2

Coarticulation models in recent speech production theories

EDDA FARNETANI and DANIEL RECASENS

Introduction

A crucial problem of speech production theories is the dichotomy between the representational and the physical aspects of speech. The former is described as a system of abstract, invariant, discrete units (the phonemes), the latter as a complex of variable, continuous, overlapping patterns of articulatory movements resulting in a variable acoustic signal, a mixture of continuous and discrete events (Fant 1968). There is general agreement in the literature that the variability and unsegmentability of the speech signal is due in great part to the universal phenomenon of coarticulation, i.e. the pervasive, systematic, reciprocal influences among contiguous and often non-contiguous speech segments. This explains why coarticulation has a central place in all recent phonetic theories.

The aim of coarticulation theories is to explain coarticulation, i.e. account for its origin, nature and function, while coarticulation models are expected to predict the details of the process bridging the invariant and discrete units of representation to articulation and acoustics. Coarticulation theories are also expected to explain how listeners overcome coarticulatory variability and recover the underlying message.

As for production, the differences among the various theories of coarticulation may concern the *nature* of the underlying units of speech, the *stage* at which coarticulatory variations emerge within the speech production process, *what* is modified by coarticulation and *why*. According to some theories coarticulatory variations occur at the level of the speech plan and thus modify the units of the plan itself, whilst other theories attribute contextual variations to the speech production system so that high level invariance is preserved. But there are divergences in the reasons why the act of speaking introduces coarticulatory variations: are these to be attributed to the inertia of speech organs

and/or to the principle of economy or to the way the speech units are temporally organized?

Two aspects of coarticulation, one mainly temporal and the other mainly spatial, appear to be crucial for testing the predictions of the models, hence the validity of the various theories:

(1) The *temporal domain of coarticulation,* i.e. how far and in which direction coarticulation can extend in time when the coarticulatory movements are free to expand, that is when the articulators engaged in the production of the key segment are not subject to competing demands for the production of adjacent segments;
(2) the *outcome of gestural conflict,* i.e. what happens when competing articulatory and coarticulatory demands are imposed on the same articulatory structures. Some models predict that coarticulation will be blocked in these cases; according to others, conflict is resolved by allowing coarticulatory changes to different degrees, depending on the different constraints imposed on the competing gestures.

Another fundamental question concerns interlanguage differences in coarticulation. It has been shown that differences across languages in inventory size and/or phoneme distribution, or in the articulatory details of speech sounds transcribed with the same phonemic symbols affect the degree and the temporal extent of coarticulation. The theoretical accounts for interlanguage differences are various: the issue is developed in detail elsewhere in this volume.

Moreover, it is well known that the spatio-temporal extent of coarticulation does not depend only on the articulatory characteristics of the key segment and of its immediate environment. Experimental research continues to uncover a number of factors that affect the degree and/or temporal extent of contextual variability in a significant way. The most relevant are:

(a) the linguistic suprasegmental structure such as stress (Benguerel and Cowan 1974; Harris 1971; Fowler 1981b, Tuller, Harris and Kelso 1982) and prosodic boundaries (Bladon and Al-Bamerni 1976; Hardcastle 1985; Abry and Lallouache 1991a, 1991b)
(b) speech rate (Lindblom 1963a, Hardcastle 1985)
(c) speaking style (Lindblom and Lindgren 1985; Krull 1989; Moon and Lindblom 1994, among others)

None of the models we are going to review deals with all the variables listed above but undoubtedly the validation of a theory or model will eventually depend also on its ability to account for these factors.

On the side of perception of coarticulated speech, there is experimental evidence that coarticulatory influences are perceived and that listeners can correctly identify the intended segment by normalizing the percept as a function of context, i.e. by factoring out the contextual influences. This is shown in a number of identification, discrimination and reaction-time experiments (e.g. Mann and Repp 1980; Martin and Bunnell 1981; Repp and Mann 1982; Whalen 1981, 1989; Fowler and Smith 1986; Ohala and Feder 1994). What exactly *normalization* means is an object of debate: do listeners use the acoustic information for recovering the invariant coarticulated gestures as propounded by the theory of direct perception (Fowler 1986; Fowler and Smith 1986) or is the sound itself and its acoustic structure the object of perception, as propounded by a number of other speech scientists (Diehl 1986; Ohala 1986)?

It is possible to draw a parallel between this issue and the so-called *motor equivalence* studies. It has been shown that speakers themselves can reduce coarticulatory variability by means of compensatory manoeuvres (Lindblom 1967; Edwards 1985; Farnetani and Faber 1992; Perkell *et al.* 1993, among others, and see Recasens, this volume, chapter 4). What is the function of compensatory movements? Do they attempt to preserve invariance at the level of articulatory targets, or rather at the acoustic level? This issue, which can be re-worded as 'what is the goal of the speaker?', besides being fundamental for production theories, is also intimately related to the theories of perception of coarticulated speech, since the goal of the speaker is presumably the object of perception.

The relation between coarticulation and perception is not the subject of the present chapter. However, from this brief summary of current debates on coarticulation, it becomes clear that, while the testing of theories by analysis and modelling of the spatio-temporal domain of coarticulation is fundamental, it is only one aspect of the issue: experiments on the behaviour of speakers and listeners such as those mentioned above, i.e. compensation and normalization, will also be crucial tests for coarticulation theories.

This review will describe, compare and comment on the most influential coarticulation theories and models, and will address in detail the issues of anticipatory coarticulation, and of gestural conflict and coarticulation resistance.

Current theories and models

Coarticulation within the theory of Adaptive Variability

Speech does not need to be invariant

At the core of the theory of adaptive variability developed by Lindblom (1983, 1989, 1990) are the concepts that the fundamental function of speech is successful communication and that the speech mechanism,

like other biological mechanisms, tends to economy of effort. As a consequence the acoustic form of speech, rather than being invariant, will always be the result of the interaction between the listener-oriented requirement of successful communication and the speaker-oriented requirement of speech economy. Adaptive variability means that speakers are able to adapt their production to the demands of the communicative situation, i.e. to perceptual demands. When the communicative situation requires a high degree of phonetic precision, speakers are able to over-articulate; when this is not needed, speakers tend to under-articulate and economize energy. In the latter situation, listeners can recover the intended message by using signal independent information (i.e. top–down information) which helps interpret the information conveyed by a more or less poor speech signal. The full range of phonetic possibilities is represented by Lindblom as a continuum from hyper- to hypo-speech (Lindblom 1990).

Within this framework, coarticulation plays a fundamental role: perceptually the hyper- to hypo-speech continuum is manifested as a gradual decrease in phonetic contrast, and articulatorily as a gradual increase in coarticulation. So, coarticulation instantiates one of the two principles that govern speech production, i.e. the principle of economy.

Vowel reduction and coarticulation: the duration-dependent undershoot model

The first version of the model (Lindblom 1963a) is based on acoustic research on vowel reduction in Swedish. In this study Lindblom showed, first that the process of vowel reduction is not categorical but continuous, and second that it is not a process towards vowel centralization,[1] but rather the effect of consonant–vowel coarticulation. He found that in CVC syllables the formant frequencies during the vowel vary as a function of vowel duration and of the consonantal context: as duration decreases, the formants tend to undershoot the acoustic target value (an ideal invariant acoustic configuration represented by the asymptotic values towards which the formant frequencies aim) and to be displaced towards the values of the consonant context; undershoot tends to be larger in the contexts exhibiting sizeable consonant–vowel transitions (i.e. large locus–nucleus distances[2]). The direction of the formant movement and its dependence on the consonant–vowel distance clearly indicated that vowel reduction is the result of consonant-to-vowel coarticulation.

Lindblom's account of the relation between target undershoot and duration was that undershoot is the automatic response of the motor system to an increase in rate of motor commands. The commands are invariable but when

they are issued at short temporal intervals, the articulators do not have sufficient time to complete the response before the next signal arrives and thus have to respond to different commands simultaneously. This induces both vowel shortening and reduced displacement of formants.

Subsequent research showed that the response to high rate commands does not automatically result in reduced movements (Kuehn and Moll 1976; Gay 1978b) and that reduction can occur also at slow rates (Nord 1986). This indicated that duration is not the only determinant of reduction. Extensive research on context-dependent undershoot in different speech styles (Lindblom and Lindgren 1985; Krull 1989; Lindblom *et al.* 1992) showed that the degree of undershoot varies across styles, indicating that speakers can functionally adapt their production to communicative and sociolinguistic demands. This is the core of Lindblom's theory of Adaptive Variability described above.

