CHAPTER 3

SPEECH PRODUCTION, LABELING, AND CHARACTERISTICS

3.1 LANGUAGE

The American Heritage dictionary defines language as "the aspect of human behavior that involves the use of vocal sounds in meaningful patterns and, when they exist, corresponding written symbols to form, express, and communicate thoughts and feelings." An alternate definition might be "any method of communicating ideas as by a system of signs, symbols, gestures, or the like; the language of algebra or computer languages such as C or FORTRAN." Thus, one might study a language in one or more fields.

3.2 LINGUISTICS

Linguistics is the study of the rules by which speech sounds are assembled in a language. Linguists study the units of language and how they function.

3.3 PHONETICS

Phonetics is the study and the classification of the sounds of speech. The discipline of phonetics includes:

- The physiology of speech production and the movements of the vocal organs (articulators) to produce sounds and words.
- Acoustic phonetics, which involves the acoustic description of speech sounds.

3.4 SOUND CLASSIFICATIONS

A basic unit of linguistics is the phone, which is an articulated sound. Articulation is the process of producing sounds by manipulating the articulators, which changes the configuration of the vocal tract. The vocal tract includes the pharynx and the oral cavity to the lips. Producing sounds by clapping one's hands or stamping one's feet is not a form of articulation, although it is a means of communication. A collection of phones is called phonemes. Phones are sounds. Phonemes are the elements of speech. Phonemes may be combined into larger units called syllables, a definition of which is not always agreed upon by linguists. A common definition of a syllable is that it has a central vowel that is preceded and succeeded

by one or more consonants, which form a subset of the language of phonemes. The definition of a phoneme is such that if one phoneme is replaced by another, then the meaning of the word may be changed. There may be restrictions in a language concerning the usage of phonemes. For example, note that English never begins a syllable with the phoneme "ng."

A combination of syllables forms a word, which is usually a combination of two to five phonemes. However, there are words with as few as one phoneme, for example, a and I, and as many as ten or more, for example, synthesizers, and Mississippi.

The ten most frequently used English words are I, the, a, it, to, you, of, and, in, and he. We usually select words from a vocabulary of 5000 to 10,000 words; however, 3000 words are adequate for conversation. Later, we will specify the ten most common phonemes of English.

Sentences are larger linguistic units, which require a grammar, or rules that determine the manner by which sequences of words are joined together. For example, the following sentence is allowed.

- The bear's ear is blue.

However, the sentence

- Bear's ear blue the is.

is not allowed. Grammar is not the only determinant of sentence word order, for example, the sentence must make "sense." To illustrate this, note that "The fox jumped the river." is both grammatically correct and sensible. However, the sentence "The idea jumped the river." does not make sense, although it is grammatically accurate.

As we progress higher in the structure of language we come to semantics, which is the study of the meaning of words. The manner by which we order words is influenced by the grammar of the language and the meaning we wish to communicate.

3.5 PROSODY

Prosodic features of speech convey information about the long-term variations of pitch, intensity, and timing. We perceive prosodic features as stress and intonation, which are difficult to measure, model, and simulate in speech synthesis. For example, we express the distinction between a question and statement and doubt by changes in intonation or changes in our fundamental frequency of voicing. A speaker's emotional state, gender, health, and other factors may also be conveyed by variations in intonation. Stress is used in conversation to indicate the importance of words. For example, the statement

- **I** will be the judge of that.

as contrasted with

- I will be the judge of **that**.

have two different meanings to both the speaker and the listener when spoken.

There is no good way to indicate stress and intonation in writing; these two factors are used nearly exclusively in speech. Lexical stress is used by a speaker to emphasize a syllable in a word, e.g., CAMpus. Stress and intonation are very effective communication

techniques and have a connection to meaning, since variations in these factors can convey information about a speaker's emotional state, gender, health, and other factors.

Stress and intonation carry information important to the quality of the speech signal and are, therefore, important for speech synthesis. Intonation is perceived as changes in the fundamental frequency of voicing, F_0. A plot of the time variation of F_0, or its reciprocal, the period of voicing, is often called a pitch contour. In actuality, as mentioned in Chapter 1, pitch is the perception of F_0. The terms pitch and F_0 are often incorrectly used interchangeably. Stress is perceived as changes in increased speaking effort as well as changes in intensity, pitch, and duration of a sound.

Stress and intonation represent two suprasegmental features, another being duration. The term suprasegmental feature refers to speech characteristics that extend beyond a short segment. For example, a suprasegmental feature, such as pitch, may extend over an entire sentence, while a vowel feature is localized within a syllable within a word. Speech sounds vary in duration. For example, some vowels are longer than other vowels, and some consonants are longer than other consonants. It is even possible for the same vowel to vary in duration depending upon the location of the vowel within a word. For example, word-ending vowels tend to be longer than word-beginning vowels.

3.6 ORTHOGRAPHY

Orthography is a system of spelling that follows a set of rules and/or common usage. Spelling provides us with a visual means of representing spoken words with written letters. However, spelling is fallible, as we know. For example, how do you pronounce "read"? It is sometimes "red" and sometimes "reed," depending on the context. The words "way" and "weigh" are pronounced the same but have different meanings.

3.7 PHONEMIC TRANSCRIPTION OR LABELING

This is a procedure to convey pronunciation information as unambiguously as possible. It is a technique to provide an allophonic transcription that is more accurate than spelling rules.

3.8 ADAPTATION, ASSIMILATION, AND COARTICULATION

The sounds one produces are influenced and altered by the neighboring sounds. One type of such influence is called phonetic adaptation. Such adaptations are the result of variations in the manner by which a speaker moves the articulators, thereby causing changes in the vocal cavities. The manner and extent to which these cavities are changed are affected by the past, present, and following phonemes in a particular utterance. The positions of the articulators and the shape of the cavities for one phoneme influence the movements of the articulators for producing neighboring phonemes. One example of adaptation comes about as our speaking rate increases. The faster we speak, the less likely the tongue will reach specific target positions for specific phonemes. In summary, adaptation is the process of varying, or changing, a phoneme due to the influence of the shape of the vocal tract for neighboring phonemes. Adaptation is the result of one articulator modifying its movement

due to phonemic context. Thus, the production of a phoneme is influenced by its neighboring vocal tract shapes. However, the acoustic manifestation of the phoneme does not change.

