Chapter 11
Articulating

Fluent articulation is probably man's most complex motor skill. It involves the coordinated use of approximately 100 muscles, such that speech sounds are produced at a rate of about 15 per second. These muscles are distributed over three anatomically distinct structures: the *respiratory*, the *laryngeal*, and the *supralaryngeal*. The respiratory system, with the lungs as its central organ, regulates the flow of air, the source of energy for speech production. The laryngeal structure, including the vocal cords, is responsible for the alternation between voicing and nonvoicing and for the modulation of pitch. The supralaryngeal structure or *vocal tract*, with the velum, the tongue, the jaw, and the lips as its major moveable parts, exercises two functions in articulation. The first is to constrict or interrupt the air flow in particular ways so as to produce fricative, plosive, and other consonants. The second is to serve as a resonator, modulating the timbre of the successive speech sounds. The timbre depends, in particular, on the shape of the oral, nasal, and pharyngeal cavities. The second section of this chapter presents a short review of these three vocal organs and of their roles in the articulatory realization of the phonetic plan.

Almost all vocal organs, from the lungs to the lips, subserve other functions than speech alone. The respiratory system's main function is breathing—the uptake of oxygen from air and the emission of waste products such as carbon dioxide and vapor. The larynx, by glottal control, protects the respiratory system from intrusions of food. The supralaryngeal structures are used in the mastication and swallowing of food. Though largely the same musculature is involved in the production of speech, the pattern of coordination is totally different. Theories of speech articulation have to account for this "speech mode" of coordination. They specify the nature of speech motor control, the way in which phonetic plans are executed by the vocal organs. The third section of this chapter reviews some of the major theories of speech motor control.

Though these theories are quite divergent, there is rather general agreement about the relatively invariant or context-free nature of phonetic or articulatory plans. It is the executive motor system that realizes the intended articulatory target depending on the prevailing context.

The Articulator is special as a processing component in that it does not map an input representation onto an output representation. Rather, it *executes* its input representation—an utterance's phonetic plan. The result is a motor pattern, not a mental representation of anything. The phonetic plan, we saw, specifies the articulatory gestures for successive syllables, with all their segmental and prosodic parameters. This plan, the Formulator's output, may become available at a rate that is not exactly tuned to the actual rate of articulation, which is the rate *specified* in the phonetic plan. As a rule, some buffering will be required to keep the phonetic plan (i.e., the motor program) available for execution. We will begin the present chapter by reviewing some work on the management of this so-called Articulatory Buffer, which forms the interface between phonological encoding and articulation.

11.1 Managing the Articulatory Buffer

The interface of phonological encoding and articulation involves a system that can temporarily store a certain amount of phonetic plan. Chapter 10 suggested that the Phonological Encoder delivers plans for phonological words as smallest units to the Articulator. As the phonetic plan becomes available to the Articulator, it can be incrementally unfolded in terms of motoneural instructions. But there are very strict temporary restrictions on the course of articulation. Sustaining a fluent, constant rate of speaking requires a storage mechanism that can buffer the phonetic plan (the speech motor program) as it develops. It can, presumably, contain a few phonological phrases. Moreover, it *must* contain a minimal amount of program in order for speech to be initiated—probably as much as a phonological word. The present section will discuss some studies that have dealt with the management of this store, in particular the work done by Klapp, Sternberg, and their colleagues.

It has long been known that when single words or digits are read aloud, the voice-onset latency, measured from the onset of the stimulus, increases with the number of syllables in the utterance (Eriksen, Pollack, and Montague 1970). Klapp, Anderson, and Berrian (1973) discovered that this *syllable latency effect* was due not to the input (visual) process but to the preparation of the articulatory response. They first replicated

Table 11.1
Pronunciation latencies for one- and two-syllable words, in msec. (Data from Klapp et al. 1973, 1976.)

Number of syllables	Word naming	Categorization	Picture naming	Simple reaction	Utterance duration
One syllable	518.4	695.6	619.3	310.8	495
Two syllables	532.8	697.4	633.3	312.5	494
Difference	14.4	1.8	14.0	1.7	−1

the syllable latency effect by having subjects pronounce visually presented one- and two-syllable words which contained the same number of letters (e.g., *clock* and *camel*). The word-naming latencies were significantly different by an average of 14 milliseconds (see table 11.1, first column). This could not have been due to a difference in word-perception times, since the difference disappeared in a semantic-categorization experiment where no articulation of the words was required. In the latter experiment, half of the subjects had to say Yes when the word was an animal name (such as *camel*) and No otherwise; the other half were instructed to say Yes when the word was an object name (such as *clock*) and No otherwise. The Yes response latencies, given in the second column of table 11.1, are virtually identical for the one-syllable and the two-syllable words. Here only the *input* words differ in the number of syllables; the subjects' utterances don't. Klapp et al. did get a syllable latency effect when the stimulus was a picture (of a clock, a camel, etc.) to be named. The latencies for this condition are given in the third column of the table. Two-syllable names took, on the average, 14 msec longer to be initiated than one-syllable names. This is the same latency difference as was found for the reading of printed words. Similar syllable latency effects have been found in the reading of digits. Reading four-syllable numbers (e.g. 27) goes with longer voice onset latencies than reading three-syllable numbers (e.g. 26); see Klapp 1974 for experimental data and further references.

Where do these latency differences arise in the preparation of the articulatory response? Do they come into being *before* the phonetic plan is delivered to the Articulatory Buffer? Or are they rather articulatory in nature? That is, do they come about when the phonetic plan for the word is retrieved from the Articulatory Buffer and "unpacked" to be executed? I will argue that much of the syllable latency effect arises before the delivery of a word's plan to the buffer. But I will subsequently discuss evidence that the size of a phonetic plan (though not necessarily its number of syllables) also affects the latency of its retrieval from the buffer. In

addition, there is a small but consistent number-of-syllables effect in the unpacking of a retrieved phonetic plan. That evidence comes from the work of Sternberg and colleagues.

It should be remembered that a speaker can prepare a phonetic plan without factually initiating the utterance. Waiting for a traffic light, the speaker can prepare to say "green" as soon as the light changes, and can keep the response in abeyance. When the light turns green, the reaction time can be as short as 300 msec. This is, then, the time needed to initiate the response. Such a response is called a *simple reaction.* Is the syllable latency effect one of phonological encoding, or one of response initiation? In order to test this, Klapp et al. used a simple reaction task. The word was presented on the screen for reading, but the speaker was told not to utter the word until a Go signal appeared, 3 seconds after stimulus onset. This gave the speaker the time to program the response, which he then kept ready in the Articulatory Buffer. When pronunciation latencies were measured from the Go signal, the numbers in the fourth column of table 11.1 were obtained. Under these circumstances there was no difference in pronunciation latency between one-syllable and two-syllable words. The syllable effect, therefore, is a real programming or phonological encoding effect, not an initiation effect.

But what is it that takes more time in the programming of a two-syllable word than in that of a one-syllable word? Were the two-syllable words in the experiments of Klapp and his colleagues simply longer than the one-syllable words, and could this be the reason that their encoding took more time? It is known from experiments with nonverbal motor reactions that longer responses require more preparation time. However, utterance duration cannot explain the syllable latency effect. In a subsequent study, Klapp and Erwin (1976) measured the utterance durations of the one- and two-syllable words of the 1973 study. The values are presented in the final column of table 11.1. There is virtually no difference. This may seem surprising in view of the syllable-dependent duration of utterances discussed in the previous chapter, but subjects may have had the tendency to make individual words about equally long when they pronounced them in a list-like fashion. That they did the same in the 1973 experiment is likely but cannot be taken for granted.

Assuming that response duration cannot have been the cause of the planning difference, Keele (1981) suggested that the difference stems from the hierarchical nature of the motor program (i.e., the phonetic plan). In particular, the Prosody Generator has to establish the relative timing of syllables in a multisyllabic word. This might involve a higher pro-

gramming load for *camel* than for *clock*. It appears from studies with nonverbal responses (e.g. tapping) that the more complex the required timing relations in a response (e.g. the tapping rhythm) the longer the response latencies (Klapp and Wyatt 1976). On this view, the syllable latency effect can be attributed to the extra time needed by the Prosody Generator to compute the durational relations between the syllables of bisyllabic words.

An alternative and simpler explanation is the one proposed in the previous chapter: In the phonetic spellout of a phonological word, syllable programs are addressed one by one, in serial order. Hence, the number of syllables in a phonological word will determine the duration of phonetic spellout. If only plans for whole words are delivered to the Articulator, monosyllabic words will become available for articulation earlier than multisyllabic ones. There is an interesting deviance from Wundt's principle here. The Articulator cannot start working as soon as a word's first syllable has been programmed; it must await the whole word before it can start executing its first syllable's phonetic program.

Let us now turn to latency studies of articulatory unpacking and execution. Sternberg, Monsell, Knoll, and Wright (1978) asked their subjects to pronounce lists of words, usually ranging in number from one to five. The words were visually presented one after another. Then, after a 4-second delay, a Go signal (an illuminated square) appeared on the screen, and the subject had to repeat the list as quickly as possible. This was therefore a simple reaction task. The subject had only to retrieve a prepared phonetic plan from the Articulatory Buffer and to initiate its execution, just as in the simple-reaction-task condition of Klapp et al. (1973). A major experimental question was whether the number of items in the buffer would affect the voice-onset latencies, measured from the Go signal.

Sternberg et al. used all sorts of lists—for instance, weekdays in normal or random order, digits in ascending or in random order, and lists of nouns. In all cases the result was essentially the same: As the number of items in the list increased, the voice-onset latency increased by about 10 msec per additional item. Initiation of pronouncing a one-word "list" took about 260 msec; for a two-word list the latency was 270 msec; for a three-word list it was 280 msec, and so on.

Sternberg and colleagues interpreted this result as a *retrieval* effect. The Articulatory Buffer, they supposed, is like a pot containing the items, each with an order number. To retrieve item 1, the speaker draws an item at random and inspects it to see whether it is item 1. If it is not, he draws

another item for inspection, and so on until item 1 turns up. At this moment the item's phonetic plan is unpacked, and the commands are issued to the neuromotor apparatus. All tested items, including the correct one, are dropped back into the pot, and the search for the next item begins. On this model, the average time to find word 1 on the list obviously depends on the number of items in the buffer. When there is only one, it is retrieved on the first draw; when there are two, retrieval requires one or two draws (1.5, on average); when there are three it takes an average of 2 draws; and so on. If a draw takes 20 msec, each additional item on the list will increase the mean voice-onset latency by 10 msec.

This model makes a further interesting prediction: that retrieving the second item will take just as much time as retrieving the first, because all the items were dropped back into the pot. Retrieving item 2 is just as complicated as retrieving item 1. In particular, it will depend in the same way on the number of items on the list. Each additional item will add 10 msec to the average retrieval time of item 2. In fact, this will hold for every item on the list. Therefore, speaking will be slower for a long list than for a short list. The speaking duration per item will increase by 10 msec for every additional word on the list. And this is almost exactly what was found. Sternberg and colleagues showed, moreover, that these increases of 10 msec, 20 msec, and so on were affecting the final parts of the words. We will return to this observation shortly.

In one experiment, Sternberg et al. compared the subjects' performances on lists of one-syllable words and lists of two-syllable words. They matched the words carefully (e.g., *bay* with *baby*, *rum* with *rumble*, and *cow* with *coward*). A first finding in this experiment was that, for all list lengths (1, 2, 3, and 4), the voice onset for lists of one-syllable words was about 4.5 msec shorter than that for lists of two-syllable words. Notice that this differs from the results of Klapp et al. given in column 4 of table 11.1. Their one-word "lists" showed the same simple reaction times for one- and two-syllable words. This difference has never been satisfactorily explained, and I will not add to the conjectures. Sternberg et al. speculated that, having *retrieved* item 1 from the buffer, the speaker has to *unpack* it further to make its constituent motor commands available for execution. This unpacking depends on type and size. A two-syllable word, for instance, involves more unpacking than a one-syllable word. And if the list begins with a two-syllable word rather than with a one-syllable word, unpacking the first syllable's plan will require a few additional milliseconds. This is because *some* unpacking of the second syllable is to be done before articulation of the first syllable can be initiated.

