SPEECH PRODUCTION

Speech is easy to take for granted. Most children acquire speech easily, as though they spontaneously take lessons from the spoken language of the adults around them. Because the acquisition of spoken language is such a robust phenomenon, some writers have concluded that it is based largely on genetic mechanisms (1). Except for children who have impaired hearing or some other unusual condition that prevents speech

acquisition, speech is the earliest form of language expression acquired. Even in languages with a signed form, speech is the preferred means of communication (2). Most people use speech so frequently, in fact, that it is one of the most common voluntary behaviors throughout the human life span. Some regard the telephone as the most valuable patent ever issued, and its importance derives from the centrality of speech to human activities. The high frequency of speech behaviors certainly reflects the importance of communication in human societies, but it also reflects the physiological ease of speech production. For most people, speech is easy and natural. It is performed even as the arms and legs are used in other tasks, and speech therefore accompanies and complements many human behaviors.

Speech is also decidedly efficient for communication, as shown by studies that compare the efficiency of different systems including typewriting, handwriting, video, voice, and their combinations (3). For example, when persons communicating in a problem-solving task have access to voice, the time to a solution is reduced by more than 50% compared to typewriting. A similar advantage of speech is evident in studies of equipment assembly tasks, for which interactive telephone speech has a threefold speed advantage over keyboard communication (4,5).

Although speech is arguably unique to humans in biological comparisons, it is increasingly shared by machines. Computer software enables machines to produce speech (speech synthesis or machine speech), to perceive speech (automatic speech recognition or machine speech recognition), and to recognize individual talkers (machine speaker recognition). Human-machine speech communication is one facet of engineering technology. If the rate of recent progress continues, it is highly likely that within a few years, it will be rather routine for a person to call a machine, which will identify the individual by his or her speech patterns (and therefore permit access to privileged data), understand an inquiry for some kind of information, retrieve the requested information, and then transmit the retrieved data as synthetic speech. The machine may even be able to detect the speaker's emotional state and to adjust its speaking style accordingly.

Progress in speech technologies, such as speech synthesis and speech recognition, depends partly on a firm understanding of human speech, especially its relationship to linguistic structures, its physiology, and its acoustics. The study of speech is also important to progress in fields, such as cognitive science, linguistics, and speech pathology.

PROPERTIES OF SPEECH

For all of its ordinariness, speech is a difficult subject to study. Part of the difficulty derives from its complexity. Some aspects of the complexity are summarized here:

1. Because speech is yoked to language, it must be understood in part through reference to linguistic structures and processes. A major question concerning speech qua language is what is the basic unit of speech that relates to language structure? Identification of this unit is a fundamental issue in speech science that continues to be investigated and debated. Although a clear consensus on the basic unit is not at hand, most speech scientists describe speech sounds with the symbols of the International Phonetic Alphabet (IPA). This system is designed to provide unambiguous transcriptions of sounds in the world's languages, and it is a convenient symbol system to describe speech. Because several of the IPA symbols are not represented on the conventional keyboard, alternative keyboard-compatible symbols have been described [e.g., the PHONASCII system of Allen (6)]. The study of the relationships between speech sounds and language structures is called *linguistic phonetics.*
2. Speech is produced by a large number of structures, principally those that constitute the respiratory system and its various valves and passages (Fig. 1). It has been estimated that over 100 muscles are used to produce speech. How is this complex motoric system regulated to produce a signal that is perceived at rates of 25 segments/sec? *Physiological phonetics* is concerned with the physiological events and processes underlying speech production.
3. Typically, speech is understood through the sense of hearing. Therefore, auditory psychophysics is fundamental to understanding how speech is perceived. Furthermore, because audition is a primary means of self-monitoring speech production, hearing must be considered a factor in regulating speech. In what ways does

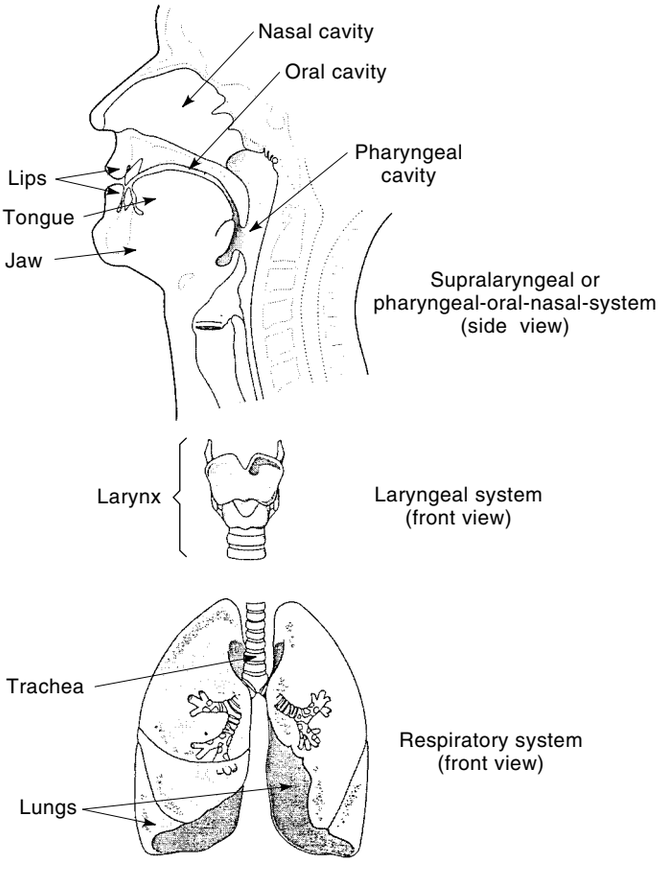

Figure 1. The major components of speech production: respiratory system, laryngeal system, and supralaryngeal (vocal tract) system.

the sense of audition determine the nature of the speech signal? *Perceptual phonetics* addresses the processes by which listeners reach phonetic decisions from the physical signal (usually acoustic, but visual information is also used).

4. Given the primacy of the acoustic signal in speech communication, the science of acoustics is essential to reveal how the movements of the speech organs are involved in generating acoustic patterns and how the acoustic patterns are perceived as a linguistic message. How is the acoustic signal generated? *Acoustic phonetics* is concerned with the acoustic signal of speech, its generation and informational content.

The study of speech, then, is an interdisciplinary undertaking. Among the disciplines that have contributed to this study are psychology, linguistics, speech science, speech pathology, physics, biophysics, and various branches of engineering. Because the concepts and techniques relevant to the study of speech come from many disciplines, it is challenging to integrate the accumulated information into a unified view. This article summarizes some of the major issues and approaches. The acoustic signal is used as a special focus because this signal mediates between the speaker and the listener (and hence between the two major aspects of speech), and because the acoustic signal of speech is sufficiently understood to be used both as a tool for investigating speech and as a foundation for practical applications, such as speech recognition by machine, speaker recognition, and speech synthesis.

THE LINGUISTIC-PHONETIC STRUCTURE OF SPEECH

It is fitting to begin with a brief discussion of linguistic aspects because "Speech is a physical and behavioral manifestation of linguistic structures" (7, p. 244). Although speech can be studied to a limited degree without recourse to the linguistic message it mediates, the ultimate concern is to understand how the properties of speech relate to the units and organization of linguistic forms. Frequently, acoustic and physiological studies of speech have simply assumed the relevance and appropriateness of linguistic units to demarcate physical events in speech production. But, as is discussed, the fit of such discrete units to the continuous variables in acoustic and physiologic analyses is often unsatisfactory.

Vowels and Consonants

A fundamental and time-honored division of speech sounds is into the two major classes of vowels and consonants. Vowels are sounds produced with a relatively open *vocal tract* (the sound-forming passage that extends from the larynx, or voice box, to the lips or nostrils). Typically, vowels are voiced, meaning that they are produced with vibratory energy generated in the larynx. However, vowels are also produced in whispered speech. In typical voiced speech, vowels are the most intense speech sounds and they form the nucleus of syllables. With some exceptions, a vowel is needed as the essential element of a syllable. Consonants are sounds produced with a relatively closed or narrowed vocal tract. The degree of constriction actually varies from moderate to completely obstructed. Compared with vowels, consonants are generally weaker and briefer. Usually, consonants combine with vowels to form syllables, but a small number of consonants serve a syllabic function (as in the second syllables of the words *button* and *battle*). The syllable is discussed in more detail later in this article.

Phonemes and Allophones

Intuition tells us that speech is made of small sound units, or segments. The general impression of speech, both to the layperson and the specialist, is that it consists of a sequence of sounds. In the study of phonetics, these units are called *phonemes* or *phonetic segments.* Phonemes are abstract units that are sufficient to contrast the words of a particular language, that is, working with an inventory of the words in a language, the phonetician seeks to determine the minimal set of sound units that distinguish the words. For example, the two words *ray* and *way* differ in their initial phonemes, /r/ versus /w/ (the slash or virgule is conventionally used to identify phoneme symbols). Minimal word pairs of this kind are especially useful in establishing phonemic differences, that is, sound differences that carry meaning in a language. But not all sound differences affect meaning. For example, the word *pop* begins and ends with the same phoneme, but this phoneme is produced in quite different ways. The initial /p/ is necessarily produced with a release of the lips and an accompanying burst of noise. The final /p/ is produced (and heard) simply as a closure of the lips, without a detectable release, that is, the phoneme /p/ is produced as a released or unreleased variant. It is in this sense that a phoneme is a class or family of speech sounds. The members of a phonemic family are called *allophones.* These are essentially alternate realizations of the abstract phonemic unit. Allophones are the pronounceable phonetic segments in a particular language. In general, a phoneme is associated with two or more allophones.