Assimilation is an excessive form of adaptation. If a phoneme changes sufficiently to become more like its neighboring phonemes, then such a change in the phoneme is called assimilation. One example can be seen in the "sh" sound at the beginning of the word "should," due to the influence of the voiced phoneme following "sh." The influence can either be anticipatory of the next phoneme or a carry over from the previous phoneme. Assimilation results in an actual change in the sound of the phoneme.

Coarticulation is defined as the movement of two articulators at the same time for different phonemes. Coarticulation can occur with or without a change in sound production. One example is for the word "two." We can pronounce this word by moving the tongue for the phoneme "t" while we simultaneously round our lips for "u." However, we can also pronounce this word with no lip rounding. On the other hand, coarticulation can result in a "smearing" of segmental boundaries between phonemes, as in assimilation, which can modify the characteristics of the phoneme. For example, for the same phoneme "k" we have two different pronunciation in the words "keel" and "cool." These two phonemes are called allophones of the phoneme "k." Another example is the phrase "keep them," which is often pronounced as "keep em." And another example is the sentence "Up and at them," which we often say as "Up an at em." The term coarticulation is often used to mean that assimilation is also occurring.

3.9 ELEMENTARY ASPECTS OF SPEECH PRODUCTION

A simplistic representation of speech production is depicted in Figure 3.1, which is from a paper by Dennis Fry (1979) and drawn by L. S. Johnson. An idea originates in the brain, which can be expressed in words, which are stored in memory. Control is communicated to the articulators, including the lungs, jaw, lips, tongue, velum, and vocal folds. In Figure 3.1 the speaker is in the process of speaking the words, "The rain in Spain falls mainly in the plain," from Pygmalion by Shaw, which later became the play, My Fair Lady, by Lerner and Loewe.

Figure 3.2 is a more scientific view of speech production, although still simplistic. Yet, this is the beginning of a model for speech production based on physiology. The acoustic signal radiated from the lips is perceived by the listener. In American English, there are on the order of 41 phonemes, more or less. Most languages have 30 to 50 phonemes.

Speech is not an orderly, precise string of phonemes. It is more like a series of phonemes that have onset and offset slopes in amplitude, which contribute to the transitions between phonemes. Thus, speech is a sequence of "blurred" phonemes that are not usually pronounced precisely, thereby making phoneme identification from waveforms or spectrograms difficult.

While it is beyond the scope of this book to provide a detailed description of the anatomy and physiology of human speech production, we do provide a brief introduction to the vocal system. The principal laryngeal function is to provide a protective closure for the respiratory system. The larynx, supported by ligaments and controlled by muscles activated by numerous nerves, is composed of soft tissues encased in a cartilaginous skeleton. A mucous membrane lines the larynx, as it does the trachea. There are two pairs of folds: the vocal folds, or cords, and the ventricular (false vocal) folds; both pairs are membranes that extend into the air passageway.

