Speech Analysis Synthesis and Perception

Third Edition

James L. Flanagan
Jont B. Allen
Mark A. Hasegawa-Johnson

2008

Contents

List of Figures

List of Tables

Chapter 2

The Mechanism of Speech Production

2.1 Physiology of the Vocal Apparatus

Speech is the acoustic end product of voluntary, formalized motions of the respiratory and masticatory apparatus. It is a motor behavior which must be learned. It is developed, controlled and maintained by the acoustic feedback of the hearing mechanism and by the kinesthetic feedback of the speech musculature. Information from these senses is organized and coordinated by the central nervous system and used to direct the speech function. Impairment of either control mechanism usually degrades the performance of the vocal apparatus[1].

The speech apparatus also subserves the more fundamental processes of breathing and eating. It has been conjectured that speech evolved when ancient peoples discovered that they could supplement their communicative hand signals with related "gestures" of the vocal tract. Sir Richard Paget sums up this speculation quite neatly. "What drove man to the invention of speech was, as I imagine, not so much the need of expressing his thoughts (for that might have been done quite satisfactorily by bodily gesture) as the difficulty of 'talking with his hands full.' It was the continual use of man's hands for craftsmanship, the chase, and the beginnings of art and agriculture, that drove him to find other methods of expressing his ideas–namely, by a specialized pantomime of the tongue and lips (?, ?)."

The machinery involved in speech production is shown schematically in Fig. 2.1. The diagram represents a mid-sagittal section through the vocal tract of an adult. The primary function of inhalation is accomplished by expanding the rib cage, reducing the air pressure in the lungs, and drawing air into the lungs via nostrils, nasal cavity, velum port and trachea (windpipe). Air is normally expelled by the same route. In eating, mastication takes place in the oral cavity. When food is swallowed the structures at the entrance to the trachea are drawn up under the epiglottis. The latter shields the opening at the vocal cords and prevents food from going into the windpipe. The esophagus, which normally lies collapsed against the back wall of the throat, is at the same time drawn open to provide a passage to the stomach.

The vocal tract proper is an acoustical tube which is nonuniform in cross-sectional area. It is terminated by the lips at one end and by the vocal cord constriction at the top of the trachea at the other end. In an adult male the vocal tube is about 17cm long and is deformed in crosssectional area by movement of the articulators; namely, the lips, jaw, longue and velum. The cross-sectional

[1] Most of us are aware of the difficulties that partially or totally deaf persons have in producing adequate speech. Even more familiar, perhaps, are the temporary difficulties in articulation experienced after the dentist desensitizes a large mouth area by an injection of anesthetic.

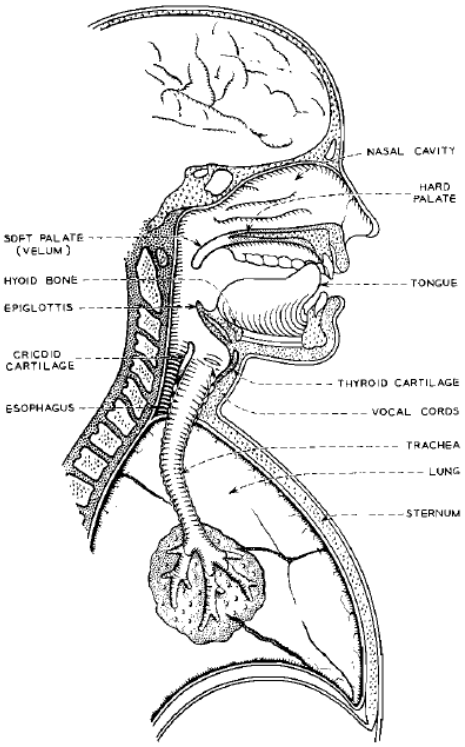

Figure 2.1: Schematic diagram of the human vocal mechanism

area in the forward portion of the tract can be varied from zero (i.e. complete closure) to upwards of 20 cc.

The nasal tract constitutes an ancillary path for sound transmission. It begins at the velum and terminates at the nostrils. In the adult male the cavity has a length of about 12 cm and a volume on the order of 60 cc, It is partitioned over part of its front-to-back extent by the nasal septum. Acoustic coupling between the nasal and vocal tracts is controlled by the size of the opening at the velum. In Fig. 2.1 the velum is shown widely open. In such a case, sound may be radiated from both the mouth and nostrils. In general, nasal coupling can substantially influence the character of sound radiated from the mouth. For the production of non-nasal sounds the velum is drawn tightly up and effectively seals off the entrance to the nasal cavity. In an adult male the area of the velar opening can range from zero to around 5 cc.