Words and Morphemes

Figure 2 shows the relationship among some important linguistic units. Words are composed of morphemes, the minimal units of meaning. Some morphemes are also words. For example, *dog* is both a word and a morpheme because it cannot be broken into smaller units of meaning. But not all morphemes are words. The suffix *s* used to form plurals such as *dogs* and *cats* is a morpheme because it signals pluralization (an aspect

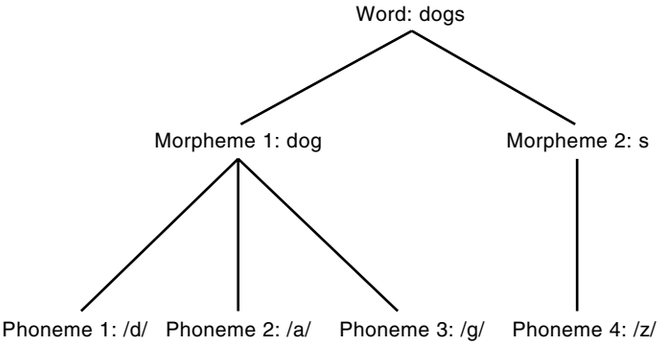

Figure 2. Hierarchy showing relationship among some major linguistic units.

of meaning) but cannot stand alone as a word. To relate morphemes to speech, a first step is to identify the phonemes that constitute a given morpheme. Phonemes are not directly pronounceable because they are abstract descriptions of sounds. Therefore, allophones are needed for a suitably detailed description of how the sounds in a word actually are produced.

Languages differ in the number of phonemes they use. English is an average language in this respect, as its 44 or so phonemes is roughly modal for the phonemic inventories of the world's languages. Identification of phonemes is a first basic step in understanding the sound patterns of a language. The next step is to describe the variations within phonemic families. Many of these variations relate to contextual adjustments, or assimilation, that is, the way a phoneme is produced is often accommodated to the neighboring sounds. These contextual adjustments account for many of the allophones (variants) within a phoneme.

The term *phonology* means the study of sound patterns in a language. It is directed toward identifying the regularities by which sounds are combined. Some authors define phonology as the study of sound patterns in words. As such, phonology is an important interface between higher linguistic levels, such as morphology, and the act of speech sound production. It is commonly assumed that the output of the phonological component of a grammar is the input to the speech production system. This assumption is highly important because it is the means by which speech as an observable motor behavior is conjoined to the covert processes of language formulation.

Syllables

The syllable is a unit frequently used to describe phonological patterns, events in speech production, and aspects of speech perception. Although a syllable is difficult to define precisely, the concept of syllable offers a number of advantages to phonology and speech science. A useful definition of syllable is in terms of its internal hierarchical structure. The onset and rhyme are the two basic parts of a syllable. The onset is the initial consonant or consonant cluster. The rhyme is the remainder of the syllable and is further analyzed as a peak (which contains the syllabic nucleus) and the coda (the final consonant or consonant cluster).

Why is the concept of syllable important? Among other things, syllabic structure constrains the possible sequences of phonemes, controls allophonic variation, predicts stress patterns in a language, and accounts for various phenomena in speech perception.

Marcus and Syrdal (8) give an example of how syllables constrain possible sequences. They point out that the phonemes of American English form over 4 billion arbitrary sequences of one to six members. However, there are only 100,000 possible monosyllabic (one-syllable) words, and only one tenth of these are actual monosyllabic words. To take the statistical analyses a bit further, syllables also differ in their frequency of occurrence. Dewey (9) determined from frequency counts of human speech that the 12 most frequently used syllables account for about a quarter of our verbal behavior, that only 70 different syllables make up half of our speech, and that 1370 syllables are sufficient for over 90% of what we say. Not surprisingly, syllables are attractive candidates as units for speech analysis and for applications in speech synthesis and machine speech recognition.

Syllables control allophonic variation because sound patterns in a language are sensitive to syllabic position and syllabic structure. First, phonemes have different privileges in syllabic formation. Some phonemes, such as the /N/ [for convenience, phonetic symbols are expressed here using the PHONASCII system (6)] at the end of the word *thing* cannot occur in syllabic-initial position. Others, such as the /h/ at the beginning of the word *hay* cannot occur in syllabic-final position. Onsets and codas have certain permissible consonant sequences within a given language. For example, in American English, the fricative /S/ (as in the word *she*) can be combined with only one other consonant (for example, /r/ as in *ray*) to form an onset. (Words such as *schlemiel* and *schlepp* are borrowed from Yiddish and do not accord with the constraints of American English.) Table 1 shows a matrix that specifies the possible two-element consonant clusters in syllabic-initial position. Only a fraction of the possible clusters are actually used in the language.

As one example of syllable-controlled allophonic variation, voiceless stops in American English are produced as aspirated allophones (meaning that they are produced with laryngeal

Table 1. A Matrix that Specifies the Possible Syllable-Initial Consonant Clusters in American English[a]

	w	l	r	p	t	k	m	n	f	O
p	−	+	+	−	−	−	−	−	−	−
b	−	+	+	−	−	−	−	−	−	−
f	−	+	+	−	−	−	−	−	−	−
t	+	−	+	−	−	−	−	−	−	−
d	+	−	+	−	−	−	−	−	−	−
O	+	−	+	−	−	−	−	−	−	−
k	+	+	+	−	−	−	−	−	−	−
g	+	+	+	−	−	−	−	−	−	−
s	+	+	−	+	+	+	+	+	?	−
S	+	+	+	?	?	−	?	?	−	−

[a]The rows show the first element of the cluster and the columns, the second. A "+" indicates that the row/column pair forms an acceptable cluster, and a "−" indicates that the pair is not an acceptable cluster. A "?" indicates some uncertainty, primarily because of rare occurrences, especially regarding borrowed non-English words containing the pair.

noise) except when they follow /s/ in an onset. In words, such as *spot, stay,* and *ski,* the voiceless stops are produced as unaspirated allophones (the laryngeal noise is absent). A number of syllable-based phonological patterns are discussed in an influential thesis by (10). Fujimura and Lovins (11) note that the majority of known allophonic variations can be described with respect to intrasyllabic contextual factors.

Syllabic structure is sufficient to predict stress patterns of words and to account for various tonal phenomena (12–15). Stress holds particular importance for word recognition in many languages. For example, in American English, about 90% of words begin with stressed syllables (16). This high percentage of stressed syllables at word onset is a useful property in speech perception because it helps to identify word beginnings in the flow of speech. Stress pattern is important also because the level of stress on a syllable is a major factor in determining syllabic duration and the properties of articulatory movement within a syllable.

Syllables appear to play a role in the timing of speech movements, both across and within syllables (17). It appears that many movements in speech are defined with respect to syllabic boundaries, so that the sequencing of movements is often described with respect to intrasyllabic regularities. Syllables therefore are useful in describing the temporal pattern of articulation. This topic is discussed further in a following section.

Phrasal Structures

Larger units such as phrases, sentences, and even discourse also have been proposed as units to analyze speech. It is only in larger units of this kind that some important phenomena are identified and described. For example, declarative statements in American English are typically produced with a vocal pitch that exhibits a general falling pattern, that is, the pitch is highest at the beginning of utterances and then falls to lower values. This *downdrift,* or falling, pattern of intonation for declarative statements contrasts with a final-rising pattern typical of interrogative statements. Phrasal structures also are relevant for the study of several issues related to stress or rhythm, which are often described relative to the syllabic sequence of utterances. For example, it has been proposed that American English has a strong-weak alternation in its syllabic pattern, so that syllables with strong stress alternate with syllables of weak stress (as exemplified in the words *telephone* and *refrigerator*).

Segmentation

If intuition is correct, then it should be possible to segment the speech signal, that is, to divide the signal into constituent pieces corresponding to phonemes, allophones, syllables, or other discrete elements. The assumption of *segmentation,* however, is not easily demonstrated, and considerable effort has been given to developing algorithms that segment a speech signal into elements useful in reconstructing a speaker's words. This objective is central to many efforts in machine speech recognition, but it also pertains quite generally to acoustic and physiological studies of speech, if only because segments provide a reasonable way of organizing and interpreting multidimensional data. There are a number of factors that contribute to the difficulty of segmentation. Many of the factors are related to the biophysical properties of human speech production. To understand these factors, some basic anatomy and physiology is needed.

THE PHYSICAL STRUCTURES OF HUMAN SPEECH PRODUCTION

A useful simplification to describe speech production is to recognize the interaction of three major subsystems: respiratory, laryngeal, and articulatory (Fig. 1). Although these subsystems interact to produce speech, their principal functions are different enough that they can be considered separately as a first step in understanding speech production. Furthermore, this tripartite approach is also advantageous for discussing speech acoustics, considered later.

The respiratory subsystem, consisting of the lungs, chest wall, and abdomen, provides the aerodynamic power of speech (18,19). The air in the lungs is the means by which air pressures and flows are generated to produce speech sounds. In American English, speech is produced on the egressive (outflowing) air stream. Therefore, speech is essentially a modulation of the expiratory phase of respiration. At least for adults, the aerodynamic requirements of conversational speech are modest and barely exceed the amount of air moved during rest breathing (about 500 mL for an adult male). Greater volumes of air are used for more forceful speech, such as in shouting. Because speech is produced on the expiratory phase of respiration, the temporal pattern of speech can be defined in terms of the *breath group* (the syllabic group formed on a single expiration). The breath group is an important unit in many formulations of prosody. It is a convenient and natural way to specify a group of syllables with a functional unity imparted by the mechanics of the respiratory-laryngeal system. For several purposes in speech analysis, the concept of the breath group is a helpful initial step in determining the linguistic-temporal structure of an utterance.

The laryngeal subsystem (or simply larynx), composed of an intricate assembly of cartilages, muscles, and related tissues, produces the basic vibratory energy of voice by valving the airstream created by respiration. The larynx is popularly called the *voice box.* Voice, or vibration of the vocal folds, is produced as air from the lungs sets the folds into a self-sustaining oscillation involving the interaction of a mucosal traveling wave and airflow through the glottis, or the opening between the vocal folds (20–22). Voice is important as a source of energy for the voiced sounds and also as a carrier for prosodic information, such as intonation. As noted previously, declarative statements in English generally have a downdrift pattern of vocal fundamental frequency. The fundamental frequency is typically highest at the onset of the utterance and then falls gradually until the end of the utterance (or breath group). Deviations from the overall falling pattern are introduced to mark stress or to produce emphasis. Vocal fundamental frequency is also varied along with loudness, speaking rate, and other dimensions of voice quality to communicate the emotional aspects of speech (23). In addition, the larynx has an important role for voiceless sounds such as the /s/ in the word *see.* For these sounds, the vocal folds are separated so that air pressure from the lungs is developed behind a point of articulatory constriction.