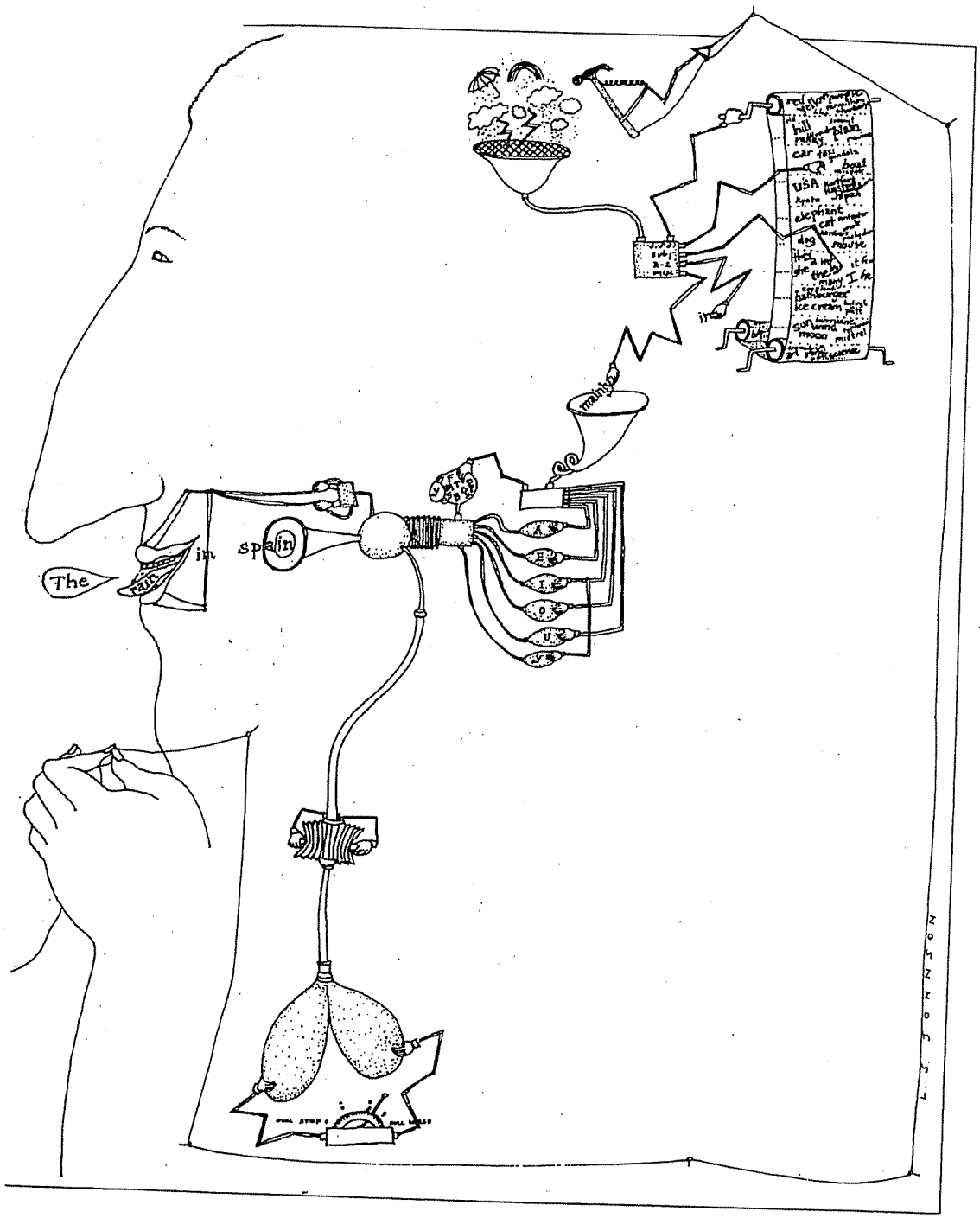


FIGURE 3.1 A representation of speech production.

Figure 3.3 shows a sketch of an individual with a laryngeal mirror (a dental mirror) located at the back of the throat, allowing a view of the vocal folds from the open mouth. Figure 3.4 is a x-ray photograph of a side view of the head of a man. This x-ray shows a laryngeal mirror located at the back of the throat, as in Figure 3.3.

Chapter 1 presented the human vocal system. The lungs are typically considered as air reservoirs, which are capable of expelling air up the trachea to the vocal folds. For voiced sounds (such as vowels), the air pressure increases until the folds are pushed apart forming a slit known as the glottis. A puff of air passes through this glottal opening, setting the vocal

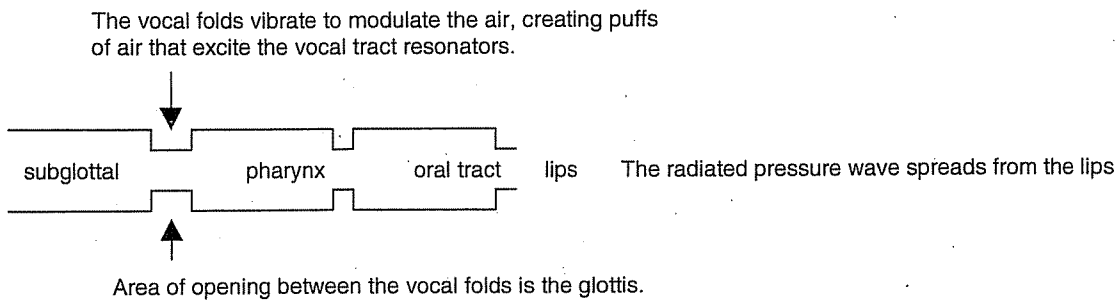


FIGURE 3.2 A model of speech production.

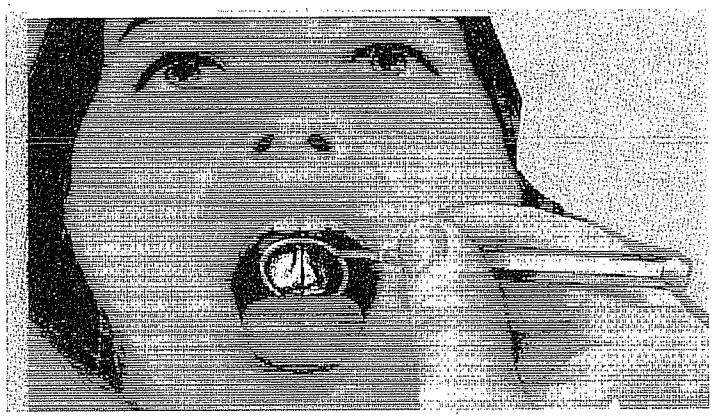


FIGURE 3.3 A view of the vocal folds using a laryngeal mirror.

FIGURE 3.4 An x-ray of the mouth and throat with laryngeal mirror.

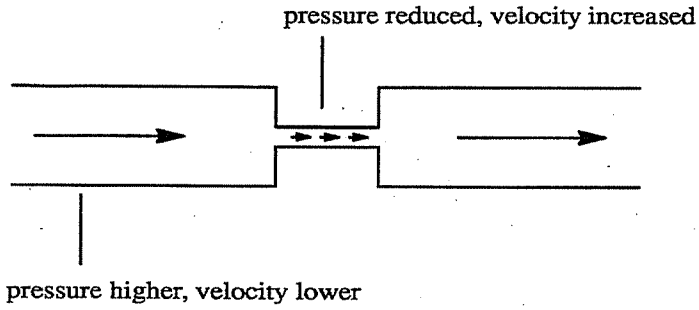
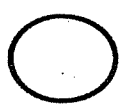


FIGURE 3.5 Bernoulli effect.

folds into vibratory motion. The myeloelastic-aerodynamic theory of phonation proposes that this vibratory motion is the result of the interplay of two forces. One force is the subglottal pressure, which causes air to push the adducted folds apart, releasing a puff of air. The other force is the Bernoulli effect (the linear combination of pressure and velocity squared is constant), which is a suction phenomenon that pulls the cords together (adducts the glottis) when the air velocity through the glottis is relatively large. If the glottis is initially abducted (opened), then the Bernoulli effect will cause the glottis to close.