The source of energy for speech production lies in the thoracic and abdominal musculatures. Air is drawn into the lungs by enlarging the chest cavity and lowering the diaphragm. It is expelled by contracting the rib cage and increasing the lung pressure. Production of vowel sounds at the softest possible level requires a lung pressure of the order of 4 cm H_20. For very loud, high-pitched sounds, on the other hand, pressures of about 20 cm H_20 or more are not uncommon. During speaking the lung pressure is maintained by a steady, slow contraction of the rib cage.

As air is forced from the lungs it passes through the trachea into the pharynx, or throat cavity. The top of the trachea is surmounted by a structure which is shown in additional detail in Fig. 2.2. This is the larynx. The cartilaginous frame houses two lips of ligament and muscle. These are the vocal cords and are denoted VC. The slit-like orifice between the cords is called the glottis. The knobby structures, protruding upward posterior to the cords, are the arytenoid cartilages, and are labelled AC. These cartilages support the fleshy cords and facilitate adjustment of tension. The principal outside cartilages of the larynx "box" are the anterior thyroid (labelled TC in Fig. 2.2)

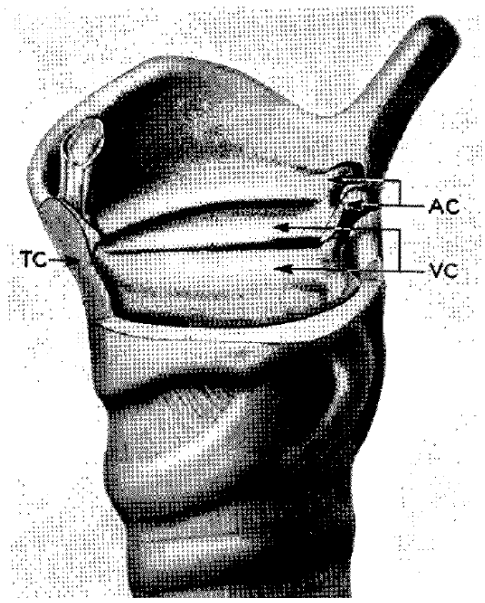

Figure 2.2: Cut-away view of the human larynx. (After Farnsworth.) VC-vocal cords; AC-arytenoid cartilages; TC-thyroid cartilage

and the posterior cricoid. Both of these can be identified in Fig. 2.1.

The voiced sounds of speech are produced by vibratory action of the vocal cords. Production of sound in this manner is called phonation. Qualitatively, the action of the vocal folds is very similar to the flapping of a flag, or the vibration of the reed in a woodwind instrument. Like a flag flapping in the wind, the vocal folds must have at least two regions that are out of phase with one another. Like the jet of air passing over the surface of a flag, the jet of air passing through the glottis has two regimes: a laminar regime, and a turbulent regime. In the laminar regime, Bernoulli's equation holds, so air pressure is inversely proportional to the square of air jet velocity. In the turbulent regime, differences in velocity are absorbed by the creation of vortices, so that air pressure remains low and constant throughout the turbulent regime. The glottis flaps from bottom to top: the lower vocal folds separate first, followed by the upper folds. While the lower folds are wider than the upper folds, air flow within the glottis is laminar, and therefore the pressure within the glottis is high, driving the folds open. When the upper folds flap open to a position wider than the lower folds, air within the glottis becomes turbulent, and therefore the pressure within the glottis drops to a low constant value. At this point, the stiffness of the vocal folds forces them back together again, and the cycle repeats. Notice that it is not necessary for the vocal folds to completely close at any point in the cycle. The "breathy voice" employed to great effect by some singers and actresses is apparently a form of phonation in which the glottis never completely closes. The mass and compliance of the cords, and the subglottal pressure, essentially determine the period of the oscillation. This period is generally shorter than the natural period of the cords; that is, the cords are driven in a forced oscillation.

The variable area orifice produced by the vibrating cords permits quasi-periodic pulses of air to excite the acoustic system above the vocal cords. The mechanism is somewhat similar to blowing a tone on a brass instrument, where the vibrating lips permit quasiperiodic pulses of air to excite the resonances of the flared horn. Over the past years the vibratory action of the vocal cords has been studied in considerable detail. Direct observations can be made by positioning a 45-degree mirror toward the back of the mouth, near the naso-pharynx. Stroboscopic illumination at the proper frequency slows or "stops" the vibratory pattern and permits detailed scrutiny.

