95d

CHAPTER 2

# Speech Production

**Carol A. Fowler**

This chapter addresses the question of how a linguistic message can be conveyed by vocal-tract activity. I make my job easier by considering linguistic messages only in their phonological aspect. So the question is not how linguistically structured *meanings* can be conveyed by vocal-tract activity, but rather, how language *forms* can be conveyed. Further, I do not discuss production of prosodic structure in speech, including its intonational and stress patterning. Despite these considerable simplifications, many difficulties remain in finding an answer.

A fundamental issue to confront right away concerns the extent of the mutual compatibilities among the different levels of description of the message that talkers, in one sense or another, embody as they speak. I consider three levels of description, those of the phonological forms of *linguistic competence* (language users' knowledge of their language), forms in a speaker's plan for an utterance, and forms in vocal-tract activity. It will become clear that a theory of speech production is profoundly shaped by the theorist's view of the compatibilities or incompatibilities among these levels (see Bock, Chapter 6, and Frazier, Chapter 1, this volume).

*Speech, Language, and Communication*
Copyright © 1995 by Academic Press, Inc. All rights of reproduction in any form reserved.

# I. PHONOLOGICAL FORMS OF LINGUISTIC COMPETENCE

Phonological theories are intended to describe part of speakers' linguistic competence. The particular domain of a phonological theory includes an appropriate description of the sound inventories of languages and of the patterning of phonological segments in words. Before the mid-seventies, phonological theories generally took one fundamental form, now known as *linear*. These theories described phonological segments in a way that implied a poor fit between consonants and vowels as known to the language user and as implemented in articulation. Since the mid-seventies, *nonlinear* phonologies have been proposed. In some implementations, nonlinear theories have considerably improved the apparent fit between phonological segments as known and as produced.

## A. A Linear Phonology

The most prominent, relatively recent, linear theory is generative phonology (Chomsky & Halle, 1968). In that class of theory (see Kenstowicz & Kisseberth, 1979, for a tutorial presentation), lexical entries for words include a specification of the word's phonological form represented as a (linear) sequence of consonants and vowels. Consonants and vowels themselves are represented as columns of feature values. For example, *bag* would have the following featural representation:

**Phonological segments**

| Features | b | æ | g |
|---|---|---|---|
| vocalic | − | + | − |
| high | − | − | + |
| back | − | − | + |
| low | − | + | + |
| anterior | + | − | − |
| coronal | − | − | − |
| voice | + |  | + |
| continuant | − |  | − |
| nasal | − |  | − |
| strident | − |  | − |
| round |  | − |  |
| tense |  | − |  |

(Vowels are not specified for the features voice, continuant, nasal, or strident; contextual influences aside, being vocalic implies +voice, +continuant, −nasal, and −strident. Consonants are not specified for rounding and tenseness.)

Features represent articulatory and acoustic attributes of phonological segments. However, for at least two reasons, the articulatory specification that they provide is abstracted away from, or is just plain unlike, the articulations that convey the segments to a listener. One reason for the abstraction has to do with a fundamental goal of most phonological theories, that of expressing as rules systematicities in the distribution of phonological properties in words. In English, syllable-initial voiceless stop consonants (/p, t, k/) are aspirated (breathy). Compare the /k/ sounds in *key*, where /k/ is syllable-initial, to the /g/-like /k/ in *ski*, where the /k/ is syllable-internal. However, the lexical entry for *key* will not include the information that the /k/ is aspirated, because aspiration is predictable by rule. In general, lexical entries for words indicate only properties of the phonological segment that are idiosyncratic to that word. Rules fill in the systematic properties of words. In articulation, of course, actions implementing all feature values of a word, whether lexically distinctive or not, must be provided.

Kenstowicz and Kisseberth (1979) offer some indications that the distinction between systematic and idiosyncratic properties of words is part of a language user's knowledge. Two such indications are speech errors and foreign accents. The authors report a speech error in which *tail spin* (phonetically, [t$^{h}$eyl spIn]) became *pail stin,* pronounced [p$^{h}$eyl stIn]. This is an exchange error, in which, apparently, /p/ and /t/ exchanged places, but aspiration (indicated by the raised [h] in the phonetic representation) remained appropriate to its context. One interpretation of this outcome is that /p/ and /t/ were exchanged before the aspiration rule was applied by the speaker. Had the exchange occurred between the pronounced forms of /p/ and /t/, the result would have been [peyl st$^{h}$In]. As for foreign accents, they can be described as the inappropriate application of the phonological systematicities of the speaker's native language to utterances in the foreign language. For example, native English speakers producing French, find it difficult to avoid aspirating syllable-initial voiceless stops in French words, even though French does not have an aspiration rule. (Accordingly, they erroneously pronounce French *pas* as [p$^{h}$a] rather than [pa].) It has not proven a straightforward matter to decide what properties of a language are sufficiently regular that they should count as systematic properties to be abstracted from lexical entries and later applied as rules. Chomsky and Halle (1968) had a rather low threshold for acceptance of a property as regular and therefore had many rules and quite abstract lexical entries.

A different reason why the lexical entries of a linear phonology provide phonological segments that are far from their articulatory implementations concerns their representation of consonants and vowels as distinct feature columns. In the lexical entry, the columns are discrete one from the other (in that they do not overlap along the abstract time axis), the features are static (i.e., they represent states of the vocal tract or of the consequent acoustic signal), and the featural representation for a segment in a feature column is context free. However, in speech production, actions relating to distinct consonants and vowels overlap in time, the vocal tract is in continuous motion, and, at least at some levels of description, its manner of implementing a given feature of a consonant or vowel is context sensitive.

## B. Nonlinear Phonologies

Development of nonlinear phonologies was not spurred by concern over this apparent mismatch between phonological competence and articulatory performance. It was fostered by failures of the linear phonologies to handle certain characteristics of some phonological systems. In particular, Goldsmith (1976) argued that an implicit assumption of linear phonologies (which he termed the *absolute slicing hypothesis*) is refuted by data from many languages. The absolute slicing assumption is that every feature's domain is the same as every other feature's domain, namely, one feature column's width. However, in fact, the domain of a feature can be less or more than one column.

In some languages, there are *complex segments,* such as pre- or postnasalized stops that behave in most respects like one segment, but they undergo a feature change (from [+nasal] to [−nasal] or vice versa) in midsegment. In other languages, nasality may span more than one column. Van der Hulst and Smith (1982) describe work of Bendor-Samuel (1960) on a language, Terena, in which the first-person possessive of a noun is expressed in a word by addition of a nasality feature to the word. The feature spans every segment beginning at the left edge of the word up to the first stop or fricative. That segment becomes a prenasalized obstruent. So, for ~~example, /ʻowoku/ (*his house*) becomes /ʻõw̃õŋgu/ (*my house;* ˜ represents~~ nasalization, and /ŋg/ is a prenasalized velar stop).

There is another way in which some features may participate in the phonological system differently from others. In particular, they may be differentially likely to undergo a phonological process in which other features participate. For example, some features evade deletion when other features of a segment are deleted. In another example from van der Hulst and Smith (taken from Elimelech, 1976), in the language Etsakọ there is a reduplication rule so that a noun X becomes each X by undergoing reduplication. For example, ówà (*house*), becomes ówǒwà (*each house;* ´ is a high tone on the vowel, ` is a low tone, and ˇ is a rising tone analyzed as a low followed by a high tone.) The first /a/ is dropped in the reduplication process, but its tone is not deleted. The tone attaches to the following /o/ forming a rising tone with the high tone already on the vowel.

These phenomena suggest that the feature column metaphor does not provide an accurate reflection of phonological competence. As an alternative, Goldsmith (1976; also see Goldsmith, 1990) proposed an analysis in which some features are represented as occupying different tiers from others; features on one tier become associated with those on others by rule. A prenasalized stop might be represented as follows (from van der Hulst & Smith, 1982) with the nasality feature occupying its own tier:

```
[+nas]  [−nas]
         −syll
         +cons
         −high, etc.
```

Compatibly, the analysis of reduplication would be represented as follows (with H and L representing high and low tones, respectively). Notice that tonal features occupy a separate tier from segmental features (indicated jointly by the symbols for each consonant or vowel):

```
H   L     H    L  H    L     H    L H  L
|   | →   |    |  |    | →   |    \|   |
o w a    o w   o w a   o w   o w a
```

In both of these examples, features that distinguish themselves from others (in the size of their domain and in their surviving application of a deletion rule, respectively) occupy tiers distinct from those that do not distinguish themselves from others. Examination of phonological processes reveals a patterning in this regard. Some features commonly participate jointly in a phonological process. Some appear never to do so. As Clements (1985) points out, this observation reveals another insufficiency of the feature column notion. The column has no internal organization; however, the set of features does. Clements' examination of the relative tendencies for subsets of features to participate jointly or not in phonological processes suggested a hierarchical organization of featural tiers. Features separating lower down in the hierarchy were those that participated jointly in more phonological processes than features that separated early in the hierarchy. To a remarkable degree, the organization of featural tiers in Clements' hierarchy mirrored patterns of anatomical independence and dependence in the vocal tract, and it is important, therefore, to realize that the model was not developed in order to mirror those patterns. Rather, as noted, it was designed to mirror patterns of featural cohesion as reflected in the features' joint participation or not in phonological processes. The fact that an organization derived this way does reflect vocal-tract organization suggests a considerably better fit between elements of the phonology and capabilities of the vocal tract than linear phonology, a matter of importance to us here. In addition and more fundamentally, perhaps, it suggests a major impact of the performance capabilities and dispositions of the vocal tract on the historical development of phonological systems.