Coarticulation and speech style

In the recent, revised model of vowel undershoot (Moon and Lindblom 1994), vowel duration is still the main factor but variables associated with speech style can substantially modify the amount of formant undershoot. The model is based on an acoustic study of American English stressed vowels produced in clear speech style and in citation forms (i.e. over-articulated versus normal speech). The results on vowel durations and F2 frequencies indicate that in clear speech the vowels tend to be longer and less reduced than in citation forms and, in the cases where the durations overlap in the two styles, clear speech exhibits a smaller amount of undershoot. A second finding is that clear speech is in most cases characterized by higher formant velocities than citation forms. This means that, for a given duration, the amount of context-dependent undershoot depends on the velocity of the articulatory movements, i.e. it decreases as velocity increases. In the revised model (where the speech motor mechanism is seen as a second-order mechanical system) the amount of undershoot is predicted by three variables reflecting the strategies available to speakers under different circumstances: duration, input force, and stiffness (the time constant of the system). Thus, in speech, an increase in articulatory force and/or an increase in speed of the system response to commands contribute to increase the movement velocity, and hence to decrease the amount of context-dependent undershoot. The relations among these variables are illustrated in figure 2.1 (from Moon and Lindblom 1994). Duration on the abscissa represents the duration of the input force; displacement on the ordinate represents the excursion of the articulator movement, high displacement means a small amount of target undershoot. It can

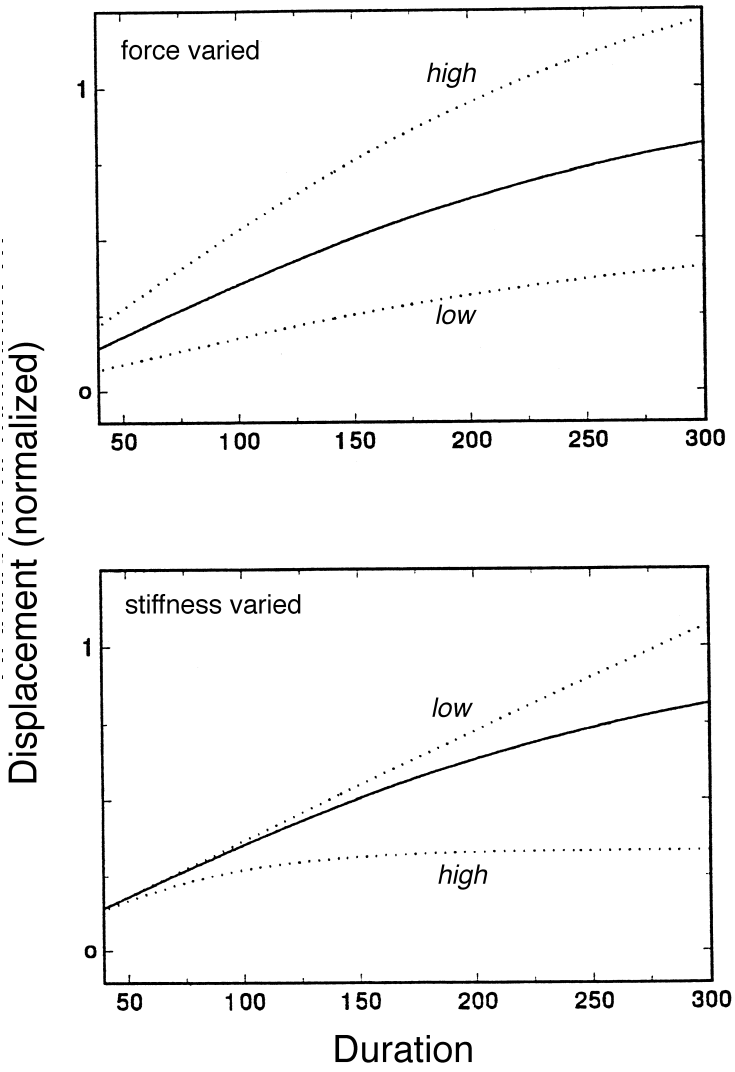

Figure 2.1 The revised model of vowel undershoot of Moon and Lindblom (1994). Movement displacement (ordinate) is a function of input force duration (abscissa) and can increase or decrease as a function of input force (top panel) and stiffness of the system (bottom panel). See text for expansion.

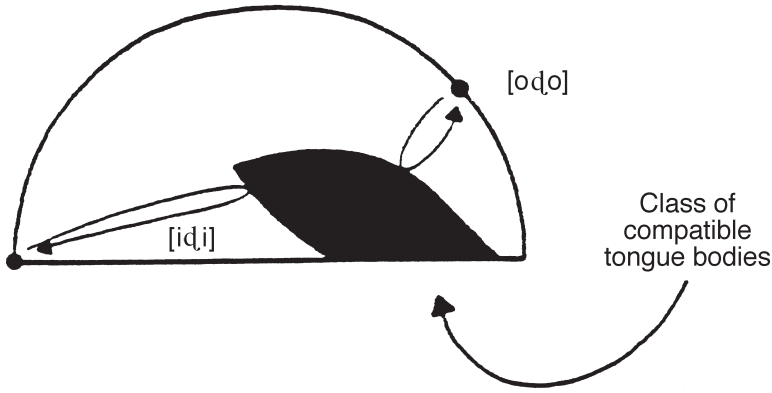

Figure 2.2 The model of tongue body coarticulation in the production of retroflex alveolars in the contexts of vowels /i/ and /o/ (filled circles). The black area contains all the tongue body contours compatible with the apical closure. The V-to-C and C-to-V trajectories reflect the minimum displacement constraints. The amount of coarticulatory effects are represented by the differences between tongue body configurations in the two contexts (from Lindblom 1983).

be seen that displacement is positively related to duration but can increase or decrease as a function of the amount of input force (top panel) or system stiffness (bottom panel).

An explanatory model of vowel-to-consonant coarticulation

In the parametric model of Lindblom, Pauli and Sundberg (1975) it is shown that V-to-C coarticulation results from a low-cost production strategy. The model is based on the analysis of VCV symmetric sequences with dental or retroflex apical stops, and vowels /i/or /o/. The model explores the reciprocal interactions between tongue tip parameters (elevation and protrusion/retraction) and tongue body parameters (position and shape) when the vocalic context is changed. Evaluation of the model is done by defining a procedure for deriving the formant values from the articulatory parameters, and by comparing the output of the model with the acoustic characteristics of natural VCV sequences, until the best match is reached. The authors show that the best match between the output of the model and natural utterances is a tongue configuration always compatible with the apical closure but characterized by a minimal displacement from the adjacent vowels. In other words, among a number of tongue body configurations compatible with the achievement of the tongue tip target, the tongue body always tends to take those requiring the least movements (see figure 2.2).

The predictions of the model are based on two kinds of constraints: first, the constraints associated with the position of the apical occlusion reduce the degrees of freedom of the tongue body (tongue tip extensibility/retractability constraints); second, the configuration of the tongue body is determined by that of the adjacent vowel, which further reduces the degrees of freedom of the system (constraints minimizing the tongue body movements, i.e. the principle of economy). The latter constraints determine the presence of coarticulation, the former its amount. Either set of constraints may be more or less severe and this is the source of the quantitative variations in coarticulation that we may expect across consonants, subjects and communicative situation.

The 'locus equation' as a measure of coarticulation

Krull (1987) was the first to use Lindblom's 'locus equations' (Lindblom 1963b) to quantify coarticulatory effects in CV syllables. For a given consonant in a variety of vocalic contexts, the locus equation reflects the relation between the formant values at the onset of CV transitions (at the consonant loci) and those at the mid-points of the following vowels, i.e. the extent to which the locus of a given consonant varies as a function of the following vowel nucleus. When such values are plotted against each other, their relation is expressed by the slope of the regression line (i.e. by the regression coefficient), which is an index of vowel-dependent variations of consonantal loci.

Using locus equations Krull showed that labial-consonant loci undergo larger coarticulatory effects than dental-consonant loci and that coarticulation is larger in spontaneous speech than in words read in isolation (Krull 1987; 1989). Since then, locus equations have been widely used for quantifying coarticulation (Duez 1992; Sussman, Hoemeke and McCaffrey 1992; Crowther 1994; Molis *et al.* 1994). A parallel approach (quantification of coarticulation through regression analysis) has been taken in articulatory studies of lingual consonants and vowels in different styles (Farnetani 1991; Farnetani and Recasens 1993).

Öhman's vowel-to-vowel coarticulation model

Öhman's model (Öhman 1966, 1967) is based on acoustic and articulatory analysis of Swedish VCV utterances produced in isolation, and on acoustic analysis of comparable utterances in American English and Russian. The major and most influential finding of Öhman's study is that in sequences V_1CV_2 where C is a stop consonant, the values of V_1C and CV_2 second formant transitions depend not only on C and the adjacent V but also on the identity of the transconsonantal vowel. Such coarticulatory effects, observed

for Swedish and American English in all the VCV sequences with intervocalic stops, indicated that vowels must be produced continuously, that is, the production of VCV utterances does not involve three linearly ordered successive gestures but rather a vowel-to-vowel diphthongal gesture on which the consonantal gesture is superimposed. The effects of a vowel on transconsonantal transitions are referred to as vowel-to-vowel coarticulatory effects. As for Russian, the VCV utterances consisted of various vowels with intervocalic palatalized /b/, /d/, /g/ and their unpalatalized (i.e. velarized) variants. The second formant data indicated that for all the consonants analysed, the variations of V_1C transitions as a function of the identity of V_2 were much smaller than those observed in American English and Swedish, and probably not above chance.

In the physiological model proposed by Öhman to account for his data, the tongue is viewed as a system of articulators executing invariant neural instructions from three independent articulatory channels: the apical (for apical consonants), the dorsal (for palatal and velar consonants) and the tongue body channel (for vowels). The variable vocal-tract shapes characterizing intervocalic consonants in different vocalic contexts are the results of the simultaneous individual motions of the three articulators. The model accounts for the different coarticulatory patterns of alveolar and velar consonants shown in the X-ray tracings for a Swedish subject: for alveolars, only the unconstricted part of the tongue (the tongue body) is modified by the adjacent vowels, while the constriction place remains invariant. Instead, for velars, the constriction place itself is modified and during closure the tongue moves continuously from V_1 to V_2: this is because independent vocalic and consonantal commands activate simultaneously adjacent tongue regions and 'the dynamic response of the tongue to a compound instruction is a complex summation of the responses to each of the components of the instruction' (Öhman 1966: 166). This implies that independent vocalic and consonantal commands on adjacent tongue regions never conflict, although the gestural trajectory from V_1 to V_2 may deviate due to the consonant command. Finally, the absence of V-to-V coarticulation in Russian is explained as a consequence of the interaction between two simultaneous and conflicting vowel commands on the tongue body during the consonants (for example, for palatalized consonants, the [i]-like command for palatalization, and the command for the adjacent vowel): coarticulation is blocked because during the consonant, 'the consonant gesture somehow overrules the vowel gesture if the latter is antagonistic to the former' (Öhman 1966: 167).

The numerical model of coarticulation developed for VCV utterances with intervocalic stops (Öhman 1967) is an algorithm predicting the appropriate

vocal-tract shapes under various conditions from two kinds of information: the idealized target shape of the consonant, and a coarticulation function derived from the vocalic context.