Taken together, the syllable latency effect seems to have a double origin. Klapp's 14-msec syllable effect is one of phonological encoding, whereas Sternberg's 4.5-msec effect is one of unpacking. Sternberg's theory of the Articulatory Buffer says that it becomes loaded with a hierarchically organized phonetic plan or motor program (see also Gordon and Meyer 1987). The units of this program are the words in the list (or, rather, the phonological phrases, as will be discussed). Each unit in the buffer consists of fully specified subprograms for its syllables and their constituent phones. The *retrieval* from the Articulatory Buffer involves complete buffer units—i.e., full phonetic plans for the words (or phrases) in the list. After retrieval of a unit, its phonetic plan or motor program has to be *unpacked* so that all its motor commands become available for execution. This takes more time for a more complex unit than for a simple unit, more time for a two-syllable word than for a one-syllable word, and perhaps—at the next level—more time for a word beginning with a consonant cluster than for one beginning with a single consonant.

An untenable alternative view would be that the Articulatory Buffer contains word-level *addresses* (equivalent to our lemma addresses) but no further phonetic plan. Upon retrieving a unit, the word's address would be opened (roughly equivalent to our spellout procedures), making its articulatory plan available. On this view the buffer would not be an articulatory one. In order to reject this theory, Sternberg, Monsell, Knoll, and Wright (1980) compared utterance latencies and durations for lists of words and lists of nonwords. If the units in the buffer are nonwords whose articulatory programs are still to be constructed after retrieval (instead of being spelled out from store), one would expect a relatively long voice-onset latency for the uttering of the first item and a relatively long duration for the uttering of the list as a whole. The lists of nonwords were phonotactically carefully matched to the lists of words. The experimental procedure was in critical respects the same as in the earlier study. The results for onset latencies and utterance durations turned out to be almost indistinguishable for word lists and nonword lists. Sternberg et al. concluded that in both cases the buffer contained fully assembled programs for all units in the list, whether words or nonwords. Hence, it is a genuine *articulatory* buffer.

According to Sternberg et al., the stages of programming (i.e., phonological encoding), retrieval from the buffer, and unpacking are, finally, followed by a *command and execution* stage. Here the motor commands are issued to the neuromotor machinery, and the response is executed. A

word's duration is its execution time. Sternberg, Wright, Knoll, and Monsell (1980) asked themselves: Is the execution time affected by the retrieval time, or is a word's execution time simply a fixed quantity? When the processes of retrieval and execution are completely disjunct in time, one would expect a short silence before each subsequent word of the list is uttered. The length of this pre-word pause would depend on the number of items in the list; there would be an additional 10 msec for each additional item.

Are these multiples of 10 milliseconds in extra retrieval time indeed projected on pauses between words? They are not. What speakers do is expand the final part of the previous word; they "cover up" the retrieval time by lengthening the execution of the utterance. Sternberg et al. carefully analyzed what happened to lists of two-syllable words. It turned out that the retrieval times were almost completely absorbed by the words' second syllables. The second syllable of the word *copper* was longer when the word appeared in the middle of a five-word list than when it occurred in the middle of a three-word list, but the first syllable was just about equally long in the two cases. The obvious interpretation is that the retrieval process takes place just before the next word is uttered, and that fluency of speech is achieved by stretching the final part of the previous word. This supports Selkirk's (1984a) notion that silent beats can be realized as much by syllable drawl as by pausing.

It was mentioned above that phonological phrases rather than words are the motor units in the Articulatory Buffer. Sternberg called them *stress groups*. This idea, which arose in the 1978 study by Sternberg et al., was based on some further experimental results. In one experiment Sternberg et al. interpolated function words between the nouns of a list. The list *bay-rum-cow*, for instance, would be presented as *bay and rum or cow*. Would this "count" as a five-item list, or as a three-item list? Analysis of the data showed that it behaved like a three-item list. Since there were three stressed words in each of the lists, the conjecture was made that the motor planning units in the Articulatory Buffer are, in fact, "stress groups." A stress group here is nothing but a small phonological phrase containing just one stressed element, for instance *and rUm*. The conjecture is, therefore, fully consonant with the notion, developed in chapter 10 above, that phonological phrases are important units of phonological encoding. It is likely that the buffer is successively filled with phonological words, but that larger phonological phrase units are formed when the buffer is heavily loaded.

Table 11.2
Phases in speech motor control.

Stage 1: Assembling the program
This is the stage of phonological encoding, with a phonetic plan as output (see chapters 8–10). The phonetic plan is a detailed motor program, delivered phonological word by phonological word. When the task requires, phonetic plans can be stored in the Articulatory Buffer. The preferred units of storage are phonological phrases.

Stage 2: Retrieving the motor programs
When the speaker decides to start a prepared utterance, its motor units (i.e., the phonetic plans for the phonological phrases) are retrieved from the Articulatory Buffer. The time needed to retrieve each unit depends on the total number of units in the buffer.

Stage 3: Unpacking the subprograms
Once retrieved, the phonetic plan for a phonological phrase has to be unpacked, making available the whole hierarchy of motor commands. The more complex a motor unit, the more time unpacking takes.

Stage 4: Executing the motor commands
At this stage the motor commands are issued to the neuromotor circuits and executed by the musculature. Syllables can be drawled to absorb retrieval latencies.

The picture emerging from these studies is summarized in table 11.2.

To what extent is this picture valid for spontaneous speech? Of course, people do reproduce lists now and then in everyday life (for instance, telephone numbers). It should, in addition, be noted that in the experiments the lists were uttered as prosodic wholes—as intonational phrases with normal declination and boundary tones (Sternberg, Wright, Knoll, and Monsell 1980). There can, moreover, be no doubt that stages 1 and 4—phonological planning and execution—are always part of normal speech. The question is, rather, how much buffering and unpacking has to be done in normal fluent speech.

Clearly, a speaker can start uttering a phonological phrase before all its details have been programmed. This is apparent from cases of prelexical hesitation. A speaker may have to stop in the midst of a phonological phrase that has begun with quite normal prosody, as in *I saw him in ... eh, in ... eh, in Vallauris.* Here the full program for the place name was, clearly, not yet assembled, let alone buffered, when it was needed for execution. Still, the preposition of the phonological phrase (*in*) was uttered normally. Execution can follow phonological encoding at a very short distance, a distance smaller than a full phonological phrase. This distance is probably the size of a phonological word (the smallest "chunk" delivered by the Phonological Encoder), and buffering will be minimal or

absent. On the other hand, grammatical and phonological encoding may occasionally go through a speedy phase, so that a greater amount of ready-made program becomes available than can be executed at a normal speaking rate. Articulatory buffering is an important facility under such circumstances.

11.2 The Vocal Organs and the Origins of Speech Sounds

The execution of a phonetic plan involves the coordinated use of a highly complex musculature. Figure 11.1 depicts the structures involved in speech production. The discussion below will follow the figure's partitioning into respiratory, laryngeal, and supralaryngeal structures.

11.2.1 The Respiratory System

In normal breathing, the lungs contain some 3 liters of air. We inhale and exhale about half a liter at a time. In speech, far more air can be exhaled at a time; 3.5 liters is not abnormal. This, of course, requires deeper inhalation. The inhalation during speech is quick; taking up no more than 15 percent of the breathing cycle (versus 40 percent in normal breathing). In speech, most of the respiratory cycle is spent on exhalation, which can easily take 10 or 15 seconds (versus 3 seconds in normal breathing).

Inhalation and exhalation are controlled by various muscles in the thorax and the abdomen. When the inspiratory muscles contract, the cavity enclosed by the ribs increases in volume, and the resulting pressure gradient causes air to flow into the lungs. During normal breathing, exhalation is mainly brought about by relaxing the inspiratory muscles. The elastic shrinking back of the thorax is enough to create the slight overpressure necessary for expiration. In speech, however, the inspiratory muscles keep being innervated during the initial phase of exhalation, holding back the air, so to say. Then they suddenly relax, and the expiratory muscles of the thorax take over to compress the volume even more. Still later during the exhalation or speaking phase the abdominal muscles may start contracting. As a result, the diaphragm is pushed upward into the thoracial cavity, decreasing its volume even more. This complex interplay of muscular activity during exhalation causes a rather constant air pressure during speech production. Still, there is a slightly decreasing slope in this pressure. It is the main cause of pitch declination in the course of an utterance (see subsection 10.3.1).

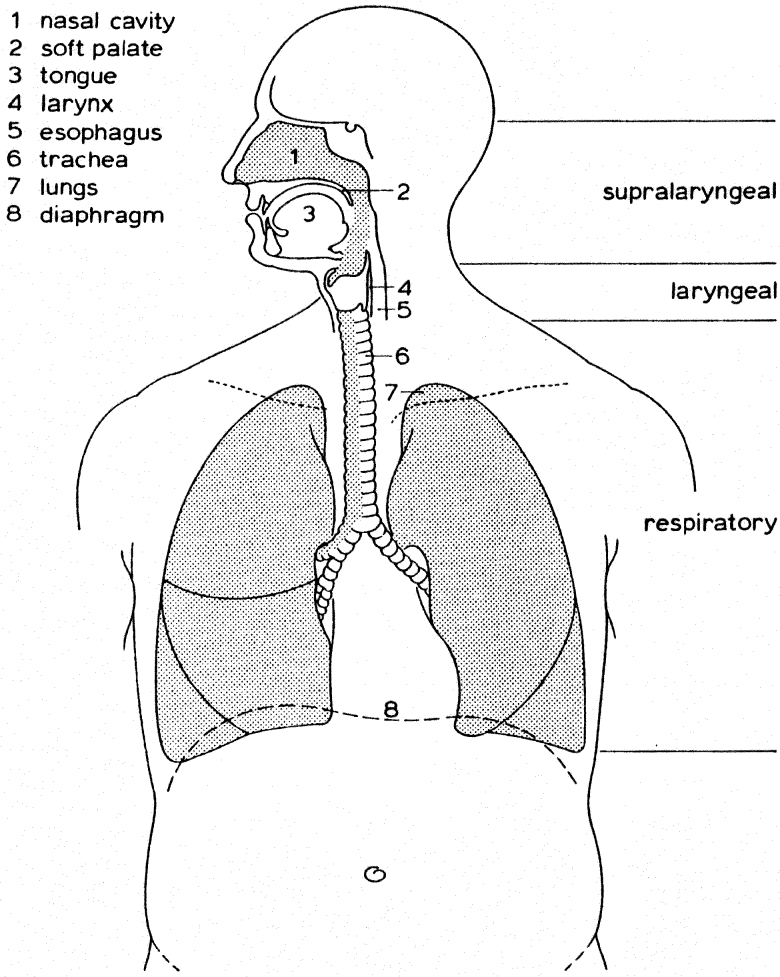

Figure 11.1
The respiratory, laryngeal, and supralaryngeal structures involved in speech production.

11.2.2 The Laryngeal System

The larynx is responsible for phonation in speech, not only for normal voicing but also for whispering and for other less common registers (such as a man's falsetto). Figure 11.2a presents a posterior view of the larynx. It sits on top of the trachea, a tube connecting it to the lungs. The larynx as a whole can be moved up and down and forward and backward by various extrinsic muscles attached to the mandible, the skull, and the thorax. These movements can easily be traced by touching the Adam's apple, the protruding part of the thyroid cartilage; they are especially pronounced during swallowing. At the top of the larynx is the epiglottis, which can cover the larynx's exit. This is done at moments of swallowing, when food is transported from the mouth to the esophagus and the stomach. At these moments two other laryngeal closures are made as well: The glottis is shut by the vocal folds, and the false vocal cords (slightly above the glottis) are moved together so as to make a firm closure. During speech and normal breathing, the false vocal cords are wide apart.

The centerpiece of the larynx is the structure around the *vocal folds* (also called *vocal cords*). Figure 11.2b gives a sagittal view of this part. The *glottis* is the area between the vocal folds. It can be opened or closed. The vocal folds, each about 2 centimeters long, can be pulled apart at the posterior side to make an angular opening. They cannot move at the anterior side, where they are both attached to the *thyroid cartilage* (directly behind the Adam's apple). But they can be drawn apart at the posterior side, because each is attached to an *arytenoid cartilage* and these two cartilages can be abducted or adducted by sets of muscles attached to them (the posterior and lateral *cricoarytenoid muscles*, respectively). The vocal folds themselves are also largely muscle tissue, except for where they touch (and maximally vibrate); these parts of the folds are ligaments. There are two kinds of muscles in the folds: (i) The *longitudinal thyromuscularis* shortens the fold when it contracts; the arytenoid cartilage, to which it is attached, is accordingly displaced in the forward direction. The antagonist muscles, which pull the cartilages back into place, are the *crycothyroid muscles*. Their contraction causes the folds to become longer and more tensed. (ii) The *vocalis* is attached to the ligament tissue, and can influence the curvature of the ligament.