Articulatory phonology (Browman & Goldstein, 1986b, 1989, 1990a,

1992) constitutes an even more radical movement away from linear phonologies and toward a phonology that should be fully compatible with vocal-tract capabilities. This phonology is articulatory in two major senses. First, its primitives are neither features of consonants and vowels nor phonemes, but, rather, gestures of the vocal tract and constellations of gestures. Gestures are "characteristic patterns of movement of vocal tract articulators or articulatory systems" (Browman & Goldstein, 1986b, p. 223). They are, in addition, like features of a linear phonology, units of contrast (i.e., a change of one feature in a linear theory or a change of one gesture in articulatory phonology can change one word of the language to another word). Phonology itself is defined (Browman & Goldstein, 1992, p. 56) as "a set of relations among physically real events, a characterization of the systems and patterns that these events, the gestures, enter into."

Articulatory phonology is explicitly articulatory in a second sense as well. Browman and Goldstein (1989) proposed the hierarchy of anatomical subsystems in Figure 1 based entirely on patterns of anatomical dependence or independence in the vocal tract, not, as Clements had, on patterns of featural cohesion in phonological processes. Browman and Goldstein's hierarchy, therefore, should constitute a language universal base on which individual languages may elaborate, but which they cannot reorganize because the dependencies are grounded in the anatomy of the speech production system.

In Figure 2, the hierarchy of Figure 1 has been rotated 90 degrees so that

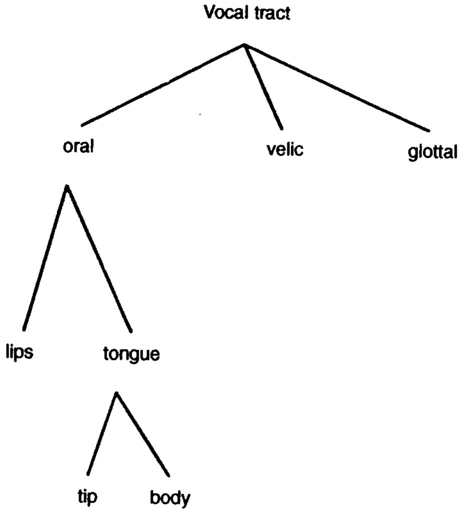

**FIGURE 1** Browman and Goldstein's hierarchy of gestural subsystems. (Redrawn from Browman & Goldstein, 1989.)

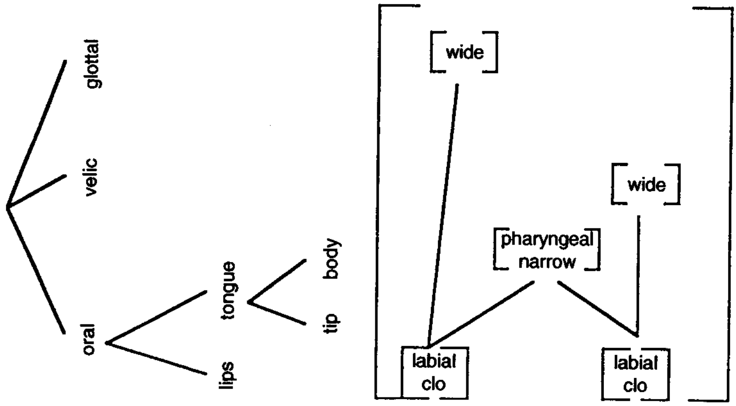

**FIGURE 2** A gestural score for the word *palm*. (Redrawn from Browman & Goldstein, 1989.)

its terminal nodes face the gestural score for the word *palm* (from Browman & Goldstein, 1989). The gestural score provides parameterizations for dynamical gestures associated with the articulatory subsystems involved in producing *palm*. In the figure, parameter values (in brackets) in the gestural score lie to the right of the corresponding parameterized articulatory subsystems in the rotated hierarchy. Thus, the parameter value [wide] to the right of the glottal subsystem signifies that, initially in the word *palm*, the glottis is open. Below that, the parameter values [labial] and [clo] opposite the lips subsystem provide the constriction location (labial) and degree (clo, for closed) of the word-initial consonant /p/. [Pharyngeal] and [narrow] characterize the constriction location and degree, respectively, of the tongue body during /a/ in *palm*. For the final consonant, /m/, the velum is lowered [wide], and the lip subsystem is parameterized as it was for /p/. In general, the gestural score for a word indicates the relative timing of successive gestures in a word; the association lines in the figure indicate gestures that are explicitly phased with respect to the others.

It is not necessary to achieve a deeper understanding of articulatory phonology to appreciate that it provides a good fit to the capabilities of the vocal tract in both central respects in which it is an articulatory phonology. First, the primitives of the phonological theory are dynamical actions, not static attributes of idealized abstract categories. Indeed, they are the patterns of movement in which the vocal tract engages during speech production. Second, articulatory phonology provides patterns of dependency among features (now gestures) that are wholly natural in reflecting patterns of anatomical dependency.

We must ask, however, whether there is any cost associated with this apparent benefit. Can the theory in fact characterize "the systems and patterns that these events, the gestures, enter into" (Browman & Goldstein, 1992)? For example, can the theory make the distinction that Kenstowicz and Kisseberth consider fundamental to a phonological theory between systematic and idiosyncratic properties of words?

Articulatory phonology does identify some systematic properties of words, for example, in rules that phase gestures one with respect to the other. To pursue the aspiration example, English voiceless stops are aspirated syllable-initially because of the way that the glottal opening gesture is phased with respect to release of the consonantal constriction. (Regulated phasings are indicated by association lines in the gestural score of Figure 2.) However, these systematic properties are not abstracted from the lexical entries for words. Accordingly, there is no distinction in the theory between words as represented lexically and as pronounced. It is fair to ask, then, how, if at all, articulatory phonology explains the two phenomena alluded to earlier, and mentioned by Kenstowicz and Kisseberth as evidence for the psychological reality of rule application by speakers, namely, errors such as [$p^h$eyl stIn] and foreign accents that preserve the systematic properties of the speaker's native language.

Articulatory phonology has an account of errors such as [$p^h$eyl stIn] that does not require an inference that a rule is applied to an abstract lexical entry: the exchange occurred between two constriction gestures (or just their location parameters) not between two phonemes. As for foreign accents, the theory has not explicitly addressed the question. However, there are at least two accounts that it could invoke. An unappealing possibility is to suppose that systematic properties are represented explicitly as rules and are represented implicitly across the set of lexical entries, and these explicit rules are applied to novel forms. An alternative account is to suppose that novel forms are pronounced by analogy with similar existing forms in the lexicon.

Clearly, theories of speech production will differ significantly depending on the theory of phonology they adopt; at the extremes, a linear theory in which phonological features of consonants and vowels cannot be implemented in a literal or analogical way in the vocal tract, or articulatory phonology in which phonological primitives *are* vocal-tract actions.

## II. PLANNING UNITS IN SPEECH PRODUCTION

### A. Speech Errors as Evidence for Planning Units

Some planning must occur in language production, because the nature of syntax is such that dependencies (e.g., subject–verb agreement) may be established between words that are far apart in the sentence to be produced.

The occurrence of anticipatory speech errors such as "we have a laboratory in our own computer" (from Fromkin, 1971) verifies the occurrence of advance planning. For the present, our interest is in the nature of the units used to represent the planned utterance. The most straightforward assumption, and the one generally made, is that they are the units of linguistic competence. Indeed, investigators have used evidence about planning units as revealed in speech errors as a way to assess the psychological reality of proposed units of linguistic competence. For example, according to Stemberger (1983, p. 43): "Speech error data argue for the psychological reality of many basic phonological units."

Salient kinds of speech errors are those in which a unit moves from an intended slot in an utterance to another slot. Anticipatory, perseverative, and exchange movement errors are illustrated, respectively, in (1)–(3) below using the phonological segment as the sampled moved unit (errors below are from Fromkin, 1973):

(1) *a reading list* → *a leading list*
(2) *leaflets written* → *leaflets litten*
(3) *left hemisphere* → *heft lemisphere*

Another common error type is a substitution of one unit for another of the same size, where the substituting segment does not appear to originate in the planned string. An example of a phonemic substitution is given in (4):

(4) *what Malcolm said* → *what balcolm said*

Phonemes and words are reported to participate frequently in speech errors, whereas syllables and, more controversially, features are reported to participate rarely. Syllables and features do serve a role in speech production planning as revealed by errors, however, even if not as planning units. Generally phonemes involved in movement errors move to the same part of a syllable as that they would have occupied in the intended utterance. Accordingly, the syllable may be seen as a frame into which planned units, syllable-position-sensitive phonemes (e.g., Dell, 1986) are inserted. As for features, increased featural similarity of two segments increases the likelihood that they will interact in an error.