Comments

Öhman's model has in common with Lindblom's the idea that coarticulatory variability does not originate from variations at the level of motor commands. In both models the instructions to the articulators are invariant but for Öhman the presence of coarticulation in VCV utterances results from the cooccurrence of consonantal and vocalic instructions, whilst for Lindblom it results from economy constraints tending to minimize the articulator displacements from a segment to the following one. Öhman's model, where the tongue body movements are organized from one vowel to the next, offers a more comprehensive account of transconsonantal effects (in both the anticipatory and the carryover direction) than Lindblom's, where the movements are organized from a target to the next, i.e. from V_1 to C and from C to V_2.

Coarticulation within featural phonology

Standard generative phonology

The position taken by standard generative phonology (Chomsky and Halle, *The Sound Pattern of English*, 1968, hereafter *SPE*) provides a clear-cut distinction between coarticulation and other context-dependent changes, such as assimilations. Coarticulation is defined as 'the transitions between a vowel and an adjacent consonant, the adjustments in the vocal tract shape made in anticipation of a subsequent motion etc.' (*SPE*: 295), while assimilations involve operations on phonological features (the minimal classificatory constituents of a phoneme), and are accounted for by phonological rules, mapping lexical representation onto phonetic representation. The speech events so represented are those controlled by the speaker and perceived by the listener, and are language-specific. Coarticulatory variations, instead, originate from the physical properties of speech, and are determined by universal rules.

The theory of feature spreading

The 'feature-spreading' theory, proposed by Daniloff and Hammarberg (1973) and Hammarberg (1976) is a clear departure from the view that coarticulatory variations originate from the speech production mechanism and are governed by universal rules. According to Hammarberg (1976) a purely physiological account of coarticulation would entail a sharp dichotomy between intent and execution, whereby mental processes would be

unaware of the capacities of the speech mechanism or the articulators would be unable to carry out the commands as specified. Such extreme dualism can be overcome if coarticulation is assigned to the phonological component and viewed as a spreading of features across segments before the commands are issued to the speech mechanism. This way the articulators will just have to execute high-level instructions.

This new account of coarticulation was supported by a number of experimental results which contradicted the idea that coarticulation is the product of inertia. The studies of Daniloff and Moll (1968) on labial coarticulation and of Moll and Daniloff (1971) on velar coarticulation revealed that the lowering of the velum in anticipation of a nasal consonant and the rounding of lips in anticipation of a rounded vowel can start two, three or even four segments before the influencing one. These patterns clearly indicated that anticipatory coarticulation cannot be the product of inertia, but rather a deliberate spread of features. Also spatial adjustments made in anticipation of subsequent segments are to be viewed as deliberate feature-spreading processes (Daniloff and Hammarberg 1973). One example is the English voiced alveolar stop, which becomes dental when followed by a dental fricative (as in the word *width).* According to the authors, such 'accommodations' have the purpose of smoothing out and minimizing transitional sounds which might otherwise occur between individual segments. These spatial adjustments, however, are not to be considered inescapable and universal phenomena: in languages like Russian, for instance, the degree of accommodation between sounds seems to be minimal, which results in the introduction of transitional vocoids between adjacent consonantal and vocalic segments. The idea of the universality of coarticulatory patterns had also been challenged by Ladefoged (1967), who observed differences between French and English in the coarticulation of velar stops with front and back vowels.

Daniloff and Hammarberg (1973) and Hammarberg (1976) propose that all anticipatory (or right-to-left) coarticulation processes be accounted for in the grammar of a language by phonological feature-spreading rules, in principle not different from other assimilatory rules. Instead, most of the carryover (or left-to-right) processes are viewed by the authors as a passive result of the inertia of the articulatory system.

The look-ahead model

Daniloff and Hammarberg (1973) borrowed Henke's articulatory model (Henke 1966) to account for the extent of anticipatory coarticulation. Another well known model considered by the authors was the 'articulatory syllable' model proposed by Kozhevnikov and Chistovich (1965). The model

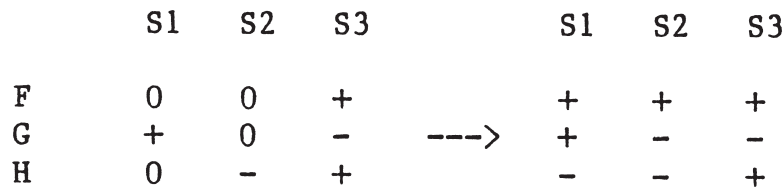

Figure 2.3 From phonological representation (left) to phonetic representation (right), through anticipatory feature-spreading rules as proposed by Daniloff and Hammarberg (1973). From Keating (1988b).

was based on durational relations among acoustic segments and on articulatory data on labial and lingual coarticulation in Russian, which suggested that the C_nV-type syllable is both a unit of rhythm and a unit of articulation. According to that model, the commands for a vowel are issued simultaneously with those for all the consonants preceding the vowel, hence a high degree of coarticulation is predicted within C_nV syllables, while little or no coarticulation is expected within other syllable types. Therefore, the Kozhevnikov and Chistovich model would fail to predict anticipatory coarticulation within VC syllables, in contradiction with the data of Moll and Daniloff (1971), showing that velar coarticulation extended from a nasal consonant across two preceding vowels.

Unlike the articulatory syllable model, Henke's model does not impose top–down boundaries to anticipatory coarticulation. Input segments are specified for articulatory targets in terms of binary features (+ or −). Unspecified features are given the 0 value. Phonological rules assign a feature of a segment to all the preceding segments unspecified for that feature. The feature-spreading mechanism is a look-ahead scanning device. Figure 2.3 (from Keating 1988b) shows the operation of the look-ahead mechanism over a sequence of three segments (S1, S2, S3) defined by three features (F, G, H).

It can be seen in the figure that no segment is left unspecified in the output phonetic representation (right-hand patterns) and that the anticipatory feature spreading is blocked only by specified segments (e.g. the spreading of feature G from S3 is blocked by S1, specified as + for that feature).

Comments

We can summarize the criticism of the 'feature-spreading' account of coarticulation by considering three important tenets of the theory:

(1) The assumption that the look-ahead mechanism assigns the coarticulated feature to all preceding unspecified segments and therefore

coarticulation extends in time as a function of the number or duration of these segments. The studies of Benguerel and Cowan (1974), Lubker (1981) and of Sussman and Westbury (1981) are consistent with the look-ahead hypothesis. Studies on anticipatory velar movements by Ushijima and Sawashima (1972) for Japanese and by Benguerel *et al.* (1977a) for French are only in partial agreement: their data indicate that the temporal extent of velar lowering in anticipation of a nasal segment is rather restricted and does not begin earlier in sequences of three than in sequences of two preceding segments. Other studies fully contradict the look-ahead model and show that the onset time of lip rounding and velar lowering is constant (Bell-Berti and Harris 1979, 1981, 1982) (see below for further discussion).

(2) The assumption that anticipatory coarticulation is blocked by segments specified for a feature that contradicts the spreading feature. Many data in the literature show that coarticulatory movements tend to start *during* rather than *after* the contradictorily specified segment, especially when few neutral segments intervene (e.g. Benguerel and Cowan 1974; Sussman and Westbury 1981, for lip rounding). Moreover, all the studies on lingual coarticulation in VCV sequences indicate that V_1 is influenced by V_2 even if it is specified for contrastive features with respect to V_2, e.g. /a/ is usually influenced by transconsonantal /i/ (see Butcher and Weiher 1976, for German; Farnetani, Vagges and Magno-Caldognetto 1985, for Italian; Magen 1989, for American English).

(3) The assumption that unspecified segments acquire the spreading feature. Data show that phonologically unspecified segments may not completely acquire the coarticulating feature, i.e. they can be non-neutral articulatorily and therefore they can be affected only to a certain degree by the influencing segment. For example, Bell-Berti and Harris (1981) showed that in American English the velum lowers during oral vowels produced in an oral context; Engstrand showed that in Swedish lip protrusion decreases during /s/ in /usu/ sequences (Engstrand 1981) and that the tongue body lowers during /p/ in /ipi/ sequences (Engstrand 1983). Such findings (and others reported below) indicate that supposedly unspecified segments may nonetheless be specified for articulatory positions.

Altogether these data indicate that a coarticulatory model based on the spreading of binary features fails to account for the graded nature of coarticulation and for its temporal changes during the segments.

The 'coarticulation resistance' model (Bladon and Al-Bamerni 1976) and the 'window' model (Keating 1985b, 1988a, 1988b, 1990a) overcome some of the inadequacies of the feature-spreading model in two different ways: the former proposes that specification in terms of binary features be accompanied by a graded coarticulatory resistance specification, the latter proposes that coarticulation be accounted for by two distinct components of the grammar, the phonological and the phonetic component.

The 'coarticulation resistance' model

The notion of coarticulatory resistance was introduced by Bladon and Al-Bamerni (1976) in an acoustic study of coarticulation in /l/ allophones in English (clear versus dark versus syllabic /l/). The study analyses the steady state frequencies of Fl and F2 in the target allophones and in a number of vowels. The results indicate that the V-to-C coarticulatory effects decrease gradually from clear to dark to syllabic /l/ and vary systematically also as a function of boundary type. These graded differences could not be accounted for by binary feature analysis which would block coarticulation in the dark (i.e. velarized) allophones, specified as [+back]. The authors propose that all the observed coarticulatory variations be accounted for in phonology by a rule assigning a coarticulatory resistance (CR) coefficient to the feature specification of each allophone and each boundary condition. According to the authors, CR coefficients are not necessarily universal, they can be language- and dialect-specific and thus account for interlanguage differences as well as for the differences in nasality between British and American English. According to the authors, the CR coefficients would be in line with the phonetic specifications in terms of numerical feature values proposed in *SPE*. A subsequent cineradiographic experiment (Bladon and Nolan 1977) showed that English apical consonants (e.g. /l/, /n/) become laminal in the context of laminal /s/, /z/ while laminal fricatives never become apical in an apical context. The data are accounted for by attaching to /s/ and /z/ feature specification a CR coefficient higher than that applied to apical consonants.