The myoelastic–aerodynamic theory originated with Helmholtz and Muller in the 19th century and was later expanded by van den Berg in the 1950s (Borden and Harris, 1980, pg. 75). The myoelastic part of the theory arises from the fact that the muscles (myo) change the elasticity and tension of the vocal folds to bring about changes in the frequency of vibration of the vocal folds. The mass of the vocal folds also affects the vocal fold vibratory frequency. As the vocal folds become shorter and thicker, they become more massive, and the frequency of vibration is lowered. As the folds become longer and thinner, they become less massive and can vibrate at a higher frequency. Elastic folds vibrate faster because they are able to "bounce" back at a more rapid rate. Tense folds vibrate faster than slack folds. The folds are stretched to make them more tense. The muscles regulate the thickness and tension of the vocal folds. The aerodynamic part of the theory says that the driving force for vibration is airflow. The air expelled from the lungs activates the vibration of the vocal folds. The Bernoulli effect, shown in Figure 3.5, is one factor that affects the vibration of the vocal folds; another is the recoil force of the vocal folds.

The length of the vocal folds for males is 17 to 24 mm, and 13 to 17 mm for females. The folds can stretch from 3 to 4 mm. For a male voice phonating at a fundamental frequency of $F_0 = 120$ Hz, the glottis is approximately 10 mm long and 4.25 mm wide, giving a glottal area of approximately 42.5 mm^2. The actual glottal area is less because the glottis is not rectangular, but oval. A truer glottal area would be about 28 mm^2. At $F_0 = 340$ Hz, the glottis is approximately 11.5 mm long and 1.65 mm wide, giving a glottal area of approximately 19 mm^2. Since the glottal area is more nearly oval than rectangular, a more accurate glottal area would be 13 mm^2. The size of the glottis in these cases can be compared to that of a dime, which is 18 mm (0.7 inches) in diameter, with an area of 254 mm^2 (see Figure 3.6).



Let this represent a dime: dia. 18 mm, area 254 mm^2

circle with dia 4 mm, area 12.56 mm^2

FIGURE 3.6 Comparison of the glottis with a dime.

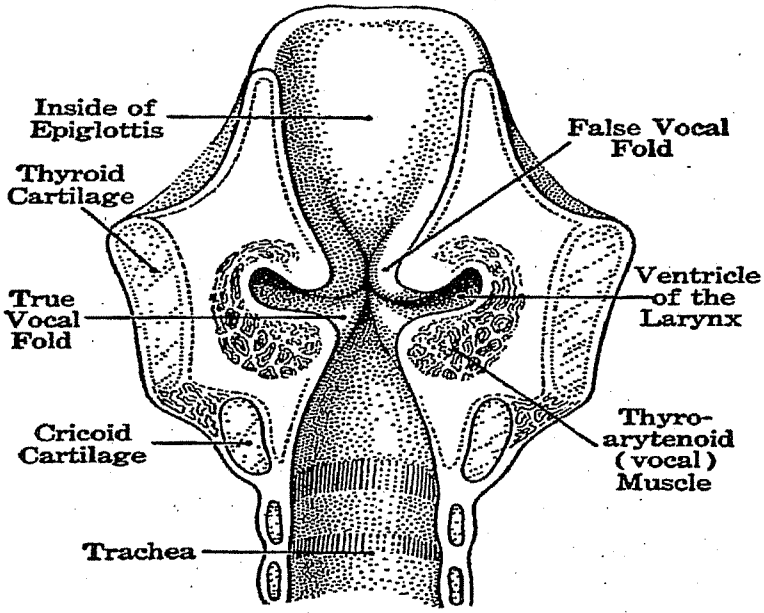


FIGURE 3.7 The vocal folds.

An artist's rendition of a frontal section of the larynx, the cartilages, and connective tissues that encloses the vocal folds, is shown in Figure 3.7 (after V. A. Anderson, 1977). Figure 3.8 is a sequence of 23 ultra high-speed film frames of the vocal folds during vibration. This sequence is to be viewed from right to left, top to bottom. The sequence begins with the vocal folds nearly closed in frame 1 and fully open at frame 4, then closed at frame 7, and so on. Since the filming was done using a laryngeal mirror, anterior is at the top and posterior is at the bottom of each frame, and the right vocal fold is at the left. Later chapters discuss this further and provide models of the vibratory motion of the vocal folds. A digitized sequence of frames of one vibratory cycle of the vocal folds along with a movie of this sequence is contained in a folder on the accompanying CDs. See the README file.

As the vocal folds vibrate, they modulate the flow of air from the lungs, creating pulses of air that are known as the volume–velocity of air flow through the folds and vocal tract. In physics courses, one usually studies the particle motion of air. For speech, we are

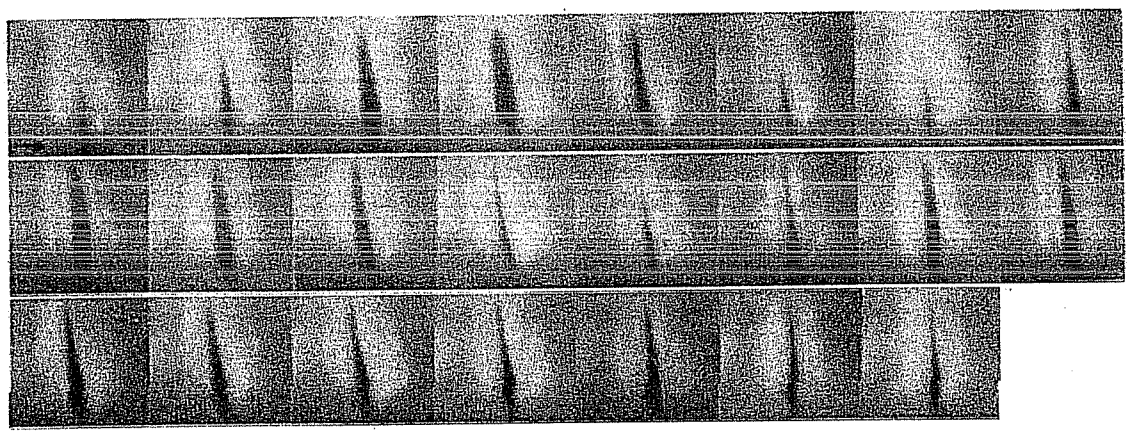


FIGURE 3.8 A sequence of ultra high-speed film frames of the vibratory motion of the vocal folds.