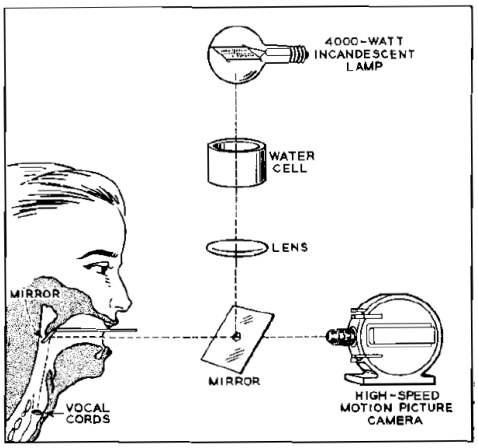

Figure 2.3: Technique for high-speed motion picture photography of the vocal cords. (After Farnsworth)

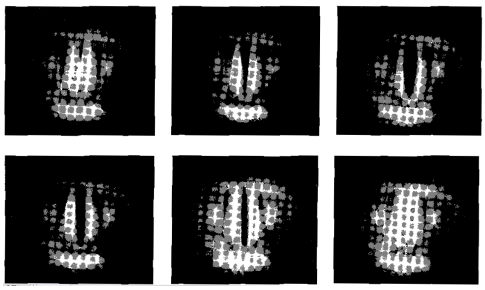

Figure 2.4: Successive phases in one cycle of vocal cord vibration. The total elapsed time is approximately 8 msec

Still more revealing and more informative is the technique of highspeed photography, pioneered by Farnsworth(?, ?), in which moving pictures are taken at a rate of 4000 frames/sec, or higher. The technique is illustrated in Fig. 2.3. The cords are illuminated by an intense light source via the arrangement of lenses and mirrors shown in the diagram. Photographs are taken through an aperture in the large front mirror to avoid obstructing the illumination. The result of such photography is illustrated in Fig. 2.4. The figure shows six selected frames in one cycle of vibration of the cords of an adult male. In this case the fundamental frequency of vibration, or voice "pitch," is 125Hz.

The volume flow of air through the glottis as a function of time is similar to (though not exactly proportional to) the area of the glottal opening. For a normal voice effort and pitch, the waveform can be roughly triangular in shape and exhibit duty factors (i.e., ratios of open time to total period) commonly of the order of 0.3 to 0.7. The glottal volume current therefore has a frequency spectrum relatively rich in overtones or harmonics. Because of the approximately triangular waveform, the higher frequency components diminish in amplitude at about 12db/octave.

The waveform of the glottal volume flow for a given individual can vary widely. In particular, it depends upon sound pitch and intensity. For low-intensity, low-pitched sounds, the subglottal pressure is low, the vocal cord duty factor high, and the amplitude of volume flow low. For high-intensity, high-pitched sounds, the subglottal pressure is large, the duty factor small and the amplitude of volume flow great. The amplitude of lateral displacement of the vocal cords, and hence the maximum glottal area, is correlated with voice intensity to a surprisingly small extent (Fletcher(?, ?)). For an adult male, common peak values of glottal area are of the order of 15 mm^2.

Because of its relatively small opening, the acoustic impedance of the glottal source is generally large compared to the acoustic impedance looking into the vocal tract, at least when the tractis not tightly constricted. Under these conditions changes in tract configuration have relatively small (but not negligible) influence upon the glottal volume flow. For tight constriction of the tract, the acoustic interaction between the tract and the vocal-cord oscillator can be pronounced.

Another source of vocal excitation is produced by a turbulent flow of air created at some point of stricture in the tract. An acoustic noise is thereby generated and provides an incoherent excitation for the vocal system. The unvoiced continuant sounds are formed from this source. Indirect measurements and theory suggest that the spectrum of the noise, at its point or region of generation, is relatively broad and uniform. The vocal cavities forward of the constriction usually are the most influential in spectrally shaping the sound.

A third source of excitation is created by a pressure buildup at some point of closure. An abrupt release of the pressure provides a transient excitation of the vocal tract. To a crude approximation the aperiodic excitation is a step function of pressure, and might be considered to huvc a spectrum which falls inversely with frequency. The closure can he effected at various positions toward the front of the tract; for example, at labial, dental, and palatal positions. The transient excitation can be used with or without vocal cord vibration to produce voiced or unvoiced plosive sounds.

Whispered speech is produced by substituting a noise source for the normally vibrating vocal cords. The source may by produced by turbulent flow at the partially closed glottis, or at some other constricted place in the tract.

2.2 The Sounds of Speech

To be a practicable medium for the transmission of information, a language must consist of a finite number of distinguishable, mutuallyexclusive sounds. That is, the language must be constructed of basic linguistic units which have the property that if one replaces another in an utterance, the meaning is changed. The acoustic manifestations of a basic unit may vary widely. All such variations, however–when heard by a listener skilled in the language–signify the same linguistic element. This basic linguistic element is called a phoneme (?, ?). Its manifold acoustic variations are called allophones.