However, the foregoing observation is, in part, why the claim that feature errors are rare is controversial. Compare and contrast:

> Single distinctive features rarely appear as exchange errors, even though many pairs of exchanged segments differ by one feature. (Shattuck-Hufnagel, 1983, p. 112)
>
> The feature paradox . . . First there is the rarity of feature errors, which, like the case with syllables, signifies a very limited role for the feature as a unit (Shattuck-Hufnagel & Klatt, 1979). However, again like the syllable, features

> play an important role in determining which phonemes can slip with which. (Dell, 1986, p. 294)
>
> [Feature] errors are less rare than has been suggested. (Fromkin, 1973, p. 17)
>
> It has been repeatedly observed that most speech production errors are single feature errors. (Browman & Goldstein, 1990b, p. 419)

When a talker produces an error such as *vactive verb* (intending *factive verb*), we cannot be sure that the error is a whole phoneme anticipation; it might be a voicing feature anticipation. (Recall also the two interpretations of [p$^{h}$eyl stIn] above.) The response to this has been that clear cases of feature errors, in which the identity of interacting phonemes is not preserved [as in Fromkin's (1971) much-cited example of *glear plue sky* for *clear blue sky*] occur rarely. In their corpus of 70 exchange errors in which interacting word-initial consonants differed by at least two features (so that feature and phoneme errors could be distinguished), Shattuck-Hufnagel and Klatt (1979) found just three errors in which the identity of the interacting consonants was not preserved.

Despite some uncertainties, speech errors generally support the claim that planning units are elementary units of linguistic competence. That is, the units that either participate actively in errors (move or substitute one for the other) or constrain the form that errors take are consistent with units proposed by linguistic theories.

Can we ask more of the errors, however? In particular, can we ask whether they can help distinguish among competing views of linguistic competence? For example, can they help determine whether phonological primitives are phonemes with featural attributes represented in competence as feature columns or else on featural tiers? Can they distinguish either of these alternatives from Browman and Goldstein's proposal that primitives are gestures and gestural constellations? Finally, can they help determine to what extent lexical representations are abstract or, alternatively, close to surface pronunciations?

The answer to all these questions is that they probably can help to address these issues, but they have not yet been used to do so to any significant extent. Consider the issue of whether primitives are featurally specified phonemes or gestures. The issue is difficult to adjudicate, because descriptions in terms of features and gestures are redundant. Thus, in an error in which interacting segments differ by a single feature (as in *vactive verb* above), the segments generally also differ by a single gesture (voicing in the example) or a parameter of a gesture (e.g., constriction location in *taddle tennis* for *paddle tennis*). Some interacting segments that differ in two features (*Fillmore's face* from the intended *Fillmore's case*) differ in both parameters, location and degree, of a single constriction gesture. Despite the redundancy, it is likely that distinctive predictions can be made about error patterns from the featural/phonemic and gestural perspectives. For example, one prediction distinguishing a gestural account from most others is that errors involving both parameters of a single constriction gesture (location and degree) will be more common, other things being equal, than errors involving one of those parameters and another gesture (constriction location and devoicing, e.g.). Such predictions have yet to be proposed and tested, however.

As for the issue of the abstraction of planned units, two interesting recent findings appear to tug the preponderance of evidence in opposite directions. Findings by Stemberger (1991a, 1991b) suggest that planned units are abstract; findings by Mowrey and MacKay (1990) suggest to them that "errors which have been consigned to the phonemic, segmental, or featural levels could be reinterpreted as errors at the motor output level" (p. 1311).

### B. Stemberger: Radical Underspecification

Generally in speech errors there is a bias for more frequent units to substitute for, or move to slots of, less frequent ones. Stemberger (1991a) pointed out two "antifrequency biases." In one, an error more frequently creates a consonant cluster than a singleton. For example, in attempts to produce sequences such as *puck plump* or *pluck pump,* speakers are more likely to produce *pluck plump* than *puck pump.* This addition bias is antifrequency because clusters are less frequent than are singletons. The second antifrequency bias, the palatal bias, was first noted by Shattuck-Hufnagel and Klatt (1979). It is a tendency for /s/ and /t/ to be replaced in errors by /š/ and /č/, respectively, even though /s/ and /t/ are the more frequent consonants in the language.

Stemberger's account of the addition bias derives from an interactive activation model of speech production. In those models, units are activated in advance of being uttered, and activated units compete. In particular, word-initial consonants compete with word-initial consonants, vowels compete with vowels, and so on. Competition may take the form of mutual inhibition. In the case of *puck plump,* the /l/ in *plump* will be active during planned production of *puck.* The /l/ will be in competition, however, with no segment in *puck.* (There is supposed to be an empty slot in *puck;* henceforth a $C_2$ slot, after the word-initial consonant and before the vowel signifying the phonotactic legality of a cluster in that position.) Because /l/ has no competitor, it will not be inhibited. If it is sufficiently activated during primary activation of *puck,* it may be selected to fill the empty $C_2$ slot in *puck* yielding *pluck.*

Stemberger uses recent linguistic proposals to suggest that the same kind of account will explain the palatal bias and can predict many other asymmetries in speech errors. These recent linguistic proposals concern *underspecification* in lexical entries for words. We encountered underspecification earlier in regard to the aspiration feature of voiceless syllable-initial stops in English words such as *key*. The proposal there was that predictable features of phonemes are not represented lexically. In a more radical approach to underspecification, some nonredundant, nonpredictable feature values are unspecified lexically. This can be done if all other values of the same feature are specified. For example, if voicing is specified lexically, then voicelessness need not be, because the lack of any voicing specification at all will signify voicelessness. Language-internal evidence, with which we do not concern ourselves here, determines which feature values are identified as underspecified.

In the case of consonantal place of articulation, the underspecified feature value is [coronal], characteristic of /s/ and /t/ among other consonants. Stemberger explains the palatal bias in abstractly the same way as he explained the tendency for errors to create consonant clusters. When /š/ or /č/, which have a place specification, compete with /s/ or /t/, which lack one, the unspecified place feature of /s/ and /t/ is vulnerable to substitution by the specified feature because the specified feature has no competitor to inhibit it.

A number of predictions can be made to test this account. The general prediction is that unspecified feature values should be subject to substitution *by* specified values more than they substitute *for* specified values, given equal opportunity. Both in spontaneous speech errors and in experimentally induced errors, Stemberger (1991a, 1991b) obtained error patterns consistent with underspecification theory and his account of asymmetrical substitution patterns. If Stemberger's account of these error patterns is correct, then phonological segments as planned (and also presumably as specified in the lexicon) are represented abstractly and in ways inconsistent with proposals of Browman and Goldstein. It remains to be seen whether an alternative account of the findings will be offered from the perspective of articulatory phonology.

### C. Mowrey and MacKay: Muscular Evidence of Speech Errors

The evidence interpreted as suggesting that errors reveal wholly unabstract planning units is provided by Mowrey and MacKay (1990). These investigators collected productions of such tongue twisters as *Bob flew by Bligh Bay* and *She sells seashells by the seashore.* A novel aspect of their procedure was that they recorded muscle (electromyographic or EMG) activity during the utterances, in particular, that of a muscle of the tongue involved in /l/ production in the first tongue twister and a muscle of the lower lip in the second tongue twister.

Across many productions, they found utterances that sounded correct and were correct as assessed by the EMG data; likewise, they heard such slips as *Bob flew bly Bligh Bay* and saw evidence of the slip in the form of tongue muscle activity during *by*. Remarkably, however, they saw errors that graded evenly between these extremes. There were utterances in which no /l/ could be heard in /b/-V words, but in which some tongue muscle activity was apparent. There were instances in which listeners disagreed about the presence or absence of an intruded /l/ and in which tongue muscle activity was clearly visible. One indisputable conclusion from their findings is that insertions and deletions are not all-or-none. However, the investigators drew stronger conclusions from their findings (Mowrey & MacKay, 1990, p. 1311). They consider the findings to show that errors can be subphonemic and even subfeatural; indeed, they can be errors involving individual muscle actions. One finding they report in support of that conclusion is the occurrence of significant labial activity during a normal sounding [s] in the second tongue twister above. Because substantial lip activity was evident during [š], they identified that activity as the likely trigger for the inappropriate activity accompanying [s]. Because the [s] sounded normal, they infer no intrusion of the alveopalatal feature of [š] on the [s]. They identify the error as subphonemic because the affected [s] is affected in just one respect, not in every respect in which it could be changed by an [š]. They identify it, further, as subfeatural, because [š] is not usually identified as having a rounding or labiality feature. In their view, the intrusion is at the motor output level. They conclude that earlier error collectors may have been misled in their interpretation of speech errors, because they only collected slips that they heard, and because they had no information on the articulatory sources of their perceptions. In their view, there is no compelling evidence yet for phonemes or features as planning units in speech production. Possibly all errors will turn out to be errors at the motor output level.

Mowrey and MacKay's conclusions are not yet compelling, however. First, there must be some organization among the muscle actions that realize a speech utterance. Research on action generally (see, e.g., Bernstein, 1967; Turvey, 1977, 1990) shows that there is no motor output level of the sort Mowrey and MacKay appear to envisage in which independent commands to muscles are issued. To explain the on-line adaptability of the motor system, including the speech system (see Section III), to variation in the context in which actions occur requires invoking the presence of low-level linkages among parts of the motor system responsible for the action. Mowrey and MacKay only recorded from one or two muscle sites at a time, and this would prevent them from seeing evidence of any organization among muscles of the vocal tract. In their example of a subfeatural error described above, they infer that only the labial activity associated with [š] intruded on [s] production, but they base that conclusion on the fact that the [s] sounded normal, a criterion that elsewhere they reject as misleading.

Mowrey and MacKay suggest that: "Units such as features, segments and phonemes may well exist; if it is found at some later time that blocks of motor commands behave as single entities, we should have good evidence of a higher level of organization" (p. 1311). There are at least two reasons for guessing that such evidence will be forthcoming. First is the evidence just alluded to that the motor system is organized in the production of intentional action, including speech. The second is evidence from the errors literature itself. There must be some explanation for the reason why the big speech errors, that is, those that error collectors have heard and recorded, pattern as they do. There are many conceivable big errors (errors in a phoneme and a half, whole syllable errors, movement errors involving single features that do not preserve the identity of the originally planned phonemes) that either do not occur or occur rarely. Others are commonly reported. This suggests that a superordinate organization of motor commands will be found, if one is sought, using Mowrey and MacKay's procedures.

