

difficult to provide a systematic account of articulatory kinematics. Some preliminary ideas about articulatory dynamics indicate that the biomechanical response properties of articulators and effort minimization strategies have roles in shaping the control of articulatory processes; on the other hand, it is very likely that perceptually-related factors are at least as important in this regard.

Many of these observations are consistent with a theoretical framework in which the goals of speech movements are defined in terms of articulatory and acoustic parameters, and the control strategies that achieve those goals are influenced by dynamical constraints and intelligibility requirements. However, these ideas depend as much on speculation as they do on established knowledge. Additional, comprehensive data on articulatory kinematics, more realistic modeling of articulatory dynamics and control, and a deeper understanding of the role of perception and mechanisms of speech understanding should shift this balance toward more of a reliance on facts.

NOTES

I am very grateful for helpful comments from: Suzanne Boyce, Peter Guiod, Frank Guenther, Eva Holmberg, Mike Jordan, Jim Kobler, Harlan Lane, Anders Löfqvist, Melanie Matthies, Clay Mitchell, Stefanie Shattuck-Hufnagel, Ken Stevens, Mario Svirsky, Alice Turk and Yi Xu. Preparation of this chapter was supported by N.I.H. Grant DC01925.

- 1 It is recognized that there are alternative frameworks (Browman and Goldstein, 1986; 1989 and Saltzman and Munhall, 1989; discussion by Löfqvist in this volume) and that too little is known currently about speech production to decide that any one theory is more valid than others.
- 2 Articulatory processes also produce visual cues for speechreading, which is beyond the scope of this chapter.
- 3 According to the somewhat different perspective taken by Articulatory Phonology (Browman and Goldstein, 1989) and task-dynamic modeling (Saltzman and Munhall, 1989) phonetic goals and articulatory processes are one and the same.
- 4 In order to be able to describe articulatory processes in a convenient way in this chapter, the terms "actions" and "articulations" are used interchangeably to refer to activities of the vocal-tract, laryngeal and respiratory systems. More often in the literature, such actions of vocal-tract and laryngeal structures are called "articulations".
- 5 Note that in contrast to "dynamics", the term "kinematics" refers to readily observable properties of movements such as displacements, distances, velocities and accelerations, without consideration of the underlying forces.
- 6 The convergence of multiple influences on individual articulators and subsystems also has to be taken into account in experimental design and analysis, particularly when studying the effect of some external perturbation or environmental change. For example, the speaker's response might well be a change in a postural setting that can be misinterpreted as a modification of a parameter value of a phonetic goal (Perkell et al., 1992b).

12 Coarticulation and Connected Speech Processes

EDDA FARNETANI

1 Context dependent variability in speech

1.1 Coarticulation

During speech the movements of different articulators for the production of successive phonetic segments overlap in time and interact with one another. As a consequence, the vocal tract configuration at any point in time is influenced by more than one segment. This is what the term "coarticulation" describes. The acoustic effects of coarticulation can be observed with spectrographic analysis: any acoustic interval, auditorily defined as a phonetic segment, will show the influence of neighboring phones in various forms and degrees. Coarticulation may or may not be audible in terms of modifications of the phonetic quality of a segment. This explains why descriptive and theoretical accounts of coarticulation in various languages became possible only after physiological and acoustical methods of speech analysis became available and widespread, that is, during the last thirty years. Recent reviews of theories and experimental data on coarticulation have been provided by Kent (1983), Harris (1983), Fowler (1980, 1985), and Farnetani (1990).

Table 12.1 shows how coarticulation can be described in terms of: 1) the main articulators involved; 2) some of the muscles considered to be primarily responsible for the articulatory-coarticulatory movements; 3) the movements that usually overlap in contiguous segments; 4) the major acoustic consequences of such overlap. As for lingual coarticulation, the tongue tip/blade and the tongue body can act quasi-independently as two distinct articulators, so that their activity in the production of adjacent segments can overlap in time.

Jaw movements are not included in the table since the jaw contributes both to lip and to tongue positioning, i.e. is part of two articulatory subsystems. Jaw movements are analyzed especially when the goal of the experiment is to establish the role of the jaw in shaping the vocal tract and thus distinguish between active and passive tongue (or lip) movements, or to investigate how

Table 12.1 Coarticulation

Articulator	Level of description		
	Myomotoric	Articulatory	Acoustic
LIPS	Orbicularis Oris/ Risorius	Lip rounding/ spreading	Changes in F1, F2 and F3
TONGUE	Genioglossus and other extrinsic and intrinsic lingual muscles	Tongue front/back, high/low displacement	Changes in F2, F1 and F3
VELUM	(Relaxation of) Levator Palatini	Velum lowering	Nasal Formants and changes in Oral Formants
LARYNX	Posterior Cricoarytenoid/ Interarytenoid, Lateral Cricoarytenoid	Vocal fold abduction/ adduction	Aperiodic/Periodic signal Acoustic duration

the jaw contributes to or compensates for coarticulatory variations (see Perkell, ARTICULATORY PROCESSES).

The present account will center on coarticulation at the supraglottal level. For laryngeal activity see Hirose, INVESTIGATING THE PHYSIOLOGY OF LARYNGEAL STRUCTURES and Ní Chasaide and Gobl, VOICE SOURCE VARIATION. For detailed accounts of the relationship between vocal tract activity and the acoustic signal, see Fant (1968) and Fujimura and Erickson, ACOUSTIC PHONETICS and Stevens, ARTICULATORY/ACOUSTIC/AUDITORY RELATIONSHIPS.

Typical examples of coarticulation in terms of muscle activity and of articulatory movements are illustrated in Figures 12.1, 2 and 3.

Figure 12.1 illustrates coarticulation at the myomotoric level. It shows electromyographic (EMG) activity of the muscles during the production of /əpɪb/ /əpɪp/ (From Hirose and Gay, 1972). It can be seen that the activity of the *orbicularis oris* for the production of the first /p/ is overlapped by the activity of the *genioglossus* for the production of the following front vowel. Moreover the activity of the laryngeal muscles responsible for abducting and adducting the vocal folds also overlap: the onset of lateral cricoarytenoid (LCA, adducting) occurs when the posterior cricoarytenoid activity (PCA, abducting) is at its peak, that is at the middle of the /p/ closure.

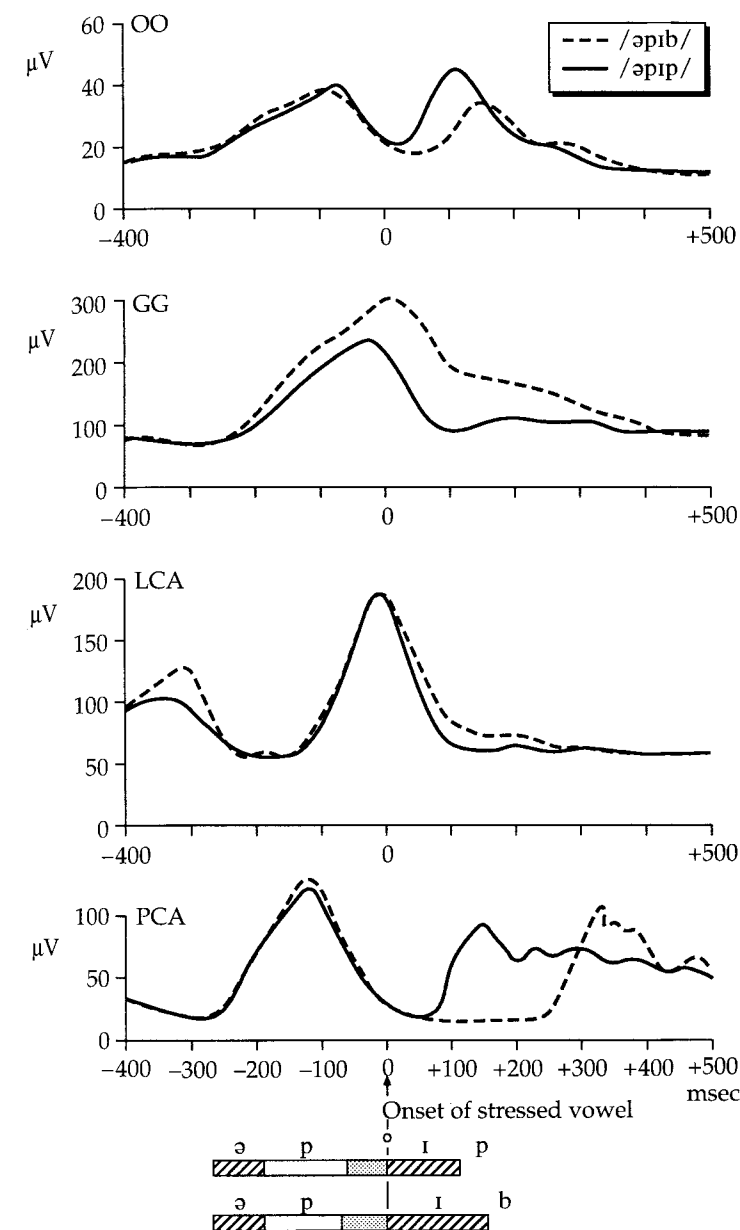


Figure 12.1 Averaged EMG curves of orbicularis oris, genioglossus, lateral and posterior cricoarytenoid muscles in utterances /əpɪp/ /əpɪb/ produced in isolation by an American subject. The line up point is the acoustic onset of /ɪ/ (from Hirose and Gay, 1972).

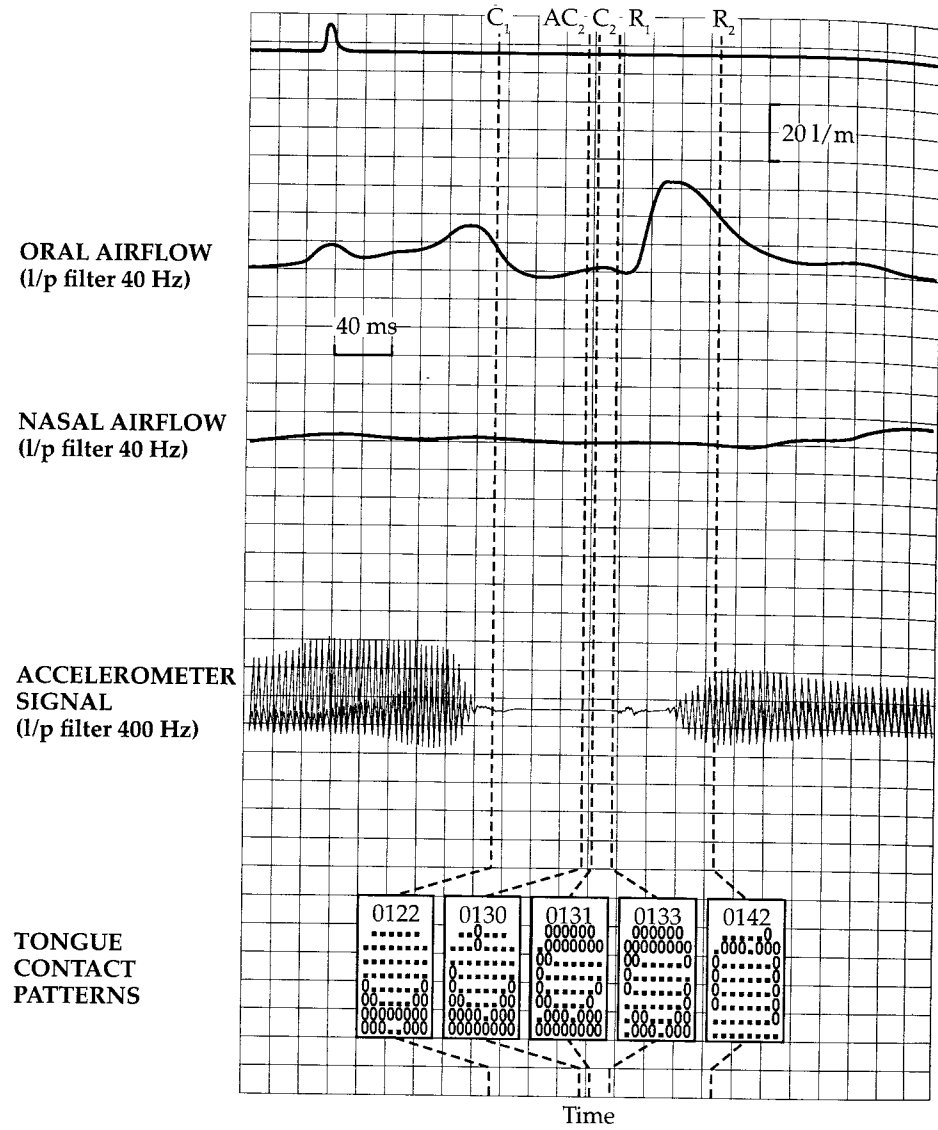


Figure 12.2 Oral and nasal flow curves, acoustic signal, and synchronized EPG activity in the production of the English word 'weakling' within phrase (from Hardcastle, 1985).