The phonemes might therefore be looked upon as a code uniquely related to the articulatory gestures of a given language. The allophones of a given phoneme might be considered representative of the acoustic freedom permissible in specifying a code symbol. This freedom is not only dependent upon the phoneme, but also upon its position in an utterance.

The set of code symbols used in speech, and their statistical properties, depend upon the language and dialect of the communicators. When a linguist initially studies an unknown language, his first step is to make a phonetic transcription in which every perceptually-distinct sound is given a symbol. He then attempts to relate this transcription to behavior, and to determine which acoustically-distinguishable sounds belong to the same phoneme. That is, he groups together those sounds which are not distinct from each other in meaning. The sounds of each group differ in pronunciation, but this difference is not important to meaning. Their difference is merely a convention of the spoken language.

Features of speech which may be phonemically distinct in one language may not be phonemic in another. For example, in many East Asian and Western African languages, changing the pitch of a vowel changes the meaning of the word. In European and Middle Eastern languages, this generally is not the case. Other striking examples are the Bantu languages of southern Africa, such as Zulu, in which tongue clicks and lip smacks are phonemes.

The preceding implications are that speech is, in some sense, discrete. Yet an oscillographic representation of the sound pressure wave emanating from a speaker producing connected speech shows surprisingly few gaps or pause intervals. Connected speech is coupled with a near continuous motion of the vocal apparatus from sound to sound. This motion involves changes in the configuration of the vocal tract as well as in its modes of excitation. In continuous articulation the vocal tract dwells only momentarily in a state appropriate to a given phoneme.

The statistical constraints of the language greatly influence the precision with which a phoneme needs to be articulated. In some cases it is merely sufficient to make a vocal gesture in the direction of the normal configuration to signal the phoneme. Too, the relations between speech sounds and vocal motions are far from unique, although normal speakers operate with gross similarity. Notable examples of the "many–valuedness" of speech production are the compensatory articulation of ventriloquists and the mimicry of parrots and myna birds.

Despite the mutability of the vocal apparatus in connected speech, and the continuous nature of the speech wave, humans can subjectively segment speech into phonemes. Phoneticians are able to make written transcriptions of connected speech events, and phonetic alphabets have been devised for the purpose. It has been argued (?, ?) that the concept of a phonetic alphabet was invented only once in human history, by the Phoenicians of Lebanon in the early first millenium B.C., but the uniqueness of this invention is obscured by the rapidity with which it was adopted worldwide. By 300 B.C., the Indus river scholar Panini had organized the phonemes of his language into a rank-three array, with dimensions specifying the manner of articulation (vowel, glide, nasal, fricative, stop), place of articulation (lips, teeth, alveolar ridge, hard palate, soft palate, uvula, pharynx), and glottal features (voiced vs. unvoiced, aspirated vs. unaspirated). Panini's organization remains the foundation of all modern phonetic alphabets, including the international standard alphabet developed by the International Phonetic Association (IPA). The international phonetic alphabet (also abbreviated IPA: the meaning of the acronym is usually apparent from context) provides symbols for representing the speech sounds of most of the major languages of the world.

Linguists transcribe speech at several different levels of precision. As specified previously, two phonemes are different only if it is possible to change the meaning of a word by interchanging the two. A transcription in terms of phonemes is called "phonemic," and is conventionally enclosed in virgules // (?, ?). On the other hand, the IPA provides notation for many subtle acoustic distinctions that are never used, in any given language, to change the meaning of a word; a transcription that specifies any of these allophonic or sub-phonemic distinctions is called "phonetic," and is conventionally enclosed in brackets []. In the remainder of this book, most transcriptions will be phonemic, but we will occsionally also make use of phonetic transcription.

Table 2.1: Vowels

Degree of constriction	Tongue hump position					
	front		central		back	
High	TIPA/i/	eve	TIPA/3^r/	bird	TIPA/u/	boot
	TIPA/I/	it	TIPA/$@^r$/	lover (unstressed)	TIPA/U/	foot
Medium	TIPA/e/	hate*	TIPA/2/	up	TIPA/o/	obey*
	TIPA/E/	met	TIPA/@/	ado (unstressed)	TIPA/O/	all
Low	TIPA/æ/	at			TIPA/A/	father

*These two sounds usually exist as diphthongs in GA dialect. They are included in the vowel table because they form the nuclei of related diphthongs. See Section 2.27 for further discussion. (See also (?, ?).)