## III. ARTICULATORY UNITS

The kinematics of the articulators during production of an utterance do not transparently reflect the units described by classic linguistic theories, such as that of Chomsky and Halle (1968). Naïve expectation would suggest that any articulators involved in the production of a phoneme should initiate their movements synchronously and end them synchronously. Movements for a next phoneme should then begin together and end together. That is not what we see in speech production. Figure 3 reveals just some of the failures of real speech utterance to satisfy those expectations. The figure displays two tokens of the utterance *perfect memory,* the first produced slowly (and represented only through the first syllable of *memory*) and the second produced more quickly. The latter utterance is transcribed phonetically as if the final /t/ in *perfect* had been omitted by the talker. Under the phonetic transcription and acoustic waveform of each utterance are the traces of pellets placed on various articulators and tracked by X ray. Symbols, ᴋ, ᴛ, and β mark constrictions for the /k/ and /t/ in *perfect* and for the /m/ in *memory*. Regarding the expectation that movements of different articulators for a common phoneme should begin and end together, notice that the velum-lowering gesture for /m/ begins well before raising of the lower lip for the same segment. It appears to reach its lowest point around the time that the labial closing gesture *begins*. As for the discreteness of successive phonemes, notice that, in both productions of the phrase, constriction gestures for /k/ and /t/ overlap, and, in the fast production, the labial gesture

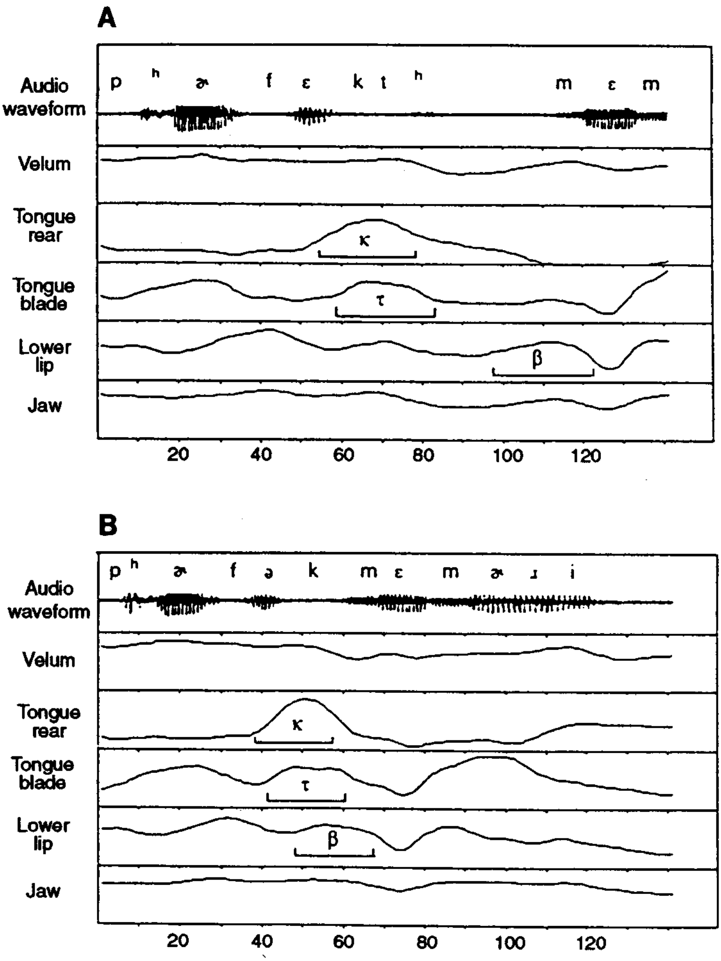

**FIGURE 3** Acoustic signal and X-ray pellet trajectories from two utterances of the phrase *perfect memory*. The first utterance (A), produced more slowly than the second (B), is represented only through the first syllable of *memory*. (From Browman & Goldstein, 1990c.)

for /m/ overlaps with that for /t/ to an extent that, although the /t/ constriction is made, it has no apparent acoustic consequences, and was not heard by the transcriber. Gestures for different phonemes overlap, even different phonemes in different words.

At least some reasons why the articulators do not reveal ostensible linguistic or planning units transparently are known. Even if the articulators were stationary when the activation of articulators for a phoneme began, articulators would not begin to move synchronously, because they have different inertias; further, the muscular support for the different articulatory movements may be differentially effective. Of course, the articulators are unlikely to be stationary in fluent speech, and different requirements to change the direction of movement may further introduce asynchronies between movement onsets. However, even factoring out such peripheral sources of asynchrony, articulatory movements would not partition temporally into discrete phonemes. Rather, movements for different consonants and vowels in an utterance overlap in time. This coarticulation means that there can be no boundaries between articulatory movements for consonants and vowels in sequence, that is, no single point in time when movements for one segment cease and those for another begin.

The apparent absence of linguistic units in articulation, of course, does not mean either that linguistic units are absent or that units of any sort are absent. Perhaps the level of description of movement in the vocal tract has been inappropriate for finding units; perhaps we should not attempt to impose on the vocal tract our preconceived ideas of what the units should be (e.g., Kelso, Saltzman, & Tuller, 1986; Moll, Zimmermann, & Smith, 1976). Rather, we first should attempt to discover the natural order in vocal-tract actions, if any, and then worry later about their relation to linguistic units.

A more distanced, less detailed, perspective on the vocal tract does suggest more organization than the one adopted above in which individual movements of individual articulators were tracked. One can describe much of what goes on in the vocal tract during speech as the overlapped, but still serially ordered, achievement and release of constrictions. Frequently, several articulators cooperate in a constriction action, and, at least in this sense, there appears to be a superordinate organization of the vocal tract into systems. If the achievement and release of a constriction can be construed as a unit of action, then perhaps there may be said to be units in articulatory behavior.

Perhaps the earliest kind of evidence suggestive of articulatory systems was provided in the literature on so-called bite-block speech. In this literature (e.g., Lindblom, Lubker, & Gay, 1979; Lindblom & Sundberg, 1971), speakers produce speech, frequently vowels, with bite blocks clenched between their upper and lower teeth so that the jaw cannot move and generally is forced to adopt either a more open position than is typical for a high vowel or a more closed position than is typical for a low vowel. The striking finding is that, with little or no practice, vowels are acoustically normal (or near normal; Fowler & Turvey, 1980) from the first pitch pulse of the vowel. This implies an equifinality in vowel production such that a given configurational target can be reached in a variety of ways. The limited need for practice suggests, too, that this flexibility is somehow already in place in the speaker; it is not learned during the course of the experimental session.

More information about sources of equifinality is obtained using a procedure pioneered by Abbs and his colleagues (e.g., Abbs & Gracco, 1984;

Folkins & Abbs, 1975). In this procedure, speakers produce target utterances repeatedly; on a low proportion of trials, and unexpected by subjects, an articulator is perturbed during production of a consonant. For example, in research by Kelso, Tuller, Vatikiotis-Bateson, and Fowler (1984), the jaw was unexpectedly tugged down during production of the final consonants of target words /bæb/ and /bæz/, produced in a carrier sentence. Within 20–30 ms of the onset of the perturbation, other articulators (the upper lip in /bæb/ and the tongue in /bæz/ began to compensate for the unusually low jaw position, such that consonantal constrictions were achieved, and consonants sounded normal. Findings are generally that the responses to perturbation are, for the most part, functionally specific. That is, articulators that would not compensate for the perturbation are not activated (but see Kelso et al., 1984, for a possible qualification); and, when articulators are perturbed that are not involved in the consonant being produced, the consonant is produced as it is on unperturbed trials (Shaiman, 1989). The short latency onset of the compensatory response implies an existing linkage among articulators, the jaw and lips during /b/ and the jaw and tongue during /z/ in the research by Kelso et al. (1984). Because the linkages are different for different consonants, they must be transiently established during speech production. Researchers in the motor skills literature refer to these linkages as "synergies" or "coordinative structures" (Easton, 1972).

The work on synergies in speech production is limited; accordingly, no catalog of them can be compiled. However, evidence for a synergy to achieve bilabial closure is well established (Abbs & Gracco, 1984; Kelso et al., 1984); some evidence supports synergies for achievement of an alveolar constriction (Kelso et al., 1984) and a labiodental constriction (for /f/; Shaiman, 1989). The partial list invites an inference that synergies exist that achieve consonantal places and manners of articulation. If the research on bite-block speech is interpreted as also revealing synergies of speech production, then the generalization can be extended to establishment of the more open constrictions characteristic of vowels.

This section began with the information that individual articulator movements do not segment temporally into familiar linguistic units. Having moved to a different level of description of vocal-tract actions, we do see unitlike systems, namely synergies, but still, these are not the discrete phonemes of classic linguistic theories. Synergies do appear to map almost transparently onto the phonetic gestures of Browman and Goldstein's articulatory phonology, however, a point we address in Section V.