Figure 12.3 (Opposite) Acoustic signal, oral and nasal flow curves and synchronized EPG curves during /ana/ /ini/ produced in isolation by an Italian subject (from Farnetani, 1986).

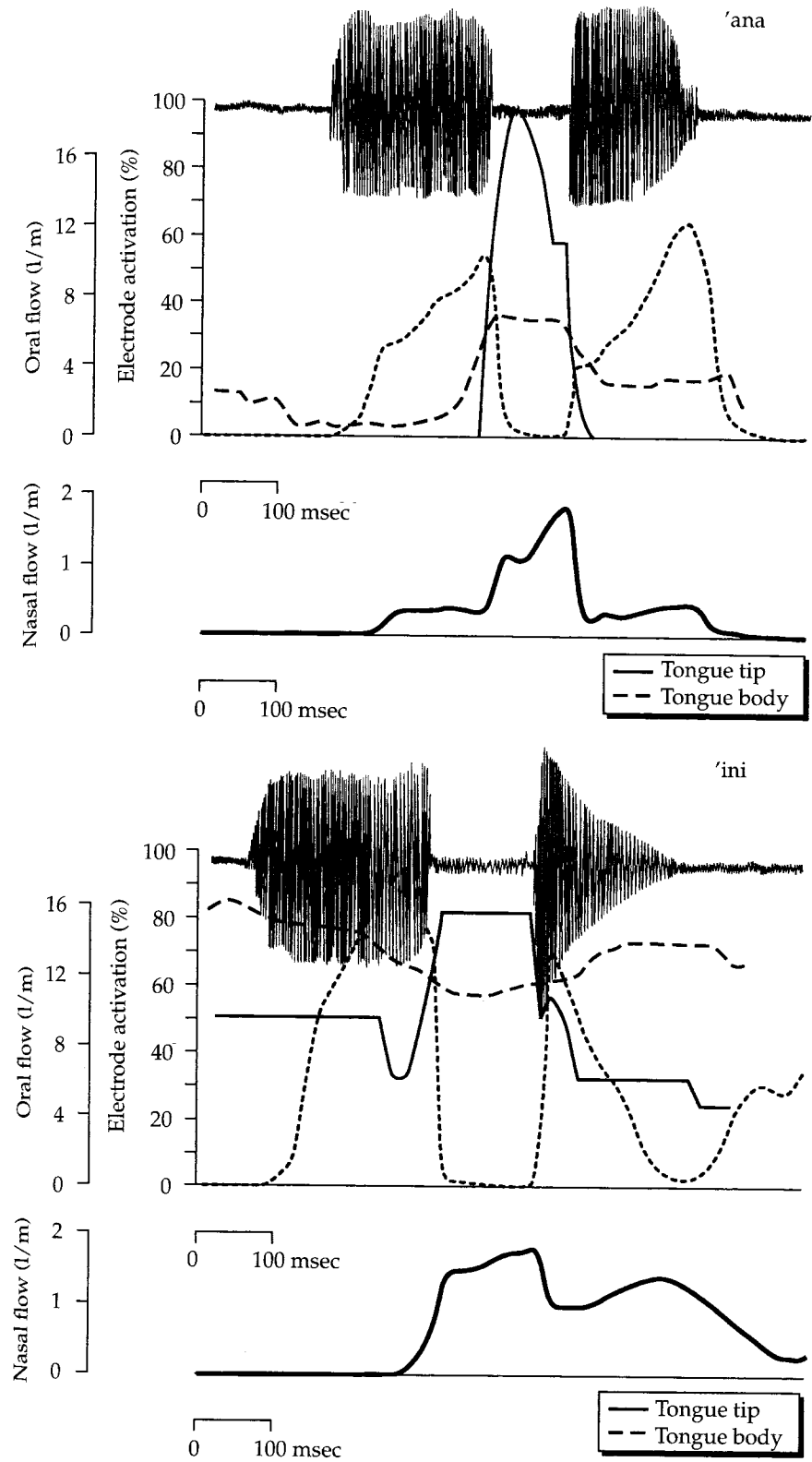


Figure 12.2 describes tongue-tip tongue-body coarticulation in the /kl/ cluster of the English word "weakling", analyzed with electropalatography (EPG) synchronized with oral and nasal airflow and the acoustic signal (from Hardcastle, 1985).

In the sequence of the EPG frames (sampled every 7.5 msec) it can be seen that the tongue body closure for the velar consonant is overlapped by the tongue tip/blade gesture for the following /l/, detectable by a light front contact as early as frame 130. The following frames show complete overlap of /k/ and /l/ closures for about 20 msec.

Figure 12.3 shows examples of velar and lingual coarticulation in sequences /'ana/ and /'ini/ in Italian, analyzed with EPG and oral/nasal flow measurements (Farnetani, 1986). From the patterns of nasal flow it can be inferred that in /'ana/ the opening of the velopharyngeal port for the production of /n/ occurs just after the acoustic onset of the initial /a/ and lasts until the end of the final /a/; in /'ini/ there is only a slight anticipation of velopharyngeal opening during the initial /i/, but after the consonant the port remains open until the end of the utterance. Thus in /'ana/ velar C-to-V coarticulation is extensive both in the anticipatory (before /n/) and in the carryover direction (after /n/), while in /'ini/ it extends mostly in the carryover direction. The two EPG curves represent the evolution of tongue-to-palate contact over time. It can be seen that during tongue tip/blade closure for /n/, tongue body contact (the percentage of electrodes contacted in the mid and the back regions of the palate) is much larger in the context of /i/ than in the context of /a/, indicating that the tongue body configuration is strongly affected by the movements associated with the vowels. These patterns describe V-to-C lingual coarticulation, that is, the effects of the vowel on the articulation of a lingual consonant.

The central theoretical issues in the studies of coarticulation concern its *origin, function, and control*. Coarticulation has been observed in all languages so far analyzed, so that it can be considered a universal phenomenon; at the same time it appears to differ among languages. Thus another important issue is how to account for *interlanguage differences in coarticulation*.

1.2 Assimilation

Assimilation refers to contextual variability of speech sounds, by which one or more of their phonetic properties are modified and become similar to those of the adjacent segments. We may ask whether assimilation and coarticulation refer to qualitatively different processes, or to similar processes described in different terms (the former reflecting an auditory approach to phonetic analysis, and the latter an instrumental articulatory/acoustic approach). The answers to this question are various and controversial. In order to illustrate the problem, we can start from the position taken by standard generative phonology (Chomsky and Halle, *The Sound Pattern of English*, 1968, hereafter SPE). SPE makes a clear-cut distinction between assimilation and coarticulation. Assimilation pertains to the domain of linguistic competence, is accounted for

by phonological rules, and refers to modifications of features (the minimal categorical-classificatory constituents of a phoneme). Hence assimilatory processes (although widespread among languages) are part of the grammar and are language-specific. Coarticulation by contrast results from the physical properties of the speech mechanism and is governed by universal rules; hence it pertains to the domain of performance and cannot be part of the grammar. Chomsky and Halle describe coarticulation as "the transition between a vowel and an adjacent consonant, the adjustments in vocal tract shape made in anticipation of a subsequent motion etc." (SPE, p. 295) Accordingly, both the *distribution* (universality vs language specificity) and the *quality* of the contextual change (mere articulatory adaptation vs intentional phonetic modification) should allow one to distinguish the two processes.

Quite often context-dependent changes involving the same articulatory structures have different acoustic and perceptual manifestations in different languages so that it is possible to distinguish what can be considered universal phonetic behavior from language particular rules. A classical example is the difference between vowel harmony, an assimilatory process present in a limited number of languages (such as Hungarian) and the process of vowel-to-vowel coarticulation, attested in many languages and probably present in all (Fowler, 1983). In other cases cross-language differences are not easily interpretable, and inferences about the nature of the underlying processes can be made only by manipulating some of the speech parameters, for example segmental durations. In a study of vowel nasalization in Spanish and American English, Solé and Ohala (1991) were able to distinguish phonological (language specific) from phonetic nasalization by manipulating speech rate. They found a quite different distribution of the temporal patterns of nasalization in the two languages as a function of rate: in American English the extent of nasalization on the vowel preceding the nasal was proportional to the varying vowel duration, while in Spanish it remained constant. They concluded that the spread of nasalization as vowel duration increases in American English, must be intentional (phonological), while the short and constant extent of nasalization in Spanish must be an automatic consequence of the speech mechanism, since it reflects the minimum time necessary for the lowering gesture of the velum. But the interpretation of contextual changes in terms of a strict dichotomy between universal and language-specific variations fails to account for the many cross-language data showing that coarticulation differs in degree across languages: a typical example is Clumeck's study on velar coarticulation, which was found to differ in temporal extent across all the six languages analyzed (Clumeck, 1976). In this case, what criterion can be used to decide in which language the patterns are unintentional and automatic, and in which they are to be ascribed to the grammar?