Classification of speech sounds is customarily accomplished according to their manner and place of production. Phoneticans have found this method convenient to indicate the gross characteristics of sounds. For example, the articulation of vowel sounds is generally described by the position of the tongue hump along the vocal tract (which is often, but not always, the place of greatest constriction) and the degree of the constriction. This classification method will be employed in the following discussion of speech sounds. The examples extend to the sounds of English speech of General American (GA) dialect.

2.2.1 Vowels

Vowels are speech sounds with no narrow constriction in the vocal tract. They are usually voiced (produced with vocal fold excitation), though they may of course be whispered. In normal articulation, the tract is maintained in a relatively stable configuration during most of the sound. The vowels are further characterized by negligible (if any) nasal coupling, and by radiation only from the mouth (excepting that which passes through the cavity walls). If the nasal tract is coupled to the vocal tract during the production of a vowel, the vowel becomes nasalized. The distinction between nasalized and non-nasalized versions of any particular vowel is phonemic in some languages (e.g., French), but not in English; thus, for example, some comedians produce an entertaining effect by nasalizing all of their vowels.

When the 12 vowels of GA speech are classified according to the tongue-hump-position degree-of-constriction scheme, they may be arranged as shown in Table 2.1. Along with each vowel is shown a key word containing the vowel.

The approximate articulatory configurations for the production of these sounds (exclusive of the two unstressed vowels) are shown qualitatively by the vocal tract profiles in Fig. 2.5 (?, ?). The physiological basis for the front-back/high-low classification is particularly well illustrated if the profiles for the vowels TIPA/i,æ,a,u/ are compared[2].

2.2.2 Consonants

Consonants are sounds produced with a constriction at some point in the vocal tract. Consonants may be further divided into four classes, based on two binary manner features: the feature **sonorant, and the feature continuant.**

[2]These profiles, and the ones shown subsequently in this chapter, mainly illustrate the oral cavity. The important pharynx cavity and the lower vocal tract are not drawn. Their shapes may be deduced from x-rays (see Figs. **??** through **??**, for example).

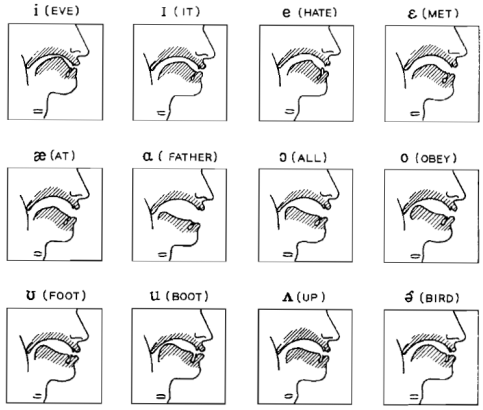

Figure 2.5: Schematic vocal tract profiles for the production of English vowels. (Adapted from Potter, Kopp and Green)

Table 2.2: All consonants may be divided into four broad manner classes, using the two binary features **sonorant** and **continuant**. The opposite of **sonorant** is **obstruent**; the opposite of **continuant** is **discontinuant**.

	Continuant	Discontinuant
Sonorant	Glides (TIPA/w,j/) and Semivowels (TIPA/l,r/)	Nasals (TIPA/m,n,N/)
Obstruent	Fricatives (TIPA/f,v,T,D,s,z,S,Z,h/)	Stops (TIPA/p,b,t,d,k,g/) and Affricates (TIPA/tS,dZ/)

The literal meaning of the word "sonorant" is "song-like." A sonorant consonant is a consonant with no increase of air pressure inside the vocal tract, either because the vocal tract constriction is not very tight (TIPA/w,j,r,l/), or because the soft palate is opened, allowing air to escape through the nose (TIPA/m,n,N/). Because there is no increase in air pressure, the voicing of a sonorant consonant is free and easy, and, for example, it is possible to sing a sonorant consonant.

A discontinuant consonant is produced with a complete closure at some point in the vocal tract. Because of this complete closure, the transition between a discontinuant consonant and its neighboring vowel is always marked by a sudden acoustic discontinuity, when the sound quality changes dramatically in a space of one or two milliseconds. A continuant consonant has no complete vocal tract closure.

Based on these two binary features, it is possible to divide all consonants into four broad manner classes, as shown in Table 2.2.