The window model

The 'window' model of coarticulation, elaborated by Keating (1985b, 1988a, 1988b, 1990a) accounts both for the continuous changes in space and time observed in speech, and for intersegment and interlanguage differences in coarticulation.

The phonological and the phonetic component of the grammar

Keating agrees that binary features and phonological rules cannot account for the graded nature of coarticulation but disagrees that such graded

variations should be ascribed to phonetic universals as automatic consequences of the speech production mechanism, since they may differ across languages (Keating 1985b). In particular she shows that contextual variations in vowel duration as a function of whether the following consonant is voiced or voiceless – an effect traditionally thought to be universal – are relevant and systematic in English and unsystematic or even absent in Polish and Czech. Therefore each language must specify these phonetic facts in its grammar. Her proposal is that all graded contextual variations, both those assumed to be phonological and those assumed to be universal and physically determined, be accounted for by the phonetic component of the grammar.

Keating's model is based on two assumptions which are a substantial departure from the feature-spreading model: (i) phonological underspecification may persist into the phonetic representation, (ii) phonetic underspecification is not a categorical but a continuous notion. The phonological representation is in terms of binary features. Unspecified features may be specified by rule or may remain unspecified. The phonological rules by which segments acquire a feature are various and may differ across languages: there are fill-in rules such as that specifying /ʃ/ as [+high] (for articulatory/aerodynamic reasons), context-sensitive rules such as that specifying Russian /x/ as [−back] before high vowels (see figure 2.4) and finally there may be assimilation rules like those proposed in the feature-spreading model (see below).

The windows

The output of phonological rules is interpreted in space and time by the phonetic implementation rules which provide a continuous (articulatory or acoustic) representation over time. For each articulatory or acoustic dimension, the feature value is associated with a range of values, called a *window.* Windows have their own duration and a width representing all the possible physical values that a target can take, i.e. the range of variability within a target. The window width depends first of all on the output of the phonological component: if features are specified, the associated window will be narrow and allow little contextual variation; if features are left unspecified, their corresponding windows will be wide and allow large contextual variation. The exact width of a window is derived for each language from information on the maximum amount of contextual variability observed in speech: all intermediate degrees between maximally narrow and maximally wide windows are then possible. By allowing windows to vary continuously in width, the model can represent the phonologically unspecified segments that offer some resistance to coarticulation, i.e. are associated with articulatory targets (see above for

experimental data). In Keating's model such segments are represented by wide, but not maximal windows, and can be defined phonetically 'not quite unspecified' (Keating 1988a: 22).

The contours

If the window represents the range of possible variability within a segment for a given physical dimension, the path or contour which connects the windows represents actual trajectories over time, i.e. the physical values over time in a specific context. Paths are interpolation functions between windows and are constrained by the requirements of smoothness and minimal articulatory effort (in agreement with the economy principle proposed by Lindblom). Even if trajectories connect one window to the next, cross-segmental coarticulatory effects are possible because windows associated with features left unspecified are wide and 'invisible' as targets; they therefore contribute nothing to the contour and allow direct interpolation between non-adjacent windows. Thus the model can account for V-to-V coarticulation and for anticipatory nasalization across any number of unspecified consonants. Specified segments, on the other hand, make their own contribution to the path, and do not allow interpolations across their windows. Figure 2.4 illustrates the phonological and the phonetic components of the model.

The example is a VCV utterance, where the two vowels are specified for [+high] and [+low]. The consonant, unspecified at the lexical level (top panel), can be specified by phonological fill-in rules (panel 2 from top), or by context-sensitive rules (panel 3 from top), or may remain unspecified (bottom). In the first case the specified feature generates narrow windows in all contexts and V-to-V interpolation is not possible. In the second case the window is narrow before the high vowel, but wide before the low vowel. In the third case the unspecified feature generates wide windows and V-to-V interpolation is possible in all contexts.

An important difference between Keating's model and other target-and-connection models proposed in the literature (e.g. Lindblom 1963a, MacNeilage 1970, among others) is that Keating's targets are not points in the phonetic space, subject to undershoot or overshoot but regions, themselves defined by the limits of all possible variations. We must remark that Lindblom, Pauli and Sundberg (1975) (see above and figure 2.2) had proposed a model similar to Keating's for tongue body coarticulation during apical stop production: the various tongue body contours compatible with the apical closure form in fact a region within which all coarticulatory variations take place.

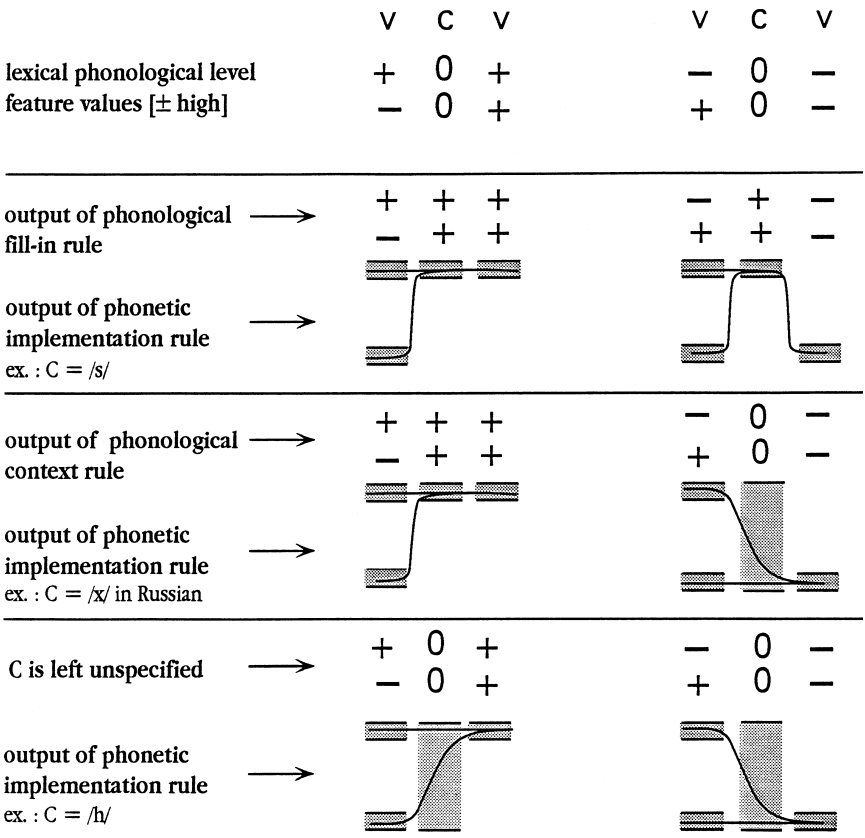

Figure 2.4 Keating's window model: from phonological representation (top panel) to phonetic parametric representation (articulatory or acoustic), through phonological specification rules plus phonetic implementation rules (panels 2 and 3 from top) or through implementation rules only (bottom panel).

Cross-language differences in coarticulation

According to Keating, interlanguage differences in coarticulation may be phonological or phonetic: the former occur when phonological assimilatory rules operate in one language and not in another, the latter occur when different languages give a different interpretation to a feature left unspecified. Speech analysis helps determine which processes are phonological and which phonetic. If phonological assimilation rules assign a contextual feature to a segment, its associated window will be narrow before that context and the contour will have a plateau-like shape (like that observed in figure 2.4, second case, a low vowel followed by C + V both specified as [+ high]).

In a language where assimilation rules do not operate, the key feature remains unspecified and the trajectories will be provided solely by interpolation between the preceding and the following context (as in the third example of figure 2.4).

In a study of nasalization, Cohn (1990) compared the nasal air-flow contour of English nasalized vowels with the contour of nasal vowels in French and of nasalized vowels in Sundanese. In French, vowel nasality is inherently phonological, while in Sundanese it is described as the output of a phonological spreading rule. Cohn found that in the nasalized vowels of Sundanese the flow patterns have plateau-like shapes very similar to the French patterns, while in English the shapes of the contours describe smooth trajectories from the [−nasal] to the [+nasal] segments. The categorical versus the gradient quality of nasalization in Sundanese versus English indicated that nasalization is indeed phonological in Sundanese and phonetic in English. Similar differences between patterns of anticipatory lip rounding were observed by Boyce (1990) between Turkish and English, suggesting that the process may be phonological in Turkish and phonetic in English. Within Keating's line of research for defining a process as phonological or phonetic, is the acoustic study of vowel allophony in Marshallese (Choi 1992), which shows, through cross-context comparisons, that Marshallese vowels are unspecified for the feature front/back, so that the allophonic variants (traditionally thought to be phonological) result from phonetic rules.

Languages may also differ in coarticulation because the phonetic rules can interpret phonological underspecification in different ways, allowing the windows to be more or less wide. In this case the interlanguage differences are only quantitative. 'Window width is to some extent an idiosyncratic aspect that languages specify about the phonetics of their sounds and features' (Keating 1988a: 22).

Comments

The novelty of Keating's theory, relating the phonological representation and the physical reality of speech through the mediation of the phonetic component of the grammar, has stimulated a number of studies which have contributed to a better understanding of coarticulatory processes. Within her line of research, the studies of Huffman (1990) and Cohn (1990) have contributed to improvements in the implementation theory and that of Choi (1992) has contributed to refining the interpolation functions in order to account for the asymmetries observed between anticipatory and carryover coarticulation.

