Journal of Phonetics (1980) 8, 113–133

Coarticulation and theories of extrinsic timing

Carol A. Fowler*

Dartmouth College, Hanover, NH 03755, U.S.A.

Received 7th August 1978

Abstract: Current accounts of coarticulation belong to a single class of theory, here called extrinsic timing theories of speech production. The accounts all assume that the dimension of time is excluded from the specification of a phonological segment in the articulatory plan for an utterance, and all of them fail to explain or predict the coarticulatory patterns of speech. Here I suggest that some of the failings are endemic to the *class* of extrinsic timing theories, and that a more adequate account must derive from an intrinsic timing theory. The essential characteristics of an intrinsic timing theory are described.

Introduction

In recent years, several articles addressing the phenomenon of coarticulation as a theoretical issue have appeared in this journal (Daniloff & Hammarberg, 1973; Hammarberg 1976; Kent & Minifie, 1977). In the most recent of these, after critically reviewing most of the extant theoretical accounts of coarticulation, Kent & Minifie conclude that none of the reviewed accounts explains coarticulatory patterns adequately. An aim of the present paper is to suggest why current theoretical accounts of coarticulation fail to explain or indeed adequately to predict coarticulatory patterns. A second aim is to sketch the form that an adequate account might take. I suggest that our current accounts are all instances of a single class of theory, and that the assumptions about a talker's control over timing in speech which characterize this class preclude its members from providing an adequate account of coarticulation. More precisely, the extant theoretical accounts of coarticulation are instances of theories of extrinsic timing control. That is, they exclude timing from representation in the talker's articulatory plan for his utterance.[2] Instead, they propose that an utterance is given coherence in time only by its actualization.

I propose that some of the inadequacies of our current attempts to explain coarticulation are endemic to this class of extrinsic timing theories, and therefore that a satisfactory account must derive from an intrinsic timing perspective.

My procedure will be first to describe the essential, defining properties of extrinsic timing theories. To provide something concrete to work with, I will describe a prototypical theory in this class, drawing in large part on the view of Daniloff & Hammarberg (1973) and of Hammarberg (1976). It is true that current theories of speech production or of coarticulation are substantially different one from the other particularly with reference to the hypo-

*Haskins Laboratories, 270 Crown St., New Haven, CT 06510, U.S.A.

[2] In an extrinsic timing theory, the articulatory plan represents the serial ordering of features, segments or syllables – i.e., it represents their ordinal relationships along the time axis – but time is not taken to inhere in, or to be essential to, the specification of these production units.

0095-4470/80/010113+21 $02.00/0

©1980 Academic Press Inc. (London) Ltd.

thesized unit of production (e.g. the articulatory syllable of Kozhevnikov & Christovich, 1965, the spatial target of MacNeilage, 1970, and the feature bundles of Daniloff & Hammarberg, 1973). Despite this variety, however, all share the assumptions of the extrinsic timing view. Since these assumptions are the concern here, it seems fair and simpler to focus where possible on a single exemplar.

Having characterized the class of extrinsic timing theories, I will specify why I think no instance can provide an adequate account of coarticulation (or of any of the other manifestations of timing control in an utterance). Kent & Minifie (1977) have catalogued many of the observations that they take to disconfirm the various extrinsic timing theories. The reader who wishes to know the grounds on which the extrinsic timing theories are individually weakened or disconfirmed is referred to that article. Here I will focus on the reasons why (I allege) the theories fail as a class.

In the concluding section of the paper I will characterize the essential properties of a theory of intrinsic timing. The characterization is not intended to constitute an explanation in detail of coarticulation but primarily to specify the theoretical perspective from which, perhaps, an adequate account may derive.

A prototypical account of coarticulation in the extrinsic timing framework

An intuitive concept of "segment" underlies our recognition that there is a phenomenon of coarticulation requiring explanation. In a sense, this is unfortunate because the problem of discovering the acoustic or articulatory correlates of our intuitive concept has been recalcitrant. Its recalcitrance stems from an evident incompatibility between the essential properties of the concept and its manifestations in speech (*cf.* Studdert-Kennedy, 1976). When we perceive speech, we hear a succession of discrete segments. The segments as perceived are discrete or separate in two senses: the successive sounds are both qualitatively and temporally distinct. But when we observe a talker's articulations, or when we look at an acoustic record of them, at most only the first distinction is preserved. One can see a talker producing different kinds of sounds: some of them occlude the vocal tract locally, some partically obstruct the passage of air, and others merely alter the global shape of the tract. Likewise, one can see the consequently different kinds of acoustic signal. What is not represented in the articulatory and acoustic records of an utterance is temporal discreteness. The different kinds of gestures go on simultaneously, and thus there are no borders perpendicular to the time axis in an articulatory or acoustic record to separate one segment from another.

For Hammarberg (1976), there is only one conclusion:

> Segments cannot be objectively observed to exist in the speech signal nor in the flow of articulatory movements. There are no invariant physical cues of segmentalness. There is no extra-human, i.e. non-subjective, way of analyzing the speech coninuum into discrete parts corresponding to the notion of a segment. . . What all this adds up to is, that the concept of segment is brought to bear *a priori* on the study of the physical–physiological aspects of language. (p. 355)

And again:

> It should be perfectly obvious by now that segments do not exist outside the human mind . . . All indications are that the segment is internally generated, the creature of some kind of perceptual–cognitive process. (p. 355)

For Hammarberg, then, the upshot is this. The mind has set of concepts of phonological segments which it *imposes* on an acoustic signal in the course of perceiving it. However, the

segments are not given in the acoustic signal nor in the articulatory gestures responsible for it; thus it takes a human mind to interpret an acoustic speech signal.[3]

Typically, a perceiver of speech hears what the talker intends him to. Thus a statement of the talker's intended utterance, no less than a statement of the hearer's percept, must invoke the concept of segment. Evidently, the talker starts with segments, but loses them somehow either in his articulatory plan for his utterance or in its actualization. The charge of a theory of coarticulation is to explain how the mental concepts of discrete phonological segments get translated into a continuous overlapping production of articulatory gestures.

For Daniloff & Hammarberg (1973; see also Hammarberg, 1976), the explanation posits an abstract description of the talker's intended utterance. The plan is a left-to-right array of discrete phonological segments – i.e. of the abstract segments which perceivers hear, but which are not given to them in the acoustic signal. Daniloff & Hammarberg call them canonical forms. "They are invariant, ideal, uncoarticulated target forms," each containing all and only that which is essential to their particular identity. Canonical forms never appear in an utterance, but they can be estimated. The best approximation occurs "when a segment is produced in isolation in a sustained manner, or when the sound is produced in a context assumed to be minimally coarticulatory" (Daniloff & Hammarberg, 1973, p. 241).

The dimensions of description of the phonological segments are assumed by Daniloff and Hammarberg to be features (but a compatible proposal is that they are spatial targets; see MacNeilage, 1970). Thus the plan for an utterance at an early stage is a left-to-right array of feature bundles. If the plan were to be executed at this stage, the abrupt changes in articulatory specification that would occur as the plan executor moved from feature bundle to feature bundle would cause transitional sounds to occur between the realizations of sounds which the speaker intended to be immediately successive. To avoid that, features are "spread" in the plan from one bundle to its neighbors so that adjacent sounds and gestures are accommodated one to the other (*cf.* Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967). In consequence, articulatory transitions between adjacent segments are smoothed, but at the expense of the canonical forms.

It is worth observing that on this view, coarticulation is not co-production of segments– i.e. it is not the overlapping production of separate ideal segments. Rather it is an adjustment of an ideal segment to its context. (*cf.* Hammarberg, 1976: "Coarticulation is, then, to be regarded as a process whereby the properties of a segment are altered due to the influences exerted on it by neighboring segments." p. 576).

[3]It is perhaps worth pointing out that Hammarberg makes two separate claims here: first that there are no invariant correlates of a segment in an acoustic or articulatory record of an utterance, and second that it takes a human mind to perceive a segment. In the quoted passage, Hammarberg evidently makes both claims, but elsewhere (see his discussion of inherent and derived properties of a segment), he seems only to make the second. It is the case that the second claim could be true without the first more radical claim also being true. For example, (see Fowler and Turvey, in press for an elaboration of this point) it probably takes a human mind to recognize something as being an example of "footwear". The essential properties of footwear – e.g. That it be foot-sized and shaped, that it have a means of attachment to the foot and so on – clearly are given in the optical signal to an eye. However, that collection of properties is only a significant collection for an animal that wears things on its feet. Other animals will not recognize that aggregate of properties as a significant collection and hence will not detect that a shoe, for example, is footwear.