Likewise, within a language, context-dependent variations may exhibit different articulatory patterns which can be interpreted either as the result of different underlying processes, or just as quantitative variations resulting from the same underlying mechanism. For example, Figure 12.4 shows the variations of the articulation of the phoneme /n/ as a function of the following

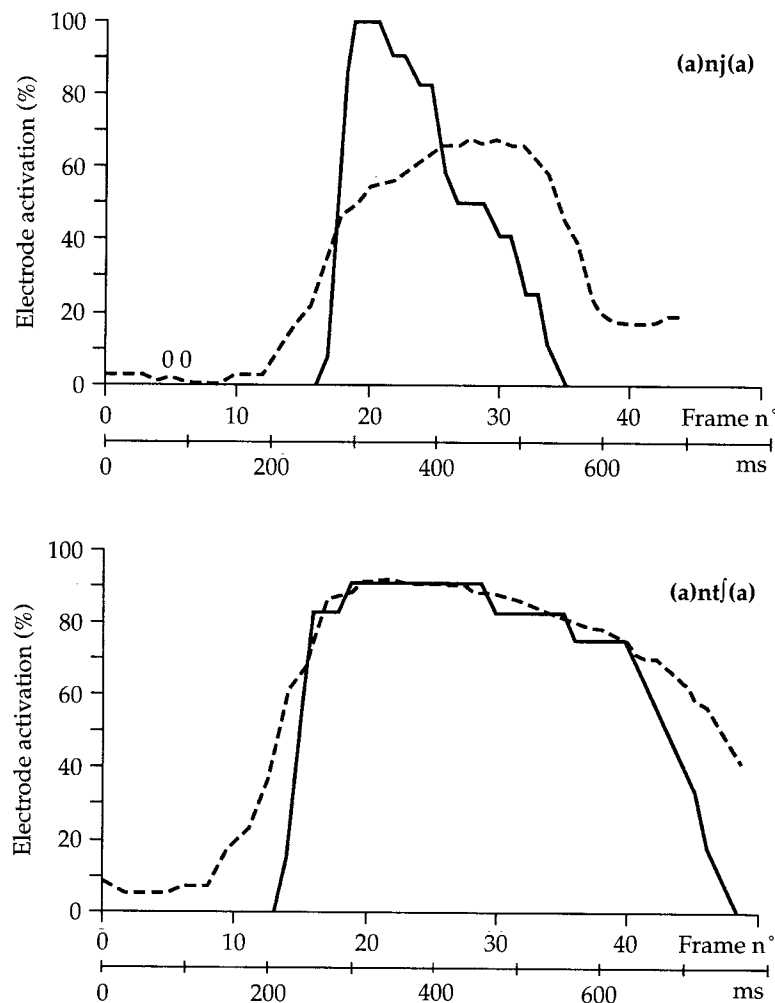


Figure 12.4 EPG curves during /anja/ and /antʃa/ produced by an Italian subject. Continuous lines: tongue-tip/blade contact; dashed lines: tongue body contact, (from Farnetani, 1986).

context, a palatal semivowel (in /anja/) and a postalveolar affricate consonant (in /antʃa/). The utterances are pseudowords produced in isolation by an Italian speaker, and analyzed with EPG.

In each graph in Figure 12.4 the two curves represent the evolution of the tongue-to-palate contact over time. We can see that in /anja/, as the tongue tip/blade (continuous line) achieves maximum front contact for /n/ closure, the tongue body (dashed line) moves smoothly from /a/ to /j/ suggesting overlapping activity of two distinct articulators, and two distinct goals. In /antʃa/, instead, the typical /n/ configuration has disappeared; the front contact

has decreased by some percentage points, and the back contact has appreciably increased; the cluster seems to be produced with one tongue movement. These differences may indicate that two distinct processes are at work in the two utterances: anti-cipatory coarticulation of /j/ on /n/ in the former, and place assimilation of /n/ to /tʃ/ in the latter. Another possible interpretation is that both patterns are instances of coarticulation: they differ because the gestures overlapping /n/ are different and impose different constraints on its articulation; the greater amount of /n/ palatalization in /antʃa/ than in /anja/ would simply be the outcome of such different constraints.

Current theories of coarticulation offer controversial views on whether there are qualitative or quantitative or even no differences between assimilatory and coarticulatory processes. It will be seen that at the core of the different positions are different answers to the fundamental issues addressed above, i.e. the domain, the function and the control of coarticulation.

1.3 Connected speech processes

In speech the phonetic form of a word is not invariable, but can vary as a function of a number of linguistic, communicative and pragmatic factors (e.g. information structure, style, communicative situation): these variations are generally referred to as *alternations*, and the phonetic processes accounting for them have been termed *connected speech processes* (Jones, 1969; Gimson, 1970). According to Gimson (p. 287), these processes describe the phonetic variations that characterize continuous speech when compared to a word spoken in isolation. The author makes a detailed list of connected speech processes in English, among them: place assimilations (within and across word boundaries), assimilations of manner and voicing, reduction of vowels to schwa in unaccented words, deletions of consonants and vowels. According to the author, the factors that contribute to modify the phonetic properties of a word are "the pressure of its sound environment or of the accentual or rhythmic group of which it forms part", and the speed of the utterance (p. 287).

Kohler (1990) proposes an explanatory account of connected speech processes in German. His analysis focuses on the difference between careful and casual pronunciation of the same items and interprets all connected speech processes as a global phenomenon of reduction and articulatory simplification. These processes include /r/ vocalization (/r/ is vocalized to [ɐ] when it is not followed by vowel), weak-form, elision and assimilation. From an analysis of the sound categories most likely to undergo such changes, he infers that connected speech processes result in large part from articulatory constraints (minimization of energy), which induce a reorganization of the articulatory gestures. He also proposes a formalization of the processes in terms of sets of phonetic rules which generate any reduced segmental pronunciation.

The two accounts, although substantially different in their perspectives, have in common the assumption that connected speech processes imply modifications of the basic units of speech (elimination and replacements of articulatory

gestures, changes in articulation places, etc.). Thus the difference between connected speech processes and the phonological assimilations described in 1.2 is that the latter occur independently of how a word is pronounced (SPE, p. 110), while the former occur in some cases (rapid, casual speech), but are absent in others.

Certain theories of coarticulation also consider connected speech processes and propose their own accounts. Experimental research in this field started only recently, but some interesting data are beginning to emerge, as will be seen below in section 2.4.4.

2 Theoretical accounts of coarticulation

2.1 Pioneering studies

That speech is a continuum, rather than an orderly sequence of distinct sounds as listeners perceive it, was pointed out by Sweet, as long ago as the last century (Sweet, 1877, cited by Wood, 1993). Sweet saw speech sounds as points "in a stream of incessant change" and this promoted the view that coarticulatory effects result from the transitional movements conjoining different articulatory targets and is reflected acoustically in the transitions to and from acoustic targets. Menzerath and de Lacerda (1933) showed that segments can be articulated together, not merely conjoined to each other (the term "coarticulation" is attributed to these authors). The pioneer acoustic analysis of American English vowels conducted by Joos (Joos, 1948) revealed that vowels vary as a function of neighboring consonants not only during the transitional periods but also during their steady state. Referring to temporal evolution of the second formant, Joos observed that "the effect of each consonant extends past the middle of the vowel so that at the middle the two effects overlap" (p. 105). In his theoretical account of the phenomenon he contests the "glide" hypothesis, which attributes coarticulation to mechanical factors, i.e. the inertia of vocal organs and muscles. According to that view, since no shift from one articulatory position to another can take place instantaneously, a transition intervenes between successive phones. Joos proposes instead the "overlapping innervation wave theory" (p. 109): each command for each segment is an invariant "wave" that "waxes and wanes smoothly"; "waves for successive phones overlap in time".

As will be seen below, these two early hypotheses on the sequential ordering of speech segments have been highly influential in the development of coarticulation theories.

2.2 Coarticulation as speech economy

2.2.1 Speech variability The theoretical premise at the basis of Lindblom's theory of speech variability is that the primary scope of phonetics is not to describe how linguistic forms are realized in speech, but to explain and derive

the linguistic forms from "substance-based principles pertaining to the use of spoken language and its biological, sociological and communicative aspects" (Liljencrants and Lindblom, 1972, p. 859).¹ Accordingly, in his theory of "Adaptive Variability" and "Hyper-Hypo-Speech" (Lindblom, 1983, 1988, 1990), phonetic variation is not viewed as a mere consequence of the inertia of the speech mechanism, but rather as a continuous adaptation of speech production to the demands of the communicative situation. Variation arises because the production strategy at the basis of any instance of speech varies, being the result of the interaction between system-oriented and output-oriented motor control. Some situations will require an output with a high degree of perceptual contrast, others will require much less perceptual contrast and will allow more variability. Thus, the acoustic characteristics of the same item will exhibit a wide range of variation reflected along the continuum of over- to under-articulation, or hyper- to hypo-speech.

2.2.2 Low-cost and high-cost production behavior What is the function of coarticulation within this framework? Coarticulation, manifested as a reduced displacement and a shift of movements towards the context, is a low-cost motor behavior, an economical way of speaking. Its pervasiveness indicates that the speech motor system, like other kinds of motor behavior, is governed by the principle of economy.

In his study of vowel reduction, Lindblom (1963) introduced the notion of *acoustic target*, an ideal context-free spectral configuration (represented for vowels by the asymptotic values towards which formant frequencies aim). Lindblom's study showed that targets are quite often not realized: his data on CVC syllables indicated that the values of formant frequencies at mid-vowel point change monotonically with changes in vowel duration. At long durations, formants tend to reach the target values; as duration decreases, the formant movements are reduced (target undershoot) and tend to shift towards the values of the adjacent consonants, as shown in Figure 12.5.

The continuous nature of the process indicated that vowel reduction is an articulatory process, largely dependent on duration, rather than a phonological process. The direction of the change towards the context, as well as the different degree of undershoot as a function of the extent of the CV transitions (the amount of vowel reduction is minimal when the consonant-to-vowel distance is small), indicated that reduction is a coarticulatory process rather than a centralization process towards a schwa-like configuration.²

Lindblom's account of the relation between duration, target undershoot and coarticulation was that reduction is the automatic response of the motor system to an increase in rate of motor commands. When successive commands on one articulator are issued at very short temporal intervals, the articulator has insufficient time to complete the response before the next signal arrives, and has to respond to different commands simultaneously. This induces both vowel shortening and reduced formant displacement. Subsequent research showed that the system response to high rate commands does not automatically result in reduced movements (Kuehn and Moll, 1976; Gay, 1978), and that reduction

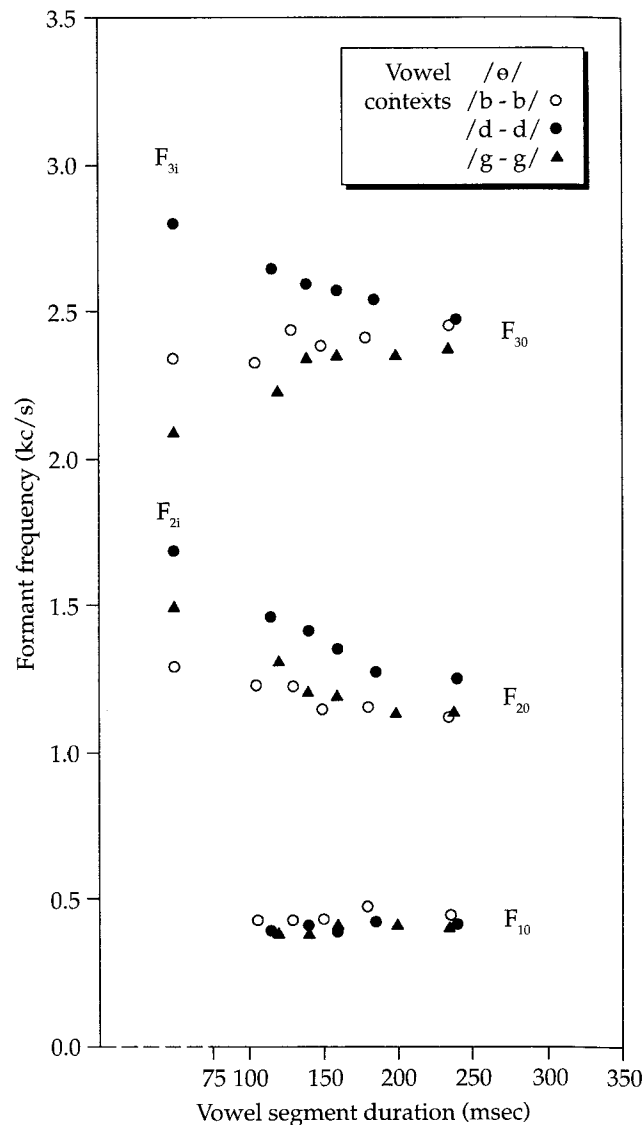


Figure 12.5 Mean F1, F2 and F3 frequencies during the steady state of the Swedish vowel /e/ plotted against vowel duration. The vowel is in the contexts of /b/, /d/ and /g/. As the vowel shortens, the F2 and F3 frequencies shift towards the formant values of the /bV/, /dV/ and /gV/ initial boundaries, F2i and F3i (from Lindblom, 1963).

can occur also at slow rates (Nord, 1986). This indicated that speakers can adapt to different speaking situations and choose different production strategies to avoid reduction/coarticulation or to allow it to occur (these were the premises of Lindblom's hyper-/hypo-speech theory, as mentioned in 2.2.1).