A serious criticism to Keating's theory comes from Browman and

Goldstein (1993), who argue that the description of speech in two separate domains, requiring two distinct types of representation related to each other only by implementation rules, renders the phonological and the phonetic levels quite distant from each other: 'it becomes very easy to view phonetic and phonological (physical and cognitive) structures as essentially independent of one another, with no interaction or mutual constraint' (Browman and Goldstein 1993: 53). In particular, Boyce, Krakow and Bell-Berti (1991b) point to the difficulty of reconciling phonological unspecified features with specific articulatory positions, which are viewed by Keating as language-specific interpretations of underspecification. For instance, data on tongue body coarticulation (e.g. Engstrand 1989, for Swedish; Farnetani 1991, for Italian) seem to speak against Keating's proposal. These studies show that there are relevant differences in V-to-C coarticulatory effects on tongue dorsum across various consonant types all unspecified for the features high and back: e.g. labial,[3] dental and alveolar consonants (excluding /s/ specified as [+high] in Keating's model) and that in both languages the degree of resistance to vowel effects tends to increase from /l/ to /d/ to /t/ (/p/ shows the lowest degree of resistance). These cross-consonant differences and cross-language similarities indicate that the various degrees of coarticulation resistance, rather than being derived from language-idiosyncratic interpretations of phonological underspecification, must result from consonant-specific production constraints going beyond language peculiarities.

Another problem concerns the effects of speech style or rate on the degree of coarticulation. In a recent neural network model of speech production proposed by Guenther (1994), targets are seen as regions laid out in orosensory coordinates, very much like Keating's windows and these regions are allowed to increase or decrease in size as a function of rate and accuracy in pronunciation. The present version of the model conceives windows as invariable in size. But if the windows associated with specified features are allowed to shrink in duration and stretch in width in order to account for informal or casual speech then the relation between phonological feature specification and window width at the level of acoustic or articulatory representation might weaken further.

Manuel's criticism of Keating's theory (Manuel 1987, 1990) addresses the topic of whether coarticulatory variations should be accounted for in the grammar of a language owing to the fact that they differ quantitatively across languages (i.e. are not universal). Manuel reports data on vowel-to-vowel coarticulation in different languages (Manuel and Krakow 1984; Manuel 1987), showing that coarticulatory variations in vowels tend to be large in languages

with a small inventory size and smaller in more crowded inventories. She proposes that cross-language quantitative differences in coarticulation be regulated by *output constraints,* which restrict coarticulatory variations in languages where a crowded vowel space might lead to acoustic overlap and perceptual confusion, i.e. they delimit the areas in the phonetic space within which coarticulatory variations can take place (see chapter 8 of this volume). Clumeck (1976) had proposed a similar principle to account for the restricted extent of anticipatory nasalization in Swedish, as compared to other languages with less crowded vowel inventories. According to Manuel, if these constraints are known for each phoneme and each language on the basis of inventory size and phoneme distribution then no language-particular rules are needed, since the amount of coarticulation should be to some extent predictable. Following Manuel's reasoning, the coarticulatory patterns of vowels in Marshallese (Choi 1992) could also be predicted from its inventory: this language has four vowel phonemes, all distributed along the high/low dimension, thus output constraints would allow minimal coarticulation along the vertical dimension and maximal coarticulation along the front/back dimension, and this is what occurs.[4] Manuel (1987) recognized, however, in line with Bladon and Al-Bamerni (1976), that language-specific coarticulation patterns may also be totally unrelated to phonemic constraints. Data on velar coarticulation (Ushijima and Sawashima 1972, for Japanese; Farnetani 1986b, for Italian) indicate that anticipatory nasalization has a limited extent in these two languages, even if they have a restricted number of oral vowels and no nasal vowels in their inventories. Similarly, the study of Solé and Ohala (1991) shows that in Spanish the extent of anticipatory velar lowering is more restricted than in American English, which has a much more crowded vowel inventory than Spanish.

Altogether, the studies just reviewed suggest that cross-language similarities and cross-language differences in coarticulation are related to different kinds of constraints: production constraints, which seem to operate both within and across languages (as shown above from the tongue body coarticulatory data in Italian and Swedish), constraints deriving from language-specific phonological structure (in agreement with Manuel) but also, in agreement with Keating, language-particular constraints, unrelated to production or phonology and therefore unpredictable. These constraints seem to have the role of preserving the exact phonetic quality expected and accepted in a given language or dialect (Farnetani 1990).

Coarticulation as coproduction and an outline of new models

The coproduction theory and articulatory phonology represent, to date, the most successful effort to bridge the gap between the cognitive and

the physical aspects of language. The coproduction theory has been elaborated through collaborative work of psychologists and linguists, starting from Fowler (1977, 1980, 1985), Fowler *et al.* (1980) and Bell-Berti and Harris (1981). In conjunction with the new theory, Kelso, Saltzman and Tuller (1986), Saltzman and Munhall (1989) and Saltzman (1991) have developed a computational model of linguistic structures, the task-dynamic model, whose aim is to account for kinematics of articulators in speech. Input to the model are the phonetic gestures, the dynamically defined units of articulatory phonology proposed by Browman and Goldstein (1986, 1989, 1990a, 1990b, 1992).

The dynamic nature of phonological units

Fowler (1977, 1980) takes a position against the speech production theories that posit phonological features as input to the speech programme and against the view that coarticulation is a phonological feature-spreading process. According to Fowler, the dichotomy between the level of representation and that of production lies in the different nature of the descriptive entities pertaining to the two levels: in all featural theories the units of representation are abstract, static and timeless and need a translation process which transforms them into substantially different entities, i.e. into articulatory movements. In this translation process, the speech plan supplies the spatial targets and a central clock specifies when the articulators have to move. An alternative proposal, which overcomes the dichotomy and gets rid of a time program separated from the plan, is to modify the phonological units of the plan: the plan must specify the act to be executed, hence the phonological units must be planned actions, i.e. dynamically specified phonetic gestures, with an intrinsic temporal dimension (Fowler 1980; Fowler *et al.* 1980; Browman and Goldstein 1986, 1989).

In order to account for coarticulation, the spreading-feature theory posits assimilatory rules by which features are exchanged or modified before being translated into targets. In the coproduction theory it is proposed that gestures are not modified when actualized in speech. Their intrinsic temporal structure allows gestures to overlap in time so gestures are not altered by the context but coproduced with the context.

According to Browman and Goldstein (1993), the new accounts of coarticulation which introduce phonetic implementation rules in addition to phonological spreading rules, e.g. Keating (1985b, 1990a), do not succeed in overcoming the dichotomy between the cognitive and the physical aspects of speech as they are still viewed as different domains, with distinct types of representation, rather than two levels of description of the same system.

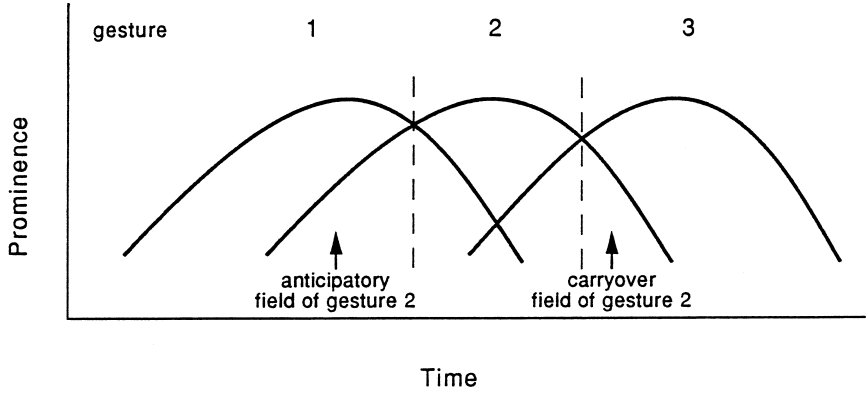

Figure 2.5 Representation of three overlapping phonetic gestures (from Fowler and Saltzman 1993). See text for details.

The gestures and their spatio-temporal organization

Figure 2.5, from Fowler and Saltzman (1993) illustrates how gestures are coproduced in speech. The activation of a gesture increases and decreases smoothly in time and so does its influence on the vocal tract shape and on the acoustic signal. In the figure, the vertical lines delimit a temporal interval (possibly corresponding to an acoustic segment) during which gesture 2 is prominent, i.e. has the maximal influence on the vocal tract shape, while the overlapping gestures 1 and 3 have a much weaker influence. Before this interval, the influence of gesture 1 predominates and the weaker influence of the following gesture gives rise to anticipatory coarticulation; analogously, after the interval of dominance of gesture 2, its influence determines the so-called carryover coarticulation. Thus both anticipatory and carryover effects are accounted for by the same principle of gesture overlap.

Gestures are implemented in speech by coordinative structures, i.e. by transient functional dependencies among the articulators contributing to the gestural goal: for instance, in the production of a bilabial stop, a functional link is established between upper lip, lower lip and jaw, so that a decrease in contribution of one articulator to the lip-closure gesture is automatically compensated for by an increase in the contribution of another. So the compensatory manoeuvres observed in speech stem from the coordination among independent articulators in achieving a gestural goal (Fowler 1977; Saltzman and Munhall 1989).

The overlap between gestures reflects their temporal coordination which is expressed in the model as intergestural phasing (i.e. the onset of a gesture occurs at a specified phase of the cycle of the preceding active gesture). The

phasing between gestures is controlled at the plan level: an increase in gestural overlap has the consequence of decreasing the measured segmental durations and of increasing the amount of coarticulatory effects. Quantitative changes in intergestural overlap can account for coarticulatory differences between slow and fast speech (Saltzman and Munhall 1989), for some of the effects of stress on the articulatory and acoustic characteristics of vowels (de Jong, Beckman and Edwards 1993) and for the effects of the number of syllables within a foot on vowel durations (Fowler 1977, 1981a).

The observed movement kinematics (displacement, velocity) is determined by the dynamic parameters that specify the gestures (i.e. force, stiffness). Also these parameters can undergo graded changes. In particular it is proposed that most of the so-called 'connected speech processes', e.g. vowel reduction, place assimilation and segment deletions are not due to categorical changes of the underlying gestures but to graded modifications, such as a decrease in gestural magnitude and an increase in temporal overlap (Browman and Goldstein 1990b, 1992; Farnetani 1997, for an overview of connected speech processes).