Likewise if it is true that it takes a human perceptual system to perceive a segment (or human knowledge system to know one), it need not also be true that the percept is based on properties that are absent in the acoustic signal. I will suggest later that the properties cannot reasonably be supposed to be absent.

The foregoing account of coarticulation differs from other accounts both in its degree of precision and in its proposed units of production. But it shares with other accounts several assumptions or claims which identify it as a member of the class of extrinsic timing theories. They are as follows:

(1) The essential properties of a segment as it is *known* to a language user are timeless.

(2) Segments in a planned sequence are discrete in the sense that (abstractly stated), their boundaries are straight lines perpendicular to the time axis, so that the terminus of one segment is the beginning of the next segment. (Feature spreading adjusts the specifications within a pair of boundaries but does not make the segments continuous. Thus, ideally, segments occupy non-overlapping time slots even after feature spreading.)

These first two assumptions exclude the dimension of time from having an essential role either in defining the phonological units themselves or their relations in a planned utterance.

(3) The plan for an utterance is distinct from its executor.

(Concomitantly, the spatial coordinates of an utterance are specified independently of its temporal coordinates. That is, the feature bundles or the spatial targets in the plan specify the successive spatial coordinates of the utterance while the executor or articulatory mechanism actualizes its temporal coordinates.)

None of these assumptions is defended by Daniloff & Hammarberg or by Hammarberg despite their unusual efforts to make explicit their assumptions and reasoning. Indeed none has been defended to my knowledge by any theorist of coarticulation or speech timing. However, the claims are addressed by Lashley (1951) in his classic paper on serial ordering. Since I will argue they are false, a digression to examine and evaluate his arguments is warranted.

Lashley: The roots of extrinsic timing theories

Many investigators and theorists recognize the relevance of the classic serial ordering issue to that of coarticulation. In that connection, it is not surprising that Lashley's well-known paper *The Problem of Serial Order in Behavior* (1951) is frequently cited when coarticulation is given theoretical treatment (e.g. Hammarberg, 1976; Kent & Moll, 1975; MacKay, 1970; Wickelgren, 1969). Nor is it surprising that his views have their counterparts in accounts of coarticulation.

Lashley presents the extrinsic timing view in the following passage:

> Since memory traces are, we believe, in large part static and persist simultaneously, it must be assumed that they are spatially differentiated. Nonetheless, reproductive memory appears almost invariably as temporal sequence, either as a succession of words or of acts... The translation from the spatial distribution of memory to temporal sequence seems to be a fundamental problem of serial order... There are indications that one neural system may be held in [a] state of partial excitation while it is scanned by another... The scanning of the spatial arrangement seems definitely to determine, in such cases, the order of procedure. (p. 521-522).

I understand Lashley to make the following argument. Some acts (in particular those like speaking in which there is a more-than-additive relationship between the sense of the act as a whole and that of its components taken separately) presuppose "mental plans." A plan must represent concurrently the act-components which will occur in succession when executed. Since the representations of the various act-components (Lashley's memory traces) persist concurrently and over some period of time, the plan is time-invariant. Thus the dimensions along which the representations may encode information about their referent act-components must be the three spatial dimensions at most. Another consequence of the plan's time invariance is that memory traces representing successive act-components must be spatially

differentiated. Only on execution of the plan are they serially ordered in time and given temporal coherence. Lashley's conclusion that time cannot be represented in the plan, then, derives from the evident fact that it cannot be given literal representation because of the necessarily static nature of the plan's components.

This is the essence of the extrinsic timing view and it survives intact in our accounts of coarticulation. But Lashley's argument, if I have characterized it fairly, is specious. He seems to conclude that because a memory trace is static–and thus at most three dimensional–it can only encode information about those three dimensions. But of course this is not a limitation. The written sentence: "John is eating an apple" is a representation in two dimensions of a four dimesional event. The information conveyed *by* a representation is not constrained by the dimensionality *of* the latter. Thus although the information conveyed by a spatial array of static memory traces could by coincidence be information about successive, static parts of an act, it need not be.

This means only that we are not bound to devise extrinsic timing theories–at least not on the grounds specified by Lashley. It does not imply that extrinsic timing views must be incorrect.

Let us turn now to what I see as the failings of the extrinsic timing theories. I will consider each of the assumptions of those theories described earlier.

Criticisms of extrinsic timing theories

Assumptions 1 and 2–Sequences as timeless; intersegmental boundaries as abutting straight lines perpendicular to the time axis

I will consider together the first two assumptions listed above–that segments are timeless and that their boundaries are abutting straight lines oriented perpendicularly to the time axis. The two assumptions are distinct in the extrinsic timing theories, but the reasons why they are incorrect overlap in large part.

In general I will argue that the concepts of a segment as, ideally, a three-dimensional layer in a four dimensional event of talking, and of talking itself as a succession of those layers must be incorrect. Furthermore, their invalidity is sufficient to prelude their constituting a model for a talker of his own utterance. That, is, they characterize neither an utterance nor a talker's plan for his utterance, nor the "canonical forms" that underlie them.

First consider some counterevidence to the proposal that a segment is timeless so that it can be specified in a plan as the (planned) convergence at an instance in time of a bundle of features or as a spatial target. Lisker (1972) provides one reason why this point of view must be incorrect. He notes that some languages have phonological segments, such as prenasalized stops or occluded nasals, which cannot be characterized as a complex of simultaneous features. These segments involve a sequence of necessarily nonoverlapping states of the velum.

This observation may have its counterpart in some sounds of English as well. Bell-Berti & Harris (1979.) found the onset of coarticulatory lip rounding for /u/ to precede the acoustically defined onset of /u/ by a fixed amount of time. That is in their data, lip rounding was not time-locked to any of the segments preceding /u/ in the plan. Rather rounding was time-locked to the remaining gestures for /u/. Hence, it would seem, /u/ is a four-dimensional, segment, which is co-produced with its neighboring segments.

In a similar vein, Kent Carney & Severeid (1974) describe the production of the first nasal /n/ in "intend". Velar closing in anticipation of the following nonnasal consonant is synchronized to the tongue gestures that produce the /n/. By the time the alveolar tongue constriction for /n/ is attained, the velum is already half way to its closed state. Whatever it is that specifies the coherence of those gestures that correspond to a given segment, it cannot be synchrony, because if it were, the /n/ in "intend" would be nonnasal.

Still focusing on the individual segment, the work of Abramson & Lisker (e.g. see Abramson, 1975) indicates that voiced and voiceless segments are distinguished on the arti-

culatorily simple (but acoustically complex) dimension of laryngeal timing. Finally, Lisker (1972) notes that the phonemes /t/ and /c/ are distinguished one from the other by their relative rates of release of the alveolar constriction. /b/ and /w/, and /d/ and /y/ are similarly distinguished (Liberman, Delattre, Gerstman & Cooper, 1956). But in order to incorporate "rate of release" as a property of a phonological segment, the phoneme's specification as a spatial target or as the convergence at a point in time of a set of features must be abandoned.

The concept of a plan as an array of feature bundles may also be evaluated on existing evidence. Recall that feature bundles in a plan are context-adjusted. The articulatory mechanism (the plan executor) successively reads out the adjacent feature bundles in the plan. Thus, the features in a bundle are triggered at once rather than over time.

A prediction may be derived from this characterization that allows it to be tested. When the articulatory mechanism executes a context-adjusted feature bundle, the component features will be realized as articulatory activity with mutually somewhat different latencies. Each latency will depend on the articulator implicated by the feature and that articulator's current activities. It may be possible to identify, either in the EMG signals for the relevant muscles or in the movement records of an utterance, the relative latencies of the different features that were simultaneously executed by the articulatory mechanism. Barring variation in carry-over coarticulation, these relative latencies should be invariant over any change in context. The latencies should reflect the invariant time that it takes each feature to be actualized when a bundle is executed. In short, the evidence should never indicate that the different members of a bundle are triggered over time rather than simultaneously.

In accordance with the feature-bundles view is evidence reported by Kent and his colleagues (Kent & Netsell, 1971; Kent, Carney & Severeid, 1974; Kent & Moll, 1975) that articulatory gestures are synchronous or time-locked over very short intervals. If this occurs universally, it supports the view of the plan as an array of feature bundles.