## IV. COMPATIBILITY OF UNITS ACROSS DESCRIPTIVE LEVELS

To what extent are units compatible across the three domains of phonological competence, planning, and articulation? Theorists and researchers have recognized two major barriers to a view that linguistic units can be commensurate with units if there are any behavioral units at all. One barrier is a view that phonological segments as known are different in kind from units as uttered because one is cognitive in nature and the other is physical:

> [P]honological representation is concerned with speakers' implicit knowledge, that is, with information in the mind . . . The hallmarks of phonetic representation follow from the fact that sounds, as well as articulatory gestures and events in peripheral auditory-system processing are observables in the physical world. Representation at this level is not cognitive, because it concerns events in the world rather than events in the mind. (Pierrehumbert, 1990, pp. 376–377)
>
> [Segments] are abstractions. They are the end result of complex perceptual and cognitive processes in the listener's brain. (Repp, 1981, p. 1462)

That problem aside, there is another barrier that, in the phonologies best known to speech researchers, phonological segments have characteristics (such as being static) that are impossible for vocal tracts to implement transparently. This has led to a view already mentioned that research should not seek, in articulatory behavior, units of linguistic description supplied by phonological theories. Rather, it should be designed, in an unbiased way, to discover natural organizational structure in vocal-tract activity (Kelso et al., 1986; Moll, et al., 1976).

For the most part, in the field of psychology, the apparent incommensurability of linguistic units and the articulatory implementation of a speech plan has been accepted, and theorists generally pick a domain in which to work: language planning and sequencing or articulation, without addressing the problem of the interface. There is some motivation for taking another look, however. Certainly, communicative efficacy would be on a more secure basis were the units in vocal-tract actions transparent implementations of units as known and planned. This is because it is only units as produced that immediately structure the acoustic signal for a perceiver.

To motivated theorists, moreover, the barriers above are not insurmountable. Regarding the first, one can challenge the ideas (following Ryle, 1949) both that cognitive things can be only in the mind and that physical things and cognitive things are mutually exclusive sets. Under an alternative conceptualization, vocal-tract action can be linguistic (and therefore cognitive) action. Regarding the second barrier, we have seen that some phonologies do offer units that are compatible with vocal-tract capabilities. It may not be necessary or even desirable for production researchers to design their research wholly unbiased by expectations derived from these new phonologies.

There is one, incomplete, model of speech production that implements phonological units of the language as vocal-tract activity. The model is incomplete in not attempting to handle all the data that a complete model will have to handle, for example, the occurrence of spontaneous speech errors. However, it is unique both in its explicit choice of a phonological theory (articulatory phonology) that eliminates incompatibilities across levels of description and in its handling the evidence described earlier that reveals the role of synergies in speech production. That model is Saltzman's "task dynamic model" (Kelso et al., 1986; Saltzman, 1986; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989).

## V. TASK DYNAMICS

The task dynamic model represents a new approach to understanding intentional, goal-directed action (see Turvey, 1990, for a review of this approach). Its novelty is in its recognition that living systems are complex physical systems that behave in some respects like other, even inanimate, complex physical systems. Those aspects may be best explained by invoking relevant physical principles and laws.

Many actions produced by synergies have characteristics in common with various kinds of oscillatory systems, and the approach that Saltzman and others take (Kelso et al., 1986; Kugler, Kelso, & Turvey, 1980; Saltzman, 1986; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989; Turvey, 1990) is to model flexible, goal-directed action as produced by one or more oscillatory subsystems. Saltzman named his model task dynamics to highlight its two salient features. In the model, actions are defined initially in functional terms, that is, in terms of the tasks they are to achieve. That is how the actions acquire and maintain their goal-directed character. In addition, tasks are defined in terms of the dynamics that underlie the actions' surface forms or kinematics.

Actions as diverse as a discrete reach to a location and bilabial closure are described in the same way in their first description, in *task space.* That is because the dynamical control regime that will implement the action is first described functionally, and both discrete reaching and bilabial closure are characterized functionally by *point attractor* dynamics. A point attractor system has point stability, that is, trajectories of the system are attracted to a point at which the system is in equilibrium. An example is a pendulum. Set in motion by a push, it comes to rest hanging parallel to the pull of gravity. In a bilabial closing gesture, a lip aperture system is attracted to a point at which the lips meet.

In the model, equations of motion that represent task achievement in task space undergo two transformations as the model "speaks," first into "body space" (with the jaw as the spatial point of origin) and next into "articulator space," where each articulator is assigned one or more directions of possible movement. In the latter transformation, a small number of dimensions in body space are rewritten as more articulatory dimensions. The redundancy of this transformation permits flexible achievement of task goals. Accordingly, the model, like subjects in the perturbation experiments of Kelso et al. (1984) and others, compensates for perturbations to an articulator. In particular, fixing the model's jaw at an unusually low position during production of bilabial closure leads to compensation, on-line, by the model's upper and lower lips.

Task dynamics has four notable features for our purposes. First, it uses the gestural scores provided by Browman and Goldstein's articulatory phonology as scripts that provide the tract variables and their dynamic parameters to be achieved in production of a word. Second, it *achieves* the task goals specified in gestural scores as vocal-tract activity. Third, the synergies in task dynamics show the equifinality characteristic exhibited by the speech system under perturbation. Fourth, synergies achieve equifinality, because they are treated as complex physical systems, sharing their dynamical characteristics with other physical systems, an approach that many investigators of action generally consider realistic. To my knowledge, this model in which the primitives of a phonological theory are realized nondestructively and in a natural way as vocal-tract action, is unique in the field of speech production.

## VI. SEQUENCING

In Section V, attention was focused on the nature of language units as known, planned, and produced. The present topic is the sequencing of these units in speech production.

### A. Coarticulation

There is no theory-neutral definition of the term *coarticulation,* and the literature reflects considerable disagreement to its referent. Accordingly, rather than attempt to offer a generic definition, I will offer three in the context of a discussion of findings that each characterization handles well or badly. The characterizations differ in the level of speech production planning or execution at which they propose that coarticulation arises and on the issue of whether coarticulation assimilates a segment to its context or, instead, is overlap of two or more essentially context-free segments.

#### 1. Coarticulation as Feature Spreading

Daniloff and Hammarberg (1973) proposed that the coarticulatory rounding of /š/ in English *shoe* (phonologically /šu/) might be viewed as the consequence of a rule that spreads /u/'s rounding feature. Such spreading of features would serve to assimilate a segment to its context and thereby to smooth articulatory transitions between segments. In the particular feature-spreading theory, known as *look-ahead* theory, ascribed to Henke (1966), in fact a feature such as rounding can spread farther than just to an immediately preceding segment. Generally, a feature will spread anticipatorily to any preceding segment (anticipatory or right-to-left coarticulation) that is unspecified for that feature. Segments are unspecified for a feature if changing the feature value does not change the segment's identity. (Recall the featural representation of *bag* in Section I.) In English, consonants are unspecified for rounding, because rounding them does not change their identity. If features spread in an anticipatory direction to any number of preceding segments that are unspecified for them, then rounding should anticipate a rounded vowel through any number of consonants that precede it up to the first occurrence of an unrounded vowel. [Carryover (left-to-right, perseveratory) coarticulation here is seen as partly due to inertia of the articulators and therefore of somewhat less interest than anticipatory coarticulation.] Supportive evidence for the look-ahead model's account of spreading of rounding was provided by Daniloff and Moll, (1968) for English and by Benguerel and Cowan (1974) for French.

Nasalization in English is complementary to rounding in being specified among consonants but not vowels. According to a look-ahead version of a feature-spreading model, the nasal feature of a nasalized consonant should spread in an anticipatory direction through any number of preceding vowels up to the first oral (i.e., nonnasal) consonant. Supportive evidence on nasality was reported by Moll and Daniloff (1971).

Because, in this theory, coarticulation is spreading of a feature, it makes the strong prediction that coarticulation will be categorial, not gradient both in space and in time. To a first approximation (but see Kent, Carney, & Severeid, 1974, for a qualification), a rounded /š/ should be as rounded as an /u/, and the rounding should always begin at the beginning of a segment to which it has spread, never in the middle. Feature-spreading theories are now generally agreed to have been disconfirmed in part because close examination of the data shows that coarticulation is clearly gradient in character and in part because experiments with improved designs have shown that spreading is less extensive than earlier research had reported. Keating's *window model* of coarticulation (Keating, 1990) and the theory of coarticulation as coproduction both address each source of disconfirmation.

## 2. The Window Model of Coarticulation

At least some of coarticulation is gradient in nature and does not respect segment boundaries. For example, Benguerrel and Cowan's (1974) frequently cited findings of anticipation of rounding from the first, rounded, vowel in *structure* in the French noun phrase *ministre structure* in fact showed gradience in time. They found anticipatory rounding during the word-initial [str] of *structure* and the word-final [str] of *ministre* as expected because the consonants are unspecified for rounding. However, they also report rounding part way through the second, unrounded, vowel of *ministre*. In a feature-spreading account, this segment should not have been rounded at all, but of central relevance here, there is no provision in the theory for a feature to be spread to part of a segment. Keating (1990) cites other examples from Arabic (Card, 1979; Ghazeli, 1977) in which effects of tongue backing (emphasis) are gradient not only in time, but also in spatial extent.

Keating's proposal retains the idea that coarticulation is assimilation of a segment to its context, and she concurs with Daniloff and Hammarberg (1973) and others that some of coarticulation is categorial and characterizable as feature spreading. However, she suggests that, in cases where coarticulation is gradient, it can be seen as a later, lower-level process rather than as one of feature spreading. Her proposal is to replace the dichotomous concept that segments can be either specified or unspecified for a feature with a graded concept that each segment is associated with target windows of some width on each articulatory dimension involved in realizing its feature values. A maximally wide window corresponds with the earlier unspecified and a minimum window width with specified. However, all window widths between the extremes are possible, as well. Coarticulation is the consequence of the speaker choosing the smoothest or most economical path through a sequence of discrete windows. The theory improves on feature-spreading theory in offering an account of gradience in coarticulatory influences.