Fricative Consonants

Fricatives are produced from an incoherent noise excitation of the vocal tract. The noise is generated by turbulent air flow at some point of constriction. In order for the air flow through a constriction to produce turbulence, the Reynold's number $R_{ee} = ud\rho/\mu$ must be larger than 1800, where u is the air particle velocity, ρ and μ are the density and viscosity of air, and d is the smallest cross-sectional width of the constriction (all expressed in consistent units, so that the Reynolds number itself is dimensionless). Since velocity is inversely proportional to the area of the constriction, small constrictions lead to high Reynolds numbers; the threshold Reynolds number for turbulence is usually reached by by constrictions of less than about 3mm width. In order to produce

Table 2.3: Fricative consonants

Place of articulation	Voiced		Voiceless	
Labio-dental	TIPA/v/	vote	TIPA/f/	for
Dental	TIPA/D/	then	TIPA/T/	thin
Alveolar	TIPA/z/	zoo	TIPA/s/	see
Palatal	TIPA/Z/	azure	TIPA/S/	she
Glottal			TIPA/h/	he

a fricative, a talker must position the tongue or lips to create a constriction with a width of 2-3mm, and allow air pressure to build up behind the constriction, so that the air flow through the constriction is turbulent. If the constriction is too wide, it will not produce turbulence; if it is too narrow, it will stop the air flow entirely. Because of the precise articulation required, fricatives are rarely the first phonemes acquired by infants learning to speak.

Common constrictions for producing fricative consonants are those formed by the tongue behind the teeth (dental: TIPA/T,D/), the upper teeth on the lower lip (labio-dental: TIPA/f,v/), the tongue to the gum ridge (alveolar: TIPA/s,z/), the tongue against the hard palate (palatal: TIPA/S,Z/), and the vocal cords constricted and fixed (glottal: TIPA/h/). Radiation of fricatives normally occurs from the mouth. If the vocal cord source operates in conjunction with the noise source, the fricative is a voiced fricative. If only the noise source is used, the fricative is unvoiced.

Both voiced and unvoiced fricatives are continuant sounds. Because a given fricative articulatory configuration can be excited either with or without voicing, the voiced and voiceless fricatives form complementary pairs called cognates. The fricative consonants of the GA dialect are listed in Table 2.3, along with typical "places" of articulation and key words for pronunciation.

Vocal tract profiles for these sounds are shown in Fig. 2.6. Those diagrams in which the vocal cords are indicated by two small lines are the voiced fricatives. The vocal cords are shown dashed for the glottal fricative (TIPA/h/).

The phoneme TIPA/h/ is a special case because, like the sonorant consonants, it requires no increase of air pressure within the vocal tract. For this reason, some phoneticians class TIPA/h/ as a glide rather than a fricative (e.g., (?, ?)). In inter-vocalic context (e.g., in the word "ahead"), the acoustic quality of TIPA/h/ may be very sonorant-like, e.g., the amplitude of voicing may not decrease at all. In other contexts, TIPA/h/ may have weakened voicing (like a typical voiced fricative), or it may be completely unvoiced (like a typical unvoiced fricative). All of these different allophones are produced and perceived interchangeably, by native speakers of English, as examples of the same underlying phoneme TIPA/h/.

Stop Consonants

Among those consonants which depend upon vocal tract dynamics for their creation are the stop consonants. To produce these sounds a complete closure is formed at some point in the vocal tract. The lungs build up pressure behind this occlusion, and the pressure is suddenly released by an abrupt motion of the articulators. Stops are distinguished from other phonemes by complete closure, followed by a characteristic acoustic "explosion" called a "transient." The transient typically lasts only one or two milliseconds, but it may be followed by a fricative burst of 5-10ms in duration, if the

Table 2.4: Stop consonants

Place of articulation	Voiced		Voiceless	
Labial	TIPA/b/	be	TIPA/p/	pay
Alveolar	TIPA/d/	day	TIPA/t/	to
Palatal/velar	TIPA/g/	go	TIPA/k/	key

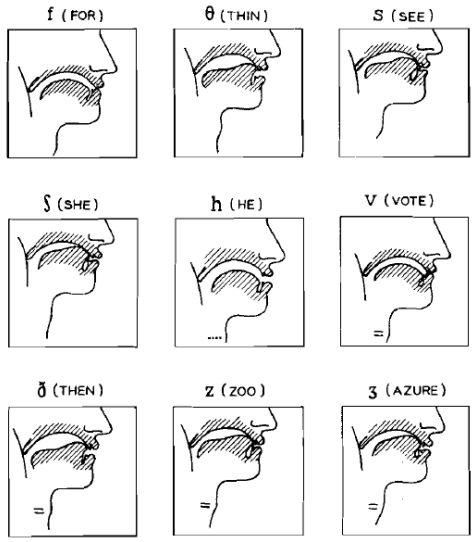

Figure 2.6: Vocal tract profiles for the fricative consonants of English. The short pairs of lines drawn on the throat represent vocal cord operation. (Adapted from Potter, Kopp and Green)

lips or tongue pass too slowly through the 2-3mm frication region.