The articulatory subsystem is the upper part of the respiratory airway, that part extending from the larynx to the

Table 2. Major Articulators for the Production of English Consonants

	Stops		Nasals	Fricatives or Affricate			Glides or Liquid
Both lips	/p/ pay	/b/ bay	/m/ may				/w/ way
Lips and teeth				/T/ thin	/D/ this	/f/ /v/ fat vat	
Tongue tip	/t/ tip	/d/ dip	/n/ nip	/s/ sip	/z/ zip		/l/ lip
Tongue blade				/S/ shin	/tS/ chin	/dZ/ gin	/j/ yawn
Tongue dorsum	/k/ cap	/g/ gap	/N/ rang				
Glottis				/h/ hay			

openings at the lips or nostrils. Here most sounds are formed using the energy of the respiratory and laryngeal subsystems (24). Articulation generally means movement. As applied to speech, articulation means movements that produce speech sounds. The movements of primary interest are those of the lips, tongue, jaw, soft palate, and pharynx and glottis. These structures are called the articulators. They are semi-independent in their movement capabilities. Therefore, they are useful in describing how speech sounds are formed. Table 2 shows how these articulators are deployed to make the consonants of American English. The articulators are not completely independent because they are mechanically coupled in various ways. For example, the lower lip and tongue are attached to the jaw and therefore are affected by jaw movement. In addition, the tongue and palate are interconnected by muscles and ligaments that impose a biomechanical linkage. Even the larynx is interconnected to the tongue and jaw. These linkages impose constraints and interactions on the speech production system. Although the biomechanical properties of speech production are only partly understood today, it is generally believed that a more thorough description of these properties will help considerably in developing articulatory synthesizers (machines that produce speech with analogs of human articulators).

To produce speech, the articulators move from position to position. Because of inertia and the biomechanical linkages described earlier, movements are noninstantaneous and interactive. X-ray motion pictures of speech reveal that the articulators are in nearly continuous motion. Furthermore, the motions are complex in respect to the number of ongoing movements and also their temporal organization. The articulatory movements for a given speech sound do not necessarily begin and end at the same time. Therefore, speech takes on a pattern of overlapping movements. This overlapping results in coarticulation, the simultaneous adjustment of the articulatory system to two or more presumed control units. Because of coarticulation, a given segment of speech is "flavored" by the phonetic context in which it occurs. For example, in American English, vowels tend to be nasalized if they precede nasal consonants, and many consonants are produced with lip rounding if they precede rounded vowels. Coarticulation has been a particular challenge in understanding the organization of speech movements, and several theories have been advanced to explain its patterns (25–27).

A speaker's sex and age determine some important characteristics of speech, especially in its acoustic signal. These differences are discussed in more detail in a following section, but for the moment it should be noted that the vocal tract grows in length roughly on the same schedule as the general skeletal system. But the vocal tract also differs in its relative configuration between children and adults and between women and men. For example, men have a proportionately longer pharynx than women or children. Individual differences in the anatomy of the speech production system contribute to the distinctiveness of different voices, and these anatomical differences are a basis for speaker identification in forensic and other applications. Individuals also differ in learned patterns of speech, and these variations hold potential for forensics. For a discussion of speaker identification, see Kent and Chial (28).

LEVELS OF OBSERVATION IN STUDIES OF SPEECH PRODUCTION

Speech is described and understood in various ways, depending on the methods and purposes of study. Figure 3 illustrates some examples. First, if the intent is to know how individual muscles participate in speech production, then detailed

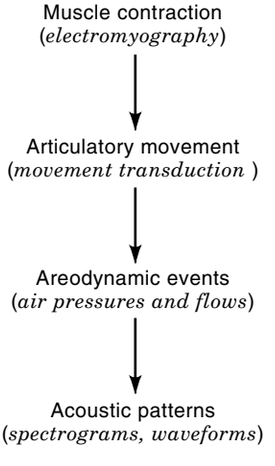

Figure 3. Levels of observation in studies of speech production.

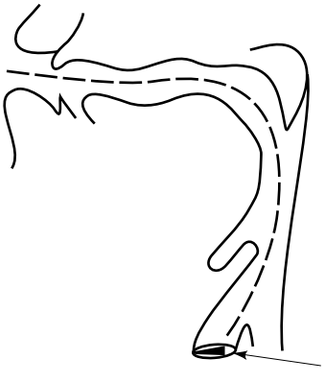

Figure 4. Midsagittal section of vocal tract, which extends as shown by the dashed line from the vocal folds (arrow) to the lips.

information about the anatomy of muscles and associated tissues is required. Second, if muscles can be neglected in favor of an articulatory characterization, then it is sufficient to identify the major movable structures and their movement possibilities. Third, if the interest is in aerodynamic variables in speech production, then only variables, such air volumes, pressures, and flows need be considered. Finally, if the goal is to understand the relationship between the vocal tract configuration and the acoustic output, then only the length and shape of the cavities need to be considered. The remainder of the article summarizes some major issues pertaining to acoustics, aerodynamics, and articulatory aspects of speech.

ACOUSTIC THEORY OF SPEECH PRODUCTION

The acoustic theory of speech production is found in an influential book by Gunnar Fant (29) published in 1960. The theory stands out in speech communication research because it provides a generally accepted account of how the speech mechanism produces sound. Most, if not all other theories of speech production are still under constant and vigorous debate. Fant's theory has undergone some minor modification since its publication nearly 40 years ago and still encounters challenges to certain of its assumptions and predictions. But the broad outline of the theory and many of its specific aspects are the bedrock of the discipline of speech production research.

A simple input-resonator-output model provides a schematic summary of the theory's account of vowel sounds. The vocal tract, defined previously as the length of tube extending from the glottis to the lips (see Fig. 4) is a resonator (or filter) that is excited by the vibrating vocal folds. More specifically, the *filter function* of the vocal tract tube *shapes* the spectrum produced by the vibrating vocal folds. The frequency locations of the multiple vocal tract resonances are determined by the specific configuration of the tube, which is a function of articulator (i.e., tongue, lips, jaw, pharynx, soft palate) positions. The spectrum of the input is also determined systematically, in this case by the specific details of vocal fold vibration.

Input (Source) Function

Vibration of the vocal folds, which serves as input to the vocal tract resonating tube, is called *phonation,* or voicing. Phonation is initiated when the vocal folds, which are no more than about 20–28 mm long in adults (30), are brought together by muscular forces but are then blown apart by an aerodynamic pressure differential across the closed glottis. Once set into motion, the vocal folds go through a series of opening and closing movements sustained by a complex interaction between aerodynamic and mechanical forces. Excellent summaries of these motions and their sustaining forces are available (22,31).

The vibratory motions of the vocal folds consist of a series of opening and closings which can occur at a wide range of frequencies. The frequency of the vibration and the vibratory mode are determined by the complex interplay of forces developed by a small set of muscles within the larynx (32,33). The human vocal folds have a remarkable range of frequencies and vibratory modes because of their unique histological structure, uncovered in a series of studies by Hirano and his colleagues. The point here is that the time- and frequency-domain characteristics of the vocal fold input function depend critically on the precise settings in the larynx. So our current characterization of the input function uses a prototypical vocal fold behavior. This behavior is referred to as *modal* phonation, the type of vocal fold vibration used in everyday communication. Vibration of the vocal folds in modal phonation involves relatively complex movements of the tissue which generate a spectrum rich in harmonics. Another well-studied phonatory mode is called *falsetto* (or *loft;* see Ref. 34), a type of vocal fold vibration with very high longitudinal tension on the vocal folds resulting in a rather simple, piston-like movement of the folds. As might be expected, the relatively simple vocal fold movements in falsetto phonation produce a relatively simple spectrum, with greatest energy at the fundamental frequently (*F*0: the primary rate of the vibration) and little energy at higher harmonics. *F*0s in falsetto phonation are typically much higher than those in modal phonation because of the high longitudinal, vocal fold tensions that define the falsetto mode. Another frequently discussed but poorly understood phonatory mode is called *vocal fry,* or *pulse* phonation. Vocal fry is produced with very low *F*0s, typically well below those associated with modal phonation, and is heard as a characteristic *popping* sound (children sometimes produce vocal fry when imitating a motorboat). As in falsetto, there is high laryngeal tension in vocal fry, but in the latter case it is a type of tension that squeezes, rather than stretches the folds. Both falsetto and vocal fry are not normal types of phonation for *chronic* voice production. When they are used chronically, at least in most Western cultures, they are considered speech pathologies. Falsetto is used, however, as a *normal* feature in singing (such as in yodeling, or in much of the Motown sound of the 1960s and the subsequent music it inspired), and vocal fry is often produced at the end of sentences as part of the paralinguistic signaling of *your turn* in normal conversation.

Figure 5(a) shows three cycles of a glottal flow waveform for an adult female whose vocal folds are vibrating at a frequency of 200 Hz in the modal type of phonation. In living humans it is essentially impossible to collect glottal flow directly, so these data must be obtained by sampling the flow through the mouth opening and then removing the resonances of the vocal tract from the collected signal (35). The residual of this inverse filtering process is the glottal flow, free of any influence from the vocal tract resonator. When the

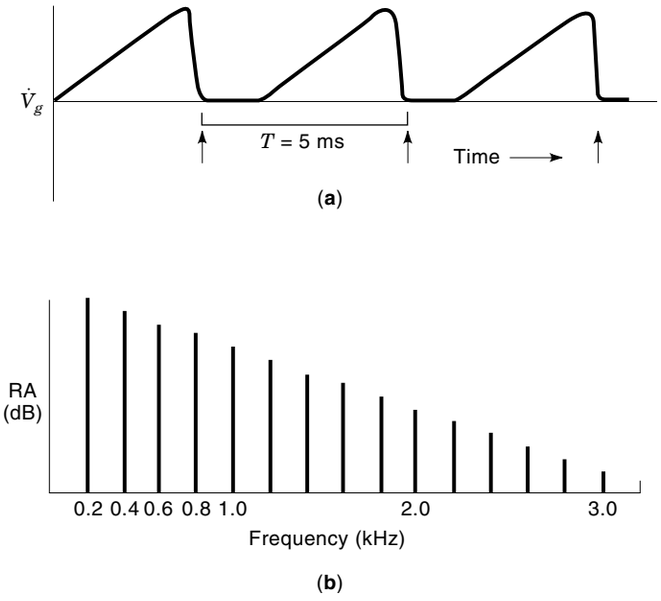

Figure 5. (a) Glottal flow ($\dot{V}_g$) waveform and (b) associated glottal spectrum for a fundamental frequency of 200 Hz.

glottal flow signal is above the baseline, the vocal folds are open, allowing flow from the trachea to the pharynx. When the signal is on the baseline, the vocal folds are closed. In this waveform the glottis is closed approximately 40% of the period of vocal fold vibration. The opening phase of the glottal flow cycle is shallower than the closing phase. The latter is a good reflection of how rapidly the vocal folds snap shut at the end of a cycle.

