In the Oxford Handbook of Psycholinguistics, (2007), (Ed.) Gaskell, M.G., Oxford University Press, pp. 489-501.

CHAPTER 29

Speech production

Carol A. Fowler

LANGUAGE forms provide the means by which language users can make an intended linguistic message available to other members of the language community. Necessarily, then, they have two distinct characteristics. On the one hand, they are linguistic entities, morphemes and phonological segments, that encode the talker's linguistic message. On the other hand, they either have physical properties themselves (e.g. Browman and Goldstein, 1986) or, by other accounts, they serve as an interface between the linguistic and physical domains of language use.

A theory of speech production provides an account of the means by which a planned sequence of language forms is implemented as vocal tract activity that gives rise to an audible, intelligible acoustic speech signal.[1] Such an account must address several issues. Two central issues are discussed here.

One issue concerns the nature of language forms that ostensibly compose plans for utterances. Because of their role in making linguistic messages public, a straightforward idea is that language forms are themselves the public behaviors in which members of a language community engage when talking. By most accounts, however, the relation of phonological segments to actions of the vocal tract is not one of identity. Rather, phonological segments are mental categories with featural attributes. We will consider reasons for this stance, relevant evidence, and an alternative theoretical perspective.

[1] By this definition, I intend to contrast the more comprehensive theories of language production from theories of speech production. A theory of language production (e.g. Levelt et al., 1999) offers an account of planning for and implementation of meaningful utterances. A theory of speech production concerns itself only with planning for and implementation of language forms.

Another issue concerns what, at various levels of description, the talker aims to achieve (e.g. Levelt et al., 1999). In my discussion of this issue, I focus here on the lowest level of description—that is, on what talkers aim to make publicly available to listeners. A fundamental theoretical divide here concerns whether the aims are acoustic or articulatory. On the one hand, it is the acoustic signal that stimulates the listener's ears, and so one might expect talkers to aim for acoustic targets that point listeners toward the language forms that compose the talker's intended message. On the other hand, acoustic speech signals are produced by vocal tract actions. The speaker has to get the actions right to get the acoustic signal right.

Readers may wonder whether this is a "tempest in a teapot." That is, why not suppose that talkers plan and control articulations that will get the signal right, so that in a sense both articulation and acoustics are controlled? Readers will see, however, that there are reasons why theorists typically choose one account or the other.

These issues are considered in turn in the following two sections.

29.1 Language forms and plans for speaking

By most accounts, as already noted, neither articulation nor the acoustic signal is presumed to implement phonological language forms transparently. Language forms are conceived of as abstract mental categories about which acoustic speech signals provide cues.

There are two quite different reasons for this point of view. One is that language forms are cognitive entities (e.g. Pierrehumbert, 1990). In particular, word forms are associated, presumably in the lexical memory of a language user,

with word meanings. As such they constitute an important part of what a language user knows that permits him or her to produce and understand language. Moreover, word forms in the lexicons of languages exhibit systematic properties which can be captured by formal rules. There is some evidence that language users know these rules. For example, in English, voiceless stop consonants are aspirated in stressed syllable-initial position. That systematic property can be captured by a rule (Kenstowicz and Kisseberth, 1979).

Evidence that such a rule is part of a language user's competence is provided, for example, by foreign accents. When native English speakers produce words in a Romance language such as French, which has unaspirated stops where English has aspirated stops, they tend to aspirate the stops. Accordingly, the word *pas*, [pa][2] in French is pronounced [p^{h}a] as if the English speaker is applying the English rule to French words. A second source of evidence comes from spontaneous errors of speech production. Kenstowicz and Kisseberth (1979) report an error in which a speaker intended to produce *tail spin*, but instead said *pail stin*. In the intended utterance, /t/ in *tail* is aspirated; /p/ in *spin* is unaspirated. The authors report, however, that, in the error, appropriately for their new locations, /p/ was pronounced [p^h]; /t/ was pronounced [t]. One account of this "accommodation" (but not the only one possible) is that the exchange of /t/ and /p/ occurred before the aspiration rule had been applied by the talker. When the aspiration rule was applied, /p/ was accommodated to its new context.[3]

A second reason to suppose that language forms exist only in the mind is coarticulation. Speakers temporally overlap the articulatory movements for successive consonants and vowels. This makes the movements associated with a given phonetic segment context-sensitive and lacking an obvious discrete segmental structure. Likewise, the acoustic signal which the movements produce is context-sensitive. Despite researchers' best efforts (e.g. Stevens and Blumstein, 1981) they have not uncovered invariant acoustic information for individual consonants and vowels. In addition, the acoustic signal, like the movements that produce it, lacks a phone-sized segmental structure.

This evidence notwithstanding, there are reasons to resist the idea that language forms reside only in the minds of language users. They are, as noted, the means that languages provide to make linguistic messages public. Successful recognition of language forms would seem more secure were the forms themselves public things.

Browman and Goldstein (e.g. 1986; 1992) have proposed that phonological language forms are gestures achieved by vocal tract synergies that create and release constrictions. They are both the actions of the vocal tract (properly described) that occur during speech and at the same time units of linguistic contrast. ("Contrast" means that a change in a gesture or gestural parameter can change the identity of a word. For example, the word *hot* can become *tot* by addition of a tongue tip constriction gesture; *tot* can become *sot* by a change in the tongue tip's constriction degree.)

From this perspective, phonetic gestures *are* cognitive in nature. That is, they are components of a language users' language competence, and, as noted, they serve as units of contrast in the language. However, cognitive entities need not be covert (see e.g. Ryle, 1949). They can be psychologically meaningful actions, in this case of a language user. As for coarticulation, although it creates context sensitivity in articulatory movements, it does not make gestures context-sensitive. For example, lip closure for /b/, /p/, and /m/ occurs despite coarticulatory encroachment from vowels that affects jaw and lip motion.

There is some skepticism about whether Browman and Goldstein's "articulatory phonology" as just described goes far enough beyond articulatory phonetics.[4] This is in part because it does not yet provide an account of many of the phonological systematicities (e.g. vowel harmony in Hungarian, Turkish, and many other languages; but see Gafos and Benus, 2003) which exist across the lexicon of languages and that other theories of phonology capture by means of rules (e.g. Kenstowicz and Kisseberth, 1979) or constraints (Archangeli,1997). However, the theory is well worth considering, because it is unique in proposing that language forms are public events.