Voicing occurs when the folds are pulled together while air pressure is built up by the respiratory system. When the pressure is sufficiently high, the folds burst apart and release a puff of air. This, in turn, causes a temporal reduction in subglottal air pressure, which makes the glottis close

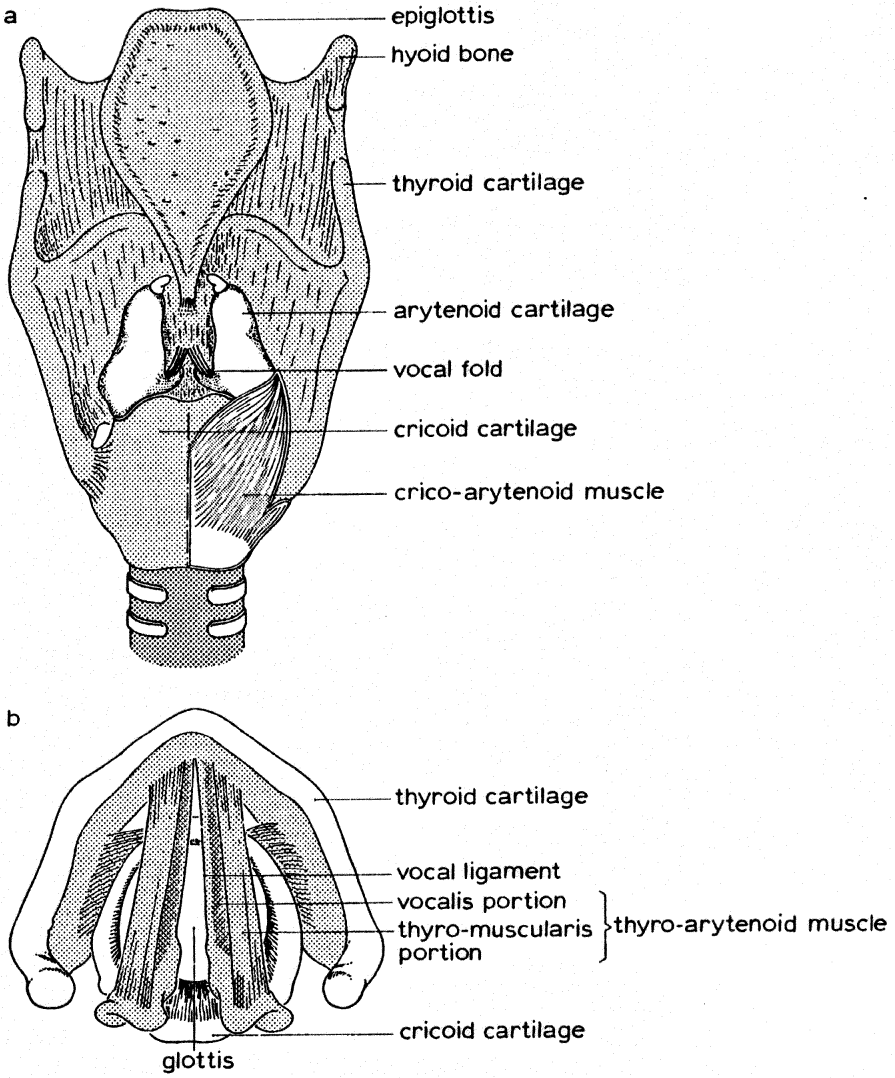

Figure 11.2
(a) A posterior view of the larynx. (b) A superior view of the vocal folds and the cartilages they are attached to. (After Calvert 1980.)

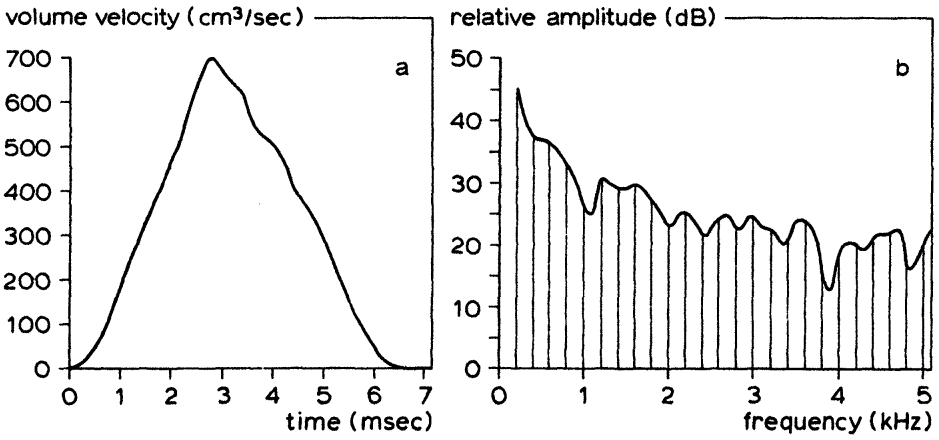

Figure 11.3
(a) Increase and decrease of air flow in a single glottal puff. (After Denes and Pinson 1963.) (b) Spectral analysis for a 200-Hz glottal puff pattern.

again. The cycle repeats itself rhythmically (this is called the "Bernoulli effect"), and the glottis releases a periodic sequence of puffs of air. The average frequency of these puffs is about 200 Hz in a woman's voice and 110 Hz in a man's. The actual frequency at any one moment is called the speech sound's *fundamental frequency*, or F_0.

During one puff, the outflow of air first increases almost linearly, then decreases again the same way. Figure 11.3a shows this pattern for a single puff. If one could listen to regular repetitions of this pattern alone, it would resemble the sharp sound of an oboe reed, not the smooth sound of a tuning fork. The latter sound is created when the air displacement is sinusoidal. Physically speaking, the sawtooth pattern of figure 11.3a when continuously repeated, can be constructed as the sum of a set of sinusoidal components: The same sound would be produced by a battery of tuning forks of the following sort: a big tuning fork vibrating at frequency F_0 (say, 200 Hz); a somewhat smaller tuning fork, precisely an octave higher (i.e., vibrating at 400 Hz); a still smaller fork, vibrating a fifth higher (i.e., at 600 Hz); and even smaller forks at 800 Hz, 1,000 Hz, 1,200 Hz, and so on. (One can ignore the very small forks vibrating at frequencies of more than 5,000 or 6,000 Hz.)

Figure 11.3b presents the sound intensity in decibels produced by each tuning fork, one bar for each fork. The intensity is high for the 200-Hz fork, and it decreases for the higher-frequency forks. Since the regular string of glottal puffs is precisely imitated by this battery of tuning forks, figure 11.3b can be seen as the *spectral analysis* of the sawtooth vibration pattern consisting of puffs such as in figure 11.3a. In other words, it is the

spectral analysis of a 200-Hz vibration produced by the glottis. The sound consists of a pure (sinusoidal) 200-Hz tone plus decreasing amounts of each of its overtones (400, 600, and 800 Hz, etc.). These higher components give the sound its sharp timbre. The fundamental frequency component (200 Hz in the present example) is also called the first *harmonic*; the first overtone (400 Hz) is the second harmonic, and so on. In figure 11.3b the intensity for each harmonic is presented by a vertical bar. The undulating line connecting the bars is called the *spectral envelope*. When the glottis vibrates at a different frequency than 200 Hz, the spectral envelope is by and large the same; only the spacing of bars varies. The spectral envelope is a useful characterization of the *timbre* of a periodic speech sound, irrespective of its pitch.

The pitch contours of speech are realized by varying F_0. The vibration frequency of the vocal folds can vary over a range of about two octaves (professional singers can do much better), but it usually doesn't surpass one octave in normal speech. This frequency is a complex function of various factors. F_0 covaries with subglottal pressure. This is, as we have seen, a main cause of pitch declination in speech. F_0 is also—and more substantially—affected by the length and the tension of the vocal cords. These two factors have opposite effects: lengthening decreases F_0, tensing increases F_0. When the crycothyroid muscles stretch the folds, there is both lengthening and tensing, but the tensing effect overrides the lengthening effect, just as when one stretches an elastic band. As a consequence, there is an increase in the fundamental frequency. The small muscles controlling the tension of the folds can adjust far more rapidly than the big inspiratory and expiratory muscles. Thus, the fine, speedy pitch movements in speech depend mainly on laryngeal muscles.

The loudness of speech is determined largely by the intensity of vocal-fold vibration. This intensity depends, in part, on subglottal pressure. The higher the pressure, the faster the flow of air and the louder the speech sound. The intensity of vibration also depends, to a substantial degree, on the size of the glottal opening. It is not strictly necessary for the glottis to be totally closed at the moments between the air puffs. The Bernouilli effect will also arise when there is a slight V-shaped opening between the folds. As a consequence, some air will escape without transmitting its energy to the folds, and the vibration is weakened. The difference in energy expenditure between loud and soft speech, measured in volumes of air displaced per unit time, is probably quite small.

In whispering, the glottis is opened so much that no periodic vibration of the folds occurs any more, but it is still so narrow that a hissing noise

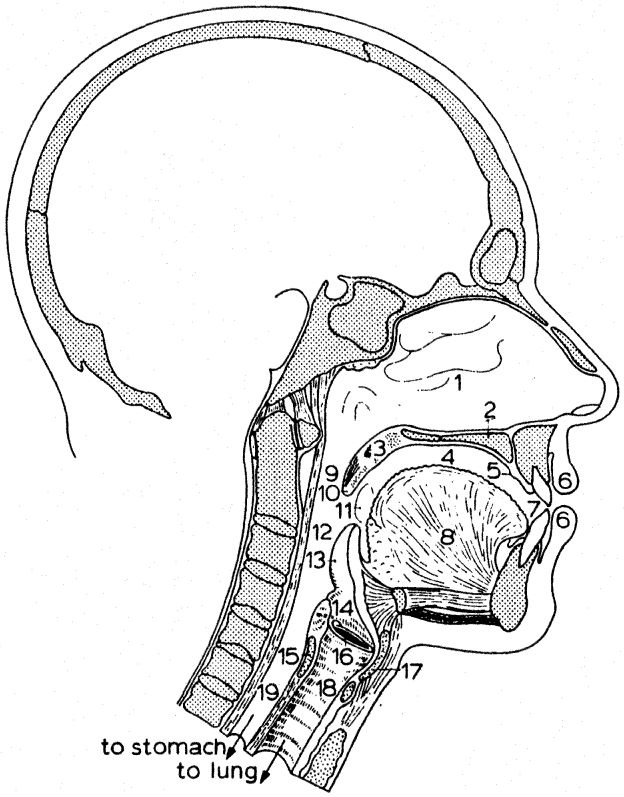

Figure 11.4
The vocal tract. (After Calvert 1980.)

results when the air passes through it. When speech is articulated this way, the voiced parts of speech are replaced by "hissed" parts.

11.2.3 The Vocal Tract

The supralaryngeal system, or vocal tract, consists of the structures between the epiglottis and the lips and nose. Figure 11.4 depicts these structures.

The vocal tract consists of three main cavities: the *pharynx* or *throat*, the *oral cavity*, and the *nasal cavity*. The pharynx and the oral cavity are flexible in shape; the nasal cavity is a rather fixed structure. The size and shape of these three cavities determine the *resonance* properties of the vocal tract. The place and manner of constricting the outflowing air stream determine the proper *articulation* of speech segments.

Resonation

Each of the three cavities can resonate with the buzzing sound produced by the vocal folds. Consider the nasal cavity. It participates in shaping the timbre of a speech sound when there is an open connection with the

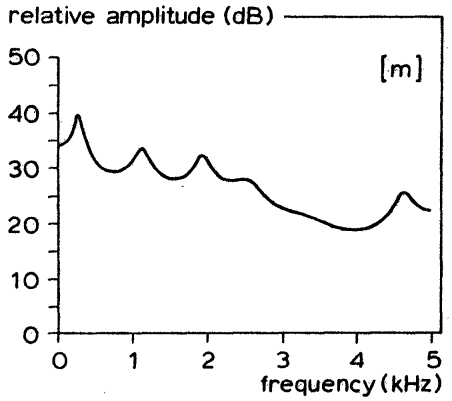

Figure 11.5
Spectral envelope for the nasal sound [m].

pharynx. This occurs when the *velum* or *soft palate*, a very flexible organ, is moved forward to open the *nasopharyngeal port*. This is its normal position in breathing. A speech sound produced with this port open has a characteristic nasal timbre. It is an articulatory feature of the consonants [m], [n], and [ŋ], and of nasalized vowels such as that in *chance* or that in French *en*.