Frequency-domain analysis of the waveform shown in Fig. 5(a) produces a glottal spectrum like that shown in Fig. 5(b). The spectrum consists of a consecutive integer series of harmonics whose amplitudes decrease at roughly 12 dB/octave. As shown by (36), the tilt of the glottal spectrum is very much a function of the slope of the closing phase. The steeper the closing phase, the less tilted (i.e., the flatter) the spectrum.

The spectrum shown in Fig. 5(b) is the input to the vocal tract resonator. The actual time of primary excitation, when the pressure wave is propagated from the vibrating vocal folds through the vocal tract, occurs at the instant of closure during each glottal cycle. This point is marked on Fig. 5(a) by the upper pointing arrows. It is important to note that the vocal tract resonances shape the spectrum generated by the vibrating vocal folds, rather than responding to the forces exerted by the puffs of air coming thorugh the folds. One of the central tenets of the acoustic theory of speech production is that the source and filter (the latter to be discussed in the next section) are independent. At a coarse level of analysis, this is true as evidenced by a speaker's ability to change the filter (the position of the articulators) without affecting the vibratory rate of the vocal folds, or to change the vibratory rate of the folds without changing the filter. It was known many years ago, however, that details of the glottal waveform were sensitive to the configuration of the vocal tract (37). Over the past 25 years there has been a good deal of work on the interactions between source and filter, and the effects of those interactions on the source spectrum and perception of voice quality; see Ref. 38 for a summary of some of this work.

Filter Function (Resonator)

The basic resonator model of the vocal tract for vowels is a tube closed at one end. The closed end is at the glottis where the excitation is supplied each time the vibrating vocal folds snap shut, and the open end is at the lips. The vocal tract responds to each of these repeating excitations with a set of damped oscillations at the resonant frequencies of the tube, determined by the configuration of the tube. The damped vocal tract oscillations induced by successive excitations overlap. The running amplitude of the resonances at any time is determined by the superimposition of new and decaying oscillations. The phase relationships of these overlapped oscillations are of little importance in understanding the phonetic behavior of vocal tract acoustics.

What is important is the way in which deformations of the vocal tract result in shifts in the resonances, called *formant frequencies.* If we take a tube closed at one end and with a uniform cross-sectional area from end to end, the resonances are determined by the quarter wavelength rule, namely, $f_r = (2n - 1) \times c/4l$, where c = the constant, speed of sound in air, l = the length of the tube, and n = the number of the resonance. In actual speech production, the vocal tract is most like a tube with uniform cross-sectional area when a schwa (symbolized phonetically as /ə/) is articulated, as in the first sound of the word *about.* If the first three formant frequencies, symbolized as $F1$, $F2$, $F3$, are measured for a schwa produced by an adult male, the agreement with the values predicted by the quarter-wavelength rule are quite good ($F1$ = 487 Hz, $F2$ = 1461 Hz, and $F3$ = 2435, assuming c = 33,140 cm/sec and l = 17 cm. The latter is selected because it is a length typical for the adult male vocal tract). In speech production, however, most vowels are produced with vocal tract tube shapes that deviate markedly from a uniform cross sectional area. For example, the vowel /i/ (as in the word *feet*) is produced with a substantial constriction toward the front of the vocal tract, whereas the vowel /ɑ/ (as in the word *hot*) is produced with a tight constriction in the back of the vocal tract. Midsagittal vocal tract shapes for schwa (/ə/), /i/, and /ɑ/, adapted from cineradiographic films traced by Perkell (39), are shown at the top of Fig. 6. Underneath each vocal tract shape is a tube resonator model simulating the cross-sectional areas of the vocal tract from the closed end (the glottis) to the open end (the lips). When a tube with uniform cross-sectional area is deformed in the ways shown for /i/ and /ɑ/ in the lower part of Fig. 6, the resonant frequencies deviate in systematic ways from those predicted by the quarter-wavelength rule. Specifically, when tubes are excited by an acoustic source, the resonances are associated with frequencies whose wavelengths are realized within the tube as standing waves. These standing waves are understood in terms of their pressure distributions, or their mirror-image velocity distributions. Regions of high pressure within the tube correspond to regions of low velocity, and vice versa. The standing wave patterns for all tube resonances are present at any instant, resulting in multiple locations of high pressures (low velocities) and low pressures (high velocities) throughout the tube. When a region of high pressure associated with a particular wavelength or resonant frequency is constricted

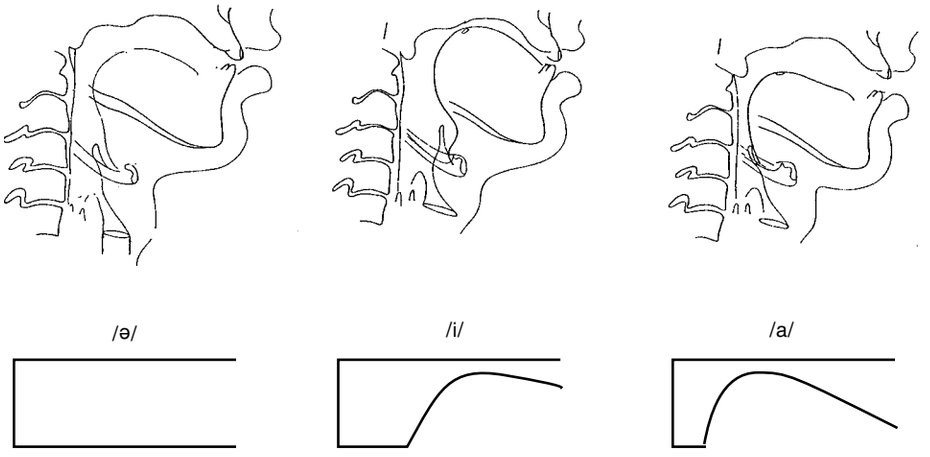

Figure 6. Vocal tract shapes for the vowels /ə/, /i/, and /a/.

within the tube, the region becomes less compliant and the resonant frequency increases. On the other hand, when a region of high velocity is constricted, the region is made more inert, resulting in a decrease in the resonant frequency. Thus deformations of the tube give rise to systematic changes in the resonant frequencies predicted by the quarter-wavelength rule for the uniform cross-sectional area.

This model of tube deformation and resulting resonant frequencies explains, in large part, the relationship between articulatory configuration and formant frequencies. Figure 7 shows the pressure distributions for the first two resonant modes of the tube, below a tracing of the vocal tract. The standing waves are shown separately, but of course are superimposed in the vibrating tube. The vertical dashed line toward the open end of the tube, labeled /i/, shows the approximate location of the tongue constriction for this vowel. This constriction occurs at a relatively low pressure (high velocity) for the first mode and at a pressure maximum for the second mode. According to the tube model, this constriction should result in an $F1$ (first mode) and $F2$ (second mode) lower and higher, respectively, than the $F1$ and $F2$ of the unconstricted tube. Similarly, the constriction location for the vowel /ɑ/ (right dashed line) occurs at a relatively high pressure for the first mode and close to zero pressure for the second mode. For this constriction, $F1$ and $F2$ should be higher and lower, respectively, than the $F1$ and $F2$ of the unconstricted tube. Table 3 reports the $F1$ and $F2$ frequencies calculated previously for a 17 cm long tube with no constrictions and data from the literature for the vowels /i/ and /ɑ/ produced by adult males. The actual formant values measured for the vowels /i/ and /ɑ/, relative to the values calculated for the tube of uniform cross-sectional area, are changed by the constrictions in the directions predicted by the acoustic model previously described. The model can be generalized and understood in articulatory terms by viewing an $F1$-$F2$ plot for the corner vowels (/i/, /u/, /ɑ/, /æ/) of American English, produced by men, women, and children (Fig. 8). This is a traditional way to plot the acoustic consequences of vowel articulations at the extremes of vowel articulation (i.e., the *corners* of vowel articulation within the vocal tract), and suggests certain broad summary statements concerning the relationship between articulatory dimensions and formant frequencies. For example, the vowels /i/ and /u/ are both called high vowels, because their major constrictions are located relatively high in the vocal tract (i.e., the tongue is close to the palate and the mandible is in a relatively high position). These vowels define the front and back boundaries of high vowels, along the dimension called *tongue advancement:* /i/ has the most forward constriction, and /u/ the most back constriction of these vowels. Note that in the $F1$-$F2$ plot, the major acoustic variation between /i/ and /u/ takes place along the $F2$ axis. The first generalization, therefore, is that changes in the tongue ad-

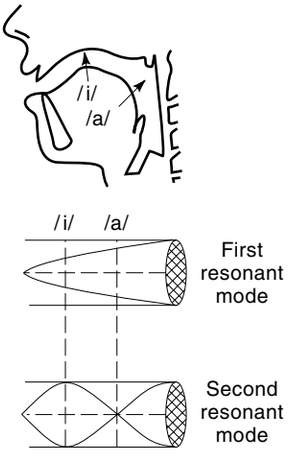

Figure 7. Pressure distributions for the first two resonant modes of an acoustic tube with constriction locations indicated for the vowels /i/ (left dashed line) and /ɑ/ (right dashed line). The horizontal dashed lines indicate atmospheric pressure, and the approximate constriction locations are shown in the vocal tract tracing above the tubes.

Table 3. A Comparison of the First Two Formant Frequencies for the Schwa[a] and the vowels /i/ and /a/

	$F1$	$F2$
/ə/	487	1461
/i/	270	2300
/ɑ/	730	1100

[a]Computed from a straight-tube model of the vocal tract.