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, such as the rate of formant frequency change, 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. overarticulated vs 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 larger formant velocity values than citation forms. This means that for a given duration, the degree of context-dependent undershoot depends on speech style and tends to decrease with an increase in velocity of the articulatory movements. In the model (where the speech motor mechanism is seen as a second-order mechanical system) it is proposed that undershoot is controlled by three variables reflecting the articulation strategies available to speakers under different circumstances: duration, as expected, input articulatory force and time constant of the system. An increase in input force and/or an increase in speed of the system response (i.e. a decrease in stiffness) contribute to increase the movement amplitude/velocity, and hence to decrease the amount of context-dependent undershoot. Thus, clear speech reflects an undershoot-compensatory reorganization of articulatory gestures.

Another experiment (Lindblom, Pauli and Sundberg, 1975) shows that a low-cost production strategy, characterized by coarticulatory variations, is the norm in natural speech. The authors analyzed apical consonants in VCV utterances. Using a numerical model of apical stop production, the authors showed that the best match between the output of the model and the spectrographic data of natural VCV utterances produced in isolation, is a tongue configuration always compatible with the apical closure but characterized by a minimal displacement from the preceding vowels. In other words, among a number of tongue body shapes that facilitate the tongue tip closure, the tongue body always tends to take those requiring the least movement and an adaptation to the configuration of the adjacent vowels.

Lindblom's hypothesis that the more speech style shifts towards the hypo-speech pole the larger will be the amount of coarticulation is confirmed by a number of studies on connected speech: Krull (1987, 1989), for Swedish; Duez (1991) for French; Farnetani (1991) for Italian.

2.2.3 Phonological adaptations Lindblom (1983) makes a clear distinction between coarticulation/target-undershoot and assimilation. The former, as seen above, is a continuous motor process, increasing in magnitude in connected

spontaneous speech, the latter is a categorical change, a language-specific grammatical rule. Assimilation is a consequence of coarticulation, an adaptation of language to speech constraints (see Ohala, *THE RELATION BETWEEN PHONETICS AND PHONOLOGY*).

2.3 Coarticulation as "creature" of the Language Grammar

The evolution of featural phonology after SPE, is marked by a gradual appropriation of coarticulation into the domain of linguistic competence.

2.3.1 The theory of feature-spreading Daniloff and Hammarberg (1973), and Hammarberg (1976) were the promoters of the "feature spreading" account of coarticulation. According to Hammarberg (1976), the view that coarticulation is a purely physiological process due to mechano-inertial constraints of the speech apparatus entails a sharp dichotomy between intent and execution, and implies that the articulators are unable to carry out the commands as specified. The only way to overcome this dichotomy is to assume that coarticulation itself is part of the phonological component. The arguments in support of this assumption are that: 1) phonology is prior to phonetics; 2) phonological segments are abstract cognitive entities, and cannot be altered by the speech mechanism (in order to be altered they would have to be physical): all that the physical mechanism can do is to execute higher level commands. Therefore the variations attributed to coarticulation must be the input to the speech mechanism, rather than its output. How? Segments have both inherent properties (the phonological features) and derived properties. These latter result from the coarticulation process, which alters the properties of a segment. Phonological rules stipulate which features get modified, and the phonetic representation, output of the phonological rules and input to the speech mechanism, specifies the relevant details of articulation and coarticulation. This departure from Chomsky and Halle's view of coarticulation was probably necessary, in face of the emerging experimental data on coarticulation: data on anticipatory lip protrusion, (Daniloff and Moll, 1968) and on velar coarticulation, (Moll and Daniloff, 1971) showed that coarticulatory movements could be initiated at least two segments before the influencing one. This indicated that coarticulation is not the product of inertia. Another reason (Hammarberg, 1976) was that coarticulation could not be accounted for by universal rules as assumed in SPE, owing to cross-language differences such as those pointed out by Ladefoged (1967) between English and French in the coarticulation of velar stops with front and back vowels.

Why does coarticulation occur? The function of coarticulation is to smooth out the differences between adjacent sounds: if phonemes were executed in their canonical forms, the speech mechanism would introduce transitional sounds between executions of contiguous segments. Coarticulatory modifications

accommodate the segments so that when they are realized, the transitions between them are minimized. Thus coarticulatory rules serve to reduce what the vocal tract introduces when we speak. Notice that, according to SPE (see 1.2), and to the "glide" hypothesis of coarticulation (see 2.1), the transitions between segments are the effects of coarticulation, while here they are what the coarticulation rules have to minimize. In Daniloff and Hammarberg's view anticipatory coarticulation is always a deliberate phonological process, while carryover coarticulation can be in part the effect of the inertia of the speech organs and in part a feed-back assisted strategy that accommodates speech segments to each other. The authors acknowledge that no clear phonological explanation in terms of feature spreading can be found for a number of contextual variations such as the lengthening of vowels before a voiced consonant, or the lack of aspiration in obstruent stops after /s/.

2.3.2 Henke's articulatory model According to Daniloff and Hammarberg (1973) the articulatory model of Henke (1966) best accounts for experimental data on the extent of coarticulation. Henke's model contrasts with another well known account of coarticulation, the articulatory syllable model proposed by Kozhevnikov and Chistovich (1965). This model was based on data on anticipatory labial coarticulation in Russian. Segments seemed to coarticulate within, but not across, C_nV sequences. The C_nV -type syllable was thus viewed as the articulatory domain of coarticulation, and its boundaries as the boundaries of coarticulation. Unlike the C_nV model, Henke's model does not impose top-down boundaries to anticipatory coarticulation. Input segments are specified for articulatory targets in terms of binary phonological features (+ or -), with features unspecified in the phonology being given a value of 0 (Moll and Daniloff, 1971). Coarticulatory rules assign a feature of a segment to all the preceding segments unspecified for that feature, by means of a look-ahead scanning mechanism. The spread of features is blocked only by a specified feature. So, the nasality feature inherent in a nasal consonant [+nasal] will be anticipated to all the preceding segments unspecified for nasality. Likewise, after the nasal, if a segment specified as [-nasal] follows, the feature [-nasal] will be applied immediately to the first unspecified segment after the nasal. Obviously a look-ahead mechanism intrinsically impedes carryover coarticulation, for which, as seen above, Daniloff and Hammarberg devised other explanations.

2.3.3 Feature specification and coarticulation: towards the concept of coarticulation resistance A number of experimental results, such as those mentioned in 2.3.1, are compatible with the hypothesis of feature spreading and the look-ahead mechanism, but many others contradict the spatial and/or the temporal predictions of the model. First, the model cannot explain the quite extensive carryover effects observed in a number of studies on V-to-V coarticulation (for example Recasens, 1989; Magen, 1989). The other disputed aspects of the theory are: 1) the adequacy of the concept of specified vs

unspecified features for blocking or allowing coarticulation to occur; 2) the hypothesis that a look-ahead mechanism accounts for the temporal extent of anticipatory coarticulation (this point will be developed below in section 2.4.3).

As for the first issue, it appears that segments specified for a contradictory feature in asymmetric V_1CV_2 type sequences can nonetheless be modified by coarticulation. Data on lip rounding in French (Benguerel and Cowan, 1974) and English (Sussman and Westbury, 1981) indicate that, in an $/iC_a u/$ sequence type, rounding movements for $/u/$ can start during $/i/$, specified as $[-round]$. Also data on lingual coarticulation indicate that tongue displacement towards a vowel can begin during a preceding cross-consonantal vowel even if this is specified for conflicting features with respect to the following vowel, for example $/a/$ vs $/i/$ (Öhman, 1966, for Swedish and English; Butcher and Weiher, 1976, for German; Farnetani, Vaggel and Magno Caldognetto, 1985, for Italian; Magen, 1989, for American English). Most interestingly, these transconsonantal V-to-V effects appear to vary in degree across languages, indicating that the *same* vowel categories in different languages are subject to different constraints that favor or disfavor coarticulatory variations (see Manuel and Krakow, 1984, comparing Swahili, Shona and English; Manuel, 1987, comparing three Bantu languages; Choi and Keating, 1991, comparing English, Polish, Russian and Bulgarian).

As for phonologically unspecified segments, some experimental data are compatible with the idea that they completely acquire a contextual feature.

Figure 12.6 illustrates how the Japanese vowel $/e/$, unspecified for nasality, acquires the nasality feature in a symmetric context as predicted by the feature spreading model. The figure shows the amount of velum height during the vowel $/e/$ in Japanese, surrounded by oral consonants (a), by nasal consonants (c), and in a mixed environment (b). It can be seen that during $/e/$ the velum is as high as for oral consonants in (a), and as low as for nasal consonants in (c): in both cases the velum height curve runs nearly flat across the vowel. In the asymmetric example (b) the curve traces a trajectory from a high to a low position during the $/e/$ preceded by $/s/$ and the reverse trajectory during the $/e/$ followed by $/d/$. The symmetric sequences show that $/e/$ is completely oral in an oral context and completely nasalized in a nasal context, indicating that this vowel has no velar target of its own and acquires that of the context. The trajectories in the asymmetric sequences do not contradict the hypothesis that this vowel has no target for velar position, but contradict the assumption that contextual features are spread in a categorical way. Accordingly, $/e/$ would have to be completely nasalized from its onset when followed by a nasal, and completely oral when followed by an oral consonant, and this does not seem to occur (see panel b).