It does not occur universally, however, as some EMG and cineradiographic data show. The decision as to when a particular gesture is to be initiated sometimes seems to be made with reference to when it has to be initiated in order to get the job done, rather than with reference to the onsets of other gestures. For example, Bell-Berti (1973) shows that activity of the levator palatini (which is necessary to raise the velum after a nasal) in /fVCmVp/ begins earlier within the nasal in anticipation of the vowels /i/ and /u/ than in anticipation of /a/. Bell-Berti suggests that this occurs, not for any mechanical reason, but because nasal coupling is less likely at a given degree of opening during the production of a low vowel such as /a/ than during the production of a high vowels /i/ and /u/. It is as if the system delays velar closure until it has to initiate it. Some cineradiographic data bear out the EMG evidence that individual articulatory gestures need not be time-locked. Borden & Gay (1975) describe tongue, jaw and lip movements during the productions of /spapə/, /stapə/ and /skapə/. Although time-locking was evident for one of three speakers producing /spapə/, it was not for the other two speakers, nor generally for any of the speakers producing /stapə/ and /skapə/.

One might seek to revise, rather than to reject the feature-bundle notion by supposing that at a finer grain of description, features are arrayed sequentially as well as in bundles in the plan (see Kent, Carney & Severeid, 1974). I think that there is a more satisfactory solution, however, which closes the separation between the properties of canonical forms and those of their actualizations.

In the discussion that follows I will make the following arguments:

(1) Any proposal in which distinctions are posited between a canonical form and its actualization (especially distinctions as difficult to bridge as the discrete/continuous, timeless/four-dimensional differences proposed in extrinsic timing theories) is an argument of last resort. This is so in part because it leads one to ask why a conceptual category should arise in evolution or ontogeny that bears so little resemblance to the actualization with which it co-evolves or co-develops. But it is also unattractive because it vastly complicates the process

of getting from canonical form to gesture and back again in perception. (As Hammarberg notes it introduces the mind/body problem. It does so because it demands translations across phases of matter-from psychological to physical and from physical to psychological). I argue that we need not yet give up on our search for the correlates of our intuitive concept of segment in a speech utterance or in its acoustic product.

(2) The separation between canonical form and actualization leads necessarily to the notion of Hammarberg's that phonological categories are brought to bear *a priori* (by a phonetician, and presumably also by a naive listener) on an acoustic signal. To the extent that this implies insufficient acoustic support for the conceptual categories that we perceive, the claim is intenable and must be gotten around.

Canonical forms

A speech production theory must be more highly valued, other things being equal, if it posits no difference between the properties of canonical forms of phonological segments (that is of segments as they are known) and those of segments described in an articulatory plan or realized in a vocal tract. That is so because any difference requires both a "translation theory" (cf Fowler, Rubin, Rémez & Turvey, in press)–that is an explanation of how the translation is effected from known segment to produced segment–and a rationalization for the evolution of concepts to be communicated that cannot be nondestructively actualized (see below). I assume that theories treat canonical forms as three dimensional and as ordinally separate from their neighbors (as of course they are not in speech itself) in part because theorists believed themselves backed into those decisions (due to the reasoning made explicit by Lashley and described earlier) and in part because they considered linguistic theory to verify these as properties of abstract linguistic segments (*cf.* McNeilage & Ladefoged, 1976). Since sequences of segments cannot be static when realized in a vocal tract, actual segments cannot share this property with segments as speaker/hearers know them. Furthermore certain cognitive demands of speech perception seem to require coarticulation (Liberman & Studdert-Kennedy, 1978), and coarticulation according to some interpretations precludes actualized segments being discrete and context-free.

However, elsewhere (Fowler *et al.*, in press), I have suggested that segments as formal lingusitic entities do not in fact have the properties "static" or "discrete". Formal linguistic theory seeks only to characterize segments as they participate in an abstract linguistic system. Because it concerns itself only with an abstract formal system, the theory is unconcerned with the way in which these formal properties are known, produced or perceived by speaker/hearers. Thus the theory assigns featural values to segments only on dimensions of description that distinguish one segment from its class members. The dimensions "static"/"dynamic", and "discrete"/"continuous" are irrelevant here because they have to do with the way in which segments are realized in some medium and that is not of concern in a formal linguistic description.

Whether or not this strategy of formal linguistic theory is useful is not of concern here. What is important is that linguistic theory does not in fact assign the values static and discrete to linguistic segments. Hence, the theorist of speech production is free to assign whatever values are most useful to him on the dimensions static/dynamic and discrete/continuous. Known and produced segments may *both* be four dimensional, (i.e. [+ dynamic]). The assignment of this value to linguistic segments does no violence whatever to their other canonical properties as assigned by linguistic theory. Similarly, instead of treating coarticulation as an adjustment of the canonical properties of a segment in acquiscence to its neighbors, it may be viewed as the *overlapping* production of successive, continuous, four dimensional segments. Thus feature spreading may be apparent, but not actual.

If segments are coproduced and if they are not temporally discrete even in intent, then why do our introspections reveal them to be distinct? Our introspections may yield an impression of discreteness for two reasons. First and foremost, the phonological segments

that are coproduced – primarily consonants and vowels – are different *kinds* of segment. The one is a rapid and local obstruction of the vocal tract, while the other is a relatively slow global change in the shape of the vocal tract. These different kinds of gesture generate different kinds of acoustic signal; crudely, clear cases of consonants provide acoustic evidence of vocal-tract constriction, while vowels reflect the formant structure of an open tract. Just as we perceive the separateness of any two acoustic events of different kinds that overlap in time (say of a singer's voice on a record and of her musical accompaniment), we can (putatively) detect the more subtle separateness of two kinds of spoken segments that overlap in time. The second reason for our perception of spoken segments as discrete may be that no two sounds are strictly concurrent. For example, a vowel and a consonant may be coproduced, as in the data on VCV production reported by Ohman (1966); but the onset of the first vowel procedes the consonantal constriction, and the second vowel persists after the local occlusion for the consonant has been released. To repeat, then, the suggestion is that segments are not temporally discrete, but rather are qualitatively separate events, and that this separateness accounts for our impression that segments in an utterance are mutually distinct.

If we accept that segments are essentially four dimensional and are coproduced, an advantage is that we need not view the velum-raising data of Kent *et al.* (1974), for example, as a paradoxical case of feature spreading. Instead, the data may indicate that a nasal segment is four-dimensional and that its properties are revealed over time. Its coherence is specified not by the convergence of its featural actualizations at a point in time, but by the continuity in time of the production of a segment of a particular kind. Likewise anticipatory lip protrusion for a rounded vowel may occur, not because the feature [+ rounding] has been spread from the rounded vowel to earlier segments, but because rounded vowels are entities of which time is an inherent dimension– they are produced in time.

A priori *categories*

Let us next consider Hammarberg's conclusion that since canonical forms are not actualized unaltered in an utterance, the concept of segment must be brought to bear *a priori* on the physical acoustic signal.

The suggestion that phonological segments are subjective categories leaves unanswered several puzzling questions. For example, how could the *a priori* concept of segment ever have arisen in evolution (i.e. if the concept cannot be learned by a child from his experiences with the acoustic signal, how could it ever have been acquired by a species based on its experience with acoustic signals)? Second, why should these fictitious concepts ever have arisen in evolution? Evolution, like ontogeny, is a process by which an organism maintains or enhances its compatibility with the properties of the world. It would appear highly disadvantageous for an organism to seek to impose on the world properties that the world does not have. Finally, if an acoustic signal offers insufficient support for the concept of segment, why do listeners so reliably perceive a talker's intended message?

It is surely more plausible to suppose that the concept of segment *has* material support. Its essential properties are manifest in the acoustic signal, although it may take a human perceptual system to detect that aggregate of properties as a significant collective. Scientists have not discovered those properties in the acoustic signal, but the reason they have not may be that they have looked for evidence of the wrong kind. They have looked for temporal discreteness when they should have looked for qualitative separateness among temporally overlapping events. And they have sought to discover abutting edges of segments perpendicular to the time axis when, perhaps, no such things are to be found.

Some theorists (e.g. Gibson, 1977; Polanyi, 1958) take issue with the dichotomy between subjective and objective events to which Hammarberg makes reference in his discussion of *a priori* concepts of segments. The dichotomy as applied to the concept of phonological segment suggests that segments as perceived and known are either objectively instantiated –

i.e., their essential properties are manifest in the acoustic signal and thus provide support for the perceiver's cognitions – or their properties have no acoustic support and the concept of segment is in Hammarberg's (1976) words "internally generated, the creatures of some kind of perceptual-cognitive process" (p. 355). But some theorists prefer a third option. The third option is that percepts and concepts are neither subjective nor objective; instead their material support *is* available in the world, but is only detected by a specially attuned organism. Consider again the example of footwear as discussed in footnote 3. Clearly the aggregate of properties – foot-shaped and -sized, protection for the foot etc., – are available in the light to an eye when that eye focuses on an instance of footwear. Thus the essential properties of the concept are "objectively" manifest. But it takes a wearer of shoes to detect that particular aggregate of properties as a significant collective.