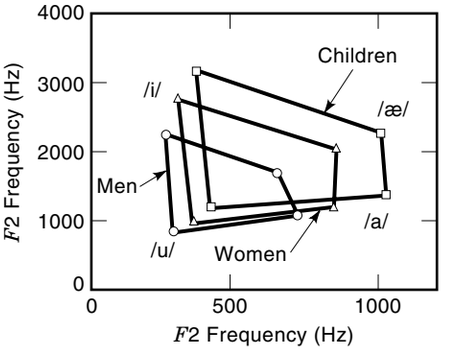

Figure 8. *F*1-*F*2 plot for the corner vowels of American English, produced by men (circles), women (triangles), and children (boxes).

vancement dimension affect primarily *F*2, with more forward constrictions resulting in higher *F*2's. Tongue advancement also has some smaller, but systematic, influences on *F*1: more forward constrictions are associated with lower *F*1's. Both of these effects are completely consistent with the predicted effects of constrictions along the pressure (flow) distributions associated with the first and second vibratory modes of a tube closed at one end. The vowel pairs /u/-/ɑ/ and /i/-/æ/ define the ends of the tongue height continuum in vowel production. The *F*1-*F*2 plot shows that a good part of the acoustic variation along the tongue height dimension is in *F*1, which suggests our second generalization: changes in tongue height affect primarily *F*1. Higher vowels (those with a less open mandible) have lower *F*1's (see Fig. 8).

In addition to the mapping between these two tongue dimensions and vowel formant frequencies, the acoustic theory of speech production also contains one other important generalization concerning articulatory-acoustic relationships. The configuration of the lips has a dramatic effect on the acoustic output of the vocal tract, and in many languages is the distinguishing feature between two vowels with nearly the same tongue configurations (e.g., Swedish has two vowels with high front tongue positions [like the tongue position in English /i/], one produced with rounded lips, the other with spread lips). Within the framework of the model described previously, rounding of the lips is understood as a lengthening of the tube, which lowers all the resonant frequencies. In the case of the vocal tract, lip rounding has the most dramatic influence on *F*2, but *F*1 and *F*3 are also affected.

Figure 9 presents an *F*1-*F*2 plot of actual formant frequency data collected by Peterson and Barney (40) from children of both genders and from adult men and women. Each plotted point represents the coordinates for a single speaker's production of a given vowel, and the ellipses enclose the majority of points associated each vowel. Within each ellipse, there is a good deal of variability in the *F*1-*F*2 values for a particular vowel. Three questions can be raised about this variability. First, why is there such a huge range of variation in the important formants for a single vowel when there is a highly deterministic relationship between articulatory configuration and formant frequencies? Second, how can so many different *F*1-*F*2 values be heard as the same vowel? And third, how can two vowels with roughly the same formant frequencies be heard as *different* vowels (e.g., note regions where ellipses from two different vowel categories overlap)? The first question is directly relevant to the theme of this article, whereas the second and third are more appropriately discussed in the article on Speech Perception. Some speech production and perception issues, however, are not easy to separate, so we consider both issues.

First, the large range of variation in *F*1-*F*2 coordinates for a given vowel is explained on the basis of *speaker differences.* The most obvious one of these differences concerns speaker gender, which is correlated with vocal tract length. Men have longer vocal tracts than women, who have longer vocal tracts than children. Much of the spread of the coordinate points reported by (40) for a specific vowel reflects the different vocal tract lengths among their pool of adult and child speakers. Somewhat more subtle speaker effects are found in dialect differences, which are often expressed in the details of vowel production. Peterson and Barney (40) made no attempt to control dialect among their speakers, which in fact is exceedingly difficult to do even when birthplace, years of residence, and other likely influences are not allowed to vary among speakers. The issue of speaker variability has special relevance for machine recognition of speech, the algorithms of which must be sufficiently flexible to recognize the kind of variability seen in Fig. 9 as within-, rather than across-, category exemplars. A recent text (39) is devoted to the issue of speaker variability as it affects various problems in speech production, perception, and machine recognition.

In Peterson and Barney's study (40), all vowels were produced in a simple and constant phonetic environment. When the same vowel is produced in different phonetic contexts, at different rates, with different levels of stress, or possibly even in different speaking "styles" (e.g., formal versus informal), there is also a good deal of variability in the formant patterns (42–45). Thus in connected speech, where it is assumed that all these factors vary simultaneously, there are multiple

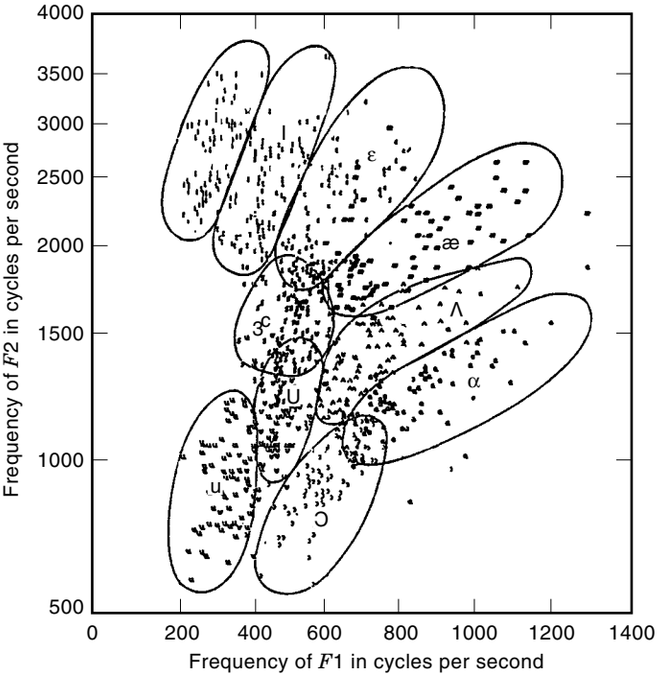

Figure 9. *F*1-*F*2 plot for the data of (40).

sources of variability affecting the formant patterns of a specific vowel.

There may even be other sources of variability in vowel formant frequency data that represent measurement issues and/or speaker-cohort effects. Hillenbrand et al. and (46) replicated Peterson and Barney's (40) experiment and found that the degree of crowding in their *F*1-*F*2 space, the extent to which some vowel categories seemed to overlap, was even greater than that reported by the latter authors. Table 4 compares the averaged formant frequencies of adult males and females reported by (40) and (46). The major difference between the two data sets is in the vowels /ɛ/, /æ/, /ɑ/, and /ɔ/. As might be expected, the measurement techniques used in these two studies, separated by forty years, were quite different and could account for some of the reported differences. Speaker-cohort effects, wherein changes in vowel pronunciation occur over time and result in subtle or easily noticeable *dialects* for speakers from different generations, may also contribute to the different findings. Surprisingly, even though the *F*1-*F*2 space was more crowded in the Hillenbrand et al. study (46), listeners' ability to *identify* the spoken vowels was equivalent to that reported by Peterson and Barney (40). This result seems unexpected if the separation of vowel categories in *F*1-*F*2 space is related more or less directly to the perceptual distinctiveness of vowels.

The issue of how within-category acoustic variation for a particular vowel is ignored by listeners for phonetic identification (i.e., identifying which vowel category has been produced) has been the subject of numerous investigations over the last 25 to 30 years. Some summaries of the problems and issues in this area are in the literature (47–50). The field has yet to embrace a particular account of how listeners accomplish this task. A partial answer to this problem, however, has been proposed by Hillenbrand and Gayvert (50) and by some other authors (47). Both Peterson and Barney (40) and Hillenbrand et al. (46) used the *standard* approach to measuring vowel formant frequencies by constructing a spectrum for a relatively short-duration (~25 ms to 50 ms) window from the temporal middle of the vowel. This acoustic representation of a vowel, however, ignores the formant frequency variation outside the measurement window, as demonstrated by Hillenbrand and Gayvert (50). These authors performed discriminant function analysis on the formant frequencies measured in the traditional way, from the temporal middle of the vowel, and also by adding into the solution the formant frequencies measured close to the onset and offset of the vowel. The discriminant function was more successful in separating the vowel categories with measurements from the three points (onset, middle, onset) compared with the single measurement (middle). Strange (47) has reported on a series of studies in which vowel identifications obtained from listeners are better when the entire vowel trajectory, rather than just a single point in the middle of the vowel, is presented. Moreover, in several studies, listeners produce accurate vowel identifications when the middle of the vowel has been edited out of the overall trajectory, leaving just the onset and offset portions of the trajectories. These kinds of acoustic and perceptual findings (47,50) suggest that the dynamic information across a vowel trajectory, rather than a static set of frequencies from the temporal midpoint of a vowel, is critical for the acoustic representation of vowels. In this view, the apparent inconsistency between the findings of Hillenbrand et al. (46)

Table 4. Mean Data on Fundamental Frequency and the First Three Formant Frequencies for Vowels Produced by Adult Male and Female Talkers[a]

	/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/
Males												
*F*0												
	136	135	—	130	127	124	129	—	137	141	130	133
	138	135	129	127	123	123	121	129	133	143	133	130
*F*1												
	270	390	—	530	660	730	570	—	440	300	640	490
	342	427	476	580	588	768	652	497	469	378	623	474
*F*2												
	2290	1990	—	1840	1720	1090	840	—	1020	870	1190	1350
	2322	2034	2089	1799	1952	1333	997	910	1122	997	1200	1379
*F*3												
	3010	2550	—	2480	2410	2440	2410	—	2240	2240	2390	1690
	3000	2684	2691	2605	2601	2522	2538	2459	2434	2343	2550	1710
Females												
*F*0												
	275	232	—	223	210	212	216	—	232	231	221	218
	270	224	219	214	215	215	210	217	230	235	218	217
*F*1												
	310	430	—	610	860	850	590	—	470	370	760	500
	437	483	536	731	669	936	781	555	519	459	753	523
*F*2												
	2790	2480	—	2330	2050	1220	920	—	1160	950	1400	1640
	2761	2365	2530	2058	2349	1551	1136	1035	1225	1105	1426	1588
*F*3												
	3310	3070	—	2990	2850	2810	2710	—	2680	2670	2780	1960
	3372	3053	3047	2979	2972	2815	2824	2828	2827	2735	2933	1929

[a]The first entry in each cell is a value from (40). The second entry is from (46).

and Peterson and Barney (40) for the temporal midpoint measurements does not have the same kind of theoretical import and may reflect no more than noise in the measurement and speaker selection process.