There are findings, however, indicating that at least some coarticulatory influences cannot be captured in a theory in which segments (or their windows) do not overlap. Öhman (1966) noticed acoustic influences of $V_2$ on closing transitions from $V_1$ to C in $V_1CV_2$ utterances. X-ray tracings in Öhman (1967) confirmed that the tongue body shape during C was different in the context of different vowels. He proposed that smooth vowel-to-vowel gestures of the tongue body occurred during speech production with consonant articulations superimposed on the diphthongal vocalic gestures (see also Barry & Kuenzel, 1975; Butcher & Weiher, 1976; Carney & Moll, 1971). Vowel-to-vowel gestures can even occur during production of consonants that themselves make demands on the vowels' primary articulator, the tongue body. Perkell (1969) found that the /k/ constriction gesture during production of /həkɛ/ consisted of a sliding tongue movement along the palate from the central location for schwa toward the more front location for /ɛ/. Although the vowel-to-vowel gestures might be seen as windows for each vowel with transitional regions between them, there appears to be no alternative to the conclusion that there is overlap between these windows and those for intervening consonants.

A third view of coarticulation holds that all of coarticulation is gradient in nature, and all of it is overlap in the production of sequences of consonants and vowels.

## 3. Coarticulation as Coproduction

In one version of this theory (Fowler, 1977; Fowler & Saltzman, 1993), coarticulation is the overlapping implementation of (to a first approximation) context-invariant synergies that realize gestures for consonants and vowels. The context sensitivity apparent in articulation (as when, e.g., lip closure for /b/ is achieved with a lower jaw posture in /ba/ than in /bi/) is seen as a peripheral blending of overlapping movements, not a revision in the plan for achieving a consonant or vowel.

Bell-Berti and Harris's (1981) *frame theory* adds to this characterization a claim that the temporally staggered onsets of the gestures for a consonant or vowel are sequenced in a temporally invariant fashion. This version of the theory appears to be in conflict with data cited earlier in favor of feature-spreading theories, findings that lip rounding and nasalization have extensive anticipatory fields that are linked to the onset of the first segment in a string that is unspecified for the feature. However, research by Bell-Berti and collaborators has shown that a missing control in those earlier investigations led to a considerable overestimation of the extent of coarticulation of lip rounding and nasalization.

Utterances in studies of anticipation of lip rounding have generally been of the form $VC_nu$, where V is an unrounded vowel and $C_n$ is a consonant string of variable length. The missing control is an utterance type, such as $VC_ni$ in which the ostensible source of any rounding during the consonant string has been eliminated. In the consonant strings of such control utterances, Bell-Berti and colleagues have found lip rounding movement or muscle activity that can be ascribed only to the consonants themselves. Using the control utterances to eliminate such spurious rounding from test utterances reveals an invariant, short-duration anticipation of rounding due to a rounded vowel, consistent with expectations from frame theory (e.g., Boyce, 1990; Gelfer, Bell-Berti, & Harris, 1989; see also Perkell & Matthies, 1992). An interpretation of the earlier findings of more extensive anticipation of rounding is that the earlier evidence was contaminated by lip movements associated with the consonants preceding the rounded vowel.

Research on anticipation of nasalization has the same history. When control utterances of the form $V_nC$ are used to eliminate vowel-related velum lowering the $V_nN$ utterances (where N is a nasal consonant), anticipation of velum lowering for N is found to be short in duration and to precede onset of oral constriction for the nasal by an invariant interval (e.g., Bell-Berti, 1980; Bell-Berti & Krakow, 1991).

Although frame theory appears to provide an account that is compatible with the available data, it is likely that some aspects of the theory will undergo modification. Bell-Berti and colleagues reported that lip rounding anticipates the acoustically defined onset of a rounded vowel by an invariant interval. However, it is unlikely that talkers time lip rounding relative to an acoustically defined landmark. It is more likely that timing or phasing is relative to the gesture achieving the oral configuration for the vowel. Second, as Bell-Berti and Harris (1981) note, it is not likely, in fact, that any real anticipation will be invariant over rate variation. Finally, Krakow's (1989) findings on anticipation of velum lowering for a nasal consonant suggest different phasing rules for pre and post vocalic consonants.

## 4. Lingual Coarticulation and Coarticulation Resistance: Another Role for Synergies?

In 1976, Bladon and Al-Bamerni introduced the term *coarticulation resistance* to describe the observation that different segments appear to resist coarticulatory influences to greater and lesser degrees. Recasens (1984a, 1984b, 1985, 1987, 1989, 1991) has done much of the work to develop the concept of coarticulation resistance. In a sequence of Catalan consonants that vary in amount of tongue dorsum contact with the palate, he found that consonants with more palatal contact show less coarticulatory influence from neighboring vowels than do consonants with less contact (Recasens, 1984a, 1987). Since vowels also use the tongue dorsum and would tend to lower the tongue away from the palate, the resistance may be seen as an act of "self-preservation" on the part of these consonants. It is interesting that consonants and vowels that strongly resist coarticulatory influences in their own domains in turn exert relatively strong influences on neighbors (see Tables II–VI in Recasens, 1987; see also Butcher & Weiher, 1976; Farnetani, 1990; Farnetani, Vagges, & Magno-Caldognetto, 1985).

A consequence of the different degrees of coarticulation resistance among the gestures of segments in their own domains and a consequence of their correspondingly different degrees of aggression outside of their domains are that production of a given consonant or vowel can appear to differ in different contexts, in particular, its coarticulatory extent will vary. A question is whether this apparent context sensitivity is "deep," that is, whether it reflects changes in a talker's articulatory plan or whether it arises in peripherally established influences on a plan's implementation.

In both Keating's windows theory and the theory of coproduction, the variation arises below the level of a speech plan. For Keating, it arises in the computation of transitions between windows for neighboring segments. A narrow window will exert a stronger influence on the transitional movement than will a wider window. In the theory of coarticulation as coproduction, there are no transitions between segments; there is only overlap. The context sensitivity is hypothesized to arise in the peripheral blending of influences on common articulators of the gestures for different consonants and vowels (Fowler & Saltzman, 1993; also see Saltzman & Munhall, 1989). In the theory, synergies are responsible for flexible achievement of invariant goals of a phonetic gesture as described earlier. They achieve these goals by establishing physiological linkages among articulators that instantiate the appropriate attractor dynamics. But the linkages can be looked at in another way from the perspective of other gestures whose influences on the vocal tract overlap with the target gestures in time. From that perspective, the linkages constitute barriers of variable strengths or resistances to the influences of those overlapping gestures. For a segment such as Catalan /j/ that requires considerable tongue dorsum contact with the palate, the linkages between jaw and tongue that bring about the contact also serve to resist influences from other gestures that would reduce the required contact. In short, in this theory, the constraints that implement synergies are at once the means by which gestural goals are achieved and the sources of resistance to coarticulatory influences that would prevent or hamper goal achievement. The same linkages that make a gesture more or less resistant to coarticulation from neighbors are sources of stronger or weaker coarticulatory influence outside their domain.

## B. Models of Sequencing in Speech Production

In a recent discussion of the task dynamic model of speech production, Saltzman and Munhall (1989) point out that, currently, the model offers an intrinsic dynamical account of the implementation of gestures in the vocal tract, but not of sequencing or phasing of gestures themselves. Rather, to generate gestural sequencing, the model uses a gestural score (see Figure 2) in which appropriate sequencing is represented explicitly. Saltzman and Munhall suggest that an approach more compatible with their general theoretical framework would incorporate an adaptation of Jordan's (1986) network model of sequence control into the task dynamic model. Jordan (1986) makes a similar suggestion.

Jordan developed his model to address the problem of serial order in action and, in particular, in speech production. This is the problem, apart from concerns about details of timing or phasing of components of a complex action, of executing the components in the required order. Jordan proposed a model, the structure of which is shown in Figure 4. In the model, a sequence consists of a succession of patterns of activation over the output units of the model. The model learns to produce sequences, by associating each with a unique pattern of activation over the plan units.

Activation at the plan and state level, and again at the level of hidden

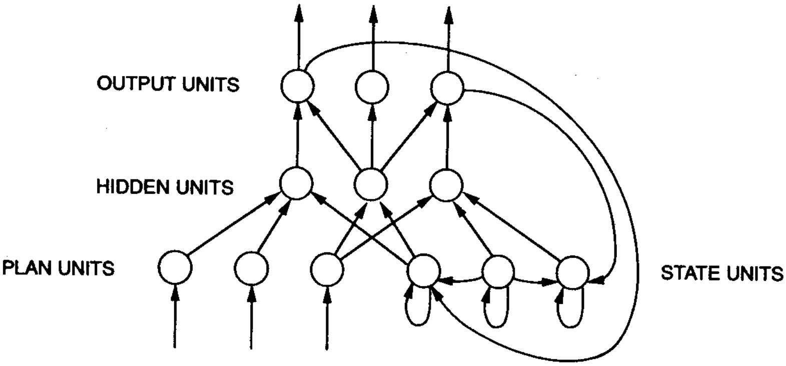

**FIGURE 4** A simple example of Jordan's (1986) network model of sequence production. (Adapted from Jordan, 1986, Figure 3.)

units, propagates along links to the next level. Activation at a level is multiplied by the weights on linkages to the next level; products converging on a node are summed and added to a bias value associated with the node. In Jordan's implementation, hidden units and output units follow a rule that, if the activation quantity is positive, a 1 is output; otherwise the output is 0. The pattern of 1s and 0s over the output at a given time represents a set of feature values for one element of the sequence.