According to gestural phonology, cross-language differences in coarticulation (be they related or not to the inventory size and phoneme distribution) are consequences of the different gestural set-up in various languages, i.e. differences in the parameters that specify the dynamics of gestures and in those that specify their phasing relations. The language-specific gestural set-up is learned during speech development by tuning gestures and their organization to language-particular values (Browman and Goldstein 1989).

Spatial coarticulation: coarticulation resistance and intergestural blending

According to Fowler and Saltzman (1993) the amount of articulatory variability induced by coproduction depends on the degree to which the temporally overlapping gestures share articulators. A case of minimal gestural interference is the production of /VbV/ sequences, where the vocalic and the consonantal constriction gestures involve two independent articulators, the tongue body and the lips respectively, and a common articulator, the jaw. Here, coproduction of the vocalic and the consonantal gestural goals takes place with minimal spatial perturbations as the engagement of the shared articulator in the two competing goals is compensated for by the tongue and/or the lips (see above). Instead, coproduction induces relevant spatial perturbations if all articulators are shared by the overlapping gestures. This occurs, for example, in the production of /VgV/ utterances, where the vocalic and the consonantal gestures share both the tongue body and the jaw: the

main effect of this interference is the variation of the consonant place of constriction as a function of the adjacent vowel (see discussion of Öhman's model above). The general idea of the coproduction theory is that gestures *blend* their influence on the common articulator, with various possible outcomes (Browman and Goldstein 1989; Fowler and Saltzman 1993).

In a sequence of two identical gestures, the outcome of their temporal overlap is a composite movement, reflecting the sum of the influences of the two gestures. This is shown in a study of laryngeal abduction/adduction movements monitored by transillumination (Munhall and Löfqvist 1992). In sequences of two voiceless consonants, two distinct glottal abduction movements were observed at slow speaking rates, while single movements of various shapes were found at moderate and fast rates. The study shows that all the various movement patterns can be simulated by summing up two distinct underlying gestures overlapping to different degrees, as seen in figure 2.6.

The figure displays the hypothesized gestural score for two distinct abduction gestures of different amplitude arranged in order of decreasing overlap (left-hand series, top to bottom) and the corresponding movements resulting from summation of the underlying gestures (right-hand series). Notice that little or no overlap between the two gestures generates two distinct glottal movements (bottom panels); the outcome of a larger gestural overlap is a single glottal movement which decreases in duration and increases in amplitude as overlap increases (mid to top panels).

The gestural summation hypothesis can account for a number of observations in the coarticulation literature, for example, the patterns of velum lowering during a vowel in CVN sequences appear to be the result of the sum of the velar lowering for the vowel and the velar lowering for the following nasal (Bell-Berti and Krakow 1991). Gestural summation could also account for the differences in tongue tip/blade contact, observed with EPG (electropalatography), in singleton versus geminated lingual consonants (Farnetani 1990), and in singleton /n/ and /t/ or /d/ versus /nt/ and /nd/ clusters (Farnetani and Busà 1994). In both studies the geminates and clusters were produced with a single tongue movement whose amplitude (in terms of extent of the contact area) was significantly larger than in singleton consonants, independently of the vocalic context.

When gestures impose conflicting demands on the same articulatory structures, gestural conflict can be resolved at the plan level, by delaying the onset of the competing gesture so that the ongoing goal can be achieved. In this case the ongoing configuration will be minimally perturbed and the movement to the next goal will be quite fast (Bell-Berti and Harris 1981).

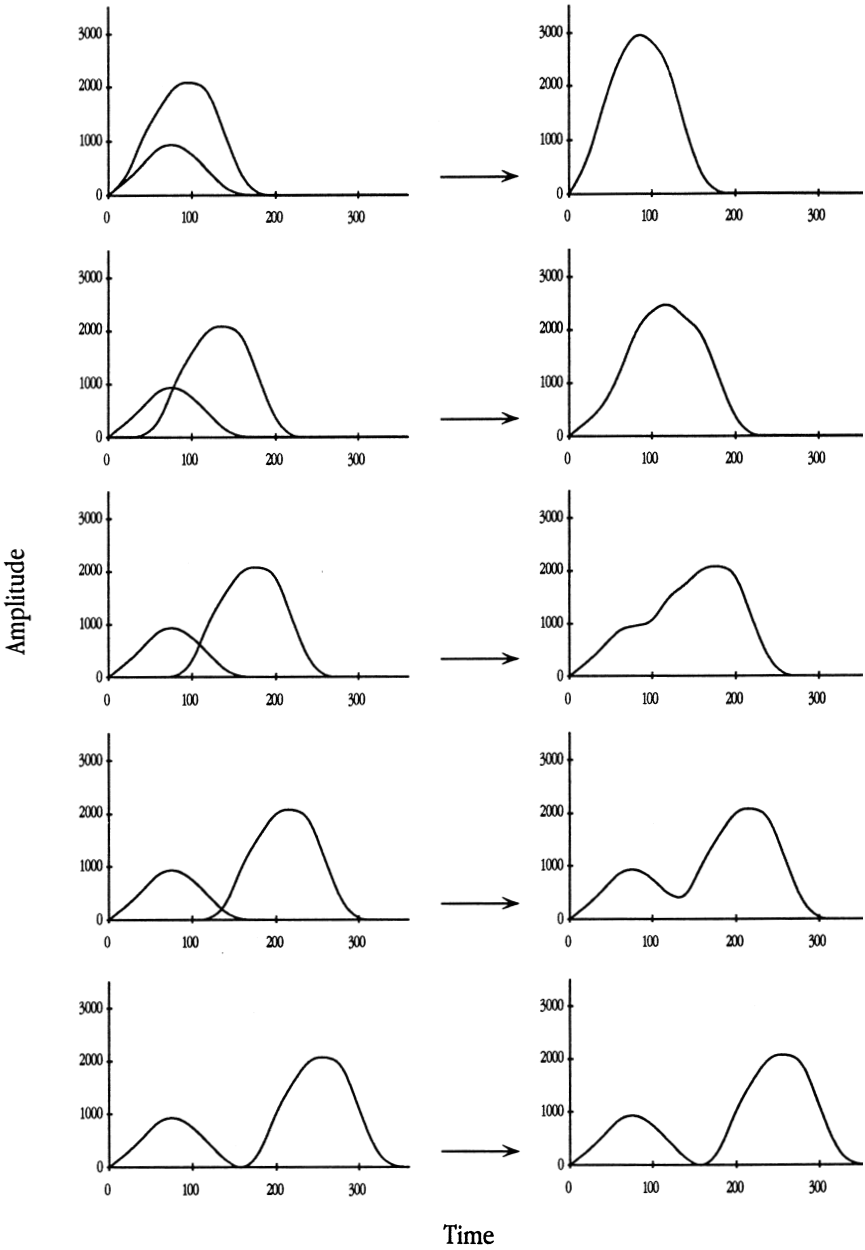

Figure 2.6 The additive blending model proposed by Munhall and Löfqvist (1992) for laryngeal gestures. The left-hand graphs of each pair show two laryngeal gestures gradually decreasing in overlap (top to bottom), the right-hand graphs show the output articulator movements resulting from gestural summation.

Alternatively, according to Saltzman and Munhall (1989) and Fowler and Saltzman (1993) no changes may be needed at the plan level and the articulatory consequences of gestural overlap 'would be shaped by the manner in which the coarticulating gesture blends its influence on the vocal tract with those of an ongoing one' (Fowler and Saltzman 1993:180). This means that the outcome of blending varies according to the gestures involved and depends in great part on the degree of resistance to coarticulation associated with the ongoing gesture.

As seen above, the notion of coarticulatory resistance was proposed by Bladon and Al-Bamerni (1976) to account for different degrees of spatial coarticulatory variations. Later studies (Bladon and Nolan 1977; Recasens 1984b, 1987; Farnetani and Recasens 1993) indicated that segments with a high degree of coarticulation resistance not only exhibit little coarticulatory variability but also induce strong coarticulatory effects on neighbouring segments. This appears to demonstrate that conflicting gestures *can* overlap: if they could not, no coarticulatory variations would be observed either in the highly resistant segment or in the weaker segment. Fowler and Saltzman (1993) propose the concept of *blending strength* to capture the relation between coarticulatory resistance and coarticulatory aggression: all the components of a given gesture have their own characteristic degree of blending strength, associated with the specific demands placed on the vocal tract so that when overlap occurs between a stronger and weaker gesture, the stronger tends to suppress the influence of the weaker and to exert greater influence on the common articulator; on the other hand, when overlap occurs between competing gestures of similar degree of blending strength, its outcome tends to be an averaging of the two influences.

Fowler and Saltzman's proposal implies that various outcomes are possible when two gestures are blended, from configurations mainly reflecting the influence of one gesture, to intermediate configurations reflecting both components of the coproduced gestures. The latter seem to occur notably between vowels, which are characterized by a moderate degree of blending strength.[5] Gestural blending is a general notion that applies to each of the articulators shared by two gestures. For example, in the production of alveolar consonants in VCV sequences, the tongue blade is controlled by the consonantal gesture alone but the jaw and the tongue body are controlled simultaneously by the consonantal and the overlapping vocalic gestures, which impose competing demands on them. Then, the observed coarticulatory V-to-C effects on the tongue body and the jaw reflect a blending of the influences of the overlapping V and C gestures, exactly like the blending of influences on the main articulator in the /VgV/ sequences mentioned above.

Comments

In order to account for the various degrees of resistance to coarticulation, the phonetic literature has used the notion 'constraints', which can be related to 'the demands placed on the vocal tract' proposed by Fowler and Saltzman (1993:180). According to Sussman and Westbury (1981), the term 'constraints' reflects resistance to coarticulation due to two possible factors: (a) aerodynamic requirements for the production of the ongoing segment (e.g. velar lowering for a nasal consonant does not occur during a preceding /s/ even if in itself it does not contradict the tongue movements for the formation of an /s/-like configuration); (b) antagonism proper, such as lip rounding versus lip spreading: according to the authors this kind of conflict does not preclude the onset of the lip-rounding neuromotor commands during /i/ because this can actively counteract the lip-spreading command. While, on the one hand, gestural phonology can account for the second type of constraints, by allowing antagonistic gestures to overlap, on the other hand, it is difficult to see how it can account for coarticulation resistance due to the aerodynamic requirement that an /s/-like lingual configuration generates the intended friction noise, unless the theory incorporates some sort of aerodynamic constraints.