This special timbre arises because the nasal cavity affects the energy spectrum of the sound produced by the vocal folds. The spectrum of the buzzing sound produced in the glottis was given in figure 11.3b. There is a string of decreasing intensity peaks, extending from F_0 to about 5,000 Hz. The nasal cavity will dampen or attenuate the energy in the high-frequency ranges and will amplify the energy in the very low range (around 200 Hz). The resulting spectral envelope is like the one given in figure 11.5, which is an analysis of the sound [m].

The nasal cavity is never the only resonator involved in shaping a speech sound's timbre. The *pharynx*, which is the mediating structure between the larynx and the nasal cavity, is necessarily involved in the production of all nasal and all non-nasal sounds. The shape of the pharynx or throat is not fixed. It can, first of all, constrict itself. This happens especially during the peristaltic movement that transports food from the mouth to the esophagus. It can, second, be raised and widened by a special set of levator muscles. And, third, its shape changes depending on the position of the soft palate. Each shape of the pharynx will have its own effect on the timbre of a speech sound.

The resonating properties of the *mouth* depend on the positions of the mandible, the tongue, the lips, and the velum. All of these are independently movable. The *mandible* can be moved up and down, forward and

backward, and sideways. It is mainly the up-and-down movement that is relevant to speech; it can drastically decrease or increase the volume of the oral cavity. When the volume is small, the higher frequencies are amplified; when it is large, the lower frequencies are more prominent.

The *tongue*, probably the most essential organ in the articulation of speech, is a highly flexible instrument. Phylogenetically, its role is to displace food within the mouth, especially during mastication, and to transport liquid and chunks of food to the pharynx. It is also the seat of an important sensory function: taste. The tongue is moved by, and largely consists of, *extrinsic* and *intrinsic* muscles. The extrinsic muscles attach to bones of the skull and the larynx. A large part of the tongue's body is formed by the *genioglossus*, an extrinsic muscle extending from the frontal cavity in the mandible to the back, the middle, and the front of the tongue. It can strongly affect the shape of the oral cavity by retracting or protruding the tongue, and by depressing or lifting it. The intrinsic muscles, which run both longitudinally and laterally through the tongue's body, can affect its finer shape in numerous ways: They can make the tongue longer and narrower, they can widen and flatten the tongue, they can move the tip up or down, and they can make the upper surface concave or convex. All these movements are relevant for resonation and for the articulation of consonants.

The *lips* can affect the timbre of speech most markedly by spreading (as in *pit*) and by rounding (as in *put*). The rounding of the lips, and their protruding, is effected by a circular intrinsic muscle around the mouth opening, the *orbicularis oris*. Other muscles can move the lips in and out, and draw the corners of the mouth up or down. The muscles of the lips and other facial muscles play an important role in the facial expression during speech communication. These expressions can provide visual backchannel signals to the interlocutor (see subsection 2.1.2).

The *velum* is the only movable part of the mouth's *palate*. The palate consists of the teeth ridge or *alveolus*, the *hard palate*, which forms the roof of the oral cavity, and the *soft palate* or velum. The velum's main function in speech, we saw, is to open and close the nasal cavity. In doing so, it also affects the shape of the mouth.

The shapes of the three vocal-tract cavities—in particular, the mouth—determine the characteristic timbres of a language's vowels. The main oral contributors to a vowel's timbre are the positions of the tongue and the lips. As far as the lips are concerned, it is especially their rounding (and their slight protruding) that matters. The tongue's contribution to the sound quality of vowels depends largely on the activity of the extrin-

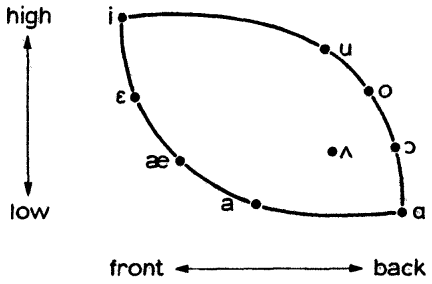

Figure 11.6
Tongue positions for eight vowels. (After Denes and Pinson 1963.)

sic muscles that regulate the position of the body of the tongue, which can vary between high (close to the palate) and low and between the front and the back of the mouth. These positions, and the English vowels that go with them, are diagrammed in figure 11.6.

The vowels in figure 11.6 differ systematically in their spectral properties. The spectral envelopes of the four vowels [ɛ], [ʌ], [i], and [æ] are given in figure 11.7. Figure 11.7a gives the spectrum for [ɛ]. It can be seen from this figure that the vocal tract resonates especially with frequencies in the ranges of 500, 1,700, and 2,400 Hz. These peaks in the spectrum are called the first, second, and third *formants*, or F_1, F_2, and F_3. (Remember that the vibration frequency of the vocal folds was called F_0.) As Stevens (1983) has pointed out, F_1 and F_2 are typically quite far apart for a front vowel such as [ɛ]. In order to appreciate this, compare the spectrum for [ɛ] with that for the back vowel [ʌ], given in figure 11.7b. Here F_1 and F_2 are quite close together. In other words, front vowels typically show a concentration of resonance in the high frequency range, whereas back vowels have their energy concentrated in the low frequency range.

High and low vowels also differ systematically in their spectral envelopes. Figures 11.7c and 11.7d give the spectra for the high vowel [i] and the low vowel [æ], respectively. Both are front vowels, with the characteristic spreading of F_1 and F_2. Their crucial difference, according to Stevens (1983), is in the position of F_1. High vowels such as [i] have their first formant in the very low frequency range, whereas the frequencies of the first formants of low vowels such as [æ] are substantially higher.

These static pictures of vowel spectra should not give the illusion that vowels do not change their spectral properties during articulation. In fluent speech the characteristic spectrum of a vowel arises only for a short moment, if at all. The degree to which the vocal tract approaches the ideal configuration for a particular vowel depends on the context in

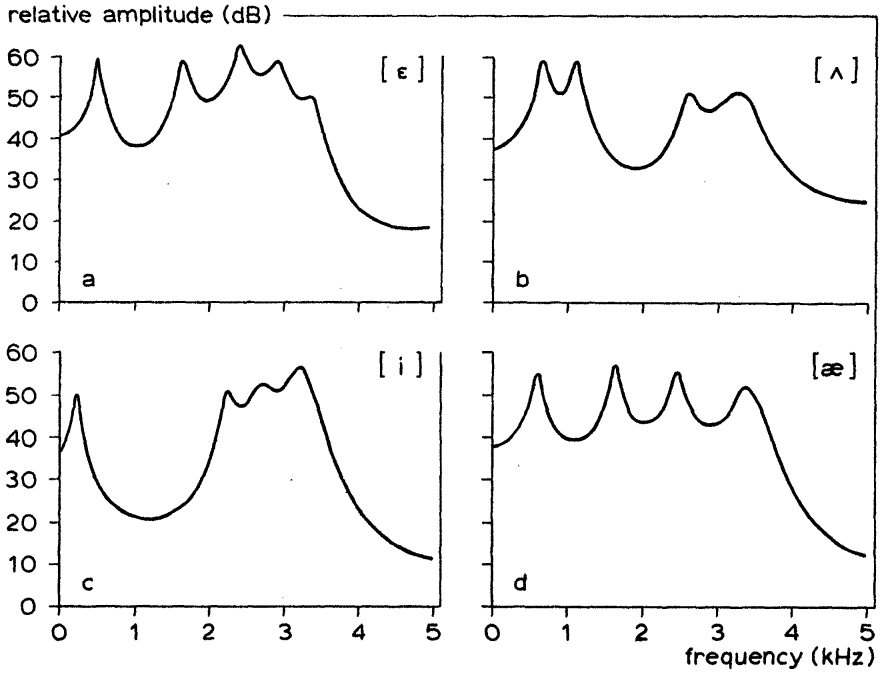

Figure 11.7
(a–d, respectively) Spectral envelopes for the vowel sounds [ɛ], [ʌ], [i], and [æ]. (After Stevens 1983.)

which the vowel appears, and on the rate of speech. There are, moreover, vowels that are characterized by a *changing* spectrum. They are called *diphthongs*. The English diphthongs are [aɪ] as in *night*, [ɔɪ] as in *toy*, [eɪ] as in *sake*, [aʊ] as in [owl], and [oʊ] as in *phone*. The pronunciation of a diphthong involves gliding the body of the tongue from one position to another.

Not only vowels, but also consonants reflect in their sound quality the configuration of shapes in the vocal tract. This is especially apparent in voiced consonants that involve periodic vibration of the vocal chords. Examples are [b], [d], [g], [v], and [z], as well as the nasal consonants [m], [n], and [ŋ]. But unvoiced consonants, such as [p], [t], [k], [f], and [s], also have their own characteristic frequency spectra, which can change rather drastically in the course of their articulation (Schouten and Pols 1979). Every speech sound has a slowly or rapidly changing timbre, which depends on the changing shape (and hence the changing resonance properties) of the vocal tract.

Articulation

The vocal tract can be constricted at different *places* and in different *manners* (see subsection 8.1.5). The most visible place of constriction is the lips; [p], [b], [m], and English [w] are produced with a *bilabial* constriction.

Constriction is also possible between the lower lip and the upper teeth. When a constriction is made in this place, the speech sound is called *labio-dental.* The phones [f] and [v] are the two labio-dentals in English; [w] is a labio-dental in various other languages.

When the place of articulation is between the tongue blade and the upper teeth, the speech sound is called a *dental*; the English consonants [ð] and [θ] are dentals. Still further back, and thus less visible, is the place of articulation for a rather heterogeneous set of speech sounds: [t], [d], [s], [z], [n], [l], [r], and [j]. They are all made with a constriction somewhere along the gums or alveolar ridge; thus, they are called *alveolars.*

When the main constriction of the vocal tract is made somewhere along the hard palate, the speech sound is *palatal*; [ʃ] and [ʒ] are cases in point. For some palatal consonants, the main constriction borders on the alveolar ridge; these speech sounds are called *palato-alveolar*, and among them are [tʃ] and [dʒ]. All speech sounds in which the tongue blade is used in the major constriction, ranging from dental to palatal, are called *coronal.*

Velars have the velum as the place where the main constriction is made. The three velar speech sounds in English are [k], [g], and [ŋ] (but [w] also often involves a secondary or even a main constriction at this place of articulation).

There are languages that involve the uvula, the fleshy clapper-like appendix of the soft palate, in certain consonants. This is, for instance, the case for French [r] and for Spanish, Dutch, and Hebrew [x]. This place of articulation is, correspondingly, called *uvular.* The deepest place of articulation is the glottis. The main constriction for [h] is just there; it is a *glottal* speech sound. All speech sounds with a main constriction behind the alveolar ridge are called *posterior* (as opposed to *anterior*).

Each place of articulation can go with different *manners of articulation.* One manner is to create a momentary but complete closure of the vocal tract at the place of articulation. The built-up air pressure is subsequently released, which creates a plosive effect. Speech sounds of this kind are called *plosives* or *stops.* Examples are English [b] and [p], [d] and [t], [dʒ] and [tʃ], and [g] and [k], which come in pairs, each consisting of a voiced and an unvoiced consonant. The distinction between *voiced and unvoiced* is, therefore, also considered to be a manner aspect of articulation. In their turn, voiced stops can have the additional manner of being *nasalized.* The English nasal stops are, we saw, [m], [n], and [ŋ].

Another manner feature involves rounding of the lips, as in the above-mentioned *put*, where the vowel [ʊ] is *rounded* as opposed to the [ɪ] in *pit*. English [w] is also rounded.

When there is no complete closure but rather a constriction so narrow that audible air turbulence is created, the manner of articulation is called *fricative*, *strident*, or *spirant*. The English fricatives also come in pairs of voiced and unvoiced consonants. They are [v] and [f], [ð] and [θ], [z] and [s], [ʒ] and [ʃ], and [ɦ] and [h]. The latter two are allophones; the voiced [ɦ] appears in intervocalic position, as in the word *Ohio*. The stop consonants [dʒ] and [tʃ] behave like fricatives after the moment when the air is released. In that sense they are hybrids of stops and fricatives; some phoneticians call them *affricatives*. Stops, fricatives, and affricatives are three varieties of *obstruents*.