It is implausible to suppose that the concept of phonological segment is wholly subjective – i.e. that it has no acoustic support. It seems more reasonable to suppose that "phonological segment" is a concept like that of footwear – it is neither subjective nor objective, but something else that spans the dichotomy.

Assumption 3: The plan as distinct from its executor

The recent literature provides both counterevidence (Sternberg, Monsell, Knoll & Wright, 1978) and counterarguments (Neisser, 1976; Fowler, 1977) to the view that plans are executed by an extrinsic articulatory mechanism. The counterevidence derives from an experimental paradigm in which subjects are asked to produce a well-learned or well-known utterance as quickly as possible following a signal. Subjects know in advance of the signal what utterance they are to produce; the signal simply indicates when they are to begin talking. Sternberg *et al.*, observe a positive linear relationship between latency to begin producing the utterance following the signal and a measure of utterance length (the number of stressed syllables in the utterance). The result holds even though the utterance is known to the subject before the signal to respond, and thus, even though he evidently has ample time to prepare for it. It also holds when the sequence to be produced is very familiar to the subject (e.g. some portions of: "one, two, three, four, five"; or "Monday, Monday, Monday, Monday, Monday"). Sternberg *et al.*, suggest that the reason why subjects fail to "take "take advantage" of the time before the presentation of the signal to devise a plan for their utterance is that plans are self-executing. When a plan for an utterance is devised, it necessarily runs off. The intrinsic timing view that I will sketch out below provides a rationalization for this result.

In addition to this limited counterevidence there are strong grounds for questioning the proposal that plans are separate from their executors. The reasons concern the obligations of a proposed model of coarticulation. A model of speech production, even if devised to explain only coarticulation, must still accommodate (or must enable elaboration that will allow it to accommodate) other manifestations of timing control. That is to say, we have sufficient grounds for eliminating a theory of coarticulation if the model of speech production which it promotes could never be made to generate well-known coarser-grained timing phenomena, such as rate effects, stress-timing and initial and final lengthening, in a natural or plausible way.

In fact, extrinsic timing theories do not adequately handle either rate effects or stress-timing. Consider rate first. Vowels produced at a fast rate are substantially shorter in duration than are their more leisurely counterparts. Their reduction in duration seems to be accomplished by a reduction in their "target" spatial coordinates (e.g. Harris, 1975). That is, the articulatory movements towards the canonical target (as estimated from the vowel produced in isolation) are less extensive for vowels spoken at a fast rate than for vowels produced at a comfortable rate. Apparently this is because they are produced with less effort or muscular force than slower vowels. See Gay & Ushijima, (1974). A typical concomitant is that their mid-vowel formant values are centralized (Lindblom, 1963).

This means of increasing speaking rate for vowels does not, and could not, have its parallel in respect to consonants. Rapidly spoken consonants are slightly shorter in duration than are leisurely consonants, and they are evidently produced by *increasing* effort or force to the relevant muscles (Gay & Hirose, 1973; Gay & Ushijima, 1974; Gay, Ushijima, Hirose & Cooper, 1974). Indeed consonants could not be produced rapidly by decreasing muscle force, because their essential articulatory properties include obstructing or totally occluding the passage of air from the lungs. If the muscle forces which effect that consonantal obstruction were diminished, a different class of segment would be produced.

This evident difference between vowels and consonants must be mimicked by a model of speech production and must be rationalized by its accompanying theory. However it is not mimicked in a natural way by a model in which feature bundles for consonants and vowels are laid out in left to right adjacency in a plan and are executed by a separate articulatory device. That device must approximately alternate, first executing a vowel at a slow rate (or equivalently first assigning a reduced-from-normal amount of muscular effort), and then reading out a consonant at a fast rate (or assigning it super-normal muscle force). But this dual strategy seems only to complicate the talker's task. Why does he not simply increase muscle force generally and produce all sounds according to the consonant strategy? Although the extrinsic timing device can generate these rate effects, it only does so in a *post-hoc* way that fails to rationalize a dual strategy.

The conclusion is similar in regard to stress-timing. There is now available fairly substantial evidence that English speakers regulate the intervals between stressed vowels in an utterance (see Fowler, 1977; also see Monsell & Sternberg, in press), and there are grounds for supposing that an extrinsic timing model cannot generate those intervals. The evidence of Morton *et al.*, (1976) of Rapp (1971) and of Allen (1972) suggests that the controlled intervals lie between the *centers* (however defined) rather than the *edges* of any linguistic units, including the phonological segments: If this is so, timing constraints can not be between feature bundles. However, even if an extrinsic timing theory could be made to generate stress-timed utterances, it can not rationalize the phenomenon. Indeed, like the different rate strategies for vowels and consonants, stress-timing would simply be an added burden for a plan executor.

However, it seems most unlikely that speech production would be easier for a talker if he were not obliged to produce the special rate effects or stress-timing: these phenomena must not be treated in a model as if they were merely arbitrary complications. Rather they are, and should be treated, as indicants of the strategy or style of control that talkers adopt when they produce an utterance. An adequate theory of timing control, then, is one which describes a style of control out of which these traces of timing control fall naturally. In the concluding section I will describe the theoretical perspective from which a more adequate account of coarticulation may perhaps be derived.

Coarticulation and intrinsic timing

Obligations of an intrinsic timing theory of coarticulation

The obligations of an intrinsic timing theory of coarticulation are at least these four:

(1) The theory must characterize the essential properties of segments as four dimensional entities.

(2) More abstractly, the theory must rationalize the classification of segments into vowels and consonants. This is evidently necessary if the model of coarticulation is to be compatible with the course-grained timing phenomena of rate and stress-timing (and if it is to capture our intuitions that consonants and vowels constitute different kinds of entities.)

(3) The theory must merge the plan and its executor by incorporating time into the plan for an utterance.

(4) The theory must rationalize coarticulatory effects, and a model derived from it must

be able to generate them.

Below I will characterize a way of conceptualizing speech production which may meet these obligations. The description is of necessity terse. For a more elaborate discussion, see Fowler, (1977) and Fowler *et al.*, (in press).

Coordinative structures

All activities that we perform are coordinated. (This is true even of our clumsiest actions. Compare them with the maximally uncoordinated convulsion.) In order for acts to be coordinated, the muscles that contribute to their realizations have also to be coordinated. If they were not, then different muscles would compete and unorganized movements would ensue. These functional organizations of muscles are called coordinative structures by Easton (1972). Their governance of acts such as locomotion (Easton, 1972; Grillner, 1975), swallowing, and chewing (Sessle & Hannam, 1975; Doty, 1968) are well documented.

Significant properties of a coordinative structure are that it generates an equivalence class of movements, that it is nested, and that it frequently is cyclic in nature.

In respect to the first property, the coordinative structure is an organization that spans several muscles and produces activities of a certain *kind*. The organization over the muscles may be described as a mapping. For example, that which regulates movement of the forearm at the elbow under some conditions may be described by the equation for a non-linear spring:

$$F = ae^{k}(l - l_{0})$$ (Feldman 1966).

The mapping has parameters, some of which are under the actor's control. In mapping above, they are k and l_0. When those parameters are given different values, different but similar movements ensue. Thus the organization described by the mapping engenders a *family* of acts.

The second property of coordinative structures – that they tend to be nested – is evidenced in the act of walking. During locomotion, small systems of muscles that govern intralimb stepping are nested within (or organized into) a large muscle system whose role is to coordinate the activities of two (or four) limbs (see Easton, 1972). Notably, the "life-span" of the small coordinative structures – that is the duration of time over which the small coordinative structures function – is shorter than that of the superordinate muscle system.

Finally, many coordinative structures, including those involved in walking, chewing, and in respiration are cyclic. That is, once a muscle system has run through its repertoire of activity (in walking, once a single step has been taken by each limb), the repertoire is reinitiated – that is, the "end" of a cycle reinitiates the sequence.

It is important to recognize the efficiency of this style of organization. Coordinative structures are self-executing organizations: once they have been marshalled, no further organizational intervention is required as they execute their special repertoire of activity. If an act is cyclic – like walking, breathing, chewing (and I will suggest vowel production) – acts of indefinite duration may be evoked by just once marshalling the requisite muscle organization.

The reader is referred to Turvey (1977) and to Greene (1972) for a verification of these properties of organized muscle systems. I will suggest below that a plan as a nesting of coordinative structures meets the obligations listed earlier for a model of intrinsic timing.

Coordinative structures in speech

First I will describe some of the coordinative structures involved in the act of speaking – both as regards its components which are considered to be involuntary and automatic and as regards those components which are considered voluntary. Next I will suggest how that proposed style of control meets the obligations of a theory of coarticulation.