Many other data indicate that phonologically unspecified segments may nonetheless exhibit some resistance to coarticulation, indicating that they are specified for articulatory targets. English data on velar movements (Bell-Berti, 1980; Bell-Berti and Krakow, 1991) show that the oral vowels are not articulatorily neutral to velar height, and have their own specific velar positions

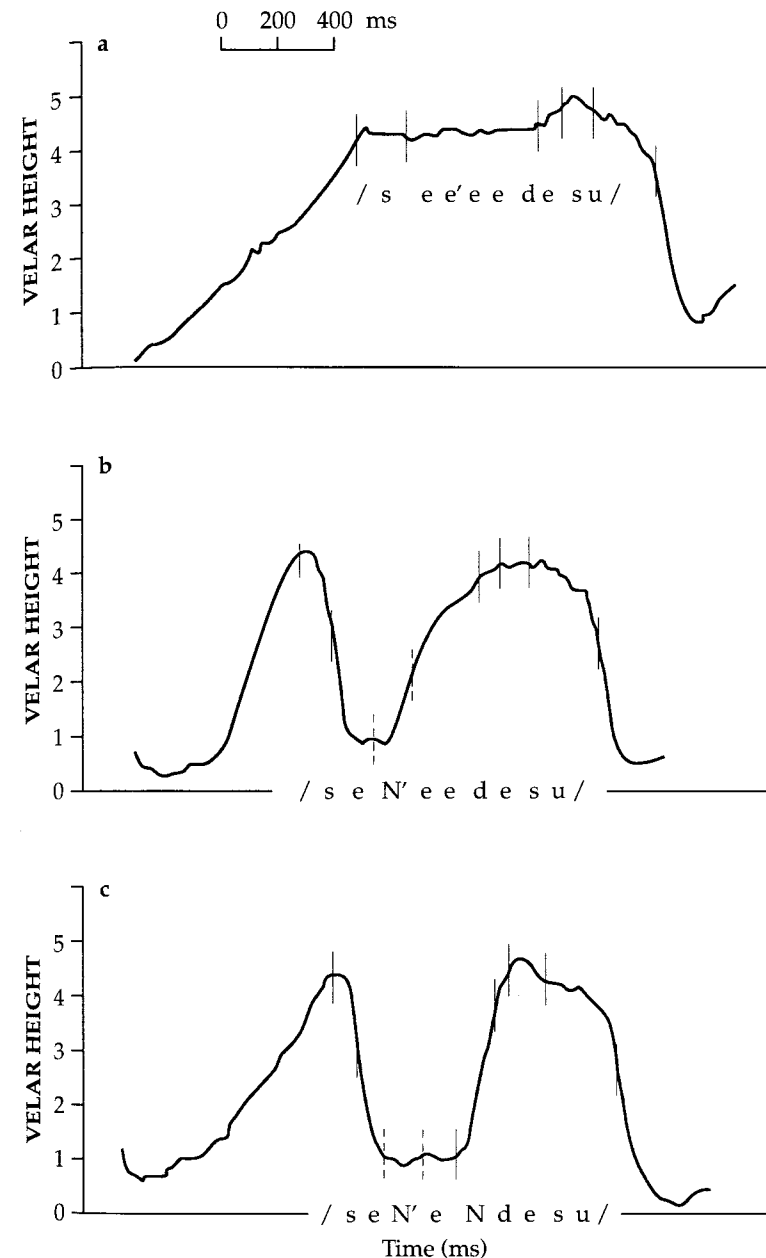


Figure 12.6 Velar movements observed with fiberscope during Japanese utterances containing the vowel $/e/$ in oral and nasal contexts (from Ushijima and Sawashima, 1972).

even in a non-nasal environment. As for lip position, Engstrand's data (1981) show that in Swedish /u-u/ sequences with intervocalic lingual consonants, protrusion of the upper lip relaxes during the consonants and the curve forms a "trough" between the vowels, suggesting that such consonants are not completely neutral to lip position. For lingual coarticulation, troughs in lingual muscles' activity during /ipi/ sequences were first observed in the 1974 study by Bell-Berti and Harris. Analogously, at the articulatory level, Engstrand (1983) observed that the vertical movement of the tongue in /ipi/ sequences shows troughs during the consonant, indicating that bilabial voiceless stops may be specified for lingual position, possibly to meet the aerodynamic requirements for the bilabial release burst to occur, and for avoiding turbulence noise during the transition to V2.

Subsequent research on lingual consonants in Catalan, Swedish and Italian (Recasens, 1983, 1984a, 1987; Engstrand, 1989; Farnetani, 1990, 1991) showed that consonants unspecified for tongue body features coarticulate to different degrees with the surrounding vowels. During the production of coronals, the amount of tongue body coarticulation tends to decrease from alveolars to postalveolars and from liquids to stops to fricatives (Farnetani, 1991, data on Italian); a similar trend is observed in Swedish (Engstrand, 1989).

Figure 12.7 is an example of how different consonants coarticulate to different degrees with the surrounding vowels /i/ and /a/ in Italian, as measured by EPG. The trajectories represent the amount of tongue body contact over time during intervocalic coronals /t/, /d/, /z/, /ʃ/, /l/ and bilabial /p/ in symmetric VCV sequences. The /i-i/ trajectories exhibit troughs of moderate degree for most consonants; /z/ shows the largest deviation from the /i/ to /i/ trajectory (see points V1 and V2), indicating that the production of this consonant requires a lowering of the tongue body from the /i/-like position. In the context of /a/ (see points V1 and V2 in this context), /p/ appears to fully coarticulate with this vowel, as it shows no contact; also /l/ strongly coarticulates with /a/. For /t/, /d/, /z/ the tongue body needs to increase the contact to about 20 per cent. During /ʃ/ it can be seen that the tongue body contact reaches the same value (between 50 per cent and 60 per cent) in the two vocalic contexts, indicating that this consonant is maximally resistant to coarticulation.

The overall data on tongue body V-to-C coarticulation indicate that no alveolar consonant fully coarticulates with the adjacent vowels, which suggests the presence of a functional and/or physical coupling between tip/blade and body. The differences in coarticulation across consonants can be accounted for by consonant specific production constraints imposed on the tongue body. Fricatives must constrain tongue dorsum position to ensure the appropriate front constriction and the intraoral pressure required for noise production; the production of stops and laterals imposes lesser constraints, and allows for a wider range of coarticulatory variations. As will be seen in 2.3.4, Keating (1988), on the basis of coarticulatory data, will propose that English /s/ be specified as [+high] for tongue body position.

The notion of coarticulation resistance was introduced by Bladon and

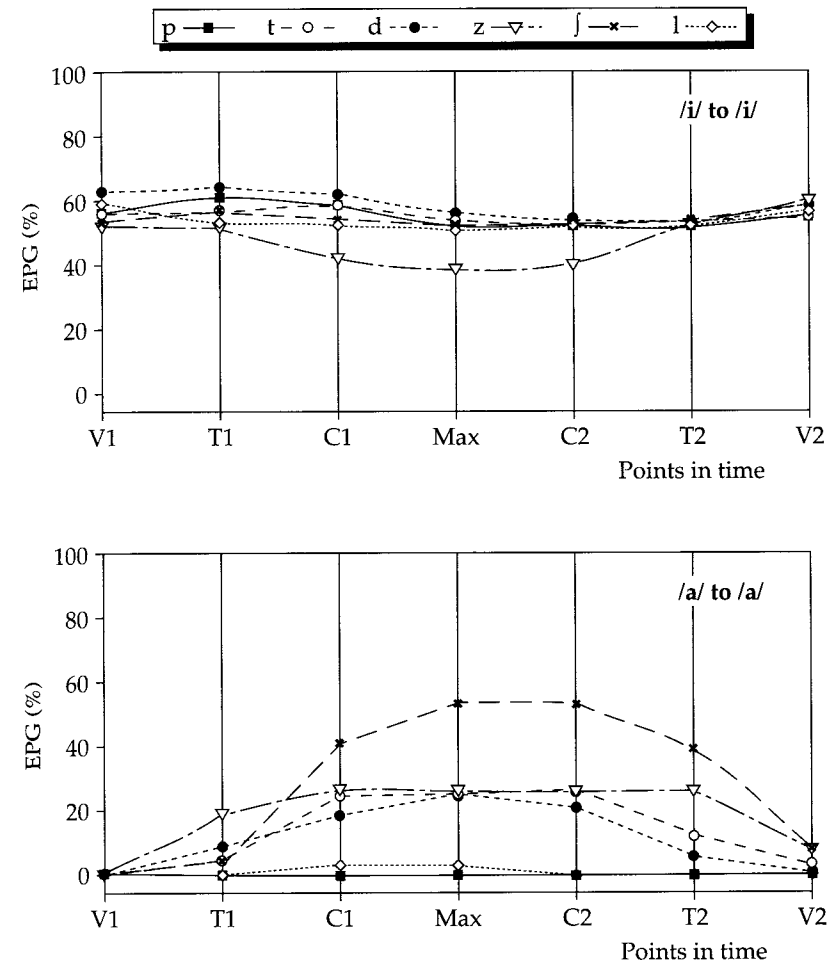


Figure 12.7 Tongue body EPG contact at various points in time during symmetric (C)VCV isolated words in Italian. V1, V2 correspond to vowel mid-points; T1, T2 to offset of V1 and onset of V2 respectively; C1, C2 correspond to onset and release of consonant closures/constrictions respectively; Max refers to the point of maximum contact during the consonant (from Farnetani, 1991).

Al-Bamerni (1976) in an acoustic study of V-to-C coarticulation in /l/ allophones (clear vs dark vs syllabic /l/). The data indicated that coarticulatory variations decrease continuously from clear to syllabic /l/. These graded differences could not be accounted for by binary feature analysis, which would block coarticulation in dark /l/, specified as [+back]. The authors propose a numerical index of coarticulation resistance to be attached to the feature specification of each allophone. A subsequent study on tongue tip/blade displacement in alveolars (Bladon and Nolan, 1977) confirmed the idea that feature specification alone cannot account for the observed coarticulatory behavior.

All these studies show that the assignment of contextual binary features to unspecified segments through phonological rules fails to account for the presence vs absence of coarticulation, for its graded nature, and for the linguistically relevant aspects of coarticulation associated with this graded nature i.e. the different degree of coarticulation exhibited by the *same* segments across languages. Explanation of these facts requires factors outside the world of phonological features: *articulatory constraints* and *aerodynamic-acoustic constraints*. Manuel (1987, 1990) proposes that interlanguage differences in coarticulation may be controlled by *perceptual output constraints* (see 2.3.4).

2.3.4 The window model of coarticulation Keating (1985, 1988a, 1988b, 1990) proposes a new articulatory model which can account for the continuous changes in space and time observed in speech, and for intersegment and interlanguage differences in coarticulation.

On one hand Keating agrees that phonological rules cannot account for the graded nature of coarticulation. At the same time she contests the assumption that such graded variations are to be ascribed to phonetic universals as automatic consequences of the speech production mechanism (Keating, 1985). One example is the duration of vowels as a function of the voiceless vs voiced consonantal context. Vowels tend to be shorter before voiceless than before voiced stops, but these patterns are not the same across languages. Keating shows that in English the differences in vowel durations are relevant and systematic, while in other languages such as Polish and Czech they are unsystematic or even absent. Therefore each language must specify these phonetic facts in its grammar.

Keating's proposal is that all graded spatial and temporal contextual variations, both those assumed to be phonological, and those assumed to be universal and physically determined, be accounted for by the *phonetic* rules of the grammar.

The windows Keating's model marks a substantial departure from the feature-spreading model, first because it assumes that phonological underspecification (i.e. unspecified features) may persist into the phonetic representation, and second because underspecification in this model is not a categorical, but a continuous notion. Input to the window model is the phonological representation in terms of binary features. Unspecified segments may be left unspecified, or can acquire specifications through a number of language specific phonological rules. For instance, English /s/ is assumed to acquire the feature [+high] through fill-in rules, owing to aerodynamic constraints requiring a high tongue body position for its production; there are context-sensitive rules, such as those that specify Russian /x/ as [-back] before high vowels; in some languages there may be assimilation rules like those proposed in the feature-spreading model (see below for expansion).

Implementation rules interpret the output of the phonological component in space and time and provide a continuous phonetic representation. For a given articulatory or acoustic dimension a feature value is associated with a range

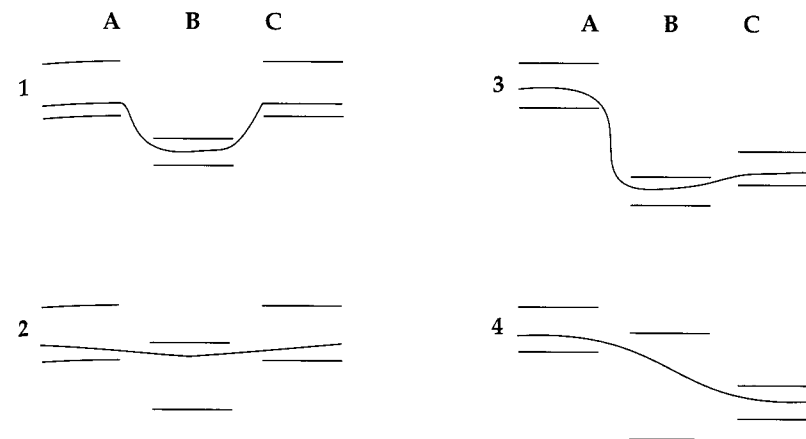


Figure 12.8 Windows and paths modeling articulator movements in three-segment sequences (selected from Keating, 1988a). The effects of narrow vs. wide windows on the interpolation contours can be observed in both the symmetric (1, 2) and the asymmetric (3, 4) sequences.

of values called a *window*. Windows have their own duration, and a width representing the entire range of contextual variability of a segment. Window widths depend first of all on phonological information: specified features are associated with narrow windows and allow for *little* contextual variation, unspecified features are associated with wide windows and allow for *large* contextual variation. Quantitative information on the maximum amount of contextual variability observed in speech will determine the exact window width for a given dimension, so that all the intermediate degrees between maximally wide and maximally narrow windows are possible.