When the constriction is still less narrow, so that no audible spiration results, very rapid changes of resonance (which resemble diphthongs) can be created. This is the *semi-vowel* manner of articulation. English [j], as in *yet*, and [w], as in *wet*, are semi-vowels, and so is [r] in many English dialects. A very special case is [l], whose manner of articulation is called *lateral*. It involves a temporary central alveolar constriction, with the air passing by laterally.

This section has reviewed the major vocal organs involved in the production of speech. More details can be found in Calvert 1980. The central issue for a theory of articulation is, of course, how a speaker's phonetic plan becomes realized as a coordinated motor activity. If the plan is roughly as suggested in the previous chapter, how are the respiratory, laryngeal, and supralaryngeal muscle systems set to bring about the intended articulatory pattern? The phonetic plan is a motor program at a still abstract level. The string of syllable gestures is parametrized for rhythmic and prosodic variables, for variables of rate and force, and for the precision of articulation to be attained. But in order for this motor program to run, its articulatory features must be realized by the musculature of the three main structures reviewed in this section. The same feature can often be realized in different but equivalent ways. There is a flexibility in motoneural execution that makes it possible to realize a particular articulatory feature in spite of varying boundary conditions. This context dependency of the motor execution of speech has puzzled phoneticians greatly. The next section will review some of the theories that have been proposed to account for this flexibility in the motor control of speech.

11.3 Motor Control of Speech

The speaker's phonetic or articulatory plan, as it eventually emerges from phonological encoding, is strongly based on syllables. A syllable's phones are articulatory gestures whose execution depends heavily on their position in the syllable and on the other phones with which they are more or less coarticulated. If indeed, as I will argue, the syllable is a unit of motor execution in speech, how much articulatory detail is specified in the phonetic plan? A complete theory of motor control will, in final analysis, have to account for the fine detail of neuromuscular activity in speech. There has been a strong tendency in the literature to include all or most of this detail in the articulatory or phonetic plan (also referred to as "motor program"), which then would prescribe the full detail of individual muscle contractions.

More recent developments in motor-control theory, however, have made this view less attractive. Studies of handwriting (van Galen and Teulings 1983; Thomassen and Teulings 1985; Kao, van Galen, and Hoosain 1986) have shown that there are astonishing invariants in the execution of a program under different modes of execution. One well-known example is the constancy in letter shape and writing speed between writing on paper and writing on a blackboard. At the level of individual muscle activity, the neuromotor patterns for these two modes of writing are totally different. But even ignoring such dramatic differences, it is a general property of motor execution that it is highly adaptive to context. Handwriting immediately adapts to the resistance of the paper, just as gait adapts to the resistance of the walking surface. In the same way, the execution of speech motor commands adapts to peripheral context—for instance, a pipe in the mouth.

It is, therefore, attractive to assume that the commands in the articulatory plan involve only the *context-free* or invariant aspects of motor execution, and that the context-dependent neuromuscular implementation of the program is left to a highly self-regulating neuromotor execution system. This system *translates* the program codes into appropriate neuromuscular activity (Gallistel 1980). Without such context-dependent translation, one must assume that a different motor program is prepared for each context of execution. Since contexts of motor execution can vary infinitely, this would involve an immense drain on information-processing resources in the planning of motor activity. In the following, this division of labor between programming and execution will be accepted. But there is a danger in this approach: The neuromuscular execution system is easily

made a *deus ex machina* that will take care of everything the theorist cannot explain. Shifting the account for a motor system's adaptability to a self-regulating executive system creates the obligation to causally explain these self-regulations. However, this distinction does exclude several theories of the nature of speech motor commands—in particular, the following two.

11.3.1 Location Programming

The theory of *location programming* assumes that the program consists of a sequence of target locations for the articulatory musculature. Each phone involves such a set of target positions. The theory is attractively simple. It is known that a muscle's target length can be encoded in the muscle spindle. The alpha/gamma loop, which is a peripheral reflex arc, will then automatically realize this target length, irrespective of the starting length of the muscle (see figure 11.8 for more detail).

To realize a phonetic segment, several muscles will be set for a particular target length. As soon as all the target positions for that phone have been reached, the next set of locations will be commanded, and so on till the end of an utterance's program. The program specifies a sequence of targets, but they are not explicitly timed; there is no *intrinsic timing* in the plan. The duration of moving from one phonetic target to the next depends only on the mechanical properties of the musculature involved, i.e., on executive factors beyond the phonetic plan. This is called *extrinsic timing*.

I argued for intrinsic timing in chapters 8 and 10. At various stages of phonological encoding, durational parameters are set that reflect the utterance's composition in prosodic units (such as phonological words and phrases, intonational phrases, and turns). These parameters are eventually transmitted to the phonetic spellout procedures and implemented in syllable programs. As a result, the phonetic plans for successive syllables are *intrinsically* timed. How this timing is realized will depend on the internal composition of the syllable. A syllable containing many consonants (such as *scratch*) will always be longer than one with few (such as *at*). Such syllable-specific durational properties are part of the stored syllable program. But they will be modified by the higher-level rhythmic parameters. Kohler (1986) provides an explicit model for this two-level intrinsic syllable timing. Though the location-programming theory gives a natural account of syllable-specific differences in duration (e.g., *scratch* takes more time than *at* because there is a longer sequence of target posi-

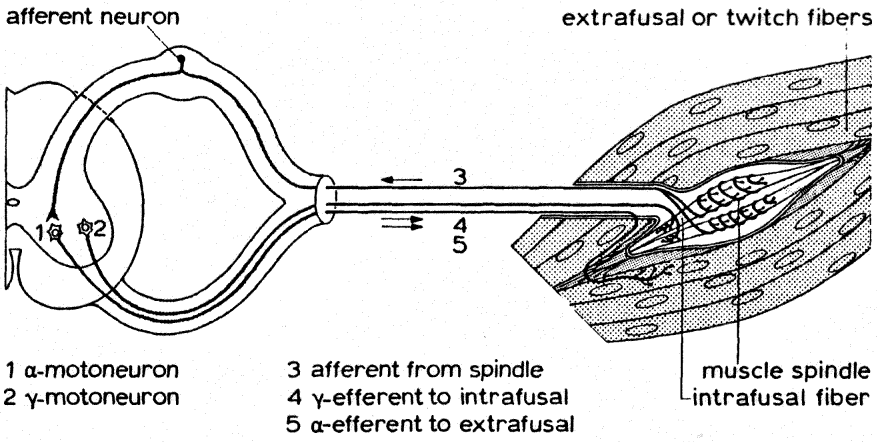

Figure 11.8
The alpha/gamma loop in muscle-tone control. A skeletal muscle's main fibers, the *extrafusal* or *twitch fibers*, are innervated by alpha motoneurons originating in the spine. Interspersed between the twitch fibers are so-called *intrafusal* fibers, whose contraction is controlled by gamma-motoneurons, which also originate in the spine. Wrapped around intrafusal fibers are *muscle spindles*. These afferent neurons fire when the intrafusal fibers are more contracted than their neighboring extrafusal fibers. This "difference information" is returned to the alpha motoneurons in the spine. It causes them to send impulses to the extrafusal fibers, which contract as a result. This causes a decrease of spindle activity, because a better match is obtained between the stretching of extrafusal and intrafusal fibers. The equilibrium point is reached when this match is complete. The alpha/gamma loop allows for fine control of the muscle's target length. The gamma system can, by contracting its low-mass intrafusal fibers, quickly "dictate" a target length to the muscle as a whole. The muscle's twitch fibers—the real mass of the muscle—will, through the alpha/gamma loop, adapt their length to the preset intrafusal fiber lengths. Measurements show that the innervation of a muscle's alpha and gamma neurons is about simultaneous. But the lightweight intrafusal muscle system, which is never loaded, reaches its target length relatively quickly. The extrafusal main body of the muscle adapts more slowly.

tions to be realized), it provides no possibility of implementing the higher-level rhythmic parameters.

The theory can, however, explain an aspect of coarticulation. The spatio-temporal route through which the articulators move in order to reach, say, a vowel's target position in a CVC syllable will depend on the previously reached target position: that of the syllable-initial consonant. Similarly, the route out of the vowel's target configuration will depend on the character of the syllable's coda. So, the vowel gesture [a] will be different in *car* and *father* because it starts and ends differently in these environments.

The theory fails to account for other aspects of context-dependent motor execution. It can, in particular, not deal with *compensatory* movements in speech articulation. If a speaker says *aba*, he moves both his jaw and his lips to realize the closure of [b]. The location-programming theory explains this by specifying the phone [b] as a set of target lengths for the musculature controlling the positions of the mandible and the lips. But it cannot explain the following: If the jaw is fixed in the open position (phoneticians do such cruel things), the lips will still get closed when the speaker is pronouncing *aba*. The speaker compensates for the jaw movement by making a more extensive closing movement with the lips. In other words, a different target position is set for the lips. Nothing in the theory predicts this. We will shortly return to this compensatory behavior, which has been a major argument for developing the notion of context-dependent motor execution.

Nobody entertained the location-programming theory in this idealized form. MacNeilage (1970) came closest, but in later papers (MacNeilage and Ladefoged 1976; MacNeilage 1980) he rejected it.

11.3.2 Mass-Spring Theory

A close relative of the location-programming theory is the *mass-spring theory*, which was developed to explain the control of limb movements (Fel'dman 1966a,b). An experimental test of this theory for pointing gestures, the results of which were essentially negative, is reported in Levelt, Richardson, and La Heij 1985. Lindblom used the theory to explain motor control in speech as early as 1963, and it was later used by Fowler and Turvey (1980) and by Fowler, Rubin, Remez, and Turvey (1980).

In its simplest form, the theory treats the agonist and the antagonist of a limb as stretchable springs that "want" to reach their normal resting position (the *zero position*). The limb will rotate around the joint, and will eventually reach a steady-state position which is determined by the equi-

librium of torques resulting from the pull of the agonist and the antagonist, as well as by gravity. In order for a limb's target location to be programmed, the zero positions of the agonist and the antagonist have to be tuned. This is done by setting a muscle's target length by means of the above-mentioned spindle system. Each muscle will then strive for its target length, until an equilibrium is reached between the pulls of the different muscles controlling the limb's movement. Normally, none of the muscle target lengths will have been reached in the limb's steady state. Here the mass-spring theory differs from the location-programming theory, according to which the muscles do reach their "tuned" target positions.

An advantage of the mass-spring theory is that there is not a *single* tuning to reach a particular steady state; rather, there is an *equivalence class.* The same steady-state position can be reached by tuning the agonist and the antagonist to become very short as can be reached by setting them both for some medium length. The muscles are more tensed in the former case than in the latter, and the limb's target position is, correspondingly, reached more quickly in the former case than in the latter. This is the way in which movement *timing* can be controlled in the mass-spring model.

Like the location-programming theory, the mass-spring theory in its simple form fails in that it cannot handle compensatory adjustment to context. When one speaks with a pipe in one's mouth, the tongue's target positions for various vowels are different than when the pipe is not there. When an adult speaker is given a biteblock between his teeth, he immediately produces quite acceptable vowels. This can only be done when the vowels are articulated with substantial deviations from their normal target positions (Lindblom, Lubker, and Gay 1979; Gay and Turvey 1979; Fowler and Turvey 1980). The articulation of consonants also adapts to such hampering circumstances (Folkins and Abbs 1975; Kelso, Tuller, and Harris 1983; Kelso, Saltzman, and Tuller 1986).

Similarly, the laryngeal system adapts immediately when the speaker looks up or down, or tilts his head, or turns it sideways. And the respiratory system is equally adaptive; it has to function quite differently in contexts of standing, sitting, and lying down. The mass-spring model does not predict compensation when a limb's (or an articulator's) movement is mechanically interfered with; the interfering force is simply added to the set of forces between which an equilibrium is established. The eventual rest or steady-state position will, as a result, be different from the case where there is unhampered movement (Levelt et al. 1985), but there will be no adequate compensation for this difference. In particular, no other

articulator will "take over," as the lips do when compensating for an immobilized jaw in the articulation of [b].