The respiratory system

Basic reflexes operate both in vegetative breathing and in speech, apparently to regulate the initiation and termination of inspiratory activity (see, for example, Kaplan, 1971). Briefly, the expiratory "center" of the brain stem receives inhibitory innervation from stretch receptors in the alveoli of the lung. The expiratory cells inhibit their inspiratory counterparts and, *ceteris paribus*, terminate their activity. Chemoreceptors in the vicinity of these brain stem centers that are sensitive, for instance, to the level of CO_2 in the blood also regulate the centers' activities. These reflexes operate like control systems in working to regulate the CO_2 level of the blood.

When viewed more macroscopically than this, vegetative respiration and speech respiration are quite different. During vegetative breathing, the activity of the inspiratory muscles is in phase with the inspiratory portion of the respiratory cycle (Lenneberg, 1967). Furthermore, except during forced expiration, that phase is typically accomplished passively – i.e. by relaxing the muscles of inspiration and by allowing the elastic recoil forces of the lungs to work unaided and unopposed. The expiratory phase occupies about 60% of the cycle.

But during speech the sequence of events is somewhat different (Draper, Ladefoged & Whitteridge, 1958; Lenneberg, 1967). In speech, the activity of the inspiratory muscles is out of phase with the act of inspiration. Their activity extends into the early part of expiration where they act to check the descent of the ribcage that characterizes passive expiration. Immediately following the decline and offset of activity in the inspiratory muscles, the internal intercostals that are muscles of active expiration, come into play. If phonation is prolonged, two other muscles that may contribute to expiration, the *rectus abdominis* and the *latissimus dorsi*, are also marshalled (Draper *et al.*, 1959).

Two results of this coordinated activity are that the proportion of the respiratory cycle occupied by the expiratory phase is about 0·87 (Lenneberg, 1967) and that subglottal pressure is maintained at a nearly constant level despite the continual decrease in the volume of air in the lungs (Draper *et al.*, 1958; Lieberman, 1967). Lenneberg refers to these coordinated activities of the muscles of inspiration and expiration as "synergisms", a term that others have used as Turvey (1977) and Easton (1972) use the term coordinative structure. In the case of this macroscopic coordinative structure, it is a device whose task is to control subglottal pressure, much as the microscopic reflexes regulate stretch and CO_2 levels in the blood. The macroscopic coordinative structure is superimposed on the smaller reflexes when an individual chooses to speak.

Notably, in the experiment of Draper *et al.*, similar sequences of muscular events occurred over a range of controlled subglottal pressures. That is, the same coordinative structure may govern an utterance over all amplitudes of production.

The laryngeal system

Evidence that the operation of autonomous laryngeal devices contribute to speech production is sparse. What evidence there is, is provided primarily by Wyke (1967, 1974). In his view:

> [T]he production of speech by human beings is, in essence, similar to a large number of other acts of daily life that the human being fashions unthinkingly out of co-ordinated variations in the tone of striated muscles (See Wyke, 1959, 1967*b*). Obvious examples of such automatic acts performed with striated muscles are walking, mastication, swallowing and breathing. In all of these situations, the muscular performance is certainly capable of voluntary initiation, modification and arrest; but its detailed production, in the circumstances of normal everyday life, is continuously adjusted by reflex mechanisms that operate at a subconscious level, and over which we

> have no voluntary control. Speech is in a similar situation, as far as phonation goes: that is to day, the production of speech although initiated voluntarily, is dependent mechanistically upon the precise subconscious integration of a large number of feed-back (servo-) reflexes which constantly adjust the tone of the large number of muscles involved in the production. (1967: pp. 2–4)

Wyke has identified three servo-systems in the larynx itself. But the role of these systems in natural speech, if any, has not been established.

Receptors in the mucosal membranes of the larynx are sensitive to changes in air pressure. Wyke (1967) suggests that the afferents from these receptors may trigger reflex adjustments in the tone of the laryngeal and respiratory musculature during speech. Evidence of this, summarized in a recent paper (Wyke, 1974), shows that when subglottal pressure increases during phonation, the tone of the vocal fold adductors is increased and that of the abductors is concomitantly decreased. The adaptive result is that the folds resist the "upward ballooning" that would otherwise follow an upward surge in subglottal pressure.

In addition, Wyke and his colleagues (summarized in Wyke, 1967) have identified mechanoreceptors in all of the laryngeal joints (thyro-epiglottic, thyrohyoid, cricoarytenoid, cricothyroid). The receptors are structurally identical with skeletal joint receptors. They respond to displacement of the joint and their effect is to change the tone of intrinsic laryngeal muscles.

Finally, mechanoreceptors in the intrinsic muscles of the larynx respond to stretch and reflexively alter the tone of the intrinsic muscles.

Superimposed on these microscopic reflexes may be more macroscopic coordinative structures that are responsible for establishing the different laryngeal "modes" of phonation (including normal phonation, whispering, falsetto, creaky voice, etc.). These modes of phonation are established by adjusting, for an extended period of time, various properties of the larynx. According to Abercrombie (1967), the kinds of long-term adjustments that languages may prescribe, or in the case of whispering or the falsetto register, that speakers may choose adopt are the following:

> [T]he glottis may be entirely in vibration, or only in part and the part that is not in vibration (usually the so-called caritilage glottis) may be firmly closed, or may be sufficiently open to allow air to pass through at that point; two parts of the glottis may be in different modes of vibration simultaneously; the whole larynx may be raised or lowered in the throat; and the parts of the larynx above the glottis may or may not be constricted. For example, what was called above "breathy" phonation is produced by part of the glottis being in vibration while the cartilage glottis is sufficiently open to allow air to pass freely through it... (1967: p. 100)

The supralaryngeal system

On a microscopic scale, Folkins & Abbs (1975) provide suggestive evidence for a closed-loop, jaw-lip system that operates during speech production. In their experiment, they asked subjects to produce the phrases "a hae 'paep again" repeatedly. On about one-quarter of the repetitions, randomly interspersed, the experimenter applied a transient disturbance to the jaw. The disturbances were applied during the course of lip-closure for the first /p/ in the phrase, and were such that they prevented the jaw from reaching its usual degree of elevation. Nonetheless, on every repetition, perturbed or not, the subject attained lip closure due to exaggerated upper and lower lip displacements, and slightly exaggerated velocities of lip closing gestures. These results are explained most simply in terms of a low-level jaw-lip control system that is responsible for lip closure.

Several researchers (for instance, Kozhevnikov & Chistovich, 1965; Ohala, Hiki, Hubler &

Harshman, 1968; MacNeilage, 1970) have observed that the velocity of jaw closure for any consonant varies directly with the distance that the jaw has to travel to attain closure. The consequence is a nearly constant duration of the closing gesture independant of its extent. It is likely that the tongue tip behaves similarly as Sussman (1972) points out. Ohman's (1967) data show nearly identical formant transition durations in /idi/ and /ada/ although the tongue tip has to travel farther in the latter case in order to attain closure. The typical interpretation of these findings is that the velocity of movement is under closed-loop control such that it is adjusted to current conditions.

The supraglottal structures as well as the respiratory and laryngeal structures participate in the coordinated movements that are involved in chewing and swallowing. These acts quite clearly are products of nestings of synergies or coordinative structures (see for instance, Doty, 1968; Sessle & Hannam, 1975). Doty (1968) has shown that selective destruction of parts of the "swallowing center" of the brain-stem leaves intact some of the "subsynergisms" of the act of swallowing while impairing others. He suggests that some of these subsynergisms resemble speech gestures and may be marshalled during speech. In particular he suggests that the movements of the soft palate during speech are "lessened" forms of those that occur during swallowing.

Others (Shohara, 1936; Bosma, 1953; Fawcus, 1969) have made similar suggestions with regard to the synergies involved in the acts of chewing, swallowing and sucking. Shohara (1936) suggests, for instance, that the tongue gestures during the production of /k/ and /g/ are versions of those involved during swallowing. According to Fawcus (1969):

> Most, if not all of these fine movements of the oralpharyngeal structures (in speech) occur during mastication in both non-speaking children and other nonspeaking animals. The harnessing of these basic movements by the CNS is the essential problem in both the normal development of speech and the procedures designed to overcome developmental failure. (1969: p.558)

But none of these claims to my knowledge is based on any hard evidence that speech gestures are the products of synergies or coordinative structures, differently organized, that participate in other orophayngeal acts. The evidence appears rather to be informal. Some speech gestures resemble those involved in other acts that utilize the same structures.

Stronger evidence for macroscopic coordinative structures in speech derives from the speech production literature itself. One way in which researchers in other domains identify a coordinative structure is by means of its muscular or gestural concomitants. Recall that a coordinative structure is a group of muscles constrained to act as a system. It generates an act whose properties are stereotyped.

The speech literature provides some examples of this.