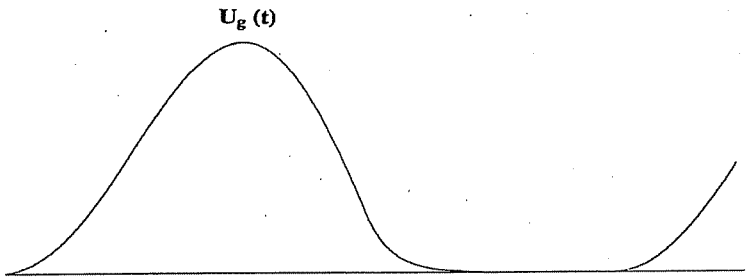


FIGURE 3.9 A glottal volume–velocity pulse.

concerned with the movement of a volume of air through the vocal tract. A sketch of a volume–velocity waveform (pulse) of air versus time is shown in Figure 3.9.

The succession of air pulses generated as result of the vibratory motion of the vocal cords sets up an acoustic field that continues to travel up the vocal tract. The phonation or sound generated has an auditory correlate (or pitch) that is directly related to the frequency of vibration and a loudness that is determined by the amplitude of the acoustic pressure wave. The frequency of oscillation of the folds is determined by their mass, length, thickness, elasticity, and compliance, as well as by the subglottal pressure.

The volume–velocity (rate of air flow), as it passes through the glottis, modified by the pharynx, mouth, and nasal cavities, is finally radiated from the lips and nostrils as voiced sounds. All voiced sounds such as all vowels and voiced consonants (e.g., b, z, and g), originate at the vocal cords. The glottal volume–velocity is considered the acoustic source waveform for voiced sounds. The other sound is called "unvoiced." In this case, the vocal cords are held apart so that air expelled from the lungs up the trachea is unaffected by glottal vibrations. There are two fundamental types of unvoiced sounds: fricative and plosive. The former is typified by the sound s, which is produced by expelling air through a constriction such as the teeth to produce turbulent air flow. The latter unvoiced sound, typified by p as in puff, is created by the rapid release of air pressure built up behind a closure such as the lips. Since fricatives and plosives can also be voiced, as in b, d, g, and z, one should be cautious of oversimplifications.

The vocal folds are held open (abducted) for unvoiced sounds. While for voiced sounds, the vocal folds are brought together (adducted). To bring the vocal folds into a vibratory mode, the subglottal pressure is greater than the supraglottal pressure. There are two major types of source excitation, voiced, where the vocal folds vibrate, and unvoiced, where the vocal folds are held open. A third type of excitation is sometimes called mixed voiced and unvoiced excitation or more simply mixed excitation. This type of excitation is a combination of voiced and unvoiced excitation, for example, for the phoneme "z," as in zebra. In this latter case, the vocal folds vibrate and there is also a turbulent sound produced at the tip of the tongue near the teeth.

Primarily the lung pressure, the tension of the folds, and their mass, affects the fundamental frequency of vibration of the vocal folds. Consequently, the average, minimum, and maximum fundamental frequency for men, women, and children differs. Table 3.1 summarizes these differences.

Primarily, the source or the shape of the volume–velocity waveform affects the quality of the voice, while the movement of the articulators affects the intelligibility of speech. This is illustrated in Figure 3.10.

Figure 3.11 shows several glottal pulse waveforms. On the left, the pulse becomes more skewed to the right for pulses 4, 3, 2, 1. The shape of pulse 4 is typical of a breathy

TABLE 3.1. Average, Minimum, and Maximum Fundamental Frequencies for Men, Women, and Children

	F_0 average (Hz)	F_0 minimum (Hz)	F_0 maximum (Hz)
Men	125	80	200
Women	225	150	350
Children	300	200	500

voice, shape 3 for a falsetto voice, shape 2 for a modal (normal) voice, and shape 1 for a vocal fry (creaky) voice. On the right in Figure 3.11, the abruptness of closure is the characteristic that is important. Pulse 1 is typical of modal voice, pulse 2 of vocal fry, pulse 3 of falsetto, and pulse 4 of breathy voice. One can see that the excitation of a breathy voice is one that has almost a sinusoidal type of volume–velocity waveform, indicating that the vocal folds never achieve full or complete closure, with some air leakage at all times. For modal voice, the excitation pulse typically occupies approximately 65% of the pitch period, while for vocal fry, the pulse is about 25% of the pitch period. We discuss voice types again later.

The vocal tract is usually considered to be the entire cavity or passageway from the glottis to the lips. The basic assumption of nearly all speech production models is that the source (glottis) is independent of the vocal tract. The vocal tract may be modeled in various ways, and the model may, in turn, be excited by numerous source waveforms. However, the assumption of independence between source and vocal tract is not always valid, for example, as in the production of certain transient sounds such as p in pot.

The primary element controlling the vocal tract is the tongue, which divides the vocal tract into two resonant cavities, the pharynx and mouth, which, in turn, determine the transmission characteristic of the vocal tract. This characteristic can also be modified by coupling into the nasal cavity under the control of the soft palate or velum. As shown in Figure 1.5 in Chapter 1, the nasal tract begins at the velum and ends at the nostrils. The length of the nasal tract is about 12 cm.

The vocal tract transmission characteristic is such that it causes certain frequency components of the laryngeal signal to pass with less attenuation than others. Examples follow later. The resonances, or peaks, in the spectrum are referred to as formants, the center frequencies of which are designated in their ascending order of appearance as F_1, F_2, and so forth. Each formant also has a bandwidth. Studies indicate that the first two or three formants generally suffice for the perceptual characterization of most voiced English vowels

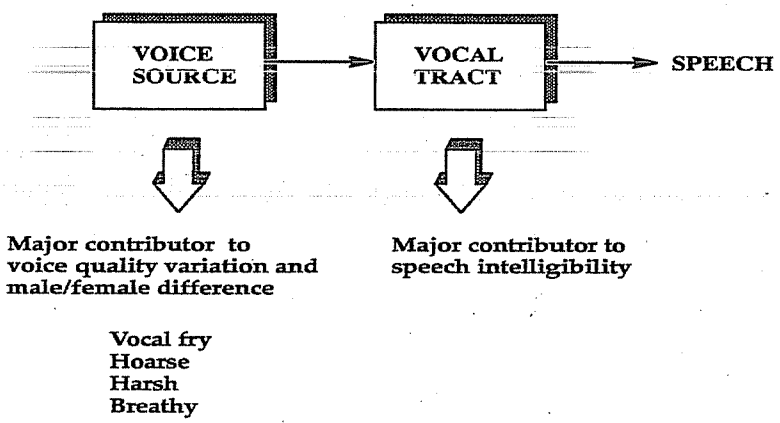


FIGURE 3.10 Quality of voice and intelligibility of speech.