THE INTERSECTION OF PHONOLOGY, SPEECH ACOUSTICS, AND SPEECH MOVEMENTS

One particular offshoot of the acoustic theory of speech production shows how the areas of speech motor control (i.e., the neurophysiological control of articulatory movements), vocal tract acoustics, and linguistics might be interrelated in a theory of the sound structure of languages. K. N. Stevens (51–53) has developed the *quantal theory* of speech production (see also Ref. 54) which states that languages of the world seek regions of the vocal tract for phoneme production where a certain amount of articulatory imprecision does not affect the acoustic output. These quantal regions in the vocal tract are favored by languages because they protect the acoustic integrity of sound categories in the face of the inability to reproduce exactly tongue, jaw, and other articulator movements each time a particular sound category is produced. Different languages have different sound inventories, but certain sounds appear in these inventories much more frequently than other sounds. For example, Stevens (51–53) points out the high frequency of occurrence of vowels such as /i/ and /ɑ/ in languages of the world, and shows by acoustic modeling how the vocal tract acoustic output for these vowels is not sensitive to some aspects of articulatory variation. There are also some tongue position data (55) consistent with Stevens' claims.

The quantal theory is attractive because it attempts to join several different levels of the sound production process, previously mentioned. Other theories of speech production, reviewed briefly here, have typically been less ambitious and focused only on a particular level of the process. The quantal theory is not without problems (53), however, and is probably more valuable for the model it provides of what a speech production theory *should* account for, than for what it actually explains.

Nasals and Nasalization

The nasal consonants /m/, /n/, and /N/ (see Table 2) are produced with complete closure of the oral vocal tract and an open velopharyngeal port. In /m/ the vocal tract seal is provided by the closed lips, in /n/ by placement of the tongue tip against the alveolar ridge, and in /N/ by contact of the tongue dorsum with the posterior part of the hard palate and anterior part of the soft palate. With the velopharyngeal port open and a complete closure in the vocal tract, the acoustic system is different from that described above for English vowels, for which the velopharyngeal port is closed. In the case of nasal consonants, the main pathway for sound transmission to the atmosphere is via the nasal cavities and through the nostrils, but the closed vocal tract chamber contributes to the acoustic output as well. Specifically, the closed vocal tract chamber acts as a "trap" for acoustic energy at frequencies determined by the its dimensions. Energy may also be trapped in the closed resonators that are formed by the sinus cavities, which branch off from the main nasal cavities (54,55). The trapped energy in the closed, side-branch resonators of the vocal tract and sinus cavities results in *antiresonances* in the spectrum of the radiated signal, or regions where little energy is allowed to pass into the atmosphere. Nasal consonants are characterized by a series of resonances related to the dimensions of the combined pharyngeal and nasal cavities, but the antiresonances are a prominent feature of nasal consonant spectra and are necessary for accurate synthesis of nasal consonants and may play some role in speech perception by providing cues to which nasal consonant is being articulated (29).

The acoustic interval during which a nasal consonant is articulated is often referred to as the nasal *murmur.* There are also cases where both the vocal tract and velopharyngeal port are open, resulting in a vowel sound that is *nasalized.* The acoustics of nasalization are somewhat distinct from those of nasal murmurs and are important for several reasons. In some languages, such as French, the difference between nasalized and non-nasalized vowels is phonemic. In English words such as *meat* and *team* the velopharyngeal port remains open for a short time following the nasal consonant (*meat*) or begins to open before the nasal consonant is actually produced (*team*); this results in portions of the vowel adjacent to the nasal consonant being nasalized, the acoustic consequences of which can be used by listeners in decoding the sound structure of a word. And finally, nasalization is important because individuals with certain speech disorders, such as those due to structural abnormalities of the speech mechanism (as in cleft palate) or to a variety of neurological diseases, may be unable to close the velopharyngeal port and will therefore have chronically nasalized vowels. The spectra of nasalized vowels have a distinct pattern in the 0 Hz to 1200 Hz range, which consists of a antiresonance in the region of 400 Hz, flanked by a lower-frequency resonance of the nasal cavities and a higher-frequency resonance which is the first formant of the oral vowel; this pattern is almost certainly a critical cue to the perception of nasalization (56).

Antiresonance also occur in the spectra of stops and fricatives, as well as in some lateral sounds (as in /l/). For a full treatment, see Fant (29).

SPEECH AERODYNAMICS AND OBSTRUENT PRODUCTION

The acoustic theory of speech production's account of vowel production is quite accurate. The account of consonant production and particularly of the class of consonants requiring a complete or nearly complete blockage of the airstream is fairly good though somewhat less precise. This class of consonants, called *obstruents,* has more complex sources and resonance patterns than vowels and is also associated with much higher frequencies. The latter point is important because vowel frequencies (typically below 4.0 kHz) are consistent with the assumption that pressure wave propagation in the vocal tract is only planar, whereas many consonants have important frequencies (above 4.0 kHz) for which this assumption is not valid. Typically, the match between theoretical and observed spectral patterns is poorer in the case of obstruents, compared with vowels.

An appreciation of the complex vocal tract behavior in obstruents is gained by considering some aspects of pressures and flows associated with sounds, such as stops (/p/, /t/, /k/, /b/, /d/, /g/) and fricatives (/f/, /θ/, /s/, /ʃ/, /v/, /ɤ/, /z/, /ʒ/). For the nonsense utterances /ɑtɑ/ and /ɑsɑ/, Fig. 10 shows

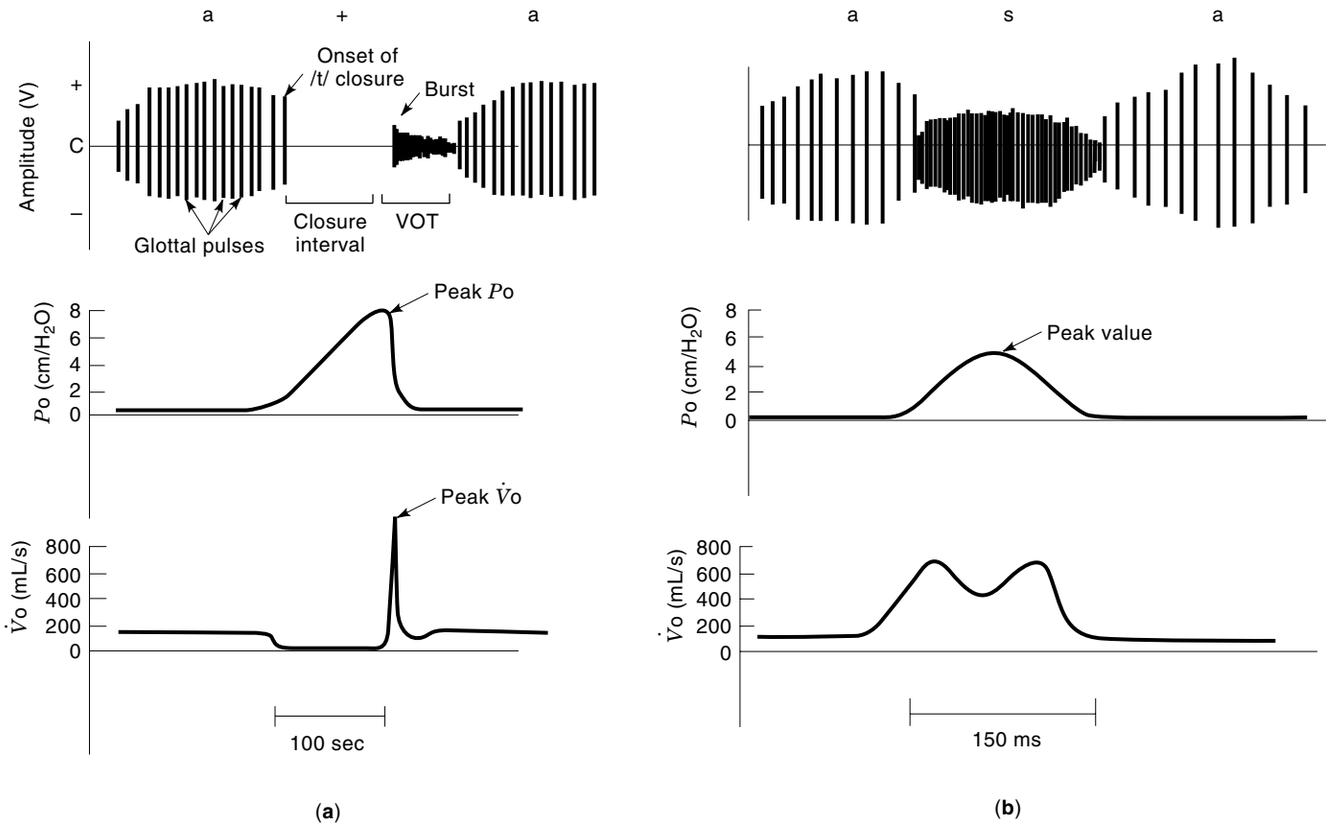

Figure 10. Acoustic waveforms, intraoral air pressure traces, and oral air flow traces for the nonsense utterances /ata/ and /asa/. See text for additional details.

acoustic waveforms and synchronized time histories for intraoral air pressure P_o and oral flow $\dot{V}_o$. The acoustic waveforms represent vocal fold vibration by a series of periodic, complex pulses. These vibrations have been filtered out of the pressure and flow waveforms. P_o is typically measured with a tube inserted through the mouth and around the teeth to protrude roughly at the point where the oral and pharyngeal cavities meet or via the nasal passages and across the nasal surface of the soft palate to hang in the upper part of the pharynx. P_o is defined as the pressure measured behind the point of constriction in the vocal tract but above the level of the glottis. In the case of voiceless stops, for which the vocal folds are open during the oral construction, P_o is essentially the same as the pressure in the lungs, because the volume of air between the lungs and the constriction is continuous and compressed uniformly, primarily by respiratory forces. For voiced stops, the vocal folds are vibrating during the constriction interval, and P_o is therefore somewhat less than lung pressure because of the significant pressure drop across the larynx. $\dot{V}_o$ is measured with a mask placed over the mouth and reflects the volume exchange through the mouth per unit time.