(1) Kent & Netsell (1971) find that the gestures of the tongue body and the lips are synchronized during the production of the word "we" in "we saw you". A figure relating the displacement of the tongue body to that of the lips over different stress patterns of the phrase (*we* saw you; we *saw* you; we saw *you*) shows that the relationship between the two variables is invariant over differences in stress. Kent & Netsell obtain a similar result for the diphthong /ɔi/ in "*con*voy" and "con*voy*", and tentatively conclude that "for sounds like /wi/ & /ɔi/ which are characterized by coordinated movements of two articulators, the stress contrast must alter both gestures or neither gesture." (p. 40).

(2) Kent, Carney & Severeid (1974) conclude on the basis of their data that in many cases "articulatory movements seem to be programmed as coordinative structures so that movements of the tongue, lips, velum and jaw often occur in highly synchronized patterns" (p. 487). Some of the data of Borden and Gay described earlier also reveals synchrony in the movements of different vocal tract structures.

(3) Of greater interest are cases when the gestures are not synchronized, but rather occur in a constrained pattern over time. Notice that evidence of this can only be obtained by

comparing similar utterances or by comparing the same utterance produced at different rates. That is, a non-synchronized pattern can only be detected if it remains invariant in different contexts. The recent data of Bell-Berti and Harris cited earlier, showing that the onset of liprounding precedes the measured acoustic onset of /u/ by a relatively fixed interval, regardless of the preceding consonantal context, seems to provide an instance of this.

Kent, Carney & Severeid provide some limited evidence that the relative timing of gestures of the tongue body, velum and lips tend to be invariant over a change in rate of speaking. For each articulator and for each of two speakers, figures are provided which superimpose the movement tracings at two different rates during the production of "soon the snow began to melt". The rates of speaking were in the ratio 2: 1. To facilitate a comparison of the movements' relative timing at the different rates, the investigators compressed the time-scale of the slower movement relative to that of the fast movement. For both speakers, and for all three articulators the tracing at the two rates of production were nearly identical. Thus the relative timing of the movements of a particular articulator and the timing relationship among articulators remained nearly invariant over a two-fold increase in rate. These findings are interesting, but questionable in that they seem to conflict with other evidence. For instance, Gay & Ushijima (1974) show that the muscle activity for consonants is of greater amplitude and that of vowels of lesser amplitude during rapid than slow speech. Of course these investigators provide only EMG data, and it is difficult to make inferences from muscle contraction to movement.

Coordinative structures that encompass the respiratory, laryngeal and supralaryngeal systems

Obviously unless the respiratory, laryngeal and supralaryngeal systems are in fact separate systems, there is not need to ask how they are coordinated. But the perspective on the action system suggested by Greene (e.g. 1972) and by Turvey (1977) is one in which coordinative structures are nested, and these three systems seem to represent a natural fractionation of the whole. Some limited evidence that the three systems or subsets of two of them are coordinated during speech is provided by the studies briefly described below.

(1) Perkell (1969) suggests that the "intent to lengthen the vocal tract" and the intent to shorten it are actualized by a coordinative structure controlling the lips, jaw, hyoid and larynx.

In his words:

> It follows that there is a physiological as well as an anatomical interaction between the lips, mandible, hyoid bone and larynx which causes these structures and the muscles connecting them to operate as a unit in shortening the vocal tract (to raise formant 1 and lower formant 2) for the three non high vowels /a, ae, ɛ/. . . the familiar lip rounding function, combined with vocal tract lengthening at the laryngeal end, comprises a physiological mechanism operating in opposition to the vocal-tract shortening function. (p. 40–41)

(2) Similarly, in the experiment of Folkins & Abbs (1975) described earlier, the investigators compared lip movements in the presence or absence of jaw movement impedance. The displacement of the lips was greater when jaw movement was restricted. However velocity of lip closure did not increase proportionately on those trials. Therefore lip closure was attained 15–25 ms later on the test trials than on the control trials. Folkins and Abbs observed that voicing offset was also delayed on those trials, suggesting to them a coordination of laryngeal and supralaryngeal structures. They also note an alternative interpretation, however. Delayed lip closure is accompanied by a delay in build-up of oral air pressure. Oral air pressure may contribute to voicing offset by reducing the transglottal airflow. Hence the delayed voicing offset may be a mechanical consequence of the delayed lip closure.

(3) The laryngeal reflex system whose receptors are sensitive to air pressure, according to Wyke (1967) has a respiratory as well as a laryngeal component. An increase in air pressure, as noted, leads to an increase in the tone of vocal fold adductors. In addition, Wyke cites some studies showing that these same receptors control unspecified alternations in the activity of respiratory muscles.

(4) Finally, Gould (1971) reports that changes in posture (e.g. standing vs. sitting curled in a chair) lead to reflexive alterations in the respiratory and laryngeal musculature.

Summary

The data describe in the four preceding sections are highly compatible with the mode of organization of the nervous system and musculature proposed by Greene and by Turvey. Although the data do not provide a clear picture of how all of the coordinative structures fit together to generate a coherent, homogeneous act of speech production, that is not surprising given that none of them was gathered with a view to supporting this theory of action. Indeed the quantity of the supportive data is remarkable in view of this.

Coarticulation in an intrinsic timing view of speech production.

The account of coarticulation to be developed here meets its obligations (as listed earlier) in these ways:

(1) Phonological segments are defined by the coordinative structures that are invoked in their realization. Since coordinative structures engender four dimensional acts, phonological segments are considered essentially or canonically four dimensional. Those phonological segments produced by the same set of coordinative structures (e.g. the class of vowels) are distinguished by the parameter values that the muscle systems are assigned. (The parameter values are equivalent, or nearly so to distinctive features).

(2) Consonants and vowels are distinguished by the coordinative structures that effect their realization. An argument can be made that all vowels are the product of a single set of coordinative structures invoked just once at the onset of an utterance. These muscle systems effect a characteristic kind of gesture (a relatively slow change in the global shape of the vocal tract) that distinguishes vowels as a class of segment. Consonants are produced by (a variety of) coordinative structures different from those that invariantly underlie vowe production. The vowel-producing system may be cyclically invoked, thereby yielding quasi-stress timing in languages such as English.

(3) The plan for an utterance is treated as identical to the coordinative structures themselves – that is with the patterning of physiological biases which organize the musculature. This decision does not preclude explaining anticipatory effects in speech production (i.e. speech errors, anticipatory shortening, anticipatory coarticulation). It does not, because, as noted, the coordinative structures are nested; the superordinate relationships are established both with respect to what the talker is doing now and with respect to what he will be doing.

(4) Coarticulation is characterized as the coproduction of four dimensional "canonical" forms.

The organization of a talker during speech

That vowels and consonants are coproduced has been suggested by Kozhevnikov & Chistovich (1965), by Ohman (1966, 1967) and by Perkell (1969). For Kozhevnikov & Chistovich, the initial consonant and the vowel of a C_oV syllable are initiated simultaneously by a talker. Given the absence in running speech, of articulatory gaps or pauses, this view implies that vowels are continuously produced. (More accurately from the perspective of Kozhevnikov & Chistovich it implies that a second vowel is initiated as soon as a preceding vowel is terminated.) The production of a consonant or a consonant cluster, then, is imposed on a background of continuous vowel production.

This view is stated more explicitly by Ohman (1966, 1967) based on acoustic data show-

ing that the formant transitions into (and out of) a medial stop consonant in a VCV vary with the identity of the following (and preceding) vowel. Again, if this is the case in a VCV, by implication vowel production may be continuous throughout the course of running speech.

Apparent coproduction of vowels and consonants may, but need not, imply that the two kinds of speech gestures are produced by different articulatory systems. A way in which coproduction might be approximated by a single articulatory system is by means of feature spreading. However, I have already argued that this view is implausible. A more plausible way, as suggested both by Ohman and by Perkell, is for the articulatory systems responsible for vowel and consonant production to be distinct.

Ohman notes that the global shape of the vocal tract during a stop closure is irrelevant to the phonemic identity of the consonant. It can vary without altering the identity of the stop for a listener. Thus if it is mechanically feasible, a "distorted" vowel gesture (Ohman, 1966) can be executed by the tongue-body during the closures for bilabial consonants, for alveolar consonants, and even, he suggests, for consonants involving the body of the tongue. His acoustic evidence bears this out as does the more recent articulatory data of Butcher & Weiher (1976) and of Barry & Kuenzel (1975).