A crucial feature of the model that allows sequences with arbitrary numbers of repeated phonemes to be produced is that output units feed back to state units, which themselves are recurrent. This means that outputs, which are functions of activation from both plan and state units, are influenced by their temporal context. (Previous outputs are exponentially weighted so that more recent outputs are represented most strongly.) Because the state reflects the history of the sequence, the first and second /l/s in *lily,* for example, are distinguished by their contexts, and the sequence can be produced without error. Aside from offering a viable solution to the problem of serial order, Jordan's network has two other interesting properties. First, learned sequences (learned trajectories through state space) can serve as limit-cycle attractors, so that compensation for errors is possible. A second interesting feature of the model is that, over learning, generalization occurs so that outputs and states that are temporally close are similar. Therefore, the model's outputs exhibit coarticulatory overlap.

In Saltzman and Munhall's projected incorporation of Jordan's model into task dynamics (described in Saltzman & Munhall, 1989), output units would produce activations for gestures, not the feature values output in Jordan's implementation. Browman and Goldstein's gestural scores would, then, be replaced by plan units that, when inserted into the network, would generate appropriately sequenced gestural activations.

Saltzman and Munhall's proposal may have an unanticipated benefit. As noted, generally, the literature bifurcates into studies of speech production (i.e., studies of articulation) and studies of language production that address plans for producing linguistic units of various grain sizes. Recently, in the literature on language production, Dell, Juliano, and Govindjee (1992) have proposed a model at the level of phoneme production in which a Jordan-like network generates the sequencing of phonemes. This convergence on a single model type from these two directions may offer some hope for the eventual development of a unified model of production.

To put the proposal of Dell et al. in context, I first briefly characterize earlier accounts of phoneme production, based, as the proposal of Dell et al. is, on spontaneous errors of speech production. Some typical single-segment speech errors were listed above in Section II. A striking characteristic of these big errors (i.e., errors audible to an error collector) is their systematicness. In substitution errors, sequences rarely violate the phonotactic constraints of a language; consonants substitute only for consonants and vowels only for vowels; VCs, a constituent of a syllable called the *rime,* are more likely to participate in a substitution than are (nonconstituent) CVs; initial consonants of a word are more likely to participate in a substitution than are postvocalic consonants. Dell et al. called these systematicities *phonological frame constraints,* because they have been interpreted as requiring an explanation in which, during production planning, abstract structural frames for an utterance are distinguished from the phonological contents that fill them.

To explain phonological frame constraints, researchers (e.g., Dell, 1986; Shattuck-Hufnagel, 1983) proposed *slot and filler* or *structure versus content* models in which building a speech plan involves creating abstract structural frames for components of an utterance and then inserting words, morphemes, and phonological segments into the frames. In Dell's (1986) model, sequencing is the consequence of gradient activation in a tiered lexical network and of successive insertion of the most highly activated lexical units of a specified type into appropriate slots in structural frames built for the planned utterance. Errors occur when the most highly activated item in the lexical network that fits the frame's requirements is not the intended item. This can happen because of the spread of activation in the network and because of the presence of some activation noise. Because of the structural frames, however, errors will be systematic. At the phonological level, in particular, most phonological frame constraints (all except the finding that initial consonants are particularly error prone) will be respected. Errors will create phonotactically acceptable sequences because generally if, for example, a vowel is required by the frame, any vowel that is selected will yield an acceptable sequence. Further, because the frame distinguishes initial (onset) consonants, vowels, and final (coda) consonants in a syllable, onset consonants will substitute only for other onset consonants, vowels for vowels, and coda consonants for coda consonants. VCs participate jointly in errors, whereas CVs do not, because VCs (syllable rimes) are units in the lexicon that spread activation to (and receive reinforcing activation from) their component phonological segments.

The model provides a successful account of sequencing in speech production in which most natural error types occur in natural relative proportions. Additionally, the model has been a source of counterintuitive or unexpected error patterns that experimental tests with human subjects have confirmed. It has some unappealing features, however, one of which is its invocation of abstract structural frames. When a lexical item, say, *bag* has been selected at its level, a syllabic frame is created at the next level down that knows that the phonological sequence will consist of a C followed by a V and then a C. But why, if that much is known, is it not already known that the sequence is /b/, /æ/, /g/? And if that is known, why must activation spread in order for selection to occur? The answer is that, as it is, the model generates errors in the proper patterns; without the frame-content distinction it would not.

Dell et al. (1992) suggest an alternative model that almost dispenses with the structural frames; they propose that the frame constraints, the very findings that led Dell (1986) and others to distinguish frames from contents in their models, can arise in the absence of explicit frames. The new model has, in general, the structure of Jordan's model described above. As in Jordan's simulation, words were assigned plans consisting of a unique pattern of 1 and 0 activations on the plan units of the model. There were 18 output units, one for each of Chomsky and Halle's (1968) phonological features. Dell et al. trained the network to generate the appropriate succession of feature values for words in several lexicons consisting of CVCs. Errors occurred in some simulations because sequence learning was incomplete and in other simulations because noise was added to linkage weights. In either case, frame constraints were evident in errors in approximately the same proportions as those in natural error corpora.

As Dell et al. explain, one reason why errors adhere to frame constraints in the absence of explicit frames is that the network is disposed to adhere to well-worn paths (i.e., well-learned sequences serve as attractors; cf. Jordan, 1986). Dell et al. refer to this as the "sequential bias." A second reason is that errors on the output units will be small and therefore will differ by just one or two features from the intended segment (the "similarity bias"). Both biases foster errors in which the output sequence is phonotactically legal. That is, the sequential bias will foster use of a learned, even if not the intended, sequence; the similarity bias will lead to selection of a segment similar to the intended one, and similar segments tend to be phonotactically permissible in similar contexts. The similarity bias will tend to ensure that consonants substitute for consonants and vowels for vowels. The sequential bias fosters a finding that VCs participate in errors more than CVs, in that, in English, VCs are more redundant that are CVs. (That is, given V, the identity of the next C is more constrained than that of the prior C.) Likewise, initial Cs are preceded in Dell et al.'s sequential coding only by a null element signifying a word edge. Consequently, they are less determinate than Cs following vowels and are relatively error prone.

This new model of Dell et al. is an exciting development in part because it eliminates the inelegance of structural frames that are abstracted from, yet tailored to, their contents. It is also exciting in the present context, because, in conjunction with Jordan's sequencing model that coarticulates, and with the possible interfacing of such a model with Saltzman's task dynamic model that generates synergistic actions to a model vocal tract, there is hope for a unified model of production that exhibits many of the central characteristics of natural speech.

The model currently has a major limitation as Dell et al. point out. It generates substitution errors [such as (4) above in Section II] but not the more common movement errors [(1)–(3) in Section II] in which segments in a planned sequence themselves interact. The authors speculate that some characteristics of movement errors may be attained if more than one plan unit (one for a planned first word at, say, 80% activation, and one for a planned next word at 20% activation) is input to the network at the same time. Even so, this would not naturally generate the spectacular exchange errors (e.g., *heft lemisphere*). Consequently, Dell et al. are not able to conclude that frames can be dispensed with. Their more moderate conclusion is that:

> The . . . model can be taken as a demonstration that sequential biases and similarity are powerful principles in explaining error patterns. In fact, we . . . argue that the general frame constraints may simply reflect similarity and sequential effects and not phonological frames, at least not directly. Moreover, we interpret the model's success in its domain as evidence that there is something right about it, [and] that phonological speech errors result from the simultaneous influence of all the words stored in the system, in addition to the rest of words in the intended utterance. (1992, p. 33)

## VII. OMISSIONS FROM THE CHAPTER: NEAR OMISSIONS IN THE FIELD

Just as theories of language production have been devised apart from theories of speech production, studies of prosody have been conducted almost independently of research on segmental production of speech (but see Beckman & Edwards, 1992, for a recent exception). Accordingly, theories of speech production have not, typically, explained the macroscopic structure in speaking that metrical or intonational patterning provides (but see Levelt, 1989, for an ambitious attempt). Perhaps a fully integrated account of speech production must await further descriptive studies of the prosodic patternings themselves that are increasingly common.

Eventually, theories of speech production must also provide an account of ordinary variability in speech. Variability in the pronunciation of a word occurs over different speaking styles from formal to casual and more generally over different communicative settings. Words may be produced with ordinary clarity, they may be "hyperarticulated" (e.g., Lindblom, 1990), as when we speak to a deaf person who must lip read, or they may be considerably reduced ("hypoarticulated") when they are redundant for one reason or another (e.g., Browman & Goldstein, 1986a; Fowler & Housum, 1987). No model of production as yet generates this gradience in production that is so natural to, and characteristic of, human speech outside the laboratory.

## Acknowledgments

Preparation of this manuscript was supported by NICHD Grant HD-01994 to Haskins Laboratories. I thank Catherine Browman and Gary Dell for comments on parts of the manuscript.

## References

Abbs, J., & Gracco, V. (1984). Control of complex gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology, 51,* 705–723.