In Recasens' review of lingual coarticulation (see chapter 4 of this volume) the different degrees of coarticulatory resistance observed in various studies on C-to-V and V-to-C effects are accounted for by *articulatory constraints* of different degree and different nature, functional and/or physical – the latter due to mechanical coupling between tongue regions (see also earlier for our account of the V-to-C coarticulatory patterns in Swedish and Italian). The data reviewed by Recasens show that production constraints are higher in the tongue regions directly involved in the formation of the constriction than in other regions and higher in laminodorsal consonants and in trills than in laminal fricatives than in apical stops, laterals, nasals and taps; for vowels, they seem to be higher in high/front vowels than in back and in low vowels. It is also shown that the degree of constraints associated with a gesture can account for the directionality and the temporal extent of coarticulation, i.e. coarticulatory anticipation depends heavily on the strength of the carryover effects of the preceding segment and vice versa. The high production constraints associated with trills (Recasens 1991b) are counter-evidence to the hypothesis that resistance to coarticulation varies inversely with gestural sonority. As for clusters, it appears that the gestural blending hypothesis accounts for some but not for all of the outcomes of gestural interference: the review shows that in Catalan, blending occurs in /ʃ/ + /s/ sequences and can be avoided if

articulator velocity is sufficient to produce the consonants at different places, as occurs in /rð/ clusters. This indicates that blending can be avoided by increasing articulator velocity (in agreement with Bell-Berti and Harris 1981). We should note that in Recasens' review, gestural blending is intended in the strict sense used by Browman and Goldstein (1989) to account for changes in consonant place of articulation.

Anticipatory extent of coarticulation

The coproduction model versus the look-ahead model

According to the coproduction theory, gestures have their own intrinsic duration. Hence the temporal extent of anticipatory coarticulation must be constant for a given gesture. Bell-Berti and Harris (1979, 1981, 1982) on the basis of experimental data on lip rounding and velar lowering, proposed the 'frame' model of anticipatory coarticulation (also referred to as the time-locked model): the onset of a movement of an articulator is independent of the preceding phone string length but begins at a fixed time before the acoustic onset of the segment with which it is associated. A number of other studies are instead consistent with the look-ahead model (see above). The cross-language study of Lubker and Gay (1982) on anticipatory labial coarticulation in English and Swedish shows that Swedish subjects use longer, more ample and less variable lip protrusion movements than English subjects; according to the authors, this is due to the perceptual importance of lip protrusion in Swedish, where lip rounding is contrastive; the relation between protrusion duration and the number of consonants indicates that three of the five Swedish speakers use a look-ahead strategy, and two a time-locked strategy. As noted earlier, two studies on velar movements (Ushijima and Sawashima 1972; Benguerel *et al.* 1977a) are in only partial agreement with the look-ahead hypothesis. These studies are based on timing measurements (like most of the studies on anticipatory coarticulation) and on observations of movement profiles as well. The study of Benguerel *et al.* (1977a) is of particular interest: the authors observed a clear velar lowering in oral open vowels in non-nasal context and, in sequences of oral vowels and nasal consonants, they could distinguish two patterns of velar movements: a lowering movement associated with the vowels (always beginning with the first vowel in the sequence), followed by a more rapid velar descent for the nasal, which caused the opening of the velar port and whose temporal extent was rather limited. Apparently this finding passed unnoticed: similar composite anticipatory movements were later observed in a number of other studies, starting with Al-Bamerni and Bladon (1982) and were given a totally different interpretation.

The hybrid model

Al-Bamerni and Bladon (1982), in a study of velar coarticulation in English, observed that speakers used two production strategies, a single velar opening gesture (one-stage pattern) and a two-stage gesture whose onset was aligned with the first non-nasal vowel and whose higher velocity stage was coordinated with the oral closing gesture of the nasal consonant. Perkell and Chiang (1986) were the first to observe two-stage patterns in lip rounding movements, which converged with Al-Bamerni and Bladon's observations: in /iC$_n$u/ utterances the onset of lip protrusion was temporally linked to the offset of /i/ (as predicted by the look-ahead model) but the onset of the second phase of the movement (detectable from the point of maximum acceleration) was linked to /u/ (as predicted by the time-locked model). The authors proposed a third model of anticipatory coarticulation, the hybrid model, a compromise between the look-ahead and the time-locked. Perkell's lip protrusion data on three English subjects (Perkell 1990) confirmed the preliminary observations of Perkell and Chiang, showing that two of the subjects used the two-stage hybrid strategy.

Boyce *et al.* (1990) give a different account of the hybrid movement profiles: they argue that the early onset of anticipatory movements cannot be attributed to a look-ahead anticipatory strategy unless it is shown that the phonologically unspecified segments preceding the key segment do not themselves have a target position for lips or velum. The data of Bell-Berti and Krakow (1991) on velar lowering show, through comparisons between vowels in nasal and oral context, that the velum lowers during vowels in oral context.[6] Accordingly, the early onset of velar lowering in the two-stage patterns is associated with a velar target position for the oral vowels, while the second stage, temporally quite stable, is associated with production of the nasal consonant. This interpretation is compatible with that given by Benguerel *et al.* (1977a) for French (see above). Therefore the two-stage patterns do not reflect the mixed anticipatory strategy proposed in the hybrid model, but two independent velar-lowering gestures, a vocalic gesture followed by a consonantal gesture; these may be more or less blended (summed) together and thus give rise to various patterns of movements, as a function of the duration of the vocalic sequence and of speech rate (see above). The velar coarticulation data for American English of Solé (1992) are in disagreement with Bell-Berti and Krakow's study. The American English speakers in Solé's study do not show any velar lowering during high and low vowels in oral context; the systematic onset of velar lowering at the beginning of the vowels preceding a nasal consonant cannot therefore be interpreted as a gesture associated with the vowels but rather as the result of a phonological nasalization rule, specific to American English.

The latest study on lip protrusion of Perkell and Matthies (1992) acknowledges that the early onset of anticipatory coarticulation is due, in part, to a consonant specific lip-protrusion gesture. Lip gestures associated with consonants were in fact observed in /iC$_n$i/ utterances and were mostly due to consonant /s/. However, in /iC$_n$u/ utterances the correlations between the rounding movement interval and the duration of consonants were found to be significant also in utterances without /s/. Moreover, also the second-stage movement, i.e. the interval from the point of maximum acceleration to the acoustic onset of /u/ was not fixed as would be predicted by the hybrid model but tended to vary as a function of consonant duration (although the low values of the correlation coefficient R^2 – from 0.09 to 0.35 – indicate that the factor of consonant duration accounts for a rather small proportion of the variance). The conclusion of this study is that the timing and the kinematics of the lip protrusion gesture are the simultaneous expression of competing constraints, that of using a preferred gestural kinematics independently of the number of consonants (as in the time-locked model) and that of beginning the protrusion gesture as soon as it is allowed by the constraint that the preceding /i/ is unrounded (as in the look-ahead model). The one or the other constraint can prevail in different subjects and even within subjects across repetitions.

The movement expansion model

A new model of anticipatory coarticulation has been recently proposed by Abry and Lallouache (1995) for labial coarticulation in French. Preliminary results on lip kinematics (Abry and Lallouache 1991a, 1991b) in /iC$_5$y/ sequences of two words, with word boundary after C$_3$, indicated that the shape of lip movement was quite stable in the session where the subject had realized the word boundary, while in the session where the phrase had been produced as a whole (and at faster rate) the lip profiles were extremely variable and exhibited both one-stage and two-stage patterns. A detailed analysis of the stable profiles indicated that there is a first phase characterized by no movement (against the predictions of the look-ahead model) followed by a protrusion movement, whose duration is proportional to that of the consonantal interval (against the predictions of the coproduction and the hybrid model).

The corpus analysed for one subject by Lallouache and Abry (1992) and for three subjects by Abry and Lallouache (1995) consists of /iC$_n$y/ utterances with a varying number of consonants, as well as /iy/ utterances, with no consonants in between. The parameters were all the kinematic points and intervals proposed by Perkell, as well as measures of the protrusion amplitude. The

overall results can be summarized as follows. First, the protrusion gesture is a one-phase movement; second, the maximum protrusion time tends to occur around the rounded vowel, i.e. the protrusion movements do not have a plateau-like shape; third, the protrusion interval as well as maximum acceleration and maximum velocity intervals correlate with the duration of consonants (the dispersion of data is much less than in Perkell and Matthies' study). Most importantly, the relation between these variables is better described by a hyperbolic function than by a linear function. This is because the protrusion movement tends to expand in longer consonant intervals but cannot be compressed in shorter intervals. In /iC_1y/ as well as in no-consonant /iy/ sequences the protrusion duration remains fixed at around 140–150 ms, indicating that the protrusion gesture heavily modifies the upper lip configuration of vowel /i/. Another finding is that there is much greater movement stability in shorter than in longer consonantal sequences. The temporal expansion of the movement does not mean that protrusion movement is linked to the unrounded vowel: it can start before, within or after it. The new model is speaker dependent and general at the same time: the amount of temporal expansion is in fact speaker specific, while the constraints on temporal compression are the same for every subject analysed (see chapter 6 for more details).

Summary and comments

The graphs in figure 2.7 illustrate the predictions of the look-ahead, the coproduction and the hybrid models, as applied to labial coarticulation. In the graphs V_1 is [− round], V_2 [+ round] and the number of unspecified consonants varies from four (left), to two (middle), to zero (right).