There are still other problems with theories that involve the programming of target lengths for muscles or target rest positions for articulators. One is that most phones cannot be characterized by just a target *position*; they are *gestures* in time. Diphthongs are diphthongs by virtue of a gliding change of the vocal tract's configuration. The place of articulation of consonants may seem to be a definable target position, but the manner of articulation requires particular temporal characteristics of the way that target position is approached or left. This is not controllable by location programming. The mass-spring theory at least allows for control of the speed at which the rest position is reached. But full control of the *trajectory* of movement also requires more than the simple mass-spring account (Saltzman and Kelso 1987).

An additional problem for a mass-spring account of articulation (which may be solvable) is to work out the equations for soft tissue such as the tongue. The theory was initially developed to account for the rotation of limbs around joints, but many of the speech articulators have peculiar damping and stiffness properties that are hard to model. Both Lindblom (1963) and Fowler et al. (1980) have considered these problems. Their solutions are quite different, however, as we will see.

In spite of the various problems of the mass-spring model (in particular, its failure to account for compensation), elements of this model will turn up again as part of another conception of motor control: the theory of coordinative structures. Before we turn to that theory, however, three other theories of articulatory motor control will be reviewed.

11.3.3 Auditory Distinctive-Feature Targets

If the motor command codes are more abstract than target positions of articulators, what is their nature? It has been suggested that their nature is *auditory* or *acoustic* (Nolan 1982; Stevens 1983; Kent 1986). According to this view, the speaker's codes are images of the intended sound structure. With the syllable as a motor unit, the speaker would code the auditory properties of the syllable's sequence of phones. Stevens (1983) suggested that these properties can best be characterized by a set of phonetic *distinctive features*. These are highly perceivable and contrastive dimensions of variation in speech sounds. Linguistically relevant contrasts are made along precisely these dimensions; /b/ contrasts with /n/ in the perceivable dimension of nasality, and with /p/ in the even more per-

ceivable dimension of voicing. The target for the realization of a phone is the set of its linguistically distinctive perceptual features.

In order to be perceptually distinctive, the features should be *acoustically* realized by the production apparatus. Consider, for instance, the acoustic condition for perceiving a stop consonant. Stevens (1983) argued that a major condition for perceiving a stop is the rapidity of spectral change. Take the perceptual distinction between [ba] and [wa]. Only the consonant in [ba] is perceived as a stop or plosive, and, as Stevens showed, this is concomitant with the rapid spectral change in going from the syllable-initial consonant to the vowel. Another of Stevens' examples is the feature of nasality, already discussed in subsection 11.2.3 (see figure 11.5). The clear perceptual distinction between nasal and non-nasal sounds is, according to Stevens, based on the presence of resonance in the very low frequency spectrum. These and other perceptually distinctive features are *acoustic goals* that the speaker tries to achieve. For each phone in a syllable, a small set of such acoustic goals must be realized.

An attraction of this theory is that it puts the goals of articulation in the ear of the listener. The relation between motor programs of speech and perceptual analysis of speech sounds is, on this view, not arbitrary but systematic. This systematicity has been stressed time and again, but it is usually interpreted in the reverse way. Lindblom (1983), for instance, argues that "languages tend to evolve sound patterns that can be seen as adaptations to the motor mechanisms of speech production." The set of possible speech sounds is not exclusively determined by perceptual distinctiveness (though Lindblom stresses, like Stevens, that this is an important criterion); the biomechanical properties of the production apparatus severely limit the class of possible sound-form distinctions languages can employ. Somewhat further back in history, Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) also made this reverse move in their motor theory of speech perception. According to this theory, the listener, in analyzing a speech sound, constructs a model of the articulatory movements by which the speaker could have produced it. The perceptual targets of analysis are speech motor patterns. A very similar view was expressed by Halle and Stevens (1964) in their analysis-by-synthesis theory of speech perception. This is exactly the reverse of the acoustic theory of speech motor production, where the goals of production are distinctive perceptual patterns.

Though one should be sympathetic to the view that there are quite systematic relations between perception and production of speech sounds (there is even important neurological support for this view; see Ojemann

1982), one should be worried when students of perception conjecture motor targets while students of production surmise perceptual targets. It is like Joseph and Mary each assuming, erroneously, that the other is taking care of the Holy Child. A major problem for an acoustic or an auditory theory of motor commands is, as MacNeilage (1980) points out, that it fails "to generate consequences for the control of actual movements." When the target is, for instance, a very low-frequency resonance component in the acoustic signal, which is perceived as nasality, why would the speaker open the nasopharyngeal port?

11.3.4 Orosensory Goals, Distinctive Features, and Intrinsic Timing

One step toward dealing with this problem was suggested by Perkell (1980; see also Stevens and Perkell 1977). Though the "distal" goals of speech motor programming are indeed sensory distinctive features, such as voicing, obstruency, and nasality, the speaker has learned how these goals can be attained by realizing more proximal goals. These are called *orosensory goals*. Each distinctive feature corresponds to a particular aspect of articulation that can be *sensed* by the speaker.

Take, for example, the feature of obstruency, which is a property of all stop consonants. The speaker has learned that the distal auditory goal of plosion can be realized by increasing the intraoral air pressure, and there are oral sensors that register this pressure. Hence, the proximal articulatory goal is to reach that air-pressure level. How that goal is attained will, in turn, depend on other features that have to be realized simultaneously. If there is, for instance, the manner feature "coronal", as in [t], the air pressure will be increased by making a constriction between the tongue and the alveolar ridge. If, however, a labial feature has to be realized, as in [p], the pressure will be increased by constricting the vocal tract at the lips. In both cases, however, the orosensory goal for making a stop consonant is the same: an increase in sensed air pressure. The goal is, therefore, still quite abstract; it does not involve the specification of concrete muscle contractions.

In order to take care of these executive aspects, Perkell assumed subsequent stages of motor control at which the abstract motor commands are reorganized and are translated into contraction patterns for the muscles. The point of departure is the abstract orosensory feature matrix, which specifies for each subsequent segment (consonant, vowel) of the utterance the orosensory features that must be realized (see figure 11.9). This is the level of abstract motor commands. There is intrinsic timing at this level; segment durations are in some way globally specified.

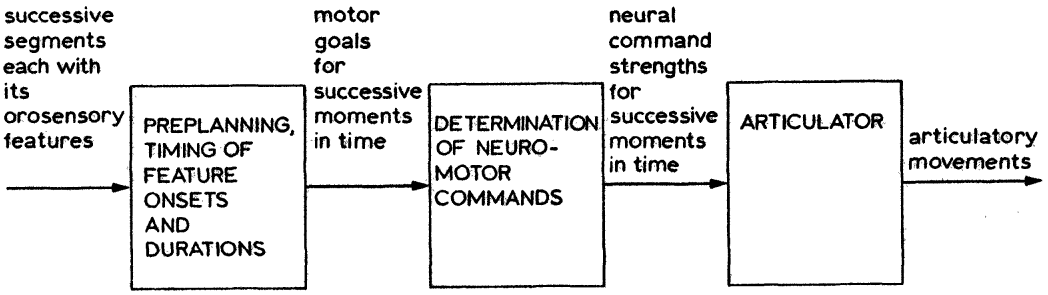

Figure 11.9
Three main components of Perkell's (1980) model of speech production.

At the subsequent stage, the motor goals for each small segment in time are specified. These goals depend in two ways on context. First, the realization of the motor goal for a feature can be dependent on other features. A vowel's duration, for instance, will depend on the place feature of the following consonant. If that consonant has an anterior place of articulation, the vowel can be shorter than if the consonant has a posterior place of articulation (as in *lob* versus *log*). Second, the realization of certain features has to be initiated quite some time before the goal is in fact reached. This requires coarticulation of features from different segments. An example is the rounding feature in vowels such as [u]. When the word *crew* is not pronounced too slowly, the lips are already rounded when the first consonant [k] appears. (Compare the pronunciation of *crew* with that of *crow*, where the vowel is unrounded.) Bell-Berti and Harris (1979) measured a fixed time anticipation of lip rounding in vowels. The reorganization at this "preplanning stage" takes care of the correct timing of the various feature onsets and durations.

These timed motor goals are, at the next stage, translated into neuromotor commands for the articulatory muscles. These neuromotor commands control the contraction patterns of the muscles. The command strengths, Perkell argues, depend not only on the motor goals but also on feedback from the articulators. This feedback should be the basis for the compensatory behaviors discussed above. Perkell discussed various kinds of feedback that may be involved here. Among them are the muscle spindle feedback loop (see figure 11.8), various kinds of orosensory feedback, and maybe some auditory feedback (as will be discussed in the next chapter). However, no account is given of how these feedback loops reshape the neuromotor command patterns so as to produce the critical compensation phenomena. The last stage depicted in figure 11.9 involves the factual speech articulation.

Perkell's answer to the question of how the distal auditory target is reached is, in short, that the speaker has learned to replace it with orosensory goals. This replacement does not occur during the moment-to-moment control of speech; the "proximal" motor commands are directly in terms of orosensory goals. The child, however, has to acquire the correct orosensory goals by auditory feedback. Also, auditory feedback is necessary to maintain the fine tuning of the orosensory goals; there is noticeable deterioration in cases of long-lasting acquired deafness.

There have been rather heated discussions in the literature about motor-control models of this type (Fowler 1980; Hammarberg 1982; Fowler 1983; Parker and Walsh 1985). Whether or not the primary or abstract level of motor commands already specifies the *timing* of motor execution was central to these discussions. According to Perkell's model, it does (to some extent); however, the related models of Moll and Daniloff (1971), Daniloff and Hammarberg (1973), and Hammarberg (1976) do not make such a claim. In the latter models, the abstract motor commands are timeless segments consisting of bundles of features; temporal features are added only at the later stages of motor execution. Fowler (1980) called such theories *extrinsic-timing* theories. She argued for *intrinsic timing*, such that the "abstract" motor command units should already be fully specified in terms of their temporal shape. This makes it unnecessary to reorganize the motor commands, as is done in Perkell's "preplanning" stage. The abstract phonetic plan for a segment already specifies its entire gestural shape over time.

But how can one account for coarticulation if Perkell's second stage is eliminated? According to Fowler, coarticulation is nothing but the simultaneous realization of phones whose temporal specifications overlap in time. The motor command for the vowel [u], for instance, specifies an early onset of rounding, which may then temporally overlap with preceding phones, as is the case in *crew*. The temporally specified motor commands for successive segments of an utterance are superimposed or added. Fowler (1983) speaks of *vector summation*. The resulting muscular innervation is some linear or nonlinear function or "sum" of the innervations proceeding from the individual commands. It will, further, depend on the biomechanics of the moving parts how the resulting muscular excursions will be related to the two (or more) input commands. Typically the "summation" is nonlinear (Fujisaki 1983). This means that the degree to which one command is realized depends on the strength of the other commands. In short: Where segmental gestures overlap in time, they are simply *co-produced*. Fowler (1986b), in response to Diehl's (1986) critique

of this co-production hypothesis, cites empirical evidence in support of vector summation. A further development of Fowler's approach can be found in Browman and Goldstein 1986.

11.3.5 Auditory Targets with Model-Referenced Control

Returning now to theories where the *proximal* targets are auditory goals, we should consider the theory proposed by Lindblom, Lubker, and Gay (1979). It contains a serious effort to deal with the problem—signaled by MacNeilage (see subsection 11.3.3)—of explaining how an auditory goal is translated in actual control of movements.

Lindblom et al. report in detail on their biteblock experiments, which show unequivocally that the formant structure of vowels is essentially unchanged when the speaker has a biteblock between his teeth. This must involve compensatory positioning of the tongue, since (as Lindblom et al. showed) without such compensation substantially different formant structures would appear in the biteblock condition. The compensation is there on the very first trial, and on the very first puff of air released by the glottis. These latter findings exclude two explanations for the biteblock results. The first one is that the speaker learns in the course of successive trials by listening to himself and by making successive approximations. No such learning over trials occurred. The second explanation, also excluded, is that the speaker uses immediate auditory feedback from what happens during the first few glottal pulses to immediately correct the vowel he is making. The almost perfect compensation is not based on auditory feedback; it is there from the very beginning of the articulation.

Lindblom et al. then go on to argue that the targets of the motor commands are *sensory*—i.e., that the target for a vowel is the way in which the vowel should be perceived. This representation is quite abstract, and it is far removed from the many equivalent ways in which this sensory impression can be produced. But these equivalent ways (for instance, with and without a biteblock) do share certain features. For vowels the shared feature is the so-called *area function*, which is a measure of how the shapes existent in the vocal tract determine the tract's resonance spectrum. The neurophysiological code for a particular sensory vowel quality is, then, in terms of the vowel's area function.