Adjacent windows are connected by "paths" (or contours), which are interpolation functions between windows, and are constrained by the requirements of smoothness and minimal articulatory effort. Paths should represent the articulatory or acoustic variations over time in a specific context. Figure 12.8 shows some selected sequences of windows and contours.

Wide windows contribute nothing to the contour and allow direct interpolation between the preceding and the following segment. For instance, in VCV sequences, if C is unspecified for tongue body position, its associated window will not be seen as a target, and will allow V-to-V interpolation (and coarticulatory effects) to occur (see sequence 4 in Figure 12.8). If instead C is specified, its associated narrow window will allow local V-to-C effects but will block V-to-V effects: in this case the contour will be constrained to take place quickly between the two narrow windows (see sequences 1 and 3 in Figure 12.8). It can be seen that the middle window of sequence 2 is not as wide as to allow direct interpolation between the first and the third segment: according to Keating, such patterns indicate that the segment is not completely unspecified. An example of "not quite unspecified" segments are English vowels with

respect to phonetic nasality. Keating shows that these vowels (phonologically unspecified for nasality) are associated with wide but not maximal windows, and in some contexts they reveal their inherent specification. According to Keating, this occurs because underspecification is categorical at the phonological level, but continuous in the surface representation, and therefore allows for a variety of window widths: in any case wide windows specify very little about a segment. On this crucial point, Boyce, Krakow and Bell-Berti (1991) argue that, if supposedly unspecified segments are associated in production with characteristic articulatory positions, it becomes hard to reconcile the demonstration of any kind of target with the notion of underspecification. The authors propose instead that phonologically unspecified features can influence speech production in another way: they may be associated with cross-speaker variability (as shown in their lip-protrusion data during unspecified consonants), and cross-dialectal variability (see below, for a compatible account proposed by Keating for cross-language phonetic differences).

Cross-language differences According to Keating, interlanguage differences in coarticulation may originate from phonology or from phonetics. Phonological differences occur when, for a given feature, phonological assimilatory rules operate in one language and not in another. Phonetic differences are due to a different phonetic interpretation of a feature left unspecified. Speech analysis will help determine which differences are phonological and which are phonetic.

In a study of nasalization in English using airflow measurement, Cohn (1993) compared the nasal flow contour in nasalized vowels in English with the contour of nasal vowels in French and of nasalized vowels in Sundanese. In French vowels nasality is phonological, 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. In nasalized English vowels, instead, the shapes of the contours describe smooth, rapid trajectories from the [-nasal] to the [+nasal] adjacent segments. The categorical vs the gradient quality of nasalization in Sundanese vs English indicates that nasalization is indeed phonological in Sundanese (i.e. the output of phonological assimilatory rules), while in English it results from phonetic interpolation rules.³

Languages may also differ in coarticulation because the phonetic rules can interpret phonological underspecification in different ways in different languages, allowing the windows to be more or less wide. In this case 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, p. 22).

Manuel (1987) disagrees with Keating's proposition that all phonetic changes have to be accounted for by grammatical rules simply because they are not universal. Referring to interlanguage differences in V-to-V coarticulation, Manuel proposes that language-particular behavior, apparently arbitrary, can

itself be deduced from the interaction between universal characteristics of the motor system and language specific phonological facts, such as the inventory and distribution of vowel phonemes. Her hypothesis is that V-to-V coarticulation is regulated in each language by the requirement that the perceptual contrast among vowels is preserved, i.e. by *output constraints*, which can be strict in some languages and rather loose in others. Languages with smaller vowel inventories, where there is less possibility of confusion, should allow more coarticulatory variations than languages with a larger number of vowels, where coarticulation may lead to articulatory/acoustic overlap of adjacent vowel spaces. This hypothesis was tested by comparing languages with different vowel inventories (Manuel and Krakow, 1984; Manuel, 1987). The results of both studies support the output constraints hypothesis. Thus, if the output constraints of a given language are related to its inventory size and to the distribution of vowels in the articulatory/acoustic space, then no particular language specific phonetic rules are needed, since different degrees of coarticulation across languages can be to some extent predictable.

Connected speech Keating does not deal with the problem of phonetic variations due to factors other than the segmental phonetic context, for instance with the phonetic differences characterizing different speaking styles. In the present version of the model, windows have no internal temporal structure allowing them to stretch or compress in time, and their width is intended to represent all possible contextual variations. So at present, samples of strongly coarticulated informal speech and samples of clear speech cannot be differentiated. A recent neural network model proposed by Guenther (1994b) allows targets (viewed as regions in the orosensory coordinates) to increase or reduce in size, as a function of rate and accuracy in production, and thus overcomes the present limitations of the window model. On the other hand, if the windows associated with specified features were allowed to stretch in width in continuous informal speech, then the relation between feature specification at the phonological level and window width at the phonetic level would become much weaker.

2.4 Coarticulation as coproduction

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

The present account centers on four topics: the nature of phonological units, coarticulation resistance, anticipatory coarticulation and connected speech processes.

2.4.1 The dynamic nature of phonological units The central point of Fowler's criticism of feature-based theories (Fowler, 1977, 1980) is the dichotomy between the abstract, discrete and timeless units posited at the level of language knowledge, and the physical, continuous and context-dependent movements at the level of performance. In other words, she contests the assumption that what speakers know about the phonological categories of their language is substantially different from the units they use when they speak. According to Fowler, all current accounts of speech production need a translation process between the abstract and the physical domain: the speech plan supplies the spatial targets to be reached, and a central clock specifies when the articulators have to move to the targets: "The articulator movements are excluded from the domain of the plan except as it is implied by the different successive articulatory positions" (Fowler, 1977, p. 99). An alternative proposal that overcomes the dichotomy between linguistic and production units 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, not only describe "an abstract summary of its significance" (Fowler et al., 1980, p. 381). These units, the gestures, must be serially ordered planned actions, specified dynamically, and context-free. It is their specification in terms of dynamic parameters (such as force, stiffness) that automatically determines the kinematics of the speech movements. Gestures have their own intrinsic temporal structure, which allows them to overlap in time when executed; the degree of gestural overlap is controlled at the plan level. So gestures are not altered by adjacent gestures, they just overlap (i.e. are coproduced) with adjacent gestures.

Figure 12.9 illustrates coproduction of gestures. The activation of a gesture increases and decreases smoothly in time, and so does its influence on the vocal tract shape. 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 maximal influence on the vocal tract shape, while the overlapping gestures 1 and 3 have weaker influences. Before and after this interval (i.e. during the implementation and relaxation period of gesture 2, respectively) its influence is less, while that of the other two gestures supervenes.

The view of gestures as intervals of activation gradually waxing and waning in time, echoes the early insight of Joos (1948) who proposed the "innervation wave theory" to account for coarticulation (cf. 2.1).

2.4.2 Coarticulation resistance Gestures are implemented in speech by coordinative structures, i.e. by transient functional dependencies among the articulators that contribute to a gesture. These constraints are established to ensure invariance of the phonetic goal; for instance, upper lip, lower lip, and jaw are functionally linked in the production of bilabial closures, so that one

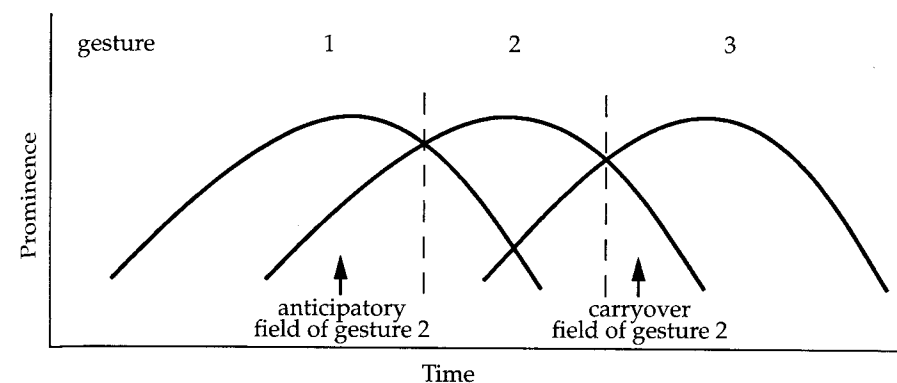


Figure 12.9 Representation of a sequence of three overlapping gestures (from Fowler and Saltzman, 1993). Notice the similarity of this representation of gestural activation and gestural overlap to actual data on EMG activity and overlap illustrated in Figure 12.1.

will automatically compensate for a decreased contribution of another due to perturbation or coarticulatory variations (see Löfqvist, *THEORIES AND MODELS OF SPEECH PRODUCTION* and Perkell, *ARTICULATORY PROCESSES*).

How are coarticulatory variations accounted for within the gestural framework? According to Fowler and Saltzman (1993) variations induced by coproduction depend on the degree to which the gestures share articulators, i.e. on the degree of spatial overlap. When subsequent gestures share only one articulator, such as the jaw in /VbV/ sequences, the effects of gestural interference will be irrelevant, and temporal overlap between vocalic and consonantal gestures will take place with minimal spatial perturbations. The highest degree of spatial overlap occurs when two overlapping gestures share the articulators directly involved in the production of gestural goals, and impose competing demands on them. In Bell-Berti and Harris (1981) it is proposed that gestural conflict be resolved by delaying the onset of the competing gesture so that the ongoing goal can be achieved, i.e. the conflict is resolved at the plan level. Browman and Goldstein (1989), Saltzman and Munhall (1989) propose that the phasing of gestures may be context free and that the output of a gestural conflict may be simply a blend of the influence of the overlapping gestures. According to Fowler and Saltzman (1993), the outcome of gestural blending depends on the degree of "blending strength" associated with the overlapping gestures: "stronger" gestures tend to suppress the influence of "weaker" gestures, while the blending of gestures of similar strength will result in an averaging of the two influences. Fowler and Saltzman's account of coarticulation resistance implies that gestures with a high degree of blending strength resist interference from other gestures, and at the same time themselves induce strong coarticulatory effects, in agreement with experimental findings (Bladon and Nolan, 1977; Recasens, 1984b; Farnetani and Recasens,

1993). On this account, the highest degree of blending strength appears to be associated with consonants requiring extreme constrictions, and/or placing strong constraints on articulator movements, while a moderate degree of blending strength appears to be associated with vowels (see a compatible proposal by Lindblom, 1983, that coarticulatory adaptability, maximal in vowels and minimal in lingual fricatives, varies as a function of the phonotactically based sonority categories).

The coproduction account of coordination and coarticulation also implies that speakers do not need a continuous feedforward control of the acoustic output and consequent articulatory adjustments. Likewise, cross-language differences do not result from on-line control of the output. Languages may differ in degree of coarticulation in relation to their inventories, but these differences are consequences of the different gestural set-up, i.e. the parameters that specify the dynamics of gestures and their overlap, which are learned by speakers of different languages during speech development.