The first two rows show the patterns of anticipatory movements as predicted by featural phonology in terms of Keating's window model, where coarticulation may result either from language-specific phonological rules (manifested in the plateau-like patterns of the first row) or from phonetic interpolation (as shown, in the second row, by the trajectories directly linking the specified V_1 and V_2 across the unspecified segments). In the third row are the patterns predicted by the coproduction model: these patterns result from underlying competing vocalic gestures, whose spatio-temporal overlap increases as the temporal distance between the two vowels decreases. At the bottom is the pattern predicted by the hybrid model, where the duration of the protrusion movement, starting at the point of maximum acceleration, should remain invariant in the shorter sequences. The studies just reviewed (for both labial and velar coarticulation) indicate that, at large temporal distances between the two specified segments (left-hand panels), all the profiles predicted by the different models can occur in speech (see earlier for the

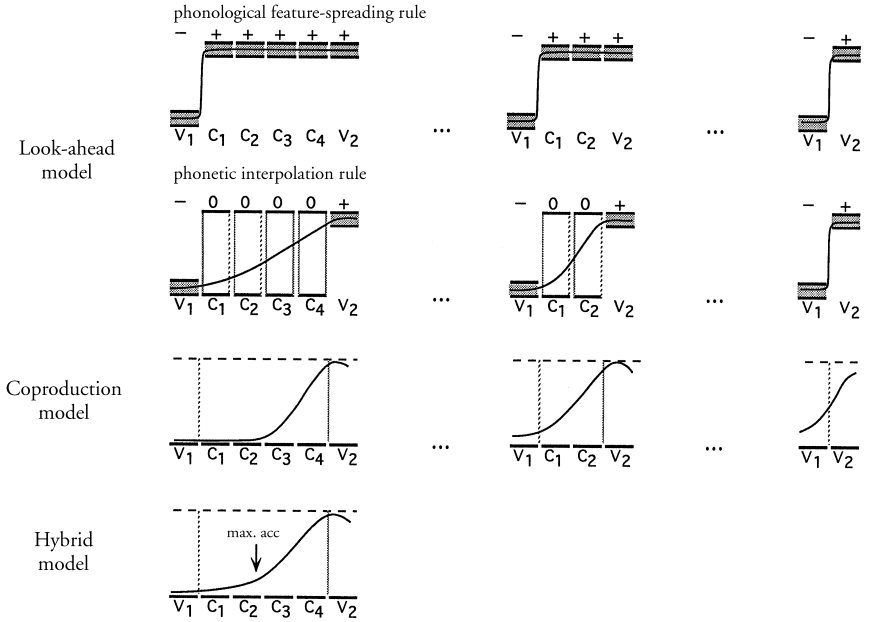

Figure 2.7 Anticipatory labial coarticulation: the predictions of the look-ahead model in terms of phonological feature-spreading rules and of phonetic rules of implementation and interpolation (first and second row respectively), the predictions of the coproduction model (third row) and of the hybrid model (bottom). From left to right: the [− *Round*] and the [+ *Round*] vowels are separated by four, two and zero consonants.

plateau patterns), even if very long interpolation patterns compatible with those of row two are not shown in the studies supporting the look-ahead model (as remarked above, the majority of these studies have analysed the timing rather than the trajectories of anticipatory movements). Moreover, the latest data on labial coarticulation (Abry and Lallouache 1995) show new movement patterns, incompatible with any of these three models. As things stand, none of the models in its strong version can account for the observed data, but models that allow expansion of the coarticulatory movements over long temporal intervals seem to offer a better account of the data than models preventing such expansion.

It can be seen that at shorter temporal intervals (e.g. in the two-consonant patterns shown in the central panels) the movement profiles are much more similar across the models. As the temporal interval decreases further (one consonant or no consonants between the vowels), the predictions are again quite different (see the right-hand panels). Feature-based models predict rapid

transition movements between adjacent segments specified for contrastive features (rows one and two), while the coproduction model (row three) predicts an increase in gestural blending, resulting in an increase in anticipation of the protrusion movement within the unrounded vowel and a decrease in the movement displacement either in V_1 or V_2 or both. In other words, the observed movement from V_1 to V_2 is not temporally compressed at the shortest inter-vowel distances. Unfortunately, very few data on vowel sequences are available in the coarticulation literature: in their study of lip rounding in French, Benguerel and Cowan (1974) observed an increase in movement velocity between adjacent unrounded and rounded vowels as compared to VCV sequences and this seems to be consistent with the look-ahead hypothesis. The more systematic data of Abry and Lallouache, on the other hand, indicate that the protrusion movement does not differ, either in duration or velocity in V_1CV_2 sequences as compared to V_1V_2 sequences, and that its onset occurs within the unrounded vowel or even before it, in the shortest V_1V_2 sequences. From these data it appears that at short V_1V_2 distances the coproduction model accounts for the data better than other models. The data on velar coarticulation of Ushijima and Sawashima (1972) may also be compatible with this hypothesis.

To sum up, although some of the discrepancies among the results of the various studies can be attributed to different experimental methodologies and/or to individual differences in the speech rate or style, the overall data indicate that each of the models succeeds only partially in accounting for speakers' behaviour.

Conclusion and perspectives

The last thirty years of theoretical and experimental research have enormously improved our knowledge of the articulatory organization and the acoustic structure of speech, and have been a rich source of novel and challenging ideas. In the introduction we raised some fundamental issues which the different theories had to confront: what is coarticulation? why is it there? and how does it operate? The review shows that the core of the different answers, the real focus of debate, is the nature of the underlying cognitive units of speech, i.e. features versus gestures (and this, as mentioned in the introduction, is also the centre of the debate on the goal of the speaker).

Gestural phonology, whose dynamic units are allowed to *overlap* in time, gives an immediate and transparent account of coarticulatory variations as well as of variations due to stress, position, rate and speaking style. Phonetic gestures are lawfully reflected in the acoustic structure of speech, from which they are recovered by listeners.

According to featural phonology, the underlying units (be they specified by articulatory or by acoustic features) are timeless and cannot overlap. So coarticulation has a different origin and, according to Keating, it stems from phonetic *underspecification* and consequently cannot occur between specified segments, unless specification itself is allowed to be modulated.

Lindblom's theory is based on acoustic features, modelled as acoustic targets: coarticulation stems from *minimum displacement constraints* of the articulatory mechanism, i.e. from the principle of economy, which is at the basis of coarticulatory variations both in the movements connecting successive targets (as occurs for tongue body movements during the production of apical consonants) and in the targets themselves (i.e. duration-dependent target undershoot) when an articulator has to respond simultaneously to different, overlapping commands, as occurs in rapid, casual speech. For Lindblom the goal of the speaker is acoustic/perceptual and itself regulates production. Lindblom's hyper–hypo theory gives an elegant account of the relation between coarticulation, timing and speech style. Although the points of departure are completely different, the predictions of Lindblom's model on the relation between duration and target undershoot are compatible with those of gestural phonology on the relation between the amount of blending and the temporal distance separating the competing gestures.

We believe that analysis of the two aspects of coarticulation brought up in the introduction (i.e. anticipatory extent of coarticulation and gestural conflict) is still a valid test for inferring whether the underlying units are features or gestures. However, the models against which the data are to be confronted require improvements and changes. On the other hand, we cannot dismiss the provoking conclusion of Perkell and Matthies (1992) that speakers are subject to two coexisting and competing constraints: that of preserving their preferred gestural routines and that of breaking them up, by allowing movement onsets to be triggered by the relaxation of constraints associated with the ongoing segments. We can then ask if there can be a rationale behind these two competing production modes, i.e. when the one or the other emerges or tends to prevail (although Perkell and Matthies' data seem to be a random mixture of the two modes): the problem may be open to future research.

These years of experimental research have also led to the overturning of some of the tenets of traditional phonetics, e.g. of the binomial that equated graded changes with universal aspects of speech, and categorical changes with language-specific aspects. Today there is general agreement that graded spatiotemporal coarticulatory variations can also be language-specific. A related issue concerns connected speech processes, traditionally thought to be cate-

gorical (i.e. phonological): Lindblom (1963a) was the first to demonstrate the continuous nature of vowel reduction. Browman and Goldstein (1989) go further and propose that the majority of such processes are continuous and, like coarticulatory processes, do not imply changes in the categorical underlying units. This challenging hypothesis has stimulated a great deal of recent research on connected speech processes as well as on rate and boundary effects (see Farnetani 1997; and Künhert and Nolan, in this volume) and we can envisage that articulatory, acoustic and perceptual experiments on continuous speech and different speech styles will be among the main themes of research over the next years.

Acknowledgments

Our thanks to Michael Studdert-Kennedy and Bill Hardcastle for insightful comments and suggestions. This work was supported in part by ESPRIT/ACCOR, WG 7098.

Notes

1 Categorical vowel reduction refers to the phonological change of unstressed vowels to the neutral vowel schwa; continuous (or phonetic) vowel reduction has been traditionally regarded as vowel centralization, i.e. a tendency of vowel formants to shift towards a more central position within the acoustic vowel space. For ample reviews of the literature on vowel reduction see Fourakis (1991), Moon and Lindblom (1994).
2 The locus of a voiced stop followed by a vowel is defined by Lindblom as being at the first glottal pulse of the vowel, i.e. onset of the CV acoustic transition.
3 Only the Italian set of consonants included the labial /p/.
4 Multiple regression analysis indicated that the contribution of vowel category (high, mid and low) to formant variation was maximal for Fl and minimal for F2, whilst the effects of consonantal context (secondary articulations) were much higher on F2 than on Fl (Choi 1992: 37–39).
5 According to the authors, blending strength varies inversely with gestural sonority, a proposal compatible with Lindblom's concept of coarticulatory adaptability, which is maximal in vowels and minimal in lingual fricatives, in line with phonotactically based sonority categories (Lindblom 1983).
6 Velar lowering in American English low vowels in oral context had been observed in a number of other studies, e.g. in Clumeck (1976).