How does a speaker create the correct vocal-tract shape whether he has a pipe, or food, or a biteblock, or new dentures, or nothing in his mouth? Lindblom et al. assume that the brain is informed about the state of the vocal tract by proprioception, in particular by tactile feedback from the mucous skin in the oral cavity (although there are other forms of orosen-

sory feedback; see Stevens and Perkell 1977). This tells the brain what the shape of the tract is like. Given the area function, the brain can then compute the sensory effect when a tract of that shape is used as a resonator. If this internally generated sensory representation is critically different from the sensory *target* of the motor command, the speaker can adapt the shape of the tract—for instance by moving the tongue—and compute the sensory representation that goes with this new constellation. When the approximation to the target is satisfactory, the vowel can be articulated and it will be correct right away.

In other words, the speaker has an internal *model* of the relevant properties of his vocal tract. By proprioceptive feedback from the actual vocal tract, he can set the parameters in this model and compute the expected sensory outcome of using the actual vocal tract of that shape. If the computed sensory outcome is not satisfactory, the actual shape can be adapted by appropriate efferent activity.

This theory is, of course, not limited to the control of vowel production. It should also be possible to internally simulate or predict the sensory effects of consonants that evolve from some articulatory gesture. And indeed, immediate adaptation phenomena have also been observed in the articulation of consonants. Folkins and Abbs (1975), for instance, interfered with jaw movement by means of some mechanical contraption. If this interference was unexpectedly applied when a word containing an intervocalic [p] had to be said, the speakers still produced complete lip closure by stretching both the upper and the lower lip more than usual. More recently, Kelso, Saltzman, and Tuller (1986) obtained similar results for the pronunciation of [b]. We will presently return to these results.

It should be obvious that the internal model that simulates the sensory results is based on extensive experience with listening to one's own speech. It is, therefore, not surprising that Oller and MacNeilage (1983) did not find full compensation in biteblock experiments with a four-year-old and a nine-year-old. Deaf people, of course, lack such a model.

Motor control of this type is called *model-referenced control* (Arbib 1981). It is a form of closed-loop control because it involves a feedback loop. The proprioceptive information is fed back to set the parameters of the internal vocal-tract model.

Fowler and Turvey (1980) pointed out that the theory of Lindblom et al. is underspecified in one major respect: When there is a relevant difference between the intended sound and the one simulated by the internal model, some adaptation of the vocal tract has to be initiated; but nothing

in the theory tells us how an *appropriate* adaptation is generated. It is most improbable that this proceeds by trial and error. It should, rather, be derived from the *character* (not the mere size) of the difference between the intended and the simulated speech sound.

At the same time, Fowler and Turvey (1980) argued that the theory is also *over*specified, i.e., unnecessarily complex. Model-referenced control could as well be conceived of as follows: The model computes, by proprioceptive feedback, the present state of the vocal tract. Given the intended sound (i.e., the code or command in the motor program), it should be able to compute a vocal-tract shape that will create the intended sound. It will then send efferent control signals to the vocal-tract musculature, which will move it from the actual to the computed configuration. This alternative, of course, requires a solution to the problem of how to effectuate a change from the actual situation to one that will produce a particular intended sound.

In spite of claims to the contrary (Kelso, Holt, Kugler, and Turvey 1980), there is nothing in the theory of Lindblom et al. that makes it inconsistent with existing data. Fowler and Turvey (1980) recognized this and stated (rather more carefully than Kelso et al.) that there may be *a priori* grounds for preferring a different theory. We will now turn to that theory.

11.3.6 Coordinative Structures

It has long been known that rather complex motor coordination is possible without much central control. The classical farmhouse demonstration is the decapitated running chicken. There are trains of motor activity, involving the coordinated use of whole sets of muscles, that can apparently run off automatically. These are traditionally call *synergisms*. Lenneberg (1967) applied this notion to speech production. One synergism that he analyzed in detail is respiration during speech. According to Lenneberg there is a special "speech mode" of respiration, which is quite different from nonspeech respiration (see subsection 11.2.1): There are long stretches of exhalation, and only short moments of inhalation (only 15 percent of the respiratory cycle time, versus 40 percent in normal breathing). There is about four times as much air displacement during the cycle in speech as in nonspeech breathing. The outflow of air has a relatively constant rate in speech, but not in normal breathing. The coordination patterns of the respiratory muscles are distinctly different in speech and nonspeech breathing.

Not only the respiratory system, but all vocal organs have other uses than speech. Mastication and swallowing involve largely the same muscles as speech, but the muscles' coordination differs substantially between the speech and nonspeech modes. A muscular organization that is set to act as an autonomous functional system is called a *coordinative structure* (Easton 1972; Turvey 1977).

A coordinative structure can be seen as a system with a severely reduced number of degrees of freedom. Each individual muscle can act in several different ways at any moment, but most of these cannot occur if the muscle functions as part of a coordinative structure. In swallowing, for instance, the oral and pharyngeal muscles contract in a strict temporal order. This organization is both automatic (it can even be released as a reflex action) and functional (it is set to perform a particular *kind* of task: transporting stuff from mouth to the esophagus). There are only a few degrees of freedom left in this system. It will, for instance, behave slightly differently in swallowing a big object than in transporting a small object or some fluid.

A coordinative structure, therefore, is not totally rigid in its action. It is set to perform an *equivalence class* of actions that will all produce the same functional *kind* of result. Which particular motor activity within the equivalence class is performed depends on the context of action. For swallowing the context has to do with the kind of food transported; for walking it has to do with the resistance between feet and floor; for eye tracking it has to do with direction and speed of the visual target, and so forth. Coordinative structures can be considered as hinges between abstract, context-free motor commands and concrete, context-adapted motor execution. The abstract command specifies the kind of act required; the coordinative structure's execution takes care of the peripheral context in which the result has to be produced.

Considered in terms of its coordinative structure, walking appears to involve a *hierarchy* of subsystems, involving the muscle systems of the arms, the legs, the feet, and the toes. A similar hierarchical organization is still largely intact in the unfortunate running chicken mentioned above. Also, coordinative structures tend to produce *cyclic* behavior. This is obvious in walking and breathing. It is at this point that mass-spring accounts become integrated with coordinative-structure theories of motor control. Cyclicity can be a quite natural result of mass-spring activity; when there is little damping of the movement, the system will overshoot its (eventual) steady state and swing back and forth for some time. As in

walking, this swinging can be functionally integrated in the coordinative structure's behavior.

Coordinative-structure accounts of speech motor control often stress its cyclic and hierarchical nature. There are two major cyclic phenomena in speech. One is the breathing rhythm, the special cyclic synergism of inspiration and expiration studied by Lenneberg (1967). The other is the syllable rhythm, which is largely a property of vocal-tract functioning.

Considered as a coordinative structure, the vocal tract in its speech mode is set to produce a string of syllables. That is the structure's characteristic kind of output. Which particular syllables are to be uttered has, of course, to be in the motor command; however, the functional organization of the vocal tract in speech motor control is such that the motor commands need not contain anything that is *common* to all syllables. Brodda (1979) and Sorokin, Gay, and Ewan (1980) have argued that the normal syllable rate in the world's languages (5–6 per second) is a consequence of the biomechanics of the vocal tract. It is the *eigenfrequency* of the system, especially of the mandible's movements. At this rate the movements absorb a minimum of muscular energy. This need not be in the motor program; it can be *extrinsically* timed.

But the *deviations* of this syllable rate must be programmed. This is *intrinsic* timing, because these temporal parameters are part of the phonetic plan. The way the extrinsic parameters are coded in the phonetic plan may be rather abstract. Shorter durations, such as in unstressed (versus stressed) syllables, can be obtained by making less extensive movements of the relevant articulators. This, in turn, may be realized by setting less extreme local targets for these articulators. This means that in the articulatory plan, duration can be, in part, *spatially* coded.

Speech motor control is also hierarchical. Take again the syllable as a programming unit. Its execution involves, on the coordinative-structure account, at least two subsystems: one for the realization of the peak and one for the execution of the less-sonorous flanking consonants. It has been argued that these involve rather separate muscular systems which can largely operate in parallel. This will be taken up in subsection 11.3.7.

The notion of coordinative structures is currently popular in theories of speech motor programming. (See especially Kelso et al. 1983, 1986; Fowler 1980, 1983, 1986a; Fowler et al. 1980.) On this view, motor commands are serially ordered instructions for phonetic acts or articulatory "tasks." Each act is specified in a context-free way. For instance, [b] involves the task of closing the lips with a certain force. But it is not fixed beforehand how much the share of jaw and lip movement should be in

accomplishing this task. In fact, there is an infinitude of different ways to accomplish it. The task itself, however, is *minimally* specified; there is no need to detail features of the phonetic act, which will be automatically taken care of at lower levels of motor execution, namely by the so-called *articulator network* (Saltzman and Kelso 1987). It will, in a highly self-organized way, adapt to the accidentals of context and find the least energy-consuming way to reach the goal. At least, such is the claim. If the jaw cannot move, the lips will compensate by moving more.

Recently, substantial progress has been made in developing the mathematics of such self-organizing motor systems (Saltzman and Kelso 1987). There are usually more articulator variables than there are task variables (two lips and one jaw can move in order to realize a single degree of lip opening). The mathematics must therefore specify how the articulator variables are *redundant* in executing the task, i.e., how the articulators' motor patterns are mutually restraining in the "articulator network." There must be a reduction of degrees of freedom in a coordinative system, as was exemplified above by the swallowing reflex. General mathematical procedures for effecting such reductions are now available.

Furthermore, the theory must specify how the articulator network can "know" the state of its articulators, i.e., their position and their direction and velocity of motion. It has been suggested that there is no feedback to be monitored in a coordinative structure's mass-spring system (Fowler, Rubin, Remez, and Turvey 1980), but this cannot be correct. Compensation behavior requires some sort of feedback. Saltzman and Kelso (1987) provide such a feedback system. Figure 11.10 presents it in the form of a

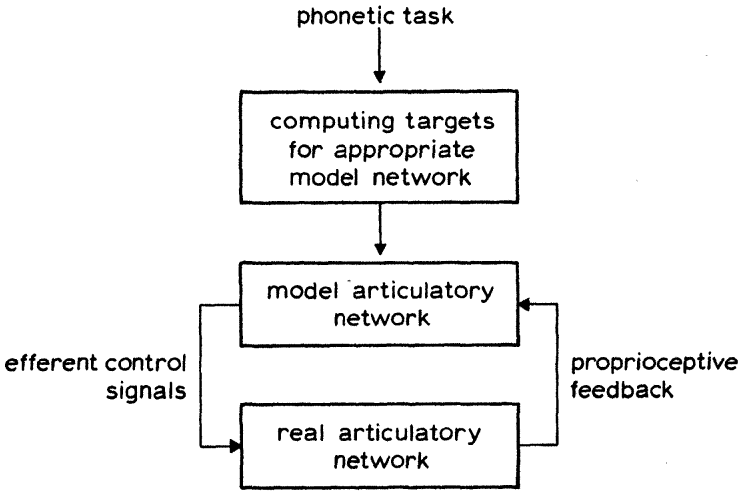

Figure 11.10
Model-referenced control in the coordinative-structures theory.

diagram. Notionally, it is not essentially different from Lindblom's model-referenced control (see above), and it is similar to Neilson and Neilson's (1987) "adaptive model theory." The articulator network consists of the real articulators and of a model of these articulators. The parameters for the model articulators' muscle tonus, position, and velocity, which constitute the model's initial state, are derived from proprioceptive feedback (see below). The task induces the *model's* behavior; the model articulators move in their mutually restrained way in such a way that the task is accomplished in the model. The spatio-temporal and force properties of the model's motions are fed on-line to the *real* articulators, which faithfully execute them. By proprioception the resulting spatio-temporal properties of these real movements are fed back to the model, which can detect any significant deviations from the model movements. If there is some significant deviation, the model will compensate in order to reach the set goal. The corrected model articulations are fed to the real articulators, which execute the compensatory move, and so on.