What makes coproduction mechanically feasible to Ohman and Perkell is that the production of vowels and consonants involve essentially different (but overlapping) sets of muscles. It is perhaps worth quoting Perkell at some length on this point:

> The behavior of the vocal tract differs in several respects for the production of vowels and consonants. The division of the observed parameters of vocal-tract behavior into classes based on the articulatory and acoustic distinctions between vowels and consonants suggest criteria that can serve as the bases for a physiological model. Many parts of the vocal tract play a role in the production of both vowels and consonants, but in general, the same organs seem to behave differently under the influence of the two differeent classes. Consonant articulations by the tongue and lips are generally observed to be faster and more geometrically complex, and they require more precision in timing than vowel articulations. . . To some extent there also seems to be an anatomical division. For example, the tongue tip is more active in consonant articulation, whereas the body of the tongue is active in articulating both consonants and vowels.
>
> The general differences in velocity, complexity, precision of movement, and in anatomy suggest that different types of muscles are generally responsible for consonant and vowel production. It is probable that articulation of vowels is accomplished principally by the larger, slower extrinsic tongue musculature which controls tongue position. On the other hand, consonant articulation requires the addition of the precise, more complex, and faster function of the smaller intrinsic tongue musculature. (p. 61).

Taken together, these observations, and extrapolations from them indicate that true coproduction occurs in speech, and that the capacity for coproduction derives from an adaptive property of speech that the two classes of articulatory gestures, consonants, and vowels, are products of different (coordinated) neuromuscular systems.

This property is adaptive for the following reason. Any two vowels are more similar in their acoustic form than any vowel is to a consonant. Their acoustic form, of course, is consequent on their manner of production. Vowels are produced as (relatively) slow alternations in the global shape of the vocal tract effected in large part by repositioning the tongue-body in the mouth. This characterization is common to vowels, but is not common to consonants and vowels. If vowels and consonants are produced by different neuromuscular systems, then vowels may be continuously produced as suggested above; those articula-

tory properties that are common to all vowels are invariant over the course of an utterance. Conceivably they are evoked anew at the initiation of each vowel. However a more parsimonious supposition is that they are evoked just once for the whole course of an utterance.

In other words the supposition is that the set of articulatory properties that is common to the vowels, and that defines them as a natural class of gestures, may be a product of a superordinate coordinative structure whose time-scale is slow relative to that of the ongoing speech gestures. If vowel production is continuous, then this hypothetical coordinative structure can be evoked just once for the course of an utterance. If instead, vowel production alternates with consonant production, then the talker cannot exploit the properties of vowels that define their equivalence. Rather, he has to re-voke the invariant (as well as the variant) properties of the class of vowels at the initiation of each new vowel. Below I will try to establish what the invariant properties of the class of vowels are, and then to characterize the coordinative structure that evidences them.

By hypothesis, five muscle systems are coordinated to produce a vowel: the first is the respiratory system characterized earlier; the second is the intrinsic musculature of the tongue which on a first approximation is set invariantly across the vowels (Perkell, 1969); two systems are responsible for adjusting the length of the vocal tract (see Perkell, 1969, p.40–41) and finally a system is responsible for moving the tongue within the oral cavity. These systems are characterized in Fowler (1977).

The last vowel system may be described (at least metaphorically) as if it were a vibratory system with its resting length as a tunable parameter. Even a simple vibratory system, for example a linear spring described by the equation

$$-F = k(l - l_0),$$

has some properties not unlike those of the articulatory system during vowel production.[4]

To describe the production of vowels, we need a system for positioning the tongue which captures those properties of the class of vowels that are equivalent across its membership. A spring-like equation with the parameter l_0 unspecified would satisfy that criterion because it represents a system which is established invariantly for every vowel. l_0 is a parameter whose particular value distinguishes one vowel from another.

In addition, we are looking for a description of a *particular* vowel that enables us to describe as equivalent the different gestures that instantiate it in different contexts. (Recall that for /ɛ/, the tongue is lowered following /i/, but raised following /a/). Possibly l_0 will work to do that if we describe it at an abstract level as the zero-state of the extrinsic tongue system. l is the actual state of the tongue system (the actual position of the tongue).

Consider what happens to a spring system when F and k are unchanged but l_0 is reduced in magnitude. To counteract the same F, the system decreases l by the same amount as l_0 was decreased. Thus l alters *in the direction of the new* l_0. (For example, let $l_0 = 10$ in arbitrary units, $F = 50$, $k = 25$. Transforming the equation above, $l = 10 - (F/k) = 10 - 2 = 8$. If now l_0 is reset to 5, $l = 5 - 2 = 3$.) Again suppose that l_0 corresponds to the zero-state of the extrinsic tongue system – to the position that the tongue would adopt if $-F = 0$, and l is the actual position of the tongue. If a talker is able to alter l_0 volitionally, as the subjects of Fel'dman could (see above), then he can change the zero-state of the extrinsic musculature. Suppose that to a particular vowel corresponds a particular value of l_0. A new vowel is initiated by changing the value of l_0. In consequence of this change in the value of l_0, the invariant shape of the tongue in the mouth alters its location. In the case of

[4]The equation for a linear spring is $-F = k(l - l_0)$ where l_0 is the spring resting length (that is the length of the spring in the absence of any force exerted on it). l is the current length of the spring; k is the stiffness parameter; and F is the force developed by the spring. Turvey (1977) suggests that all coordinative structures may be vibratory systems, though not necessarily linear.

the vowel /ɛ/ its l_o is less than that for the vowel /i/ and greater than for /a/. When l_{oe} is substituted for l_{oa}, the tongue moves upwards; when l_{oe} is substituted for l_{oi}, the tongue moves downwards. Although the gesture is different for /ɛ/ in different contexts, the parameter l_{oe} is the same. And therefore it is incorrect to equate the view of the extrinsic muscle system of the tongue as a spring system with a feature-bundles or targets view. The reason why this is so can be stated in two ways. First, l_o is just a parameter of a system. /ɛ/ is the functioning of the system *when the spring function is assigned the parametric value* l_{oe}, but /ɛ/ is not identical with the parameter itself. Equivalently it is as incorrect to equate l_{oe} with /ɛ/ as it is to equate some curve with its asymptote. l_{oe} is just the limiting shape of the vocal tract (as controlled by the position of the tongue) towards which /ɛ/ invariantly aims.

Now consider how this proposal handles some coarticulatory phenomena.

Coarticulatory effects are due to the coproduction of consonants and vowels and of stressed and unstressed vowels. Consider first the coarticulatory effects of vowels on consonants. Lip rounding precedes the measured acoustic onset of a rounded vowel and therefore coarticulates with consonants that precede a vowel. This occurs, we suppose, not because the feature [+rounding] has attached itself in the plan to the preceding consonants, but rather because the vowel /u/ is coproduced with them. Vocal tract lengthening via lip rounding is, then, an observable correlate of co-production. Another correlate, reported by MacNeilage & DeClerk (1969), is that the tongue-body configuration for the vowel may be attained during the closure for a consonant (see also Butcher & Weiher, 1976; Barry & Kuenzel, 1975).

Similarly, velar opening precedes the gesture of the primary articulator for a nasal consonant. This occurrence probably parallels that of liprounding for the vowels. That is, it may participate in a coordinative structure, the component gestures of which are patterned over time rather than being synchronized. Velar opening occurs during a vowel, then, because vowels and consonants are coproduced. Parallel to the finding of Bell-Berti & Harris described earlier, one might expect velar opening to precede the major articulatory gestures for an /m/ or /n/ by a relatively fixed duration, and not to be synchronized consistently with the gestures towards a vowel target.

Coarticulatory vowel-to-vowel effects may again be explained as owing to coproduction. The left-to-right and right-to-left transconsonantal effects observed by Fowler (1977) of stressed vowels on F_2 of an intervening unstressed vowel is explained most naturally as coproduction. The production of an unstressed vowel is superimposed on a trajectory of the shape of a vocal tract from one stressed vowel to another (*cf.* Martin, 1972).

Concluding remarks

This section has provided a sketchy and speculative view of intrinsic timing in speech production. However, I believe that it can account for coarticulatory and other timing effects more plausibly and adequately than the views developed within the extrinsic timing framework. Its major advantage is in incorporating the dimension of time into the specification of a phonological segment with the consequence that the ideal or canonical form is considered to be executed unaltered in an utterance (*cf.* Fowler *et al*, in press).

I thank Michael Studdert-Kennedy for criticizing an earlier draft of the paper. This work was supported by NIH Grant NS13617 to Haskins Laboratories.

References

Abercrombie, D. (1967). *Elements of General Phonetics.* Chicago: Aldine

Abramson, A. (1975). Laryngeal timing in consonant distinctions. Paper presented at the *Eighth International Congress of Phonetic Sciences*, Leeds, England, August, 17–23.