[2] Slashes (e.g. /p/) indicate phonological segments; square brackets (e.g. [p]) signify phonetic segments. The difference is one of abstractness. For example, the phonological segment /p/ is said to occur in two varieties—the aspirated phonetic segment [p^h] and the unaspirated [p].

[3] An alternative account, which does not implicate rule use, is that *pail stin* reflects a single feature or gesture error. From a featural standpoint, place of articulation features of /p/ and /t/ exchange, stranding the aspiration feature.

[4] See articles in the 1992 special issue of the journal *Phonetica* devoted to a critical analysis of articulatory phonology.

Spontaneous errors of speech production have proved important sources of evidence about language planning units. These errors, produced by people who are capable of producing error-free tokens, appear to provide evidence both about the units of language that speakers plan to produce and about the domain over which they plan. Happily, the units which participate in errors have appeared to converge with units that linguistic analysis has identified as real units of the language. For example, words participate in errors as anticipations (e.g. *sky is in the sky* for intended *sun is in the sky*; this and other errors from Dell, 1986), perseverations (*class will be about discussing the class* for intended *class will be about discussing the test*), exchanges (*writing a mother to my letter* for *writing a letter to my mother*), and non-contextual substitutions (*pass the salt* for *pass the pepper*). Consonants and vowels participate in the same kinds of error. Syllables do so only rarely; however, they serve as frames that constrain how consonants and vowels participate in errors. Onset consonants interact only with onset consonants; vowels interact with vowels; and, albeit rarely, coda consonants interact with coda consonants. Interacting segments tend to be featurally similar to one another. Moreover, when segments move, they tend to move to contexts which are featurally similar to the contexts in which they were planned to occur. Segments are anticipated over shorter distances than words (Garrett, 1980), suggesting that the planning domains for words and phonological segments are different.

Historically, most error corpora were collected by individuals who transcribed the errors that they heard. As noted, the errors tended to converge with linguists' view of language forms as cognitive, not physical entities (e.g. Pierrehumbert, 1990). As researchers moved error collection into the laboratory, however, it became clear that errors occur that are inaudible. Moreover, these errors violate constraints on errors that collectors had identified.

One constraint was that errors are categorical in nature. If, in production of *Bob flew by Bligh Bay*, the /l/ of *Bligh* were perseverated into the onset of *Bay*, producing *Blay*, the /l/ would be a fully audible production. However, electromyographic evidence revealed to Mowrey and MacKay (1990) that errors are gradient. Some produce an audible token of /l/; others do not, yet show activity of a lingual muscle indicating the occurrence of a small lingual (tongue) gesture for /l/.

A second constraint is that errors result in phonologically well-formed utterances. Not only do vowels interact only with other vowels in errors, and onsets with onsets and codas with codas, but also sequences of consonants in onsets and codas tend to be permissible in the speaker's language. Or so investigators thought before articulatory data were collected in the laboratory. Pouplier (2003a; 2003b) used a mid-sagittal electromagnetometer to collect articulator movement data as participants produced repetitions of pairs of words such as *cop–top* or *sop–shop*. Like Mowrey and MacKay (1990), she found errorful articulations (for example, intrusive tongue tip movement toward a /t/ articulation during *cop*) in utterances that sounded error-free. In addition, however, she found that characteristically intrusions were not accompanied by reductions of the intended gesture. This meant that, in the foregoing example, constriction gestures for both /t/ and /k/ occurred in the onset of a syllable, a phonotactically impermissible cluster for her English speakers.

What do these findings imply for theories of speech production? For Pouplier and colleagues (Pouplier, 2003b; Goldstein et al., forthcoming), planning units are intended sequences of vocal-tract gestures that are coordinated in the manner of coupled oscillators. In the literature on limb movements, it has been found that two modes of coordination are stable. Limbs (or hands or fingers) may be oscillated in phase or 180 degrees out of phase (so that extension of one limb occurs when the other limb is flexing). In tasks in which, for example (Kelso, 1984; see also Yamanishi et al., 1980), hands are oscillated about the wrist at increasing rates, in-phase movements remain stable; however, out-of-phase movements become unstable. Participants attempting to maintain out-of-phase movements slip into phase. Pouplier and colleagues suggest that findings of intrusive tongue tip gestures in the onset of *cop* and of intrusive tongue body gestures in *top* constitute a similar shift from a less to a more stable oscillation mode. When *top–cop* is repeated, syllable onsets /t/ and /k/ each occur once for each pair of rime (/ap/) productions giving a 1:2 coordination mode. When intrusive /t/ and /k/ gestures occur, the new coordination mode is 1:1; that is, the new onset is produced once for each one production of the syllable rime. A 1:1 coordination mode is more stable than a 1:2 mode.

A question is what the findings of gradient, phonotactically impermissible errors imply about the interpretability of error analyses based on transcribed, rather than articulatory, corpora. Certainly these errors occur, and certainly they were missed in transcription corpora.

However, does it mean that categorical consonant and vowel errors do not occur, that planning units should be considered to be intended phonetic gestures (Pouplier) or even commands to muscles (Mowrey and MacKay), not the consonants and vowels of traditional phonetic analysis?

There are clearly categorical errors that occur at the level of whole words (recall *writing a mother to my letter*). It does not seem implausible, therefore, that categorical phonetic errors also occur. It may be appropriate (as in the model of Levelt et al., 1999) to imagine levels of speech planning, with consonants and vowels of traditional analyses serving as elements of plans at one level, giving way to planned gestures at another.

Findings that error corpora in some ways misrepresent the nature of spontaneous errors of speech production, however, have had the positive consequence that researchers have sought converging (or, as appropriate, diverging) evidence from experiments that elicit error-free speech. For example, Meyer (1991) found evidence for syllable constituents serving as "encoding" units in language production planning. Participants memorized sets of word pairs consisting of a prompt word produced by the experimenter and a response word produced as quickly as possible by the participant. Response words in a set were "homogeneous" if they shared one or more phonological segments; otherwise they were "heterogeneous." Meyer found faster responses to words in homogeneous compared to heterogeneous sets if response words shared their initial consonant or initial syllable, but not if they shared the syllable rime (that is, the vowel and any following consonants). There was no further advantage over responses to heterogeneous words when the CV of a CVC syllable was shared in homogeneous sets as compared to when just the initial C was shared. There was an advantage over responses to words sharing the initial consonant of responses to words sharing the whole first syllable. These findings suggest, as errors do, that syllable constituents are among the planning units. They also suggest that encoding for production is a sequential "left-to-right" process.