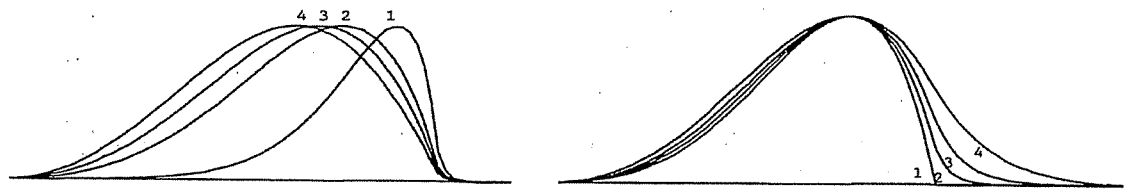


FIGURE 3.11 Glottal pulse waveforms for various voice types.

and consonants and, therefore, of speech. The higher formants are important for natural sounding or good quality speech. A similar formant structure is also observed for unvoiced sounds. The most significant frequency range for speech is approximately 250 to 3000 Hz.

The formants occur at the natural frequencies of the human vocal tract, which is actually a nonuniform acoustical tube. To a first-order approximation, we can model the vocal tract as a uniform acoustical tube closed at one end and open at the other. Recall from physics that the natural frequencies (standing waves) of a tube closed at one end and open at the other are determined by the wavelength, which in this case is $\lambda = 4L$. Thus, the natural frequencies, F_n, of this model are

$$F_n = (2n + 1)c/4L, \quad \text{for } n = 0, 1, 2, 3, 4, 5, \ldots \tag{3.9.1}$$

where c is the velocity of sound in air ($\cong$34,000 cm/sec) and L is the length of the tube, which for the adult male is approximately 17 cm. Making these substitutions, we have $F_n = (2n + 1)500$ Hz. Thus, for this idealized model in the range from 0 to 5 kHz, there are five equally spaced formants. Figure 3.12 shows the first three formants from 500 Hz to 2.5 kHz. In actuality, the vocal tract is not uniform, and the formant locations as well as their bandwidths depend on the particular configuration of the vocal tract. For continuous speech, the speaker is continuously changing the vocal tract as well as the glottal excitation.

3.10 VOICE VERSUS SPEECH

The situation depicted in Figure 3.10 is overly simplistic. While it is true that the shape of the source excitation waveform does affect the quality and type of vocal characteristics, it is not true that this is exclusively the case. The type of voice is also affected by the movement and placement of the articulators, by the rate of speaking, as well as by other factors. So the configuration of the vocal tract does affect vocal characteristics. As an example, New England and New York dialects are often called clipped because native

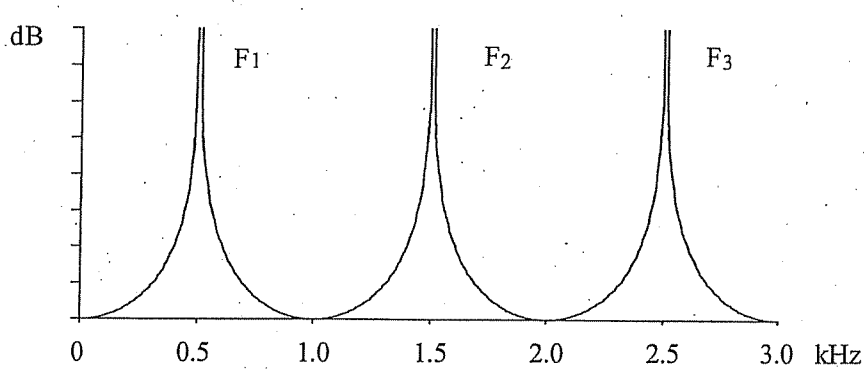


FIGURE 3.12 An idealization of the formants for a tube closed at one end and open at the other.

New Englanders speak rapidly, often truncating some phonemes. They often habitually make phoneme substitutions, and/or insertions, and/or omissions. On the other hand, native residents of southern states in the United States often speak slowly, with drawn-out vowels. These speaking styles come about through habit and are due to variations in speaking rate and in articulatory movements. Another example is accents. A specific sentence can be spoken in standard American English as well as with a Spanish or German or other accent. In this case, the perceived differences can be due again to differences in the movement of the articulators, with perhaps some differences in the shape of the excitation waveform as well. Perhaps the major point to remember is that voice or vocal characteristics are different from speech. This can be illustrated by thinking of voice and speech synthesis. The primary goal of speech synthesis is to produce speech that is highly intelligible with good vocal qualities that sounds like either a "standard" male or female voice. Speech synthesis for most commercial applications does not require the mimicking of a specific vocal characteristic or voice type. The primary goal of voice synthesis is to reproduce a specific voice, such as that of former President John F. Kennedy, or former actor John Wayne, or Mickey Mouse; or one may want to create a new voice. Mel Blanc was known as the man of a thousand voices because he created the voices of Bugs Bunny, Porky Pig, Daffey Duck, Tweety Bird, Sylvester the Cat, and many others. One goal of voice synthesis is to be able to create new vocal characteristics much like a composer may create a new musical composition. In summary, voice and speech synthesis are different, even though most engineering authors do not make this distinction, although they should. Thus, the characteristics of voice are not the same as the characteristics of speech, although they may be related. The parameters we use to model a voice differs from those we use to model speech.