2.4.3 Anticipatory extent of coarticulation 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. Compatibly, 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 an articulator movement is independent of the preceding phone string length and begins at a fixed time before the acoustic onset of the segment with which it is associated. Other studies, however, are consistent with the look-ahead model (see 2.3.2) and indicate that the onset of lip rounding or velar lowering is not fixed, but extends in the anticipatory direction as a function of the number of neutral segments preceding the influencing segment (see, for lip rounding coarticulation, Daniloff and Moll, 1968; Benguerel and Cowan, 1974; Lubker 1981; Sussman and Westbury, 1981; and for velar coarticulation, Moll and Daniloff, 1971). Yet, other results on velar coarticulation in Japanese (Ushijima and Sawashima, 1972; Ushijima and Hirose, 1974) and in French (Benguerel, Hirose, Sawashima and Ushijima, 1977a) indicate that velar lowering for a nasal consonant does not start earlier in sequences of two than in sequences of three oral vowels preceding the nasal. An important finding of Benguerel et al. (1977a), apparently disregarded in the literature, was the distinction between the velar lowering associated with the oral segments preceding the nasal, and a subsequent more rapid velar lowering for the nasal which causes the opening of the velar port. Al-Bamerni and Bladon (1982) made similar observations on velar coarticulation in CV_nN sequences in English: the speakers seemed to use 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 nasal consonant. Perkell and Chiang (1986) and Perkell (1990) were the first to observe two-stage patterns in lip rounding movements, which converged

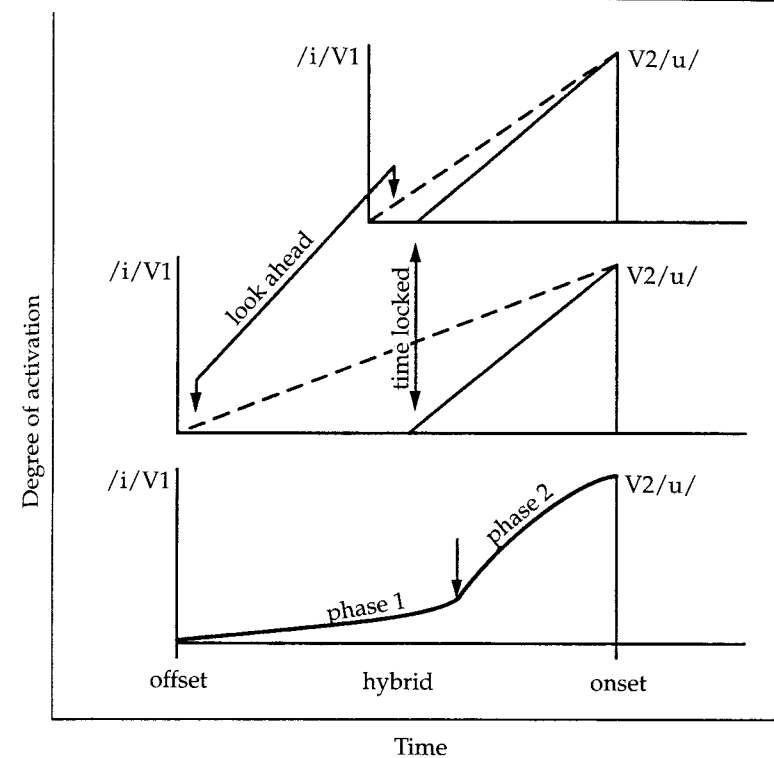


Figure 12.10 Schematic representation of the three models of anticipatory lip rounding coarticulation proposed by Perkell (see text for description).

with Al-Bamerni and Bladon's observations: in /iC_nu/ utterances there was a gradual onset of lip protrusion linked to the offset of /i/, followed by an increase in velocity during the consonants with an additional protrusion closely linked with /u/, and quite invariant. The authors interpreted the composite movements of the two-stage patterns as a mixed coarticulation strategy, and proposed a third model of anticipatory coarticulation, the *hybrid model*. Accordingly, the early onset of the protrusion movement would reflect a look-ahead strategy, while the rapid increase in protrusion at a fixed interval before the rounded vowel would reflect a time-locked strategy. Figure 12.10 compares the three models of anticipatory coarticulation. Perkell's data on three English subjects (Perkell, 1990) indicated that two of the three subjects used the two-stage pattern, compatible with the hybrid model.

Boyce, Krakow, Bell-Berti and Gelfer (1990) argue that many of the conflicting results on the extent of anticipatory coarticulation stem from the assumption that phonologically unspecified segments are also articulatorily neutral (see 2.3.3 on this problem); a number of studies have attributed the onset of lip rounding or velar lowering to anticipatory coarticulation without testing first whether or not the phonologically neutral segments had specific target

positions for lips or velum. The authors also argue against the hypothesis put forth by Al-Bamerni and Bladon that the occurrence of one-stage and two-stage patterns of anticipatory coarticulation might occur randomly. The data by Bell-Berti and Krakow (1991) on velar lowering show, through comparisons between vowels in nasal vs oral context, that the early onset of velar lowering in the two-stage patterns is associated with the characteristic velar positions of the oral vowels, while the second stage, temporally quite stable, is associated with the production of the nasal consonant; therefore the two-stage patterns do not reflect a mixture of two coarticulation strategies, but simply a vocalic gesture followed by consonantal gestures. Moreover the study shows that the patterns of velar movements do not change randomly, but depend on speech rate and the number of vowels in the string, the two-movement patterns prevailing in longer utterances and the one-movement patterns in shorter utterances. These variations, according to the authors, simply reflect different degrees of blending between the vocalic and the consonantal gestures, which sum together into a single movement when they are temporally adjacent.

In a subsequent study on four American speakers, Perkell and Matthies (1992) tested whether the onset of the initial phase of lip protrusion in /iC_nu/ utterances is related to consonant specific lip targets, as proposed by Boyce et al. (1990), and whether the second phase, from the maximum acceleration event, is indeed stable, as predicted by the hybrid and coproduction models, or is itself affected by the number of consonants. In agreement with Boyce et al. (1990), the movement patterns in the control /iC_ni/ utterances showed consonant-related protrusion gestures (especially for /s/), and in /iC_nu/ utterances it was found that lip movements begin earlier when the first consonant in the string is /s/, thus confirming the hypothesis of consonant contribution to the onset of lip movements. The analysis of the second-phase movement, i.e. of the /u/ related component of lip protrusion, revealed that the interval between the acceleration peak and the onset of /u/ was not fixed, but tended to vary as a function of consonant duration for three of the subjects (although the correlations were very low, with R² ranging from 0.057 to 0.35). According to the authors, the timing and the kinematics of this gesture reflect the simultaneous expression of competing constraints, that of using the same kinematics (as in the time-locked model), and that of starting the protrusion gesture for /u/ when it is permitted by the relaxation of the retraction gesture for /i/ (as in the look-ahead model). The variability between and within subjects would reflect the degree to which such constraints are expressed. Also recent data on three French subjects (Abry and Lallouache, 1995) indicate that the lip protrusion movement (from the point of maximum acceleration to that of maximum protrusion) varies in duration as a function of the consonant interval. However, its duration does not decrease from /iC_ny/ to /iy/ utterances, as would be predicted by the look-ahead model; in other words the lip protrusion movement can expand in time, but cannot be compressed (see Farnetani, in press, for a more detailed account of the "movement expansion" model proposed by these authors).

The possibility that the slow onset of the protrusion movement occurring around the offset of /i/ may reflect a passive movement due to the relaxation of the retraction gesture of /i/, rather than an active look-ahead mechanism has not yet been explored. Sussman and Westbury (1981), did observe that in /iC_nu/ sequences the onset of lip protrusion occurred before the onset of *orbicularis oris* activity, and suggested that the movement might be the passive result of the cessation of *risorius* activity, and simply reflect a return of lips to a neutral position. We believe that this problem should be investigated further.

All cross-language studies indicate that anticipatory coarticulation varies across languages: Clumeck (1976) observed that the timing and amplitude of velar lowering varies across the six languages analyzed; Lubker and Gay (1982) showed that anticipatory lip protrusion in Swedish is much more extensive than in English; a study by Boyce (1990), on lip rounding in English and Turkish, indicated that while the English patterns were consistent with the coproduction model, the plateau-like protrusion curves of Turkish marked lip rounding coarticulation as a phonological process (in agreement with Cohn's interpretation of nasality in Sundanese – see 2.3.4).

2.4.4 Connected speech processes

Articulatory model According to Browman and Goldstein (1990b, 1992) gestural phonology can give an explanatory and unifying account of apparently unrelated speech processes (coarticulation, allophonic variations, alternations) that in featural phonology require a number of separate and unrelated phonological rules. Here the phonological structure of an utterance is modeled as a set of overlapping gestures specified on different tiers (the vocal tract variables, see Figure 12.11). Quantitative (gradient) variations in overlap, or quantitative variations in gestural parameters can account for a large number of allophonic variations as a function of stress and position, as well as for the alternations observed in connected speech. Connected speech processes such as assimilations, deletions, reductions (or weakenings) can be accounted for by an *increase in gestural overlap* and a *decrease in gestural amplitude*. In casual rapid speech, subsequent consonantal gestures can so far overlap as to hide each other when they occur on different tiers, or to completely blend their characteristics when they occur on the same tier. Hiding gives rise to perceived deletions and/or assimilations, while blending gives rise to perceived assimilations. For example, the deletion of /t/ in a rapid execution of the utterance "perfect memory" is only apparent; Xray trajectories reveal the presence of the /t/ gesture, overlapped by the following /m/ gesture. Figure 12.11 shows a schematic gestural representation of part of the utterance "perfect memory" spoken in isolation (a), and within a fluent phrase (b).

In the figure the extent of each box represents the duration (or activation interval) of a gesture. It can be seen that within each word the gestures always overlap, but in version (b) the labial closure for the initial /m/ of word 2

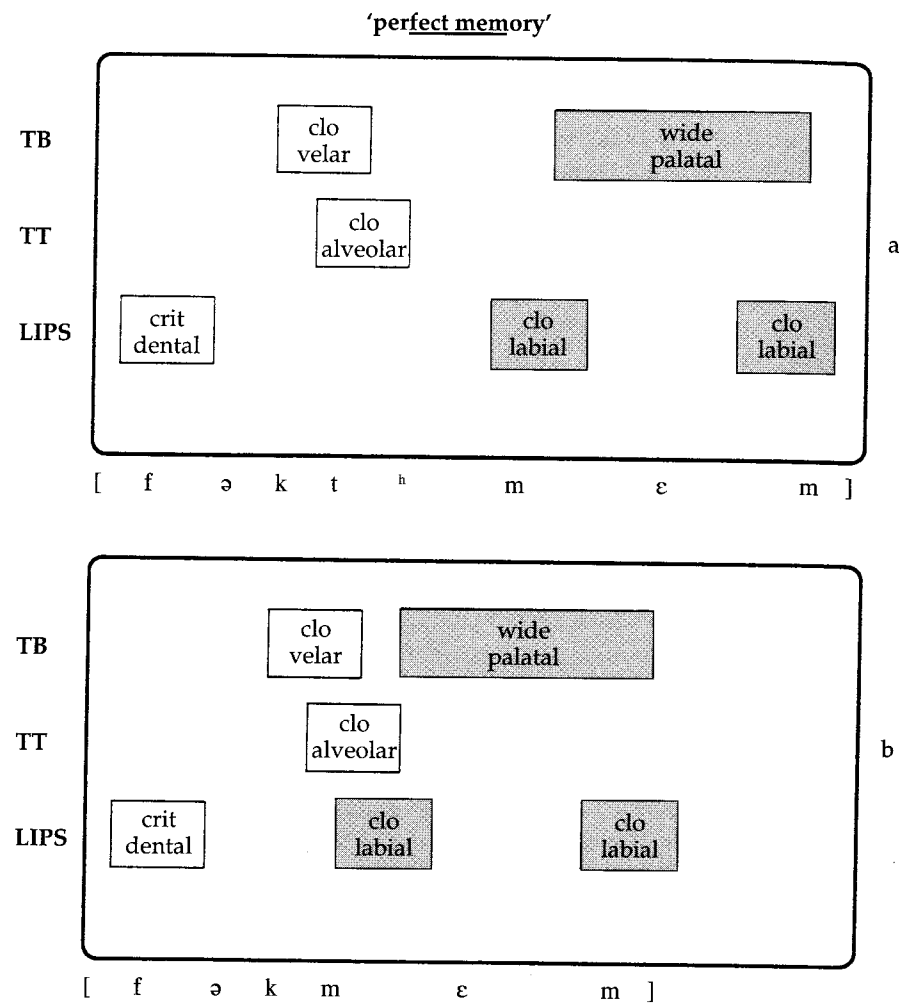


Figure 12.11 Representation of the phonological structure of utterance 'perfect memory' produced in isolation (a) and in continuous speech (b). Vocal tract variables are from top: tongue body, tongue tip and lips (from Browman and Goldstein, 1989).