The closure interval for /ɑtɑ/ and the fricative interval for /ɑsɑ/ are important in understanding how the aerodynamic events shown in Fig. 10 are transformed into acoustic sources. During the closure interval for the stop (Fig. 10, left), pressure rises behind the complete constriction as shown in the P_o trace. The peak pressure is typically about 6 cm/H_2O to 10 cm/H_2O (1 cm/H_2O to 1.5 cm/H_2O less in the case of voiced stops) and is reached after 60 ms to 100 ms, at which time it is released suddenly when the constriction is broken by muscular mechanisms (e.g., of the lips for stops /p/ and /b/ or the tongue for stops /t/, /d/, /k/, /g/). The $\dot{V}_o$, which was relatively low (~130 mL/s) during the vowel preceding the closure interval and caused entirely by the valving action of the vibrating vocal folds, drops to zero during the closure interval and remains null until the pressure is released. The large $\dot{V}_o$ peak at the release of P_o reflects the sudden release of the stop constriction. When the vocal tract is fully constricted for the closure interval, there is obviously no sound source and thus no vocal tract output, as suggested by the interval of no energy on the acoustic trace. The sudden release of the P_o provides a transient acoustic source, called *shock excitation,* that excites the vocal tract cavities with a broadband, aperiodic spectrum. In the acoustic record, this event usually lasts no more than about 10 ms to 15 ms and is called the *burst* (see acoustic record, Fig. 10 left). Stop bursts have been the object of a large research effort, largely because their spectral content reflects place-specific (labial (/p/, /b/) versus lingua-alveolar (/t/, /d/) versus dorsal (/k/, /g/) shaping of the shock excitation source, which has often been assumed an important cue for stop consonant identity in theories of speech perception (55,59–61).

The fricative interval in the acoustic trace of /ɑsɑ/ (Fig. 10, right) shows an aperiodic waveform flanked by the two periodic vocalic events. Fricatives are produced by forming a nar-

row, but not complete, constriction between two vocal tract structures. In the case of /s/, the tongue blade is pushed against the maxillary arch and grooved slightly in the middle, creating a narrow passageway between the tongue surface and the hard palate. This constriction sits between two cavities, the back one relatively large and the front one relatively small. As the narrow constriction is formed, the P_o rises and forces high air flow through the grooved passageway. When the flow exits the constriction into the front cavity, it emerges as a narrowly focused jet of gradually expanding laminar streams. The boundaries of this jet are defined by a qualitatively different pattern of flow in the form of rotating air molecules. Thus the jet is surrounded by rotational flow, also called turbulent flow. For some fricatives (like /s/) there is another region of turbulence, induced when the teeth act as an obstacle in the pathway of the laminar jet.

The acoustic result of turbulent flow is aperiodic energy with spectral characteristics that depend largely on the nature of the constriction from which the flow emerges and the obstacles encountered along the flow pathway. This energy serves as the source for fricatives, and it is shaped by the cavities of the vocal tract. An excellent treatment of source characteristics in fricative production is found in Shadle (62).

Fricative acoustics are important because they may provide insight to a class of sounds learned by children at an age later than stops, and also because they are gaining increased attention in theories of speech production. Now we turn to a specific case of how fricative acoustics are used to explore certain aspects of speech production.

THEORIES OF SPEECH PRODUCTION: AN ABSTRACT

Theories of speech production have often been structured around a sequence of processes leading to the movements of the respiratory apparatus, larynx, and upper articulators and the sound resulting from these movements. One level in this process, very close to the *output* movements and sound, is called the *phonemic* level. At this level the class of sound to be produced (such as a /p/, /s/, or /u/) is represented, but *not* the movements (or, in a more composite view, the vocal tract shape) required for its production. For production of words such as *seat* and *suit,* this phonemic representation is sequential (i.e., /sit/ and /sut/) only in the sense of ordering the sound categories to be produced. In other words, the representation shows which sound in these words comes first, next, and last, but provides no information on how the transition is made from one sound category to another. A primary problem in many theories of speech production has been to discover how this discrete representation of successive sound categories is transformed into the continuous movements of the speech mechanism.

As suggested earlier, speech scientists who have studied the movements of the articulators find a lack of clear correspondence between movement boundaries and these phonemic classes. In a word like *seat,* for example, it is very difficult to identify a set of movements belonging exclusively to the /s/, or the /i/, or the /t/. The articulators are in constant motion throughout for a word like *seat,* and contributing further to this confusion is the typical observation that the vocal tract movements at the beginning of the word, in the vicinity of the /s/, are different if the /i/ is replaced by an /u/, making the word *suit.* In this particular example, one difference between the two /s/'s is rather straightforward. The /s/ in *seat* is produced with the lips spread apart, whereas the /s/ in *suit* is produced with the lips rounded. However, there is another, perhaps more subtle difference, which is a vowel-specific adjustment of the tongue configuration during the /s/ constriction. The tongue position for an /s/ constriction is modified in accordance with the tongue position requirements for the vowel. In *seat* the constriction is slightly more forward in the vocal tract, because of the forward articulation of the /i/, whereas in *suit* the constriction is moved posteriorly, more like the tongue requirements for the /u/.

All of these context-conditioned changes in the /s/ articulation are seen in the acoustic output of the vocal tract. The context-conditioned varieties of a given sound class are the allophones discussed earlier in the chapter. As noted earlier, the term *coarticulation* describes the modification of the articulatory characteristics of one sound by the articulatory characteristics of another sound. Coarticulation is the central focus of many speech production theories.

How have speech production theorists attempted to explain coarticulation? One broad class of theories, sometimes called *translation theories,* holds that the phonemes represented at one level of the process are *translated* into a set of motor commands that cause the articulators to move into the positions required for a particular sound production, but that certain special processes adjust the motor commands depending on the context in which a sound is produced. The most noteworthy version of this explanation is sometimes called a *look-ahead operator,* wherein the ordered phonemes are scanned from left to right to determine which articulatory characteristics of later phonemes can be incorporated into earlier phonemes [see review of this notion in (26)]. For example, in the words *seat* and *suit* the formulation of the motor commands involve the specific characteristics required for the /s/ and also involve a forward scan that identifies the lip configuration feature for the following vowel (spread lips for /i/, round lips for /u/) and essentially superimpose that lip setting on the /s/. In early versions of this kind of translation theory, features could migrate from one segment to another, as in the example given here, provided there was no conflict between the features pulled from upcoming segments and those produced for the current segment. In our *seat–suit* example, an /s/ can be produced satisfactorily with any configuration of the lips, so a spread or rounded lip configuration can migrate successfully from a later occurring phoneme (in this case, the vowel /i/ or /u/) to the current phoneme (/s/), produced primarily with the tongue and requiring no special labial configuration. In a word such as *boot,* where the consonant and vowel both require labial articulations, the migration of the vowel lip features may corrupt the consonant production and would therefore be inhibited by the forward scanning process.

An alternate explanation of the coarticulation phenomenon discards the idea of translation between phonemic units and motor commands, and holds that the units to be produced are represented by the articulatory gestures themselves, not by abstract phonological entities, such as phonemes (17,63). In this view a particular sound category is represented as a set of gestures for the tongue, jaw, soft palate, and so forth, where the intergesture timing, the precise way in which the gestures are phased relative to one another, determines the exact nature of the sound category. Moreover, the succession of two or more sound categories is realized by the phasing of

the gesture sets for the successive sounds. For example, in the word *suit* the phasing of the rounding gesture for /u/ is roughly cotemporal with the phasing of the tongue gesture for /s/, but in the word *boot* the rounding gesture is likely to be delayed relative to the lip closing gesture for /b/. Although this is similar to the idea of feature migration described previously, there are some important differences between the two views of coarticulation. Phonemes as representational units have no temporal value, which therefore requires that a great deal of ad hoc machinery be interposed between the representation and the observed behavior of the moving articulators. After all, speech production behavior is inherently dynamic behavior, and a useful theory should be able to deal with the temporal interplay between the moving structures more directly. Gestures as representational units solve this problem, because the units are defined across time, with specified onsets and offsets. This view is also appealing because when gestures for successive sound categories compete for the movement of the same articulator (as in the case of *boot,* for which the lip gestures for the /b/ and /u/ *compete* for some time interval) the theory may explain fine details of movement timing. This is simply not possible when phonemes are the representational units. The primary shortcoming of the gesture formulation is that the phasing of gestures within a sound category and across successive sound categories is determined in a completely ad hoc manner.

Finally, it should be noted that one very useful outcome of the gesture approach is the view of the articulators as a collective, rather than a set of independent agents requiring independent control. Some early research efforts were concerned with the actions during speech of individual articulators, such as the tongue, lips, or jaw, and speech production models were sometimes conceived with independent control parameters for individual articulators or even individual muscles comprising an articulator. The gesture perspective has caused a reconceptualization of the control parameters for articulatory behavior, in particular pursuing the idea of vocal tract *goals,* not muscles or articulators, as the variables requiring control. For example, in the case of the fricative /s/, the controlled variables are the place and degree of constriction, which are realized by many different combinations of muscular behavior and articulatory positions. The term *coordinative structure* describes the collective of articulatory behaviors organized to achieve goals, such as constriction location and degree (64). Coordinative structures are not read-only devices ready for output at appropriate times, but rather more like adaptable and evolving software, organized and reorganized on the fly to meet the changing goals of articulatory behavior as speakers generate utterances. One of the great challenges of continuing work in speech production theory is to gain a deeper understanding of coordinative structures and to understand their role in speech development and dysfunction (65–67).