3.11 SOURCE-FILTER MODEL

The source-filter model, due to Fant (1960), is shown in Figure 3.13. This model of sound production is linear and assumes superposition holds. The source provides the excitation,

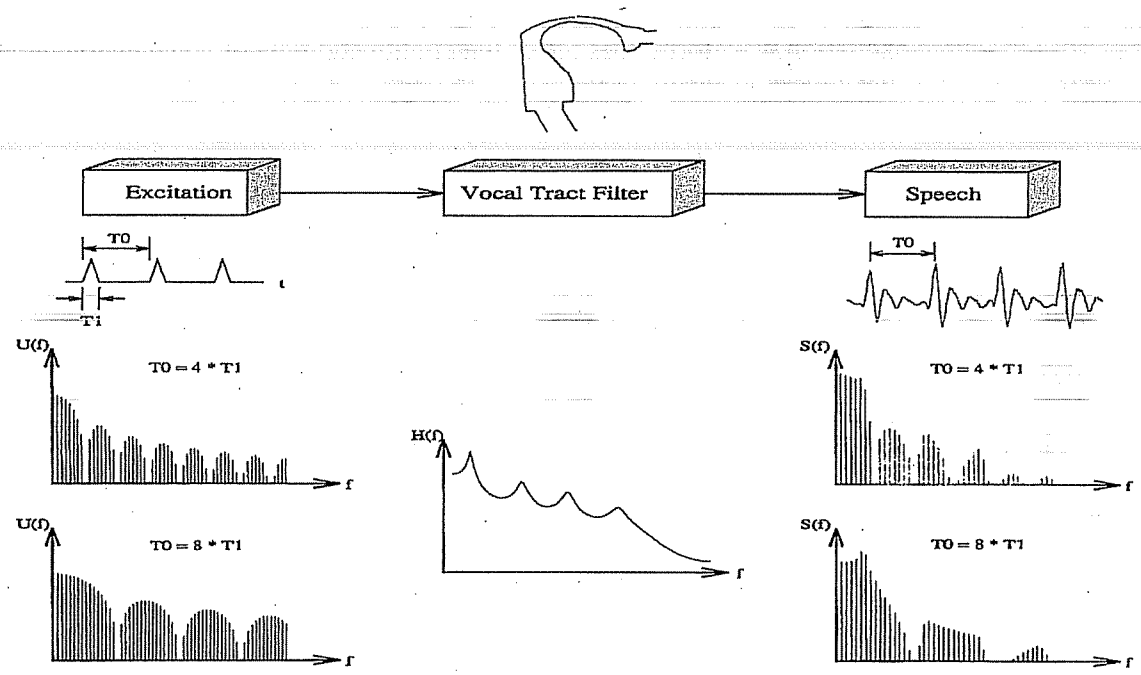


FIGURE 3.13 The source-filter model of speech production.

which is shaped spectrally by the vocal tract filter. Since the source is periodic for vowels, its spectrum consists of the fundamental and its harmonics, or in other words, a line spectrum. This line spectrum is filtered to produce speech. The filter changes with time to simulate the effect of changes in the vocal tract shape. So the vocal tract filter is the primary factor in producing various sounds. We will elaborate on this model in succeeding chapters.

3.12 CLASSIFICATION OF PHONEMES

For a spoken language, the acoustic signal radiated from the lips consists of a string of sound elements called phonemes. There are 30 to 50 phonemes for most languages, while for American English there are about 41 phonemes. There are various ways to classify phonemes.

3.12.1 Acoustic

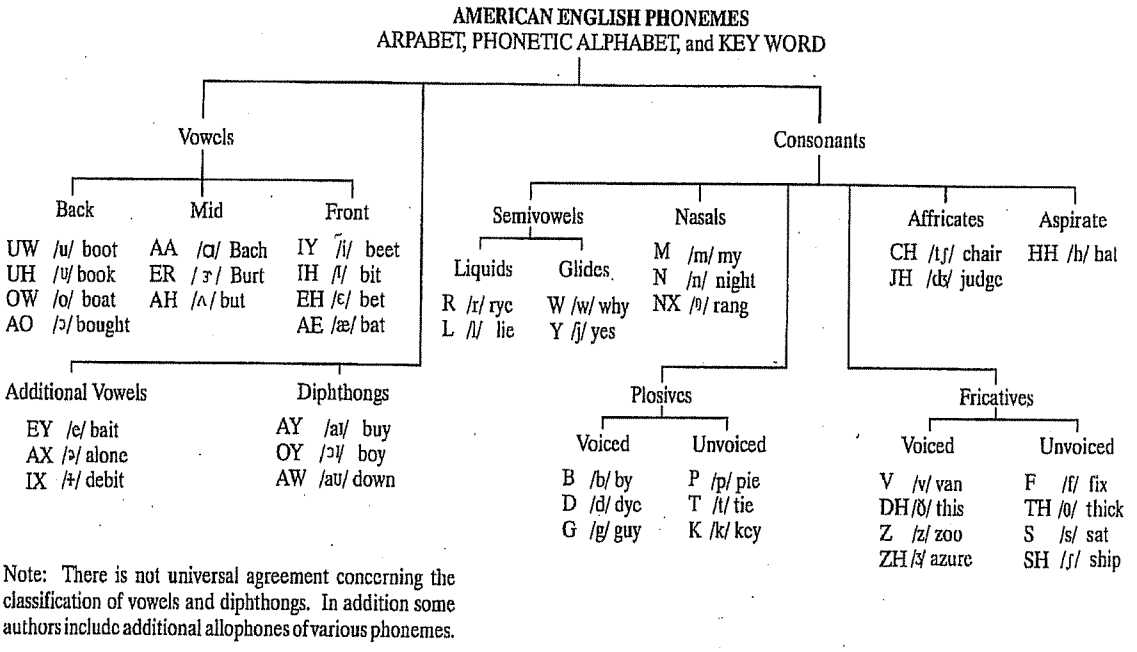
One classification is according to the type of excitation, for example, voiced and unvoiced.