overlaps and hides the alveolar gesture for final /t/ of word 1. According to the authors, hidden gestures may be extremely reduced in magnitude or completely deleted. "Even deletion, however, can be seen as an extreme reduction, and thus as an endpoint in a continuum of gestural reduction, leaving the underlying representation unchanged" (Browman and Goldstein, 1990b, p. 366).

An example of within-tiers blending is the palatalization of /s/ followed by a palatal in the utterance "this shop": the articulatory analysis should show a

smooth transition between the first and the second consonant, not the substitution of the first with the second. Finally, in CVCV utterances with unstressed schwa as first vowel, the gestures on the consonantal tier can overlap the schwa gesture so far as to completely hide it, giving the impression of deletion of the unstressed syllable.

Data As for experimental data, an increase in coarticulation and reduction in rapid, fluent speech has been shown in a number of acoustic and articulatory studies (see 2.2.2) and this is consistent with both gestural theory and Lindblom's hyper-/hypo-speech account.

The proposition that position dependent allophonic variations are continuous rather than categorical changes, is supported by experimental data on contrast neutralization (in generative phonology contrast neutralization is accounted for by rules that delete the feature(s) responsible for the contrast). Beckman and Shoji (1984), in an acoustic-perceptual experiment on vowel contrast neutralization in devoiced syllables in Japanese, show that contrast is not completely neutralized: listeners in fact are able to recover the underlying vowels /i/ and /u/, possibly from coarticulatory information present in the preceding consonant. Port and O'Dell (1985), in an acoustic-perceptual study on neutralization of voicing contrast in word-final obstruents in German, showed that voicing is not completely neutralized and that listeners are able to distinguish the voiced from the voiceless consonants with better-than-chance accuracy.

Also the majority of English data on alveolar-velar place assimilation in connected speech is consistent with the proposition that the nature of changes is gradient. EPG studies on VC₁#C₂V sequences, where C₁ is an alveolar stop, (Kerswill and Wright, 1989; Wright and Kerswill, 1989; Nolan, 1992) show an intermediate stage between absence of assimilation and complete assimilation, which the authors refer to as residual alveolar gesture. It is also shown that the occurrences of partial and complete assimilations increase from careful/slow speech to normal/fast speech. Most interestingly, a perceptual study reveals that the rate of correct identification of C₁ decreases (as expected) from unassimilated to assimilated alveolars, but never falls to 0, suggesting that also in the cases of apparently complete assimilation (i.e. absence of alveolar tongue-to-palate contact), listeners can make use of some residual cues to the place distinction. The data are in agreement with the hypothesis that in English the assimilation of alveolars to velars is a continuous process. This is confirmed in a recent investigation on clusters of alveolar nasals + velar stops (Hardcastle, 1994). Some other data, however, (Barry, 1991; Nolan and Holst, 1993) challenge some of the assumptions of gestural phonology. The cross-language study by Barry (1991) on English and Russian alveolar-velar clusters confirms that assimilation in English is a graded process. In Russian, instead, assimilation never occurs when C₁ is an oral stop; when C₁ is a nasal, assimilation may be continuous or categorical, depending on syllabic structure. The data of Nolan and Holst (1993) on /s#f/ sequences, do show intermediate

articulations between two-gesture and one-gesture patterns, as predicted by gestural phonology. Accordingly, the one-gesture or static patterns should reflect complete spatio-temporal overlap, i.e. show a blending of the /s/ and /ʃ/ influences and a duration comparable to that of a single consonant. Contrary to this hypothesis, the preliminary results indicate that the static patterns have the spatial characteristic of a typical /ʃ/, and are 16 per cent longer than an initial /ʃ/. Recent EPG and durational data on Italian clusters of /n/ + oral consonants of different places and manners of articulation, seem to indicate that both categorical and continuous processes may coexist in a language, the occurrence of the one or the other depending on cluster type and individual speech style. Moreover, the finding that in Italian the alveolar-velar assimilation in /nk/ clusters is always categorical, indicates, in agreement with Barry (1991), that the assimilatory process for the same cluster type may differ qualitatively across languages (Farnetani and Busà, 1994).

3 Summary

This excursus on the problem of contextual variability shows, on one hand, the incredible complexity of the speech production mechanism, which renders the task of understanding its underlying control principles so difficult. It shows, on the other hand, the enormous theoretical and experimental ongoing progress, as reflected in continuously evolving and improving models, and in increasingly rigorous and sophisticated research methodologies.

We started with the questions of the origin, function and control of coarticulation, and the consequent question of the relation between coarticulation and other context dependent processes. At the moment there is no single answer to these questions. For generative phonology, assimilations, connected speech processes and coarticulation are different steps linking the domain of competence with that of performance, with no bottom-up influences from the physical to the cognitive structure of the language. For both the theory of "adaptive variability" and the theory of gestural phonology the origin of coarticulation lies in speech (in its plasticity and adaptability for the former, in its intrinsic organization in time for the latter). Both theories assume that the nature of speech production itself is at the root of linguistic morpho-phonological rules, which are viewed as adaptations of language to speech processes, sometimes eventuating in historical sound changes. However, there is a discrepancy between the two theories on the primacy of production vs perception in the control of speech variability. Gestural phonology considers acoustics/perception as the effect of speech production, whilst the theory of "adaptive variability" sees acoustics/perception as the final cause of production, hence perception itself contributes to "shape" production.

Two general control principles for speech variability have been repeatedly advocated: economy (by Lindblom and by Keating) and output constraints (advocated by Lindblom for the preservation of perceptual contrast across

styles within a language and extended by Manuel to account for interlanguage differences in coarticulation). Manuel's proposal is supported by a variety of V-to-V coarticulation data and by other cross-language experiments (for example, the lip protrusion data of Lubker and Gay, 1982). But other data do not seem to be consistent with this account: anticipatory nasalization has a limited extent not only in French, which has nasal vowels in its inventory (Benguerel et al., 1977a), but also in Italian (Farnetani, 1986), and Japanese (Ushijima and Hirose, 1974), which have a restricted number of oral vowels and no contrastively nasal vowels. Probably other principles are at work besides that of perceptual distinctiveness – among them, that of preserving the exact phonetic quality expected and accepted in a given language or dialect.

If we confront the various articulatory models with experimental data, it seems that the overall results on coarticulation resistance are more consistent with the gestural model than with others, although certain patterns of coarticulation resistance (Sussman and Westbury, 1981; Engstrand, 1983) could be better explained if aerodynamic/acoustic constraints, in addition to articulatory constraints, were taken into account. The challenging hypothesis of gestural phonology that connected speech processes are not substantially different from coarticulation processes (i.e. are continuous and do not imply qualitative changes in the categorical underlying units) is supported by a large number of experimental results. However, recent data, based on both spatial and temporal parameters, indicate that assimilation can also be a rule-governed categorical process.

As for anticipatory coarticulation, no model in its present version can account for the diverse results within and across languages: the review shows that languages differ both quantitatively and qualitatively in the way they implement this process. English and Swedish seem to differ quantitatively in lip rounding anticipation (Lubker and Gay, 1982); while the plateau-patterns observed in some languages (Boyce, 1990; Cohn, 1993) suggest that the process is phonological in some languages and phonetic in others.

Most intriguing in the data on anticipatory coarticulation are the discrepancies among the results for the same language (on nasalization in American English: cf. Moll and Daniloff, 1971 vs. Bell-Berti and Harris, 1980, vs. Solé and Ohala, 1991). Such discrepancies might be due to different experimental techniques, or the different speech material may itself have conditioned the speaking style or rate and hence the coarticulatory patterns. The discrepancies might also reveal actual regional variants, suggesting ongoing phonetic changes, yet to be fully explored.

NOTES

I thank Michael Studdert-Kennedy for substance of this work, and Björn Lindblom, Daniel Recasens and Bill

Hardcastle for their valuable suggestions.

This research was supported in part by ESPRIT/ACCOR, WG 7098.

- 1 This position is shared by many other speech scientists (see Ohala, THE RELATION BETWEEN PHONETICS AND PHONOLOGY).
- 2 See Fourakis (1991) for a recent review of the literature on vowel reduction.
- 3 Notice the different definitions of the term *phonetic* given by Keating and

by Cohn on the one hand, and by Solé and Ohala (1991) on the other: according to the former phonetics pertains to competence, it is knowledge of the phonetic aspect of a language; for the latter phonetics pertains to the speech mechanism, therefore what is phonetic is unintentional and automatic. Thus the actual data on nasalization in English (Cohn's and Solé and Ohala's) may be not so diverse as the two different accounts would suggest.

13 Theories and Models of Speech Production

ANDERS LÖFQVIST

*"The purpose of models is not to fit the data
but to sharpen the questions"*

Samuel Karlin

(11th R.A. Fisher Memorial Lecture,
Royal Society, 20 April 1983)

1 The speech signal and its description

For the purpose of this chapter, it is convenient to view speech as audible gestures. A speaker creates variations in air pressure and air flow in the vocal tract by making valving actions with different parts of the vocal tract: the glottis, the velum, the tongue, the lips, and the jaw. The changes in pressure and flow give rise to the acoustic signal that we hear when perceiving speech. Most of the variations in the acoustic signal are made intentionally by the speaker to convey linguistic information. Other properties convey what is called paralinguistic information, such as attitudes and emotions, social and geographical dialect characteristics. In addition, there are properties reflecting biological features of the speaker such as sex and age. The resulting acoustic signal is thus shaped by contributions from many different sources that are all overlaid on each other. The fact that listeners can usually identify these different sources suggests that they are recoverable from the acoustic signal.

In describing speech and language, it is common to use one of two modes that can be referred to as the linguistic and the dynamic mode (see Pattee, 1977, for a further elaboration of this distinction). In the linguistic mode, the units of language are described without a temporal domain. For example, most phonological descriptions use a set of symbols that can be arranged in different ways to produce different messages. Although the primitives used for this type of analysis vary depending on the theoretical framework being adopted, the units are commonly described as being discrete and serially ordered. The dynamic mode is used for describing articulatory and acoustic properties of speech. Here, the focus is on the time-varying properties of articulatory movements and/or the spectral characteristics of the speech signal.