Stops in English come in voiced/unvoiced cognate pairs, as do fricatives. Cognate pairs are distinguished in two ways. First, the vocal folds may continue to vibrate during the closure interval of a voiced stop. Closure voicing is often heard in carefully produced speech, but rarely in casual speech. Instead, most speakers of GA English signal that a stop is voiced by allowing the vocal folds to begin vibrating immediately after stop release. An unvoiced stop, by contrast, has a period of aspiration following release, during which the vocal folds are held open and turbulence is produced at the glottis. The acoustic effect is exactly what one would achieve by producing an unvoiced stop followed immediately by an TIPA/h/.

The cognate pairs of stops, with typical places of articulation, are shown in Table 2.4. Articulatory profiles for these sounds are shown in Fig. 2.7. Each position is that just prior to the pressure release.

Nasal Consonants

The nasal consonants, or nasals, are sonorant consonants; they are normally voiced in GA English, although unvoiced allophones might be heard in some contexts (e.g., some speakers will devoice the TIPA/n/ in "fishnet"). A complete closure is made toward the front of the vocal tract, either by the lips, by the tongue at the gum ridge, or by the tongue at the hard or soft palate. The velum is opened wide and the nasal tract

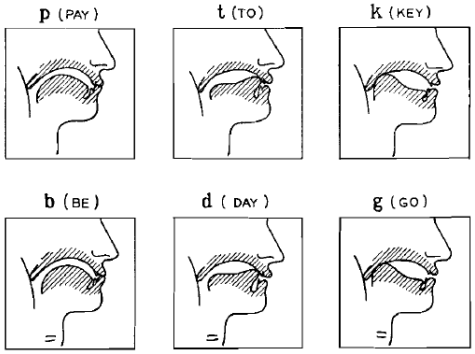

Figure 2.7: Articulatory profiles for the English stop consonants. (After Potter, Kopp and Green)

Table 2.5: Nasals

Place		
Labial	TIPA/m/	me
Alveolar	TIPA/n/	no
Palatal/velar	TIPA/N/	sing (no initial form)

provides the main sound transmission channel. Most of the sound radiation takes place at the nostrils. The closed oral cavity functions as a side branch resonator coupled to the main path, and it can substantially influence the sound radiated. Because there is no increase of the air pressure in the mouth, nasals are classed as sonorant consonants; because there is a complete closure within the vocal tract, they are discontinuant. The GA nasal consonants are listed in Table 2.5, and their vocal profiles are illustrated in Fig. 2.8.

Glides and Semivowels

Two small groups of consonants contain sounds that greatly resemble vowels. These are the glides TIPA/w,j/ and the semivowels TIPA/r,l/ (?, ?). Both are characterized by sonorant voicing, no effective nasal coupling, and sound radiation from the mouth. All four phonemes may be optionally devoiced (as in "which" or "rheum"); speakers of GA English usually consider voiced and devoiced allophones to be examples of the same underlying phoneme.

The glides TIPA/w/ and TIPA/j/, respectively, may be interpreted as extreme examples of the vowels TIPA/u/ and TIPA/i/—the former involves an extreme lip constriction, the latter an extreme palatal tongue constriction. In both cases, the constriction is a dynamic one, released gradually into the following vowel.

Figure 2.8: Vocal profiles for the nasal consonants. (After Potter, Kopp and Green)

Table 2.6: Glides and semi-vowels

Place		
Palatal	TIPA/j/	you
Labial	TIPA/w/	we (no final form)
Palatal	TIPA/r/	read
Alveolar	TIPA/l/	let

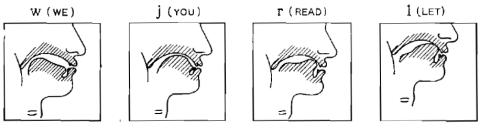

Figure 2.9: Vocal tract configurations for the beginning positions of the glides and semivowels. (After Potter, Kopp and Green)

The semivowels, by contrast, may be produced either dynamically or in a relatively static configuration; in fact, either of these two consonants may be produced as the nucleus of a syllable in English (e.g., in the words "bird" and "bull"; when produced as a syllable nuclei, these phonemes may be transcribed as TIPA/3^{r}/ and TIPA/l̩/, respectively). Both sounds are most reliably identified by a unique acoustic pattern: TIPA/r/ is the only sound in English with a third formant below 2000Hz, and TIPA/l/ is one of the few sounds in English with a third formant above 3000Hz. Both sounds are typically produced in syllable-initial position with both a tongue body constriction and a tongue tip constriction; in syllable-final position, both sounds are optionally produced with only a tongue body constriction. The tongue tip constriction for TIPA/r/ is curled back ("retroflex"), and the tongue body constriction is tightly bunched in the middle of the hard palate. The tongue tip constriction for TIPA/l/ is made with the tip touching the gum ridge like a TIPA/d/, but open on the left and/or right ("lateral"); the tongue body constriction is near the uvula. These are the only retroflex and lateral phonemes in English, but other languages have other phonemes (in some cases, stops and fricatives) with similar tongue tip positions.