BIBLIOGRAPHY

1. S. Pinker, *The Language Instinct,* New York: William Morrow, 1994.
2. D. J. Umiker-Sebeok and T. A. Sebeok, (eds.), *Aboriginal Sign Languages of the Americas and Australia,* New York: Plenum, 1978.
3. R. B. Ochsman and A. Chapanis, The effects of 10 communication modes on the behaviour of teams during cooperative problem solving, *Int. J. Man-Mach. Studies,* **6**: 579–620, 1974.
4. P. R. Cohen, The pragmatics of referring and the modality of communication, *Computat. Linguistics,* **10**: 97–146, 1984.
5. S. L. Oviatt and P. R. Cohen, The contributing influence of speech and interaction on human discourse patterns, in J. W. Sullivan and S. W. Tyler (eds.), *Intelligent User Interfaces,* Reading, MA: Addison-Wesley, 1991.
6. G. D. Allen, The PHONASCII system, *J. Int. Phonetic Assoc.,* **21**: 11–17, 1988.
7. O. Fujimura, Methods and goals of speech production research, *Language Speech,* **33**: 195–258, 1990.
8. S. M. Marcus and A. K. Syrdal, Speech: Articulatory, linguistic, acoustic, and perceptual descriptions, in A. Syrdal, R. Bennett, and S. Greenspan (eds.), *Applied Speech Technology,* Boca Raton, FL: CRC Press, 1995.
9. G. Dewey, *Relative Frequency of English Speech Sounds,* Cambridge, MA: Harvard Univ. Press, 1923.
10. D. Kahn, *Syllable-based generalizations in English phonology,* Ph.D. dissertation, Massachusetts Institute of Technology. New York: Garland, 1980.
11. O. Fujimura and J. Lovins, Syllables as concatenative phonetic units, in A. Bell & J. B. Hopper (eds.), *Syllables and Segments,* Amsterdam: North Holland, 1978.
12. J. A. Goldsmith, *Autosegmental and Metrical Phonology,* Oxford, UK: Blackwell, 1990.
13. M. Y. Liberman and A. Prince, On stress and linguistic rhythm, *Linguistic Inquiry,* **8**: 249–336, 1977.
14. E. O. Selkirk, The role of prosodic categories in English word stress, *Linguistic Inquiry,* **11**: 563–605, 1980.
15. E. Williams, Underlying tone in Margi and Igbo, *Linguistic Inquiry,* **7**: 463–484, 1976.
16. A. Cutler and D. M. Carter, The predominance of strong initial syllables in the English vocabulary, *Comput. Speech Language,* **2**: 133–142, 1987.
17. C. P. Browman and L. M. Goldstein, Some notes on syllable structure in articulatory phonology, *Phonetica,* **45**: 140–155, 1988.
18. T. J. Hixon and collaborators, *Respiratory Function in Speech and Song,* Boston: College Hill Press, 1987.
19. G. Weismer, Speech breathing, in R. G. Daniloff (ed.), *Speech Science,* San Diego: College Hill Press, 1985.
20. G. S. Berke and B. R. Gerratt, Laryngeal mechanics: An overview of mucosal wave mechanics, *J. Voice,* **7**: 123–128, 1993.
21. R. C. Scherer, Laryngeal function during phonation, in J. S. Rubin, G. Korovin, R. T. Sataloff, and W. J. Gould (eds.), *Diagnosis and Treatment of Voice Disorders,* New York: Igaku-Shoin Medical, 1995.
22. I. Titze, *Principles of Voice Production,* Englewood Cliffs, NJ: Prentice-Hall, 1994.
23. I. R. Murray and J. L. Arnott, Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *J. Acoust. Soc. Amer.,* **93**: 1097–1108, 1993.
24. R. D. Kent, *The Speech Sciences,* San Diego: Singular, 1997.
25. C. A. Fowler and E. Saltzman, Coordination and coarticulation in speech production, *Language Speech,* **36**: 171–195, 1993.
26. R. D. Kent and F. D. Minifie, Coarticulation in recent speech production models, *J. Phonetics,* **5**: 115–133, 1977.
27. A. Smith, The control of orofacial movements in speech, *Critical Rev. Oral Biol. Med.,* **3**: 233–267, 1992.
28. R. D. Kent and M. R. Chial, Talker identification, in D. L. Faigman, D. Kaye, M. J. Saks, and J. Sanders (eds.), *Modern Scientific Evidence: The Law and Science of Expert Testimony,* St. Paul, MN: West, 1997.

29. G. Fant, *Acoustic Theory of Speech Production,* The Hague: Mouton, 1960.
30. R. D. Kent and H. K. Vorperian, *Development of the craniofacial-oral-laryngeal anatomy,* San Diego, CA: Singular, 1995.
31. D. Broad, The new theories of vocal fold vibration, in N. J. Lass (ed.), *Speech and Language: Advances in Basic Research and Practice,* New York: Academic Press, 1979.
32. M. Hirano and Y. Kakita, Cover-body theory of vocal fold vibration, in R. G. Daniloff (ed.), *Speech Science,* San Diego: College Hill Press, 1985.
33. J. Kahane, Anatomy and physiology of the organs of the peripheral speech mechanism, in N. J. Lass, L. V. McReynolds, J. L. Northern, and D. E. Yoder (eds.), *Handbook of Speech-Language Pathology and Audiology,* Toronto: B. C. Decker, 1988.
34. H. Hollien, On vocal registers, *J. Phonetics,* **2**: 125–143, 1974.
35. M. Rothenberg, A new inverse-filtering technique for deriving the glottal air flow waveform during voicing, *J. Acoust. Soc. Amer.,* **53**: 1632–1645, 1973.
36. G. Fant and Q. Lin, Frequency domain interpretation and derivation of glottal flow parameters, *Speech Transmission Laboratory Q. Progress Status Rep.,* **2–3**: 1–21, 1988.
37. J. L. Flanagan, *Speech Analysis Synthesis and Perception,* Berlin: Springer-Verlag, 1972.
38. G. Fant, The voice source in connected speech, *Speech Commun.,* **22**: 125–139, 1997.
39. J. S. Perkell, *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study,* Cambridge, MA: MIT Press, 1969.
40. G. E. Peterson and H. L. Barney, Control methods used in a study of the vowels, *J. Acoust. Soc. Amer.,* **24**: 175–184, 1952.
41. K. Johnson and J. W. Mullennix, *Talker Variability in Speech Processing,* San Diego: Academic Press, 1997.
42. M. Fourakis, Tempo, stress, and vowel reduction in American English, *J. Acoust. Soc. Amer.,* **90**: 1816–1827, 1991.
43. A. S. House and K. N. Stevens, Perturbations of vowel articulations by consonantal context: An acoustical study, *J. Speech Hearing Res.,* **6**: 111–128, 1963.
44. S.-J. Moon and B. Lindblom, Interaction between duration, context and speaking style in English stressed vowels, *J. Acoust. Soc. Amer.,* **96**: 40–55, 1994.
45. M. A. Picheny, N. I. Durlach, and L. D. Braida, Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech, *J. Speech Hearing Res.,* **29**: 434–446, 1986.
46. J. Hillenbrand et al., Acoustic characteristics of American English vowels, *J. Acoust. Soc. Amer.,* **97**: 3099–3111, 1995.
47. W. Strange, Evolving theories of vowel perception, *J. Acoust. Soc. Amer.,* **85**: 2081–2087, 1989.
48. J. D. Miller, Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Amer.,* **85**: 2114–2134, 1989.
49. T. M. Nearey, Static, dynamic, and relational properties in vowel perception, *J. Acoust. Soc. Amer.,* **85**: 2088–2113, 1989.
50. J. Hillenbrand and R. Gayvert, Vowel classification based on fundamental frequencies and formant frequencies, *J. Speech Hearing Res.,* **94**: 668–674, 1993.
51. K. N. Stevens, The quantal nature of speech: Evidence from articulatory-acoustic data, in P. B. Denes and E. E. David (eds.), *Human Communication: A Unified View,* New York: McGraw–Hill, 1972.
52. K. N. Stevens, On the quantal nature of speech, *J. Phonetics,* **17**: 3–46, 1989.
53. K. N. Stevens, Articulatory-acoustic-auditory relations, in W. J. Hardcastle and J. Laver (eds.), *The Handbook of Phonetic Sciences,* Oxford, UK: Blackwell, 1997.
54. M. Mrayati, R. Carre, and B. Guerin, Distinctive regions and modes: A new theory of speech production, *Speech Commun.,* **7**: 257–286, 1988.
55. J. S. Perkell and W. L. Nelson, Variability in production of the vowels /i/ and /a/, *J. Acoust. Soc. Amer.,* **77**: 1889–1895, 1985.
56. J. Dang, K. Honda, and H. Suzuki, Morphological and acoustical analysis of the nasal and paranasal cavities, *J. Acoust. Soc. Amer.,* **96**: 2088–2100, 1994.
57. J. Dang and K. Honda, Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation, *J. Acoust. Soc. Amer.,* **100**: 3374–3383, 1996.
58. S. Hawkins and K. N. Stevens, Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels, *J. Acoust. Soc. Amer.,* **77**: 1560–1575, 1985.
59. S. E. Blumstein and K. N. Stevens, Acoustic invariance in speech production: Evidence from the measurement of the spectral characteristics of stop consonants, *J. Acoust. Soc. Amer.,* **66**: 1001–1017, 1979.
60. K. Forrest et al., Statistical analysis of word-initial voiceless obstruents: Preliminary data, *J. Acoust. Soc. Amer.,* **84**: 115–123, 1988.
61. D. Kewley-Port, Time-varying features as correlates of place of articulation in stop consonants, *J. Acoust. Soc. Amer.,* **73**: 322–335, 1983.
62. C. H. Shadle, The acoustics of fricative consonants, Unpublished Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1985.
63. C. P. Browman and L. M. Goldstein, Towards an articulatory phonology, in C. Ewan and J. Anderson (eds.), *Phonology Yearbook 3,* Cambridge, UK: Cambridge Univ. Press, 1986.
64. J. A. S. Kelso, E. Saltzman, and N. Tuller, The dynamical perspective on speech production: Data and theory, *J. Phonetics,* **14**: 29–59, 1986.
65. F. H. Guenther, Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production, *Psychological Rev.,* **102**: 594–621, 1995.
66. G. Weismer, K. Tjaden, and R. D. Kent, Articulatory characteristics in motor speech disorders: Relationships to models and theories of normal speech production, *J. Phonetics,* **23**: 149–164, 1995.
67. A. Lofqvist, Theories and models of speech production, in W. J. Hardcastle and J. Laver (eds.), *The Handbook of Phonetic Sciences,* Oxford, UK: Blackwell, 1997.

RAY D. KENT
GARY WEISMER
University of Wisconsin-Madison

SPEECH QUALITY IMPROVEMENT. See SPEECH ENHANCEMENT.