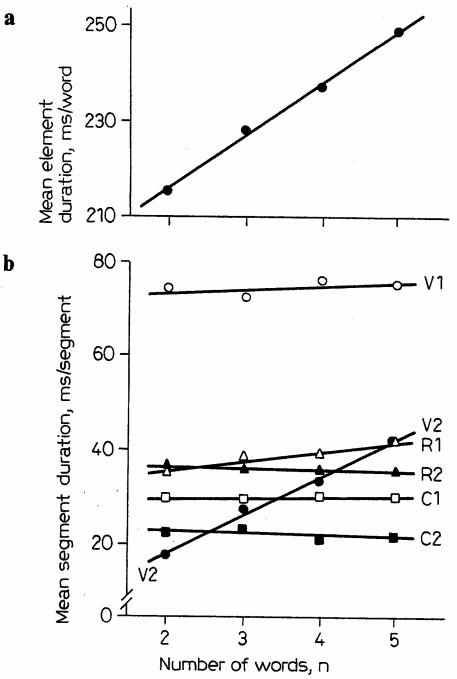

Fig. 10. Mean element duration **(a)** (with fitted function 194+11n ms) and mean segment durations **(b)** of two-syllable spoken words. Results are averaged over 3 subjects and 2 days; about 100 (50) observations per point for lists of $n > 2$ ($n = 2$) words.

The longest segment is the first vowel (V1). Its duration is essentially independent of utterance length. Most of the length effect, on the other hand, is in the second vowel (V2). This finding confirms the third property: intermittency of the length effect. (That duration of the stressed vowel reflects most of the effect of utterance length in sequences of monosyllabic words provides indirect evidence that V1 in initially stressed disyllabic words satisfies the 'expandability' criterion (section 4.3).) There are also small effects in the first closure (C1) and first release (R1). Note that because of the 'wrap-around' property (the second syllable

of one word is followed by the first syllable of the next), these three segments are in fact contiguous. (Thus V2 is followed by C1, which is followed by R1.) This result confirms the fourth property suggested by the model, together with the conjecture that in this kind of performance the word or stress group, rather than the syllable, is the action unit.

If the length effect is localized at the end of the unit (as suggested by our model together with appropriate assumptions about the coupling of articulation with the selection and command processes), then the effect of utterance length on the first closure and release suggests that unit boundaries in this kind of performance are not the same as stress-group (here word) boundaries. Thus, whereas the stress group is described as C1, R1, V1, C2, R2, V2, the action unit (for units other than the first or last) must be described as V1, C2, R2, V2, C1, R1.

5. Caveats and Summary

The work described above contains anomalies and unanswered questions. Anomalies include the following: (1) Variation of parameter values across experiments is large. We have some ideas about the causes, but they are unlikely to apply to every case and have not yet been validated. (2) There is some evidence that the same mechanism cannot explain the full effects of both length and serial position [Sternberg et al., 1980, section VIII]. This reopens the question why averaging over serial positions produces such orderly data. (3) We have seen some evidence of an effect of utterance length of the usual size on the duration of the final word in the utterance. Given the model we are led to conclude that a selection process occurs even when no further word is to be produced. If the evidence pointing in this direction is strengthened, we may have to modify the model, to incorporate a final 'stop' subprogram associated with the cessation of speech and the termination of the series of selections. The number of subprograms controlling the utterance would then be m+1 rather than m, where m is the number of units; this would alter the interpretation of the constant terms in the presently formulated latency and duration functions. (4) As we have seen in section 3.2, unstressed words *(and, minus)* can be interpolated between successive stressed words with no increase in the number of units. When *by* is the interpolated word, however *(1 by 2 by...)*, this invariance fails, and the slope of the element-duration function increases drastically. We do not know what distinguishes *by* from the other words. One possibility is that the vowel in *by* spoken in the context *1 by 2 by 3* cannot be sufficiently reduced for it to fall under the scope of the stress group dominated by the adjacent digit name. It is not clear, however, why this was not equally true for the same vowel in *1 minus 2 minus 3*.

Examples of open questions are: (1) How does the model deal with coarticulation and advance positioning? (2) What is the relation of this kind of 'speeded' performance to natural speech? One conflicting effect that has been claimed is that segment durations become shorter with increasing utterance length in natural speech, rather than being prolonged [Fowler, 1981; Huggins, 1978; Lehiste, 1980]. (One possibility is that a length-dependent subprogram-selection process also operates under 'natural' conditions, but that its effects are masked because it occurs concurrently and

in parallel with prolongation of articulation to satisfy prosodic requirements. Thus it may be that only when speech is produced under unnatural time pressure does the subprogram-selection process reveal itself by constraining the articulation rate.) One surprising similarity is that despite instructions to produce these rapid utterances in a monotone, subjects' pitch contours exhibit a normal declination effect [Sternberg et al., 1978, section VI]; this finding perhaps justifies identifying these rapid utterances with intonation groups. (3) Is the modulation of articulatory units by utterance-spanning prosodic features (such as intonation contour) represented in the corresponding subprograms?