The glides and semivowels for the GA dialect are listed, according to place of articulation, in Table 2.6. Their profiles, for the beginning positions, are given in Fig. 2.9.

Combination Sounds: Diphthongs and Affricates

Some of the preceding vowel or consonant elements can be combined to form basic sounds whose phonetic values depend upon vocal tract motion. An appropriate pair of vowels, so combined, form a diphthong. The diphthong is vowel-like in nature, but is characterized by change from one vowel position to another. For example, if the vocal tract is changed from the TIPA/e/ position to the TIPA/I/ position, the diphthong TIPA/eI/ as in say is formed. Other GA diphthongs are TIPA/Iu/ as in new, TIPA/OI/ as in boy; TIPA/AU/ as in out, TIPA/AI/ as in I, and TIPA/oU/ as in go.

As vowel combinations form the diphthongs, stop-fricative combinations likewise create the two GA affricates. These are the TIPA/tS/ as in chew and the TIPA/dZ/ as in jar.

Each of these combination sounds is perceived to be a phoneme by typical speakers of GA English. For example, in games where subjects are asked to reverse the order of

phonemes in a word (turning "scram" into "marks," for example), an affricate or diphthong will be treated as a single phoneme (e.g., turning "chide" into "daytch" rather than "dyasht"). The acoustic signal also gives us one reason to treat a combination sound as if it were a single phoneme: the average duration of a combination phoneme is shorter than the average total duration of its component phonemes (e.g., the average duration of TIPA/tS/ is shorter than the sum of the average durations of TIPA/t/ and TIPA/S/). In most other respects, however, a combination sound has exactly the same articulatory and acoustic characteristics as a sequence of two separate phonemes, e.g., a TIPA/tS/ is produced with an unvoiced alveolar closure that looks (e.g., if viewed using MRI) and sounds exactly like a TIPA/t/ closure, followed by an unvoiced palatal fricative that looks and sounds exactly like an TIPA/S/. The standard IPA notation for these sounds writes them as the sequence of two phonemes (e.g., TIPA/t/ followed by TIPA/S/) in order to emphasize their articulatory and acoustic decomposibility.

2.3 Quantitative Description of Speech

The preceding discussion has described the production of speech in a completely qualitative way. It has outlined the mechanism of the voice and the means for producing an audible code which, within a given language, consists of distinctive sounds. However, for any transmission system to benefit from prior knowledge of the information source, this knowledge must be cast into a tractable analytical form that can be employed in the design of signal processing operations. Detailed inquiry into the physical principles underlying the speech-producing mechanism is therefore indicated.

The following chapter will consider the characteristics of the vocal system in a quantitative fashion. It will treat the physics of the vocal and nasal tracts in some depth and will set forth certain acoustical properties of the vocal excitations. The primary objective–as stated earlier–is to describe the acoustic speech signal in terms of the physical parameters of the system that produced it. Because of physiological and linguistic constraints, such a description carries important implications for analysis-synthesis telephony.

References

Dewey, G. (1923). *Relative frequency of English speech sounds.* Cambridge, Massachusetts: Harvard University Press.

Fletcher, H. (1922). The nature of speech and its interpretation. *Bell System Technical Journal*, *1*, 129-144.

Levelt, W. J. M. (1989). *Speaking: from intention to articulation.* Cambridge, MA: MIT Press.

Licklider, J. C. R., Stevens, K. N., & Hayes, J. R. M. (1954). *Studies in speech, hearing and communication. final report, contract W-19122ac-1430.* Cambridge, Mass.

Miller, G. A., & Nicely, P. E. (1955). Analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.*, *27*, 338-352.

Pierce, J. R., & Karlin, J. E. (1957). Information rate of a human channel. *Proc. I.R.E.*, *45*, 368.

Pollack, I., & Ficks, L. (1954). Information of elementary multidimensional auditory displays. *J. Acoust. Soc. Am.*, *26*, 155-158.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois.

Webster, J. C. (1961). Information in simple multidimensional speech messages. *J. Acoust. Soc. Am.*, *33*, 940-944.