Let us see how this works out for the jaw-lip compensation. The task is to close the lips for [b]. The mathematics of the coordinate system distributes the closing movement in some optimal (energy-preserving) way over jaw and lip movements. These movements are performed in the model, and the real articulators follow faithfully until the jaw is (experimentally) braked in its course of motion. Proprioception reports this deviant motion pattern back to the model, which then computes a different distribution of movement for the three articulators such that the same effect—lip closure—is obtained. These corrected movement patterns are transmitted to the real articulators, which execute them.

This, at least, is the story when the model has no advance proprioceptive information about the imminent obstruction. What the model should do in such a case, according to Lindblom and MacNeilage (1986), is give the initial reaction of *exerting more force* on the jaw so that the apparent resistance is overcome. If the feedback is still deviant, the other articulators take over full compensation. Lindblom and MacNeilage noted the finding by Abbs and Cole (1982) that initially more force is indeed exerted on the perturbed articulator. This also precedes the compensation in the other articulators.

But with a biteblock in the mouth, there is at least some proprioceptive feedback *before* the movement is initiated. There is high tonus in the masseter muscles, but no corresponding movement of the mandible. The initial parameters of the model will be set accordingly, and compensation

can be immediate. And this is indeed what is found in the biteblock experiments.

The elementary articulatory elements of a coordinative structure are mass-spring systems, but they are joined together to operate as a whole that can perform a particular phonetic task. The coalition of articulators can be a different one for each task. Any single articulator can participate in several different coordinative structures. As we saw, the jaw is part of the coordinative structure that takes care of pronouncing [b]. It is also part of the structure that takes care of producing [z]. But these are two distinct coordinative structures. Kelso, Tuller, Vatikiotis-Bateson, and Fowler (1984) performed compensation experiments for both [b] and [z]. When the task was to pronounce [b] and the jaw's movement was restrained, there was increased, compensatory lip movement. However, when the task was to pronounce [z] and the jaw's movement was similarly braked, no compensatory movement occurred in the lips. This shows that the jaw-lip coupling is specific to the phonetic task at hand. A coordinative structure is a functional, not a "hard-wired," coalition of articulators.

A dooming conclusion, however, is that each recurring phonetic segment requires its own coordinative structure. One would, of course, like to see some higher-order *organization* of elementary coordinative structures (the hierarchical control structure promised by theorists in this camp). A good candidate for a second-order level of organization is one that takes care of producing a language's possible syllables. The motor control of syllables may be hierarchically organized, as is walking or typing. Syllabic organization will be the subject of the next subsection.

This short review of the theories of speech motor control makes one thing abundantly clear: There is no lack of theories, but there is a great need of convergence. Theories differ both in the nature of the commands they conjecture and in their modeling of motor execution. One point of convergence among the theories, however, is the view that, whereas speech motor commands are relatively invariant, the executive motor system can take care of adaptations to the immediate context of execution without being instructed to do so. A related inevitable development is toward accounts involving model-referenced control. Finally, most theorists sympathize with the notion of the syllable as a unit of speech motor execution.

11.3.7 The Articulation of Syllables

The syllable is in many ways an *optimal* articulatory motor unit. It allows the consonants to be co-produced with the peak without too much articu-

latory interference. The reason is, as Öhman (1966), Perkell (1969), Fowler, Rubin, Remez, and Turvey (1980), and Lindblom (1983) have argued, that consonants and vowels often involve disjunct articulators.

The production of most vowels depends on the positioning of the tongue body by means of the extrinsic muscles—in particular, the genioglossus. Most consonants that involve the tongue at all (in particular, the coronals) are articulated by means of the intrinsic muscles, which affect the shape rather then the body position of the tongue. Whereas moving the tongue body around is a relatively slow process, most consonantal articulators can be adjusted fairly rapidly. To the extent that the muscles involved in the production of consonants and vowels are indeed disjoint, the movements can be co-produced without interference. On this view, articulation can be seen as the production of a stream of syllable peaks, with consonantal articulation superimposed.

If Brodda (1979) and Lindblom (1983) are correct in supposing that syllable rate is largely determined by the *eigenfrequency* of the moving parts involved in vowel production (the mandible and the tongue body), this steady stream of syllable peaks is the outcome of an energy-saving production mechanism. Indeed, Fowler et al. (1980) conjecture that there is a *continuous* stream of vowel articulation, which is handled by a relatively independent muscular organization. This is the "carrier wave" for the articulation of consonants. One should, of course, be careful not to overstate the disjunction of consonant and vowel articulators. There are various articulators, such as the jaw, which are actively involved in both consonant and vowel production, as Fowler (1986b) admits.

But the economy of the syllable organization is not only based on the co-producibility of consonants and vowels. There is also a tendency to distribute consonants over the syllable in such a way that consonant articulation will involve a minimum of spatial excursion of articulators, and that means a minimum of energy expenditure. The languages of the world, according to Lindblom's claim, tend to arrange the segments of a syllable in such a way that adjacent segments involve *compatible* articulators. This not only holds between the peak and its flanking consonants; it may also hold between adjacent consonants. Compatibility is present in consonant clusters such as [sp], [sk], [pl], [skr], and [spl]; it is not in clusters like [pb], [kg], and [tfd]. Clusters of the latter kind hardly ever appear in syllables.

Also, as was discussed in sections 8.1 and 9.6, syllables are "hills of sonority." The more *sonorant* consonants in a syllable tend to be positioned closer to the peak than the less sonorant ones. The sonorants [r]

and [l], for instance, tend to directly precede or follow the peak, whereas [t] or [s] can be separated from the peak by one or two intermediate consonants. The syllable *tram* is a rather more likely one than *rtam*, and *slam* is more likely than *lsam*.

A syllable, in other words, is a production unit designed for optimal co-articulation of its segments. As a consequence, a maximum of perceptual distinctiveness is produced by means of a minimal amount of articulatory effort.

Summary

This chapter has dealt with one of man's most complex motor skills: the fluent articulation of speech. Articulation is the motor execution of a phonetic plan. Before it becomes articulated, a phonetic plan can be temporarily buffered. The first section of the chapter discussed the management of this so-called Articulatory Buffer, which presumably compensates for fluctuating differences between the rate of formulating and the rate of articulating.

In order for speech to be initiated, some minimal amount of phonetic plan (probably a phonological word) must have been assembled and delivered to the Articulator. The work of Klapp et al. showed that, correspondingly, onset latency is longer for two-syllable words than for one-syllable words. A condition for this so-called syllable effect is that the speaker cannot have prepared and buffered the articulatory response. The phonetic spelling out of a word's syllables is a serial process. Hence, preparing a two-syllable word for articulation takes longer than preparing a one-syllable word.

Sternberg and colleagues extensively studied the mechanism of retrieving the program from the Articulatory Buffer and unpacking it. The time needed to retrieve an articulatory program is a linear function of the number of items (probably phonological phrases) in the buffer. A further finding was that unpacking a two-syllable word takes slightly more time than unpacking a monosyllabic word. A likely interpretation of the latter phenomenon is that a word's articulation cannot start right after its first syllable has been unpacked. Rather, part of the second syllable must have been made available as well. It may be a necessary condition for making a fluent articulatory liaison between a word's syllables. It was, finally, found that the whole duration of the execution of an utterance's phonetic plan depends not only on the number of its syllables but also on the total number of items (words or short phrases) in the buffer. The more items

there are in the buffer, the harder it is to retrieve each of them. The extra time needed to retrieve the next item is often gained by drawling the current item's last syllable. In fluent speech, the role of the Articulatory Buffer may be rather limited. Still, articulatory buffering is not just a laboratory effect.

Next we turned to the vocal organs involved in speech production. They are organized in three major structures. The respiratory system controls the steady outflow of air during speech; it provides the source of acoustic energy. It has its own mode of functioning during speech – a mode that differs markedly from normal nonspeech breathing. The laryngeal system, with the vocal folds as its central part, controls voicing and loudness in speech. During voicing it generates a periodic train of air puffs, which provide the wide frequency spectrum on which resonation builds. The supralaryngeal system or vocal tract contains the chambers in which resonation develops, in particular the nasal, oral, and pharyngeal cavities. Their shapes determine the timbre of vowels and consonants. The vocal tract, moreover, is the main contributor to the proper articulation of speech segments. There are different places where the vocal tract can be constricted, and there are different manners in which these constrictions are made or released. The combinations of places and manners of articulation provide a rich variety of possible speech sounds; each language uses only a subset of these.

The third section of the chapter concerned the organization of speech motor control. The major questions concerned the nature of a motor command and the way in which such a command is executed. These questions are not independent; most theories of speech motor control couple the answers in some way or another.

A first theory considers motor commands to be target positions for articulators. This location-programming theory was criticized because it is not sufficiently abstract. Target positions are highly context dependent. A segment's articulation depends not only on the immediately abutting segments but also on incidental contextual factors, such as food in the mouth, clenched teeth, and tilt of the head. Since contextual variation is unlimited, preparing a phonetic or articulatory program would become an unduly complex affair. A similar critique applies to a simple mass-spring account of motor execution. On that account, an "intended" position will not be reached when external forces interfere with the movement of articulators. Each such context would require a different motor command.

All other theories agree on the abstract, relatively invariant nature of motor commands in speech (the articulatory plan). It is only the executive apparatus that adapts the motor commands to the prevailing physical context.

We then reviewed theories according to which the abstract commands are to reach certain auditory goals. On these theories, speakers build sensory images of the sounds they intend to produce. Lindblom's theory, in particular, goes into much detail about how such an image can guide motor activity. There is model-referenced control. This means that the speaker has an internal model of his own vocal apparatus. For each configuration of the model, the resulting sensory image can be derived and compared against the goal image. The notion of model-referenced control is especially useful for dealing with so-called compensation phenomena. When an articulator is hampered in its movements, another can "take over" so that the intended sound is nevertheless produced.

Other theories, though sympathetic to the notion of auditory targets as distal goals, take motor commands to involve more proximal goals. Perkell suggests the existence of orosensory goals, i.e., goals definable in terms of the sensory experience of one's own vocal tract. It is claimed that auditorily important distinctive features, such as obstruency and nasality, have close orosensory correlates.

In the coordinative-structures theory, finally, it is supposed that the executive system consists of a hierarchy of task-oriented structures. Each such coordinative structure is a group of muscles temporarily set to function as a unit. The phonetic plan is a string or hierarchy of articulatory tasks, and each articulatory task requires a different coalition of cooperating muscles. Within such a coalition, the muscular innervations are coupled by some function that severely limits their degrees of freedom but which guarantees a particular kind of result. That result is obtained whatever the initial states of the individual muscles, or whatever the limiting external conditions, such as a pipe in the mouth obstructing jaw movement. Model-referenced control is an essential aspect of coordinative structures, and so is a mass-spring account of movement control. The executive system has a great deal of intelligence, on the coordinative-structures account of articulation. Though much work is still to be done to decipher this "wisdom of the body," recent developments in the mathematics of coordinative control show that there is life in the enterprise.

The final section of the chapter was devoted to the articulation of syllables. It was argued there that syllables are important articulatory units.

Not only is it likely that major aspects of a syllable's phonetic plan are stored (and retrieved during phonetic spellout); it is also likely that syllables are optimally organized to facilitate high-speed co-articulation of their segments. This minimization of articulatory effort goes with a maximization of perceptual distinctiveness, even at high rates of speaking.

Chapter 12
Self-Monitoring and Self-Repair

Speakers monitor what they are saying and how they are saying it. When they make a mistake, or express something in a less felicitous way, they may interrupt themselves and make a repair. This is apparent not only in spontaneous conversations, but in all kinds of discourse. Here are three examples. The first was reported by Schegloff, Jefferson, and Sacks (1977):

(1)
A: And he's going to make his own paintings.
B: Mm hm,
A: And – or I mean his own frames
B: Yeah

In spite of B's *mm hm*, a sign of acceptance, A became aware that she had said *paintings* instead of *frames*, and corrected this on her next turn.

The second and third examples are pattern descriptions obtained in an experiment reported in Levelt 1982a,b. The subjects were asked to describe patterns such as the one shown in figure 12.1. They were told that their descriptions would be tape recorded and given to other subjects, who were to draw the patterns on the basis of these recordings (see also subsection 4.4.2 and figure 4.5). One subject was in the process of indicating the connection between the yellow node and the pink node and said

(2) And above that a horizon –, no a vertical line to a pink ball

This is much like the repair of an error in example 1, but here the speaker was very quick in effectuating the repair; the trouble item (*horizontal*) was not even completed. Another subject, going from the yellow node to the blue one, said

(3) To the right is blue – is a blue point