In summary we review the issues raised at the outset in relation to the control of a fluent utterance considered as a rapid action sequence. First, we mentioned the idea discussed by Lashley [1951] of the advance planning of whole sequences, with such plans embodied in motor programs. Second, we discussed the idea of hierarchical structure – the existence of action *units* that contain more than one distinguishable *action*. We have discussed methods and findings that bear on both these ideas. By analyzing the timing of rapid utterances we have found evidence for advance planning of whole utterances and have provided some elaboration of the idea of an utterance program. Our model provides new precision about what an *action unit* might be in the execution of such a program, and our measurements indicate that the stress group is the fundamental action unit. Furthermore, we have shown instances where more than one 'action' is executed as part of the same unit, confirming the hypothesis of hierarchical control.

Acknowledgements

Research conducted in the Human Information-Processing Research Department of Bell Laboratories, and at the University of Chicago where it was supported by NSF Grant BNS-8121372. We thank M. V. Mathews for support, O. Fujimura and M. Y. Liberman for encouragement, and C. A. Fowler and O. Fujimura for comments on the manuscript.

References

Ashby, G. F.: Deriving exact predictions from the cascade model. Psychol. Rev. *89:* 36–44 (1982).

Baddeley, A. D.: Working memory (Oxford University Press, Oxford 1986).

Baddeley, A. D.; Thomson, N; Buchanan, M.: Word length and the structure of short-term memory. J. verbal learn. verbal Behavior *14:* 575–589 (1975).

Book, W. F.: The psychology of skill with special reference to its acquisition in typewriting (University of Montana, Missoula, 1908).

Bryan, W. L; Harter, N: Studies on the telegraphic language: the acqusition of a hierarchy of habits. Psychol. Rev. *6:* 345–375 (1899).

Butsch, R. L.: Eye movements and the eye-hand span in typewriting. J. educ. Psychol. *23:* 104–121 (1932).

Butterworth, B.: Evidence from pauses in speech; in Butterworth, Speech and talk. Language production, vol. 1, pp. 155–176 (Academic Press, London 1980).

Collard, R.; Povel, D.-J.: Theory of serial pattern production: tree traversals. Psychol. Rev. *89:* 693–707 (1982).

Cutler, A.; Norris, D.: The role of strong syllables in segmentation for lexical access. J. exp. Psychol. Hum. Perception Performance *14:* 113–121 (1988).

Dell, D. S.: The representation of serial order in speech: evidence from the repeated phoneme effect in speech errors. J. exp. Psychol. Learn. Memory Cognition *10:* 222–233 (1984).

Fowler, C. A.: A relationship between coarticulation and compensatory shortening, Phonetica *38:* 35–50 (1981).

Fromkin, V. A.: The non-anomalous nature of anomalous utterances. Language *47:* 27–52 (1971).

Fromkin, V. A.: Errors of linguistic performance: slips of the tongue, ear, pen, and hands (Academic Press, New York 1981).

Fujimura, O.: Fundamentals and applications in speech production research; in Proc. XIth ICPhS, Tallin, Estonia, USSR, vol. 6, pp. 10–27 (Academy of Sciences of the Estonian SSR 1987).

Gallistel, C. R.: The organization of action: a new synthesis (Erlbaum, Hillsdale 1980).

Gordon, P. C.; Meyer, D. E.: Control of serial order in rapidly spoken syllable sequences. J. Memory Lang. *26:* 300–321 (1987).

Greene, P. H.: Problems of organization of motor systems; in Rosen, Snell, Progress in theoretical biology, vol. 2, pp. 303–338 (Academic Press, New York 1972).

Huggins, A. W. F.: Speech timing and intelligibility; in Requin, Attention and performance VII, pp. 279–297 (Erlbaum, Hillsdale 1978).

Johnson, N. F.: Organization and the concept of a memory code; in Melton, Martin, Coding processes in human memory, pp. 125–159 (Winston, Washington 1972).

Keele, S. W.: Movement control in skilled motor performance. Psychol. Bull. *70:* 387–403 (1968).

Keele, S. W.: Sequencing and timing in skilled perception and action: an overview; in Allport, MacKay, Prinz, Scheerer, Language perception and production: relationships between listening, speaking, reading and writing, pp. 463–487 (Academic Press, London 1987).

Keele , S. W.; Summers, J. J.: The structure of motor programs; in Stelmach, Motor control: issues and trends, pp. 109–142 (Academic Press, New York 1976).

Klatt, D. H.: Linguistic uses of segmental duration in English: acoustic and perceptual evidence. J. acoust. Asoc. Am. *59:* 1208–1221 (1976).

Klein, R.: Nonhierarchical control of rapid movement sequences: a comment on Rosenbaum, Kenny, and Derr. J. exp. Psychol. hum. Perception Performance *9:* 834–836 (1983).

Knoll, R. L.; Sternberg, S.: Local invariance in hierarchical models of sequence production. Psychonomic Society Meeting, Seattle 1987.

Lashley, K. S.: The Problem of serial order in behavior; in Jeffress, Cerebral mechanisms in behavior, pp. 112–136 (Wiley, New York 1951).