Allen, G. (1972). The location of rhythmic stress – beats in English: An experimental study I. *Language and Speech* **15**, 72–100.
Barry, W. & Kuenzel, H. (1975). Co-articulatory airflow characteristics of intervocalic voiceless plosives. *Journal of Phonetics* **3**, 163–282.
Bell-Berti, F. & Harris, K. (1979). Anticipating coarticulation: Some implications from a study of liprounding, *Journal of the Acoustical Society of America*, **65**, 1268–1270.
Bosma, J. F. (1953). A correlated study of the anatomy and motor activity of the upper pharynx by cadaver dissection and by cinematic study of patients after maxillofacial surgery. *Annals of Otology, Rhinology and Laryngology* **62**, 51–72.
Butcher, A. & Weiher, E. (1976). An electropalatographic investigation of coarticulation in VCV sequences. *Journal of Phonetics* **4**, 59–74.
Daniloff, R. G. & Hammarberg, R. E. (1973). On defining coarticulation, *Journal of Phonetics* **1**, 239–248.
Doty, R. W. (1968). Neural organization of deglutition. In *Handbook of Physiology* (C. F. Code ed.) Vol. IV Sec. 6 pp. 1861–1902. Washington: Am. Physiol. Soc.
Borden, G. J. & Gay, T. (1975). Durations of articulator movement for /s/ -stop clusters. *Haskins Laboratories Status Reports on Speech Research* **SR-44**, 147–161.
Draper, M., Ladefoged, P. & Whitteredge, D. (1958). Respiratory muscles in speech. *Journal of Speech and Hearing Research* **2**, 6–27.
Easton, T. (1972). On the normal use of reflexes. *American Scientist* **60**, 591–599.
Fawcus, B. (1969). Oropharyngeal function in relation to speech. *Developmental Medicine and Child Neurology* **11**, 556–560.
Feldman, A. G. (1966). Functional tuning of the nervous system with control of movement or maintenance of a steady posture-II. Controllable parameters of the muscles. *Biophysics* **11**, 565–578.
Folkins, J & Abbs, J. H. (1975). Lip and jaw motor control during speech: responses to resistive loading of the jaw. *Journal of Speech and Hearing Research* **18**, 207–220.
Fowler, C. A. (1977). *Timing Control in Speech Production.* Bloomington, Indiana University Linguistics Club.
Fowler, C. A., Rubin, P., Remez, R. & Turvey, M. T. (in press) Implications for speech production of the general theory of action. In *Speech Production.* (B. Butterworth, ed.) New York: Academic Press.
Fowler, C. & Turvey, M. T. (in press). Observational perspective and descriptive level in perceiving and acting. In *Cognition and the symbolic processes II* (W. Weimer & D. Palermo, eds) Hillsdale, N. J. : Erlbaum.
Gay, T. & Hirose, H. (1973). Effect of speaking rate on labial consonant production: A combined electromyographic/high speed motion picture study. *Phonetica* **27**, 44–56.
Gay, T. & Ushijima, T. (1974). Effect of speaking rate on stop consonant-vowel articulation. *Haskins Laboratories Status Reports on Speech Research* **SR-39/40**, 213–217.
Gay, T., Ushijima, T., Hirose, A. & Cooper, F. S. (1974). Effect of speaking rate on labial consonant-vowel articulation. *Journal of Phonetics* **2**, 47–63
Gibson, J. J. (1977). The theory of affordances. In *Perceiving, acting and knowing: Toward an Ecological Psychology.* (R. Shaw & J. Bransford, eds) pp. 67–82. Hillsdale, NJ: Erlbaum.
Gould, W. J. (1971). Effect of respiratory and postural mechanism upon action of the vocal cords. *Folia phoniatrica* **23**, 211–224.
Greene, P. H. (1972). Problems of organization of motor systems. In *Progress in theoretical biology*, V. 2. (R. Rosen & F. Snell, eds) New York: Academic Press, 304–335.
Grillner, S. (1975). Locomotion in vertebrates. *Physiological Reviews* **55**, 247–304.
Hammarberg, R. (1976). The metaphysics of coarticulation. *Journal of Phonetics* **4**, 353–363.
Harris, K. (1975). Vowel duration change and its underlying physiological mechanisms. Paper presented at the 50th session of the American Speech and Hearing convention.
Kaplan, H. M. (1971). *Anatomy and Physiology of Speech.* 2nd en, New York: McGraw-Hill.
Kent, R., Carney, P. & Severeid, L. (1974). Velar movement and timing: Evaluation of a model for binary control. *Journal of Speech and Hearing Research* **17**, 470–488.
Kent, R. D. & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics* **5**, 115–117.
Kent, R. D. & Moll, K. L. (1975). Articulatory timing in selected consonant sequences. *Brain and Language* **2**, 304–323.
Kent, R. & Netsell, R. (1971). Effects of stress contrast on certain articulatory parameters. *Phonetics* **24**, 23–44.
Kozhevnikov, V. A. & Chistovich, L. A. (1975). *Speech Articulation and Perception, J. P. R. S.* **30**, 543.
Lashley, K. (1951). The problem of serial order in behavior. In *Cerebal Mechanisms in Behavior.* (L. A. Jeffress, ed.) pp. 506–528. New York: Wiley.
Lenneberg, E. (1967) *Biological foundations of Language*, New York: Wiley.

Liberman, A. M., Cooper, F. S., Shankweiler, D. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review* 74, 431–461.
Liberman, A., Delattre, P., Gerstman, L. & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology* 52, 127–137.
Liberman, A. & Studdert-Kennedy, M. (1978). Phonetic perception. In *Handbook of Sensory Physiology*, (R. Held, H. Leibowitz, & H. L. Teuber eds.) Vol. VIII, "Perception". Heidelberg: Springer-Verlag.
Lieberman, P. (1967). *Intonation, Perception, and Language.* Cambridge, Mass: MIT Press.
Lisker, Leigh (1972). On time timing in speech. In *Current trends in linguistics*, Vol. XII. (T. Sebeok, ed.) pp. 2387–2418. The Hague: Mouton.
Mackay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia* 8, 323–350.
MacNeilage, P. (1970). Motor control of serial ordering of speech. *Psychological Review* 77, 182–196.
MacNeilage, P. & J. L. DeClerk (1969). On the motor control of coarticulation in CVC monosyllables. *Journal of the Acoustical Society of America* 45, 1217–1233.
MacNeilage, P. & Ladefoged, P. (1976). The production of speech and laguage. In *Handbook of perception* Vol. 7: *Language and Speech.* (M. P. Friedman & E. C. Carterette, eds) pp. 75–120. New York: Academic Press.
Martin, J. (1972). Rhythmic (hierarchical) vs serial structure in speech and other behavior, *Psychological Review* 79, 487–509.
Morton, J., Marcus, S. & Frankish, C. (1976). Perceptual centres (P-centres). *Psychological Review* 83, 405–408.
Neisser, U. (1976). *Cognition and Reality.* San Francisco: Appleton-Century–Crofts.
Ohala, J., Hiki, S. Hubler, S. & Harshman, R. (1968). Photoelectric methods of transducing lip and jaw movements in speech. UCLA *Working Papers in Phonetics* 10, 135–144.
Ohman, S. (1966). Coarticulation in VCV utterance: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151–168.
Ohman, S. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America* 41, 310–320.
Perkell, J. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study.* Cambridge, Mass.: MIT Press.
Polanyi, M. (1958. *Personal knowledge.* Chicago: University of Chicago Press.
Rapp, K. (1971). A study of syllable timing. Papers from the Institute of Linguistics, University of Stockholm, November, 14–19.
Sessle, Barry J. & Hannam, Alan G. (eds) (1975). *Mastication and swallowing: biological and clinical correlates.* Toronto: University of Toronto Press.
Shohara, H. (1936). The genesis of the articulatory movements of speech. *Quarterly Journal of Speech* 21, 343–348.
Sternberg, S., Monsell, S., Knoll, R & Wright, C. (1978). The latency and duration of rapid movement sequences: Comparison of speech and typewriting. In *Information Processing in Motor Control and Learning.* (G. Stelmach, ed.) New York: Academic Press.
Sussman, H. (1972). What the tongue tells the brain. *Psychological Bulletin* 77, 262–272.
Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In *Perceiving, Acting and Knowing: Towards and Ecological Psychology.* (R. Shaw & J. Bransford, eds) pp. 211–265 Hillsdale, N. J.: Erlbaum.
Wickelgren, W. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behaviour. *Psychological Review* 76, 1–15.
Wyke, B. (1967). Recent advances in the neurology of phonation: phonatory reflex mechanisms in the larynx. *British Journal of Communication Disorders* 2, 2–14.
Wyke, B. (1974). Laryngeal myotactic reflexes and phonation. *Folia Phoniatrica* 26, 249–264.