3.12.2 Articulation

Another classification is by the manner and place of articulation. The typical places of articulation include labial (lips), dental (teeth), labiodental (lips and teeth), alveolar (gums), palatal (palate), velar (soft palate), and glottal (glottis).

3.12.3 Phonemes

Yet another classification is made among vowels and consonants. These phoneme classes contain various subclasses. Vowels tend to be steady, at least over a relatively long time interval, while consonants tend to be brief. There are several schemes for labeling the phonemes. The International Phonetic Alphabet (IPA) is one standard of labels or symbols. An abbreviated IPA list is shown in Figure 3.14. Also shown in this figure is an abbreviated list of the



Note: There is not universal agreement concerning the classification of vowels and diphthongs. In addition some authors include additional allophones of various phonemes.

FIGURE 3.14 A list of phonemes using the IPA and upper case ARPAbet alphabets.

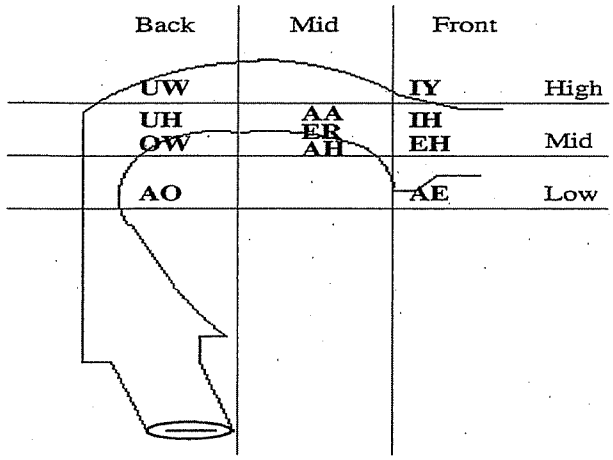


FIGURE 3.15 Some examples of vowel place of articulation.

upper case ARPAbet, or the ARPA (Advanced Research Projects Agency) alphabet, which we shall use, since it is easily typed and can therefore be electronically communicated. Figure 3.14 contains 41 phonemes. It is common practice to bracket phonetic transcriptions with forward slashes, for example, /AA/ as in Bach.

3.13 VOWELS (14)

For our purposes, we consider 14 vowels, which are all voiced. The place of articulation for these vowels varies as shown in Figures 3.14 and 3.15. One example is the vowel /IY/, as in beet, which is a high, front, unrounded (no lip protrusion) vowel. /AO/, as in bought, is a low, back vowel, while /UW/, as in boot, is a high, back, rounded vowel.

3.13.1 Schwa Vowel

A schwa is often called a degenerate vowel, which is a vowel to which other vowels tend to when one articulates rapidly or carelessly during fluent speech. The initial vowel in the word "ahead" is a schwa. A schwa tends to occur when the tongue hump does not have time to move into a correct position, but instead assumes a neutral position in the vocal tract, thereby causing the vocal tract to approximate a uniform tube. The schwa vowel tends to be short in duration and weak in amplitude.

3.13.2 Tense and Lax Vowels

Some vowels and diphthongs (in American English) are intrinsically longer than others and are made by the tongue reaching a more extreme position. The vowels with more extreme tongue adjustments and longer duration are termed tense vowels, more for their function in the language than for their method of production. Some examples appear in open syllables such as see, saw, lah, sew, Sue, say, sow, soy, cue.

Shorter vowels, which can appear in closed syllables (syllables ending with a consonant), but not in open syllables, are called lax vowels because they are produced with less extreme movements. Some examples appear in closed syllables such as sing, strength, sang, song, sung.

3.13.3 Diphthongs (3)

A diphthong is a gliding monosyllabic phoneme starting at one articulatory position for one vowel and moving toward the position of another vowel. There is also a change in vowel resonance.

3.13.4 Jitter and Shimmer

While vowels and diphthongs are considered to be steady, the nature of their production is such that there is a variation between successive periods, which is called jitter, as well as a change in the amplitude, which is called shimmer. These characteristics are illustrated in Chapter 1.

3.13.5 Semivowels: Liquids (2) and Glides (2)

Liquids move the tongue toward the alveolar ridge. Glides are similar to diphthongs, but a glide has a faster transition, with the tongue tip raised and the oral cavity is constricted.

3.14 CONSONANTS (17)

Consonants consist of several subclasses. For example, using the manner of articulation, there are **plosives or stops** (6), which are generated by blocking the air flow in the oral cavity to build up pressure, then the pressure is suddenly released. There are three voiced and three unvoiced plosives. The **fricatives** (8) are often classified by the type of excitation, that is, voiced and unvoiced. There are four voiced and four unvoiced fricatives, the latter form a constriction in the oral tract to produce turbulence (friction). The **nasals** (3) are generated by lowering the velum to connect the nasal cavity to the pharynx and the oral tract. The mouth is also closed. Nasals have a resonance in the nasal cavity, which introduces zeros in the overall system transfer function. The acoustics are such that a nasal murmur is within the 200 to 300 Hz range for males. The resonance or formant is a bit lower for /M/ and /N/ than for /NX/ because there is a progressively decreased volume of the oral cavity as the closure moves back in the mouth. Sometimes, diphthongs and semivowels (liquids, and glides) are grouped with consonants. Figure 3.16 shows the place of articulation for the plosives, fricatives, nasals, and semivowels.

3.15 AFFRICATES (2)

An affricate is a stop with a fricative release. One is voiced and one is unvoiced.

3.16 ASPIRATE (1)

An aspirate is produced by a steady airflow through the open vocal folds.

3.17 CONTINUANTS

These phonemes have a fixed, non-time varying vocal tract, which is excited by a source. Vowels, fricatives (voiced and unvoiced), semivowels, aspirates, and nasals, are all sustained, unlike stops.