Lehiste, I.: Interaction between test word duration and length of utterance; in Waugh, van Schooneveld. The melody of language, pp. 169–176 (University Park Press, Baltimore 1980).

Leonard, J. A.; Newman, R. C.: Formation of higher habits. Nature *203:* 550–551 (1964).

Levin, H.: The eye-voice span (MIT Press, Cambridge 1979).

Lindblom, B. E. F.: Spectrographic study of vowel reduction. J. acoust. Soc. Am. *35:* 1773–1781 (1963).

MacKay, D. G.: The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. Psychol. Rev. *89:* 483–506 (1982).

MacNeilage, P. F.: Motor control of serial ordering of speech. Psychol. Rev. *77:* 182–196 (1970).

Marteniuk, R. G.; Romanow, S. K. E.: Human movement organization and learning as revealed by variability of movement, use of kinematic information, and Fourier analysis. In Magill, Memory and control of action, pp. 167–197 (North-Holland, Amsterdam 1983).

McClelland, J. L.: On the time relations of mental processes: An examination of sytems of processes in cascade. Psychological Review, *86:* 287–330 (1979).

Miller, G. A.; Galanter, E.; Pribram, K. H.: Plans and the structure of behavior (Holt, Rinehardt, & Winston, New York, 1960).

Monsell, S.: Programming of complex sequences: evidence from the timing of rapid speech and other productions; in Heuer, Fromm, Generation and modulation of action patterns, pp. 72–86 Springer, Berlin 1986).

Monsell, S.; Nelson, E. M.: Control of rapid speech: limits on utterance programming capacity. Psychonomic Society Meeting, San Diego 1983.

Monsell, S.; Sternberg, S.: Speech programming: a critical review, a new experimental approach, and a model of the timing of rapid utterances. Part 1. Bell Laboratories Technical Memorandum (Bell Laboratories, Murray Hill 1981).

Nakatani, L. H.; O'Connor, K. D.; Aston, C. H.: Prosodic aspects of American English speech rhythm. Phonetica *38:* 84–106 (1981).

Namikas, G.: Vertical processes and motor performance; in Magill, Memory and control of

action, pp. 95–113 (North Holland, Amsterdam 1983).

Perkell, J. S.: Phonetic features and the physiology of speech production; in Butterworth, Speech and talk. Language production, vol. 1, pp. 135–173 (Academic Press, London 1980).

Povel, D.-J.; Collard, R.: Structural factors in patterned finger tapping. Acta psychol. *52:* 105–123 (1982).

Roberts, S.: Evidence for distinct serial processes in animals: the multiplicative-factors method. Anim. Learn. Behav. *15:* 135–173 (1987).

Rosenbaum, D. A.: Hierarchical versus nonhierarchical models of movement sequence control: a reply to Klein. J. exp. Psychol. hum. Perception Performance *9:* 837–839 (1983).

Rosenbaum, D. A.; Kenny, S. B.; Derr, M. A.: Hierarchical control of rapid movement sequences. J. exp. Psychol. hum. Perception Performance *9:* 86–102 (1983).

Saltzman, E.: Levels of sensorimotor representation. J. math. Psychol. *20:* 91–163 (1979).

Schweickert, R.: Separable effects of factors on speed and accuracy: memory scanning, lexical decision, and choice tasks. Psychol. Bull. *97:* 530–546 (1985).

Shattuck-Hufnagel, S.: Sublexical units and suprasegmental structure in speech production planning; in MacNeilage, The production of speech, pp. 109–136 (Springer, New York 1983).

Sternberg, S.: The discovery of processing stages: extensions of Donders' method; in Koster, Attention and performance II. Acta psychol. *30:* 276–315 (1969).

Sternberg, S.: Stage models of mental processing and the additive-factor method. Behavioral Brain Sci. *7:* 82–84 (1984).

Sternberg, S.; Monsell, S., Knoll, R. L.; Wright, C. E.: The latency and duration of rapid movement sequences: comparisons of speech and typewriting; in Stelmach, Information processing in motor control and learning, pp. 117–152 (Academic Press, New York 1978).

Sternberg, S.; Wright, C. E.; Knoll, R. L.; Monsell, S.: Motor programs in rapid speech: additional evidence; in Cole, The perception and production of fluent speech, pp. 507–534 (Erlbaum, Hillsdale 1980).

Szentagothai, J.; Arbib, M. A.: Conceptual models of neural organization (MIT Press, Cambridge, 1975).

Townsend, J. T.: Uncovering mental processes with factorial experiments. J. math. Psychol. *28:* 363–400 (1984).

Vorberg, D.; Hambuch, R.: Timing of two-handed rhythmic performance; in Gibbon, Allan, Timing and time perception. Ann. N. Y. Acad. Sci *423:* 390–406 (1984).

Wetzel, M. C.; Howell, L. G.: Properties and mechanisms of locomotion; in Towe, Luschei, Motor coordination. Handbook of behavioral neurobiology, vol. 5, pp. 567–625 (Plenum Press, New York 1981).

Saul Sternberg
Department of Psychology
University of Pennsylvania
3815 Walnut Street
Philadelphia, PA 19104–6196 (USA)

Index autorum