Barry, W., & Kuenzel, H. (1975). Co-articulatory airflow characteristics of intervocalic voiceless plosives. *Journal of Phonetics, 3,* 263–282.

Beckman, M., & Edwards, J. (1992). Intonational categories and the articulatory control of duration. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 359–376). Tokyo: IOS Press.

Bell-Berti, F. (1980). Velopharyngeal function: A spatio-temporal model. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice* (pp. 291–316). New York: Academic Press.

Bell-Berti, F., & Harris, K. (1981). A temporal model of speech production. *Phonetica, 38,* 9–20.

Bell-Berti, F., & Krakow, R. (1991). Anticipatory velar lowering: A coproduction account. *Journal of the Acoustical Society of America, 90,* 112–123.

Bendor-Samuel, J.-T. (1960). Some problems of segmentation in the phonological analysis of Terena. *Word, 16,* 348–355.

Benguerel, A., & Cowan, H. (1974). Coarticulation of upper lip protrusion in French. *Phonetica, 30,* 41–55.

Bernstein, N. (1967). *The coordination and regulation of movement.* London: Pergamon Press.

Bladon, A., & Al-Bamerni, A. (1976). Coarticulation resistance in English /l/. *Journal of Phonetics, 4,* 137–150.

Boyce, S. (1990). Coarticulatory organization for lip rounding in Turkish and English. *Journal of the Acoustical Society of America, 88,* 2584–2595.

Browman, C., & Goldstein, L. (1986a). Dynamic processes in linguistics: Casual speech and sound change. *Perceiving Acting Workshop Review (Working Papers of the Center for the Ecological Study of Perception and Action), 1,* 17–18.

Browman, C., & Goldstein, L. (1986b). Towards an articulatory phonology. *Phonology Yearbook, 3,* 219–252.

Browman, C., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6,* 201–252.

Browman, C., & Goldstein, L. (1990a). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18,* 299–320.

Browman, C., & Goldstein, L. (1990b). Representation and reality: Physical systems and phonological structure. *Journal of Phonetics, 18,* 411–425.

Browman, C., & Goldstein, L. (1990c). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology: I. Between the grammar and the physics of speech* (pp. 341–376). Cambridge, UK: Cambridge University Press.

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49,* 222–234.

Butcher, A., & Weiher, E. (1976). An electropalatographic investigation of coarticulation in VCV sequences. *Journal of Phonetics, 4,* 59–74.

Card, E. (1979). *A phonetic and phonological study of Arabic emphasis.* Doctoral dissertation, Cornell University, Ithaca, NY.

Carney, P., & Moll, K. (1971). A cinefluorographic investigation of fricative consonant-vowel coarticulation. *Phonetica, 23,* 193–202.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook, 2,* 225–252.

Daniloff, R., & Hammarberg, R. (1973). On defining coarticulation. *Journal of Phonetics, 1,* 239–248.

Daniloff, R., & Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing Research, 11,* 707–721.

Dell, G. (1986). A spreading-activation theory of retrieval in speech production. *Psychological Review, 93,* 283–321.

Dell, G., Juliano, C., & Govindjee, A. (1992). *Structure and content in language production: A theory of frame constraints in phonological speech errors* (Cognitive Science Technical Report No. UIUC-BI-CS-91-16). Urbana-Champaign: University of Illinois.

Easton, T. (1972). On the normal use of reflexes. *American Scientist, 60,* 591–599.

Elimelech, B. (1976). A tonal grammar of Etsakọ. *UCLA Working Papers in Phonetics, 35.*

Farnetani, E. (1990). V-C-V lingual coarticulation and its spatiotemporal domain. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 93–130). Dordrecht: Kluwer.

Farnetani, E., Vagges, K., & Magno-Caldognetto, E. (1985). Coarticulation in Italian /VtV/ sequences: A palatographic study. *Phonetica, 42,* 78–99.

Folkins, J., & Abbs, J. (1975). Lip and jaw motor control during speech. *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., & Housum, J. (1987). Talkers' signalling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26,* 489–504.

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36,* 171–195.

Fowler, C. A., & Turvey, M. T. (1980). Immediate compensation for bite block speech. *Phonetica, 37,* 307–326.

Fromkin, V. (1971). The nonamalous nature of anomalous utterances. *Language, 47,* 27–52.

Fromkin, V. (1973). *Speech errors as linguistic evidence.* The Hague: Mouton.

Gelfer, C., Bell-Berti, F., & Harris, K. (1989). Determining the extent of coarticulation: Effects of experimental design. *Journal of the Acoustical Society of America, 86,* 2443–2445.

Ghazeli, S. (1977). *Back consonants and backing coarticulation in Arabic.* Doctoral dissertation, University of Texas, Austin.

Goldsmith, J. (1976). *Autosegmental phonology.* Bloomington: Indiana University Linguistics Club.

Goldsmith, J. (1990). *Autosegmental and metrical phonology.* Oxford: Basil Blackwell.

Henke, W. (1966). *Dynamic articulatory models of speech production using computer simulation.* Doctoral dissertation, MIT, Cambridge, MA.

Jordan, M. (1986). *Serial order: A parallel distributed process.* (ICS Report 8604). San Diego: University of California, Institute for Cognitive Science.

Keating, P. (1990). The window model of coarticulation: Articulatory evidence. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology: Between the grammar and the physics of speech* (pp. 451–470). Cambridge, UK: Cambridge University Press.

Kelso, J. A. S., Saltzman, E., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics, 14,* 29–59.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally-specific cooperation following jaw perturbation during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 812–832.

Kenstowicz, M., & Kisseberth, C. (1979)). *Generative phonology.* New York: Academic Press.

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.

Krakow, R. (1989). *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures.* Doctoral dissertation, Yale University, New Haven, CT.

Kugler, P., Kelso, J. A. S., & Turvey, M. (1980). On the concept of coordinative structures as dissipative structures. I. Theoretical lines of convergence. In G. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 3–47). Amsterdam: North-Holland.

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 403–439). Dordrecht: Kluwer.

Lindblom, B., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics, 7,* 147–161.

Lindblom, B., & Sundberg, J. (1971). Acoustical consequences of lip, tongue, jaw and larynx movement. *Journal of the Acoustical Society of America, 50,* 1166–1179.

Moll, K., & Daniloff, R. (1971). Investigation of the timing of velar movements during speech. *Journal of the Acoustical Society of America, 50,* 678–684.

Moll, K., Zimmermann, G., & Smith, A. (1976). The study of speech production as a human neuromotor system. In M. Sawashima & F. S. Cooper (Eds.), *Dynamic aspects of speech production* (pp. 71–82). Tokyo: University of Tokyo.

Mowrey, R., & MacKay, I. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88,* 1299–1312.

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39,* 151–168.

Öhman, 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, MA: MIT Press.

Perkell, J., & Matthies, M. (1992). Temporal measures of labial coarticulation for the vowel /u/. *Journal of the Acoustical Society of America, 91,* 2911–2925.

Pierrehumbert, J. (1990). Phonological and phonetic representations. *Journal of Phonetics, 18,* 375–394.

Recasens, D. (1984a). V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study. *Journal of Phonetics, 12,* 61–73.

Recasens, D. (1984b). Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America, 76,* 1624–1635.

Recasens, D. (1985). Coarticulatory patterns and degrees of coarticulation resistance in Catalan CV sequences. *Language and Speech, 28,* 97–114.

Recasens, D. (1987). An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences. *Journal of Phonetics, 15,* 299–312.

Recasens, D. (1989). Long range coarticulatory effect for tongue dorsum contact in VCVCV sequences. *Speech Communication, 8,* 293–307.

Recasens, D. (1991). An electropalatal and acoustic study of consonant-to-vowel coarticulation. *Journal of Phonetics, 19,* 177–196.

Repp, B. (1981). On levels of description is speech research. *Journal of the Acoustical Society of America, 69,* 1462–1464.

Ryle, G. (1949). *The concept of mind.* New York: Barnes and Noble.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators. In H. Heuer & C. Fromm (Eds.), *Generation and modeling of action patterns* (pp. 129–144). New York: Springer-Verlag.

Saltzman, E., & Kelso, J. A. S. (1987). Skilled action: A task-dynamic approach. *Psychological Review, 94,* 84–106.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1,* 333–382.

Shaiman, S. (1989). Kinematic and electromyographic respones to perturbation of the jaw. *Journal of the Acoustical Society of America, 86,* 78–88.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production parsing. In P. F. MacNeilage (Ed.), *The Production of speech* (pp. 109–135). New York: Springer-Verlag.

Shattuck-Hufnagel, S., & Klatt, D. (1979). Minimal uses of features and markedness in speech production: Evidence from speech errors. *Journal of Verbal Learning and Verbal Behavior, 18,* 41–55.

Stemberger, J. P. (1983). *Speech errors and theoretical phonology: A review.* Bloomington: Indiana University Linguistics Club.

Stemberger, J. P. (1991a). Apparent anti-frequency effects in language production: The addition bias and phonological underspecification. *Journal of Memory and Language, 30,* 161–185.

Stemberger, J. P. (1991b). Radical underspecification in language production. *Phonology, 8,* 73–112.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 211–266). Hillsdale, NJ: Erlbaum.

Turvey, M. T. (1990). Coordination. *American Psychologist, 45,* 938–953.

van der Hulst, H., & Smith, N. (1982). An overview of autosegmental and metrical phonology. In H. van der Hulst & N. Smith (Eds.), *The structure of phonological representations.* Part 1 (pp. 1–46). Dordrecht: Foris.