Sevald et al. (1995) obtained converging evidence with errors data suggesting that syllables serve as planning frames. They asked participants to repeat pairs of non-words (e.g. KIL KILPER or KIL KILPNER) in which the initial monosyllable either did or did not match the initial syllable of the disyllable. The task was to repeat the pair as many times as possible in four seconds. Mean syllable production time was less when the syllable structure matched. Remarkably, the advantage of matching syllable structure was no less when only syllable structure, but not syllable content, matched (e.g. KEM TILFER vs. KEM TILFNER). In the foregoing examples, it looks as if the advantage could be due to the fact that there were fewer phonetic segments to produce in the matching condition. However, there were other items in which the length advantage was reversed.

29.2 Speakers' goals as acoustic targets or vocal tract gestures

A next issue is how intended sequences of phonetic entities are planned to be implemented as actions or their consequences that are available to a listener. In principle, this issue is orthogonal to the one just considered about the nature of planned language forms. As just discussed, these forms are variously held to be covert, cognitive representations or public, albeit still cognitive, entities. Either view is compatible with proposals that, at the lowest level of description, talkers aim to achieve either acoustic or gestural targets. In the discussion below, therefore, the issue of whether language forms are covert or public in nature is set aside. It may be obvious, however, that, in fact, acoustic target theorists at least implicitly hold the former view and gesture theorists the latter.

Guenther et al. (1998) argue against gestural targets on several grounds and argue for acoustic targets. One ground for rejecting gestural targets, such as constriction location and degree, concerns the feedback information that speakers would need to implement the targets sufficiently accurately. To know whether or not a particular constriction has been achieved requires perceptual information. If, for example, an intended constriction is by the lips (as for /b/, /p/, or /m/), talkers can verify that the lips are closed from proprioceptive information for lip contact. However, Guenther et al. argue that, in particular for vowels, constrictions do not always involve contact by articulators, and therefore intended constrictions cannot be verified. In addition, they argue, to propose that talkers intend to achieve particular constrictions implies that talkers should not be able to compensate for experimental perturbations that prevent those constrictions from being achieved. However, some evidence suggests that they can. For example, Savariaux et al. (1995) had talkers

produce vowels with a tube between their lips that prevented normal lip rounding for the vowel /u/. The acoustic effects of the lip tube could be compensated for by lowering the larynx (thereby enlarging the oral cavity by another means than rounding). Of the eleven participants, one compensated fully for the lip tube. Six others showed limited evidence of compensation.

A third argument for acoustic targets is provided by American English /r/. According to Guenther et al., /r/ is produced in very different ways by different speakers or even by the same speaker in different contexts. The different means of producing /r/ are acoustically very similar. One account for the articulatory variability, then, is that it is tolerated if the different means of production produce inaudibly different acoustic signals, the talker's production aim. Finally, Guenther et al. argue that ostensible evidence for constriction targets—that, for example, invariant constriction gestures occur for /b/ and other segments—need not be seen as evidence uniquely favoring gestural targets. Their model "DIVA" (originally "directions in orosensory space onto velocities of articulators"; described below) learns to achieve acoustic-perceptual targets, but nonetheless shows constriction invariance. However, there is also evidence favoring the alternative idea that talkers' goals are articulatory not acoustic. Moreover, the arguments of Guenther et al. favoring acoustic targets can be challenged.

Tremblay et al. (2003) applied mechanical perturbations to the jaw of talkers producing the word sequence *see–at*. The perturbation altered the motion path of the jaw, but had small and inaudible acoustic effects. Even though acoustic effects were inaudible, over repetitions, talkers compensated for the perturbations and showed after-effects when the perturbation was removed. Compensation also occurred in a silent speech condition, but not in a non-speech jaw movement condition. These results appear inconsistent with a hypothesis that speech targets are acoustic.

There is also a more natural speech example of preservation of inaudible articulations. In an investigation of an X-ray microbeam database, Browman and Goldstein (1991) found examples of utterances such as *perfect memory* in which transcription suggested deletion of the final /t/ of *perfect*. However, examination of the tongue tip gesture for the /t/ revealed its presence. Because of overlap from the bilabial gesture of /m/, however, acoustic consequences of the /t/ constriction gesture were absent or inaudible. As for the suggestion that constriction goals should be unverifiable by feedback when constricting articulators are not in contact with another structure, to my knowledge this is untested speculation.

As for the compensation found by Savariaux et al (1995; see also Perkell et al., 1993), Guenther et al. do not remark that the compensation is markedly different from that associated with certain other perturbations in being, for most participants, either partial or absent. Compensations for a bite block (which prevents jaw movement) are immediate and nearly complete in production of vowels (e.g. Lindblom et al., 1979). Compensations for jaw and lip perturbations during speech (e.g. tugging the jaw down as it raises to close the lips for a /b/) are very short in latency, immediate, and nearly complete (e.g. Kelso et al., 1984). These different patterns of compensation are not distinct in the DIVA model. However, they are in speakers. The difference may be understood as relating to the extent to which they mimic perturbations which occur naturally in speech production. When a speaker produces, say, /ba/ versus /bi/, coarticulation by the following low (/a/) or high (/i/) vowel will tug the jaw and lower lip down or up. Speakers have to compensate for that to get the lips shut for bilabial /b/. That routine compensation for coarticulation may underlie fast and functional compensations which occur in the laboratory (Fowler and Saltzman, 1993). However, it is a rare perturbation outside the laboratory that prevents lip rounding. Accordingly, talkers may have no routines in place to compensate for the lip tube, and have to learn them. In a gestural theory, they have to learn to create a mirage—that is, an acoustic signal that mimics consequences of lip rounding

As for /r/, ironically, it has turned out to be a poster child for both acoustic and articulatory theorists. Delattre and Freeman (1968), whom Guenther et al. cite as showing considerable variability in American English articulation of /r/, in fact remark that in every variant they observed there were two constrictions, one by the back of the tongue in the pharyngeal region and one by the tongue tip against the hard palate. (Delattre and Freeman were only looking at the tongue, and so did not remark on a third shared constriction, rounding by the lips.) Accordingly, whether one sees variability or invariance in /r/ articulations may depend on the level of description of the vocal tract configuration deemed relevant to talkers and listeners. In Browman and Goldstein's articulatory phonology (e.g. 1986; 1995), the relevant level is

that of constriction locations and degrees, and those are invariant across the /r/ variants.

Focus on constrictions permits an understanding of a source of dialect variation in American English /r/ that is not illuminated by a proposal that acoustic targets are talkers' aims. Among consonants involving more than one constriction—for example, the nasal consonants (constrictions by lips, tongue tip or tongue body, and by the velum), the liquids, /l/ (tongue tip and body) and /r/ (tongue body, tip, and lips), and the approximant /w/ (tongue body and lips)—a generalization holds regarding the phasing of the constriction gestures. Prevocalically, the gestures are achieved nearly simultaneously; postvocalically, the gesture with the more open (vowel-like) constriction degree leads (see research by Sproat and Fujimora, 1993; Krakow, 1989; 1993; Gick, 1999). This is consistent with the general tendency in syllables for the more sonorant (roughly more vowel-like) consonants to be positioned closest to the vowel. (For example, the ordering in English is /tr/ before the vowel as in *tray*, but /rt/ after the vowel as in *art*.) Goldstein (pers. comm., 15 Aug. 2005) points out that, in two dialects of American English, one spoken in Brooklyn and one in New Orleans, talkers produce postvocalic consonants in such a way that, for example, *bird* sounds to listeners somewhat like *boyd*. This is understandable if talkers exaggerate the tendency for the open lip and tongue body constrictions to lead the tip constriction. Together, the lip and tongue body configurations create a vowel sound like /ɔ/ (in *saw*); by itself, the tip gesture is like /i/ (in *see*). Together, the set of gestures yield something resembling the diphthong /ɔ^{i}/ as in *boy*.

In short, there are arguments and there is evidence favoring both theoretical perspectives—that targets of speech production planning are acoustic or else are gestural. Deciding between the perspectives will require further research.

29.2.1 Theories of speech production

As noted, theories of speech production differ in their answer to the question of what talkers aim to achieve, and a fundamental difference is whether intended targets are acoustic or articulatory. Within acoustic theories, accounts can differ in the nature of acoustic targets; within articulatory theories, accounts can be that muscle lengths or muscle contractions are targets, that articulatory movements are targets, or that coordinated articulatory gestures are targets. I will review one acoustic and one articulatory account. I chose these accounts because they are the most fully developed theories within the acoustic and articulatory domains.

29.3 The DIVA theory of speech production

In this account (e.g. Guenther et al., 1998), targets of speaking are normalized acoustic signals reflecting resonances of the vocal tract ("formants"). The normalization transformations create formant values that are the same for men, women, and children even though acoustic reflections of formants are higher in frequency for women than for men and for young children than for women. Because formants characterize vowels and sonorant consonants but not (for example) stop or fricative consonants, the model is restricted to explanation of just those classes of phones.

Between approximately six and eight months of age, infants engage in vocal behavior called babbling in which they produce what sounds like sequences of CV syllables. In this way, in DIVA, the young model learns a mapping from articulator positions to normalized acoustic signals. Over learning, this mapping is inverted so that acoustic-perceptual targets can underlie control of articulatory movements. In the model, the perceived acoustic signal has three degrees of freedom (one per normalized formant). In contrast, the articulatory system has the seven degrees of freedom of Maeda's (1990) articulatory model. This difference in degrees of freedom mean that the inverted mapping is one to many. Accordingly, a constraint is required to make the mapping determinate. Guenther et al. use a "postural relaxation" constraint whereby the articulators remain as close as possible to the centers of their ranges of motion. This constraint underlies the model's tendency to show near-invariance of constrictions despite having acoustic-perceptual rather than articulatory targets.

In addition to that characteristic, the model compensates for perturbations—not, however, distinguishing those that humans do well and poorly.

29.4 The task dynamic model

Substantially influenced by the theorizing of Bernstein (1967), Turvey (1977) introduced a theory of action in which he proposed that the minimal meaningful units of action were

produced by synergies or coordinative structures (Easton, 1972). These are transiently established coordinative relations among articulators—those of the vocal tract for speech—which achieve action goals. An example in speech is the organized relation among the jaw and the two lips that achieves bilabial constriction for English /b/, /p/, or /m/. That coordinative relation is not in place when speakers produce a constriction which does not include lip closure (e.g. Kelso et al., 1984). The coordinative relation underlies the ability of speakers to compensate for jaw or lip perturbations in the laboratory, and presumably to compensate for coarticulatory demands on articulators shared by temporally overlapping phones outside the laboratory.

Saltzman and colleagues (e.g. Saltsman and Kelso, 1987; Saltzman and Munhall, 1989; see also Turvey, 1990) proposed that synergies are usefully modeled as dynamical systems. Specifically, they suggested that speech gestures can be modeled as mass-spring systems with point attractor dynamics. In turn those systems are characterized by equations that reflect how the systems' states undergo change over time. Each vocal tract gesture is defined in terms of "tract variables." Variables include lip protrusion (a constriction location) and lip aperture (constriction degree). Appropriately parameterized, the variables achieve gestural goals. The tract variables have associated articulators (e.g. the jaw and the two lips) that constitute the synergy that achieves that gestural goal. In one version of the theory, a word is specified by a "gestural score" (Browman and Goldstein, 1986) which provides parameters for the relevant tract variables and the interval of time over which they should be active. In a more recent version (Saltzman et al., 2000) gestural scores are replaced by a central "clock" that regulates the timing of gesture activation. The clock's average "tick" rate determines the average rate of speaking. As we will see later, local clock slowing can mark the edges of prosodic domains.

These systems show the equifinality characteristic of real speakers which underlies their ability to compensate for perturbations. That is, although the parameters of the dynamical system for a gesture have context independent values, gestural goals are achieved in a context-dependent manner so that, for example, as in the research by Kelso et al.(1984), lip closure for /b/ is achieved by different contributions from the lips and jaw on perturbed and unperturbed trials. The model compensates for perturbations which speakers handle without learning, but not for those such as in the study by Savariaux et al., which speakers require learning to handle, if they handle them at all.

29.5 Coarticulation

A hallmark of speech production is coarticulation. Speakers talk very quickly, and talking involves rapid sequencing of the particulate atoms (Studdert-Kennedy, 1998) which constitute language forms. Although the atoms are discrete, their articulation is not. Much research on speech production has been conducted with an aim to understand coarticulation. Coarticulation is characterized either as context-sensitivity of production of language forms or as temporally overlapping production. It occurs in both an anticipatory and a carryover direction. In the word *stew*, for example, lip rounding from the vowel /u/ begins near the beginning of the /s/. In *use*, it carries over during /s/.

Thirty years ago, there were two classes of accounts of coarticulation. In one point of view (e.g. Daniloff and Hammarberg, 1973) coarticulation was seen as "feature spreading." Consonants and vowels can be characterized by their featural attributes. For example, consonants can be described as being voiced or unvoiced, as having a particular place of articulation (e.g. bilabial for /b/, /p/, and /m/) and a particular manner of articulation (e.g. /b/ and /p/ are stops; /f/ is a fricative). Vowels are front, mid, or back; high, mid, or low, and rounded or unrounded. Many features which characterize consonants and vowels are contrastive, in that changing a feature value changes the identity of a consonant or vowel and the identity of a word that they, in part, compose. For example, changing the feature of a consonant from voiced to unvoiced can change a consonant from /b/ to /p/ and a word from *bat* to *pat*. However, some features are not contrastive. Adding rounding to a consonant does not change its identity in English; adding nasalization to a vowel in English likewise does not change its identity.

In feature spreading accounts of coarticulation, non-contrastive features were proposed to spread in an anticipatory direction to any phone unspecified for the feature (i.e. for which the feature was non-contrastive). Accordingly, lip rounding should spread through any consonant preceding a rounded vowel; nasalization should spread through any vowel preceding a nasal consonant. Carryover coarticulation was seen as inertial. Articulators cannot stop on a dime. Accordingly lip rounding might continue during

a segment following a rounded vowel. There was some supportive evidence for the feature spreading view of anticipatory coarticulation (Daniloff and Moll, 1968).

However, there was also disconfirming evidence. One was a persistent finding (e.g. Benguerel and Cowan, 1974) that indications of coarticulation did not neatly begin at phonetic segment edges, as they should if a feature had spread from one phone to another. A second kind of evidence consisted of reports of "troughs" (e.g. Gay, 1978; Boyce, 1990). These were findings that, for example, during a consonant string between two rounded vowels, the lips would reduce their rounding and lip muscle activity would reduce, inconsistent with an idea that a rounding feature had spread to consonants in the string.

A different general point of view was that coarticulation was "coproduction" (e.g. Fowler, 1977)—i.e. temporal overlap in the production of two or more phones. In this point of view, for example, rounding need not begin at the beginning of a consonant string preceding a rounded vowel, and a trough during a consonant string between two rounded vowels would be expected as the rounding gesture for the first vowel wound down and before rounding for the second vowel began. Bell-Berti and Harris (1981) proposed a specific account of coproduction, known as "frame" theory, in which anticipatory coarticulation began a fixed interval before the acoustically defined onset of a rounded vowel or nasalized consonant.

For a while (Bladon and Al-Bamerni, 1982; Perkell and Chiang, 1986), there was the congenial suggestion that both theories might be right. Investigators found evidence sometimes that there was a start of a rounding or nasalization gesture at the beginning of a consonant (for rounding) or vowel string preceding a rounded vowel or nasalized consonant. Then, at an invariant interval before the rounded or nasalized phone, there was a rapid increase in rounding or nasalization as predicted by frame theory. However, that evidence was contaminated by a confounding (Perkell and Matthies, 1992). Bell-Berti and colleagues (e.g. Boyce et al., 1990); Gelfer et al., 1989) pointed out that some consonants are associated with lip rounding (e.g. /s/). Similarly, vowels are associated with lower positions of the velum compared to oral obstruents. Accordingly, to assess when anticipatory coarticulation of lip rounding or nasalization begins requires appropriate control utterances, to enable a distinction to be made between lip rounding or velum lowering due to coarticulation and that due to characteristics of phonetic segments in the coarticulatory domain. For lip rounding, for example, rounding during an utterance such as *stew* requires comparision with rounding during a control utterance such as *stee* in which the rounded vowel is replaced by an unrounded vowel. Any lip rounding during the latter utterance indicates rounding associated with the consonant string, and needs to be subtracted from lip activity during *stew*. Likewise, velum movement during a CV_nN sequence (that is, a sequence consisting of an oral consonant followed by *n* vowels preceding a nasal consonant) needs to be compared to velum movement during a CV_nC sequence. When those comparisons are made, evidence for feature spreading evaporates.

Recently, two different coproduction theories have been distinguished (Lindblom et al., 2002). In the account proposed by Ohman (1966), vowels are produced continuously. In a VCV utterance, according to the account, speakers produce a diphthongal movement from the first to the second vowel. The consonant was superimposed on that diphthongal trajectory. In the alternative account (e.g. Fowler and Saltzman, 1993), gestures for consonants and vowels overlap temporally. Any vowel-to-vowel overlap is temporal overlap, not production of a diphthongal gesture.

Evidence favoring the view of Fowler and Saltzman is the same kind of evidence that disconfirmed feature spreading theory. As noted earlier, speakers show troughs in lip gestures in sequences of consonants that intervene between rounded vowels. They should not if vowels are produced as diphthongal tongue gestures, but they are expected to if vowels are produced as separate gestures that overlap temporally with consonantal gestures.

29.5.1 Coarticulation resistance

Coarticulation has been variously characterized as a source of distortion (e.g. Ohala, 1981)—i.e. as a means by which articulation does not transparently implement essential phonological properties of consonants and vowels—or even as destructive of those properties (e.g. Hockett, 1955).

However, these characterizations overlook the finding of "coarticulation resistance"—an observation first made by Bladon and Al-Bamerni (1976), but developed largely by Recasens (e.g. 1984a; 1984b; 1985; 198); see also Farnetani, 1990). This is the observation that phones resist coarticulatory overlap by neighbors to the extent that the neighbors would interfere with achievement of the phones' gestural goals. For example,

Recasens (1984a) found decreasing vowel-to-vowel coarticulation in Catalan VCV sequences when the intervening consonant was one of the set: /j/ (a dorso-palatal approximant), /ɲ/ (an alveolo-palatal nasal), /ʎ/ (an alveolo-palatal lateral), /n/ (an alveolar nasal). In the set, the consonants decreasingly use the tongue body to achieve their place of articulation. The tongue body is a major articulator in the production of vowels. Accordingly, it is likely that the decrease in vowel-to-vowel coarticulation in the consonant series occurs to prevent the vowels from interfering with achievement of the consonants' constriction location and degree. Recasens (1984b) found increasing vowel-to-consonant coarticulation in the same consonant series.

Compatible data from English can be seen in Figure 29.1. Figure 1a shows tongue body fronting

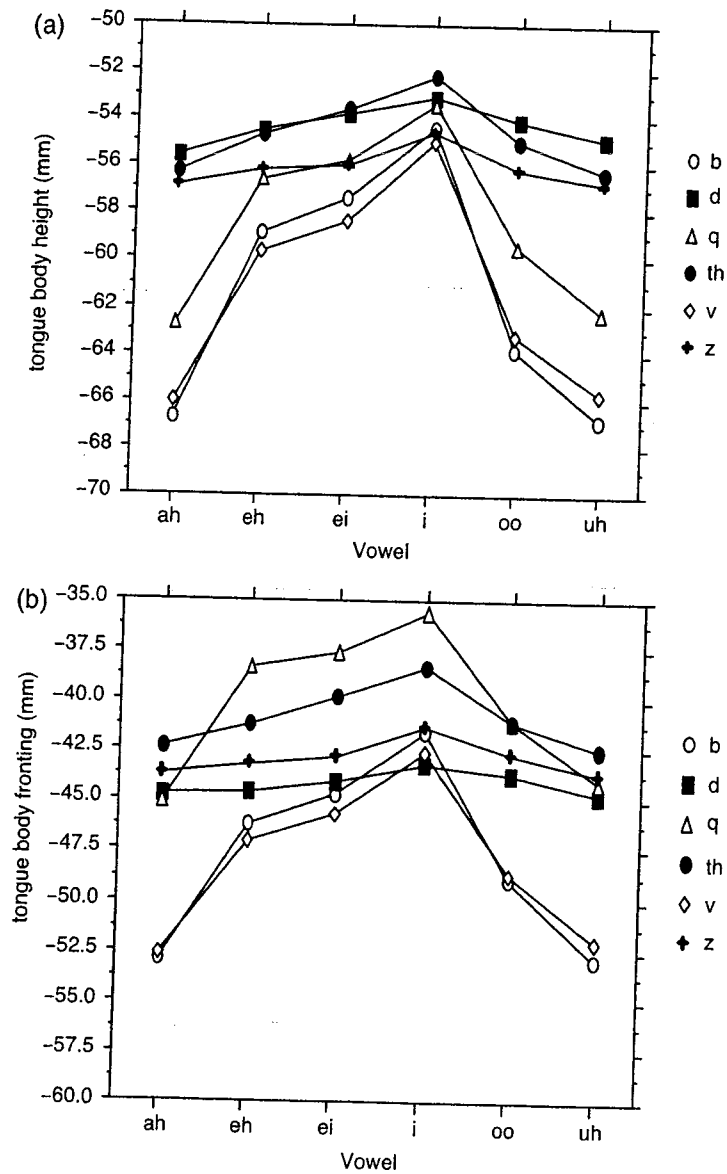

Figure 29.1 Tongue body height (a) and fronting (b) during production of three high and three low coarticulation resistant consonants produced in the context of six following stressed vowels. Measures taken in mid consonant closure.

data from a speaker of American English producing each of six consonants in the context of six following vowels (Fowler, 2005). During closure of three consonants (/b/, /v/, and /g/), there is a substantial shift in the tongue body height depending on the following vowel. During closure of the other three consonants (/d/, /z/, and /ð/, there is considerably less. Figure 1b shows similar results for tongue dorsum fronting. /b/, /v/, and, perhaps surprisingly, /g/ show less resistance to coarticulation for this speaker of American English than do /d/, /z/ and /ð/. The results for /b/ and /v/ most likely reflect the fact that they are labial consonants. They do not use the tongue, and so coproduction by vowels does not interfere with achievement of their gestural goals. The results for /g/, the fronting results at least, may reflect the fact that there is no stop in American English that is close in place of articulation with /g/ that might be confused with it were /g/'s place of articulation to shift due to coarticulation by the vowels.

29.5.2 Other factors affecting coarticulation

Frame theory (Bell-Berti and Harris, 1981) suggests a fixed extent of anticipatory coarticulation, modulated perhaps by speaking rate. However, the picture is more complicated. Browman and Goldstein (1988) reported a difference in respect to how consonants are phased to a tautosyllabic vowel depending on whether the consonants were in the syllable onset or in the coda. Consonants in the onset of American English syllables are phased so that the gestural midpoint of the consonants aligns with the vowel. In contrast, in the coda, the first consonant is phased invariantly with respect to the vowel regardless of the number of consonants in the coda.

For multi-gesture consonants, such as /l/ (Sproat and Fujimura, 1993), /r/, /w/ (Gick, 1999), and the nasal consonants (Krakow, 1989), the gestures are phased differently in the onset and coda. Whereas they are nearly simultaneous in the onset, the more open (more vowel-like) gestures precede in the coda. This latter phasing appears to respect the "sonority hierarchy" such that more vowel-like phones are closest to the vowel.

29.6 Prosody

There is more to producing speech than sequencing consonants and vowels. Speech has prosodic properties including an intonation contour, various temporal properties, and variations in articulatory "strength."

Theorists (see Shattuck-Hufnagel and Turk, 1996 for a review) identify hierarchical prosodic domains, each marked in some way phonologically. Domains include intonational phrases, which constitute the domain of complete intonational contours, intermediate phrases marked by a major ("nuclear") pitch accent and a tone at the phrase boundary, prosodic words (lexical words or a content word followed by a function word as in "call up"), feet (a strong syllable followed by zero or one weak syllables), and syllables. Larger prosodic domains often, but not always, set off syntactic phrases or clauses.

Intonation contours are patterns of variation in fundamental frequency consisting of high and low pitch accents, or accents that combine a high and low (or low and high) pitch excursions, and boundary tones at intonational and intermediate phrase boundaries. Pitch accents in the contours serve to accent information that the speaker wants to focus attention on, perhaps because it is new information in the utterance or because the speaker wants to contrast that information with other information. A whole intonation contour expresses some kind of meaning. For example, intonation contours can distinguish yes/no questions from statements (e.g. *So you are staying home this weekend ?*) Other contours can express surprise, disbelief or other expressions.

Because intonation contours reflect variation in fundamental frequency (f0), their production involves laryngeal control. This laryngeal control is coarticulated with other uses of the larynx, for example, to implement voicing or devoicing, intrinsic f0 (higher f0 for higher vowels), and tonal accompaniments of obstruent devoicing (a high tone on a vowel following an unvoiced obstruent).

Prosody is marked by other indications of phrasing. Prosodic domains from intonational phrases to prosodic words tend to be marked by final lengthening, pausing, and initial and final "strengthening." These effects generally increase in magnitude with the "strength" of the prosodic boundary (where "strength" increases with height of a phrase in the prosodic hierarchy). Final lengthening is an increase in the duration of articulatory gestures and their acoustic consequences before a phrase boundary. Strengthening is a quite local increase in the magnitude of gestures at phrase edges (e.g. Fougeron and Keating, 1997). Less coarticulation occurs across stronger phrase boundaries, and accented vowels resist vowel-to-vowel coarticulation (Cho, 2004).

These marks of prosodic structure serve to demarcate informational units in an utterance. However, we need to ask: why these marks? Final lengthening and pausing are, perhaps, intuitive. Physical systems cannot stop on a dime, and if the larger prosodic domains involve articulatory stoppings and restartings, then we should expect to see slowing to a stop and, sometimes, pausing before restarting. However, why strengthening? Byrd and Saltzman (2003) provide an account of final lengthening and pausing that may also provide some insight into at least some of the occurrences of strengthening. They have extended the task dynamic model, described earlier, to produce the timing variation that characterizes phrasing in prosody. They do so by slowing the rate of time flow of the model's central clock at phrase boundaries. Clock slowing gives rise to longer and less overlapped gestures at phrase edges. The magnitude of slowing reflects the strength of a phrase boundary. Byrd and Saltzman conceive of the slowing as a gesture (a "π gesture") that consists of an activation wave applied to any segmental gesture with which it overlaps temporally. π gestures span phrase boundaries, and therefore have effects at both edges of a phrase. Because clock slowing has as one effect, less overlap of gestures, a consequence may be less truncation of gestures due to overlap and so larger gestures.

Acknowledgments

Preparation of the manuscript was supported by NICHD grant HD-01994 and NIDCD grant DC-03782 to Haskins Laboratories.

References

Archangeli, D. (1997) Optimality theory: an introduction to linguistics in the 1990s. In D. Archangeli and D. T. Langendoen (eds), *Optimality Theory: An Overview*, pp. 1–32. Blackwell, Malden, MA.

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

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

Bernstein, N. (1967) *The Coordination and Regulation of Movement*. Pergamon, London.

Bladon, A., and Al-Bamerni, A. (1982) One-stage and two-stage temporal patterns of coarticulation. *Journal of the Acoustical Society of America*, 72: S104.

Bladon, A., and Al-Bamerni, A. (1976). Coarticulation resistance in English /l/. *Journal of Phonetics*, 4: 137–50.

Boyce, S. (1990) Coarticulatory organization for lip rounding in Turkish and in English. *Journal of the Acoustical Society of America*, 8: 2584–95.

Boyce, S., Krakow, R., Bell-Berti, F., and Gelfer, C. (1990) Converging sources of evidence for dissecting articulatory movements into gestures. *Journal of Phonetics*, 18: 173–88.

Browman, C., and Goldstein, L. (1986) Towards an articulatory phonology. *Phonology Yearbook*, 3: 219–52.

Browman, C., and Goldstein, L. (1988) Some notes on syllable structure in articulatory phonology. *Phonetica*, 45: 140–55.

Browman, C., and Goldstein, L. (1991) Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. Beckman (eds), *Papers in Laboratory Phonology*, vol. 1: *Between the Grammar and the Physics of Speech*, pp. 341–76. Cambridge University Press, Cambridge.

Browman, C., and Goldstein, L. (1992) Articulatory phonology: an overview. *Phonetica*, 49: 155–80.

Browman, C., and Goldstein, L. (1995) Dynamics and articulatory phonology. In R. Port and T. van Gelder (eds), *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 175–93. MIT Press, Cambridge, MA.

Byrd, D., and Saltzman, E. (2003) The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31: 149–80.

Cho, T. (2004) Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, 32: 141–76.

Daniloff, R., and Hammarberg, R. (1973) On defining coarticulation. *Journal of Phonetics*, 1: 239–48.

Daniloff, R., and Moll, K. (1968) Coarticulation of lip rounding. *Journal of Speech and Hearing Research*, 11: 707–21.

Delattre, P., and Freeman, D. (1968) A dialect study of American r's by x-ray motion picture. *Linguistics*, 44: 29–68.

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

Easton, T. (1972) On the normal use of reflexes. *American Scientist*, 60: 591–9.

Farnetani, E. (1990) V-C-V lingual coarticulation and its spatiotemporal domain. In W. J. Hardcastle and A. Marchal (eds), *Speech Production and Speech Modeling*, pp. 93–130. Kluwer, The Netherlands.

Fougeron, C., and Keating, P. (1997) Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, 101: 3728–40.

Fowler, C. A. (1977) *Timing Control in Speech Production*. Indiana University Linguistics Club, Bloomington.

Fowler, C. A. (2005) Parsing coarticulated speech: effects of coarticulation resistance. *Journal of Phonetics*, 33: 195–213.

Fowler, C. A., and Saltzman, E. (1993) Coordination and coarticulation in speech production. *Language and Speech*, 36: 171–95.

Gafos, A., and Benus, S. (2003) On neutral vowels in Hungarian. Paper presented at the 15th International Congress of Phonetic Sciences, Barcelona.

Garrett, M. (1980) Levels of processing in speech production. In B. Butterworth (ed.), *Language Production*, vol. 1: *Speech and Talk*, pp. 177–220. Academic Press, London.

Gay, T. (1978) Articulatory units: segments or syllables? In A. Bell and J. B. Hooper (eds), *Syllables and Segments*, pp. 121–31. North-Holland, Amsterdam.

Gelfer, C., Bell-Berti, F., and Harris, K. (1989) Determining the extent of coarticulation: effects of experimental design. *Journal of the Acoustical Society of America*, 86: 2443–5.

Gick, B. (1999) The articulatory basis of syllable structure: a study of English glides and liquids. Ph.D. dissertation, Yale University.

Goldstein, L., Pouplier, M., Chen, L. Saltzman, E., and Byrd, D. (forthcoming) Action units slip in speech production errors. *Cognition*.

Guenther, F., Hampson, M., and Johnson, D. (1998) A theoretical investigation of reference frames for the planning of speech. *Psychological Review*, 105: 611–633.

Hockett, C. (1955) *A Manual of Phonetics*. Indiana University Press, Bloomington.

Kelso, J. A. S. (1984) Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology*, 246: 1000–1004.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., and Fowler, C. A. (1984) Functionally-specific articulatory cooperation following jaw perturbation during speech: evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, 10: 812–32.

Kenstowicz, M., and Kisseberth, C. (1979) *Generative Phonology*. Academic Press, New York.

Krakow, R. (1989) The articulatory organization of syllables: a kinematic analysis of labial and velar gestures. Ph.D. dissertation, Yale University.

Krakow, R. (1993) Nonsegmental influences on velum movement patterns: syllables, segments, stress and speaking rate. In M. Huffman, and R. Krakow (eds), *Phonetics and Phonology*, vol. 5: *Nasals, Nasalization and the Velum*, pp. 87–116. Academic Press, New York.

Levelt, W., Roelofs, A., and Meyer, A. (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22: 1–38.

Lindblom, B., Lubker, J., and Gay, T. (1979) Formant frequencies of some fixed mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, 7: 147–61.

Lindblom, B., Sussman, H., Modaressi, G., and Burlingame, E. (2002) The trough effect in speech production: implications for speech motor programming. *Phonetica*, 59: 245–62.

Maeda, S. (1990) Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W. Hardcastle and A. Marchal (eds), *Speech Production and Speech Modeling*, pp. 131–49. Kluwer Academic, Boston, MA.

Meyer, A. (1991) The time course of phonological encoding in language production: phonological encoding inside a syllable. *Journal of Memory and Language*, 30: 69–89.

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

Ohala, J. (1981) The listener as a source of sound change. In C. Masek, R. Hendrick, R. Miller, and M. Mille (eds), *Papers from the Parasession on Language and Behavior*, pp. 178–03. Chicago Linguistics Society, Chicago.

Ohman, S. (1966) Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America*, 39: 151–68.

Perkell, J. and Chiang, C. (1986) Preliminary support for a 'hybrid model' of anticipatory coarticulation. In *Proceedings of the 12th International Congress of Acoustic*, pp. A3–A6.

Perkell, J. and Matthies, M. (1992) Temporal measures of labial coarticulation for the vowel /u/. *Journal of the Acoustical Society of America*, 91: 2911–25.

Perkell, J., Matthies, M., Svirsky, M., and Jordan, M. (1993) Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: a pilot 'motor equivalence' study. *Journal of the Acoustical Society of America*, 93: 2948–61.

Pierrehumbert, J. (1990) Phonological and phonetic representations. *Journal of Phonetics*, 18: 375–94.

Pouplier, M. (2003a) The dynamics of error. Paper presented at the 15th International Congress of Phonetic Sciences, Barcelona.

Pouplier, M. (2003b) Units of phonological encoding: empirical evidence. Ph.D. dissertation, Yale University.

Recasens, D. (1984a) Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America*, 76: 1624–35.

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

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.

Ryle, G. (1949) *The Concept of Mind*. Barnes & Noble, New York.

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

Saltzman, E., Lofqvist, A., and Mitra, S. (2000) 'Clocks' and 'glue': global timing and intergestural cohesion. In M. B. Broe and J. Pierrehumbert (eds), *Papers in Laboratory Phonology*, vol. 5: *Acquisition and the Lexicon*, pp. 88–101. Cambridge University Press, Cambridge.

Saltzman, E., and Munhall, K. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1: 333–82.

Savariaux, C., Perrier, P., and Orliaguet, J. P. (1995) Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: a study of the control space in speech production. *Journal of the Acoustical Society of America*, 98: 2428–42.

Sevald, C. A., Dell, G., and Cole, J. (1995) Syllable structure in speech production: are syllables chunks or schemas? *Journal of Memory and Language*, 34: 807–20.

Shattuck-Hufnagel, S., and Turk, A. E. (1996) A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25: 193–247.

Sproat, R., and Fujimura, O. (1993) Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics*, 21: 291–311.

Stevens, K., and Blumstein, S. (1981) The search for invariant correlates of phonetic features. In P Eimas and J Miller (eds), *Perspectives on the Study of Speech*, pp. 1–38. Erlbaum, Hillsdale, NJ.

Studdert-Kennedy, M. (1998) The particulate origins of language generativity: from syllable to gesture. In J. Hurford, M. Studdert-Kennedy, and C. Knight (eds), *Approaches to the Evolution of Language*, pp. 202–21. Cambridge University Press, Cambridge.

Tremblay, S., Shiller, D., and Ostry, D. (2003) Somatosensory basis of speech production. *Nature*, 423: 866–9.

Turvey, M. T. (1977) Preliminaries to a theory of action with reference to vision. In R. Shaw and J. Bransford (eds), *Perceiving, Acting and Knowing: Toward an Ecological Psychology*, pp. 211–66. Erlbaum, Hillsdale, NJ.

Turvey, M. T. (1990) Coordination. *American Psychologist*, 45: 938–53.

Yamanishi, J. Kawato, M., and Suzuki, R. (1980) Two coupled oscillators as a model for the coordinated finger tapping by both hands. *Biological Cybernetics*, 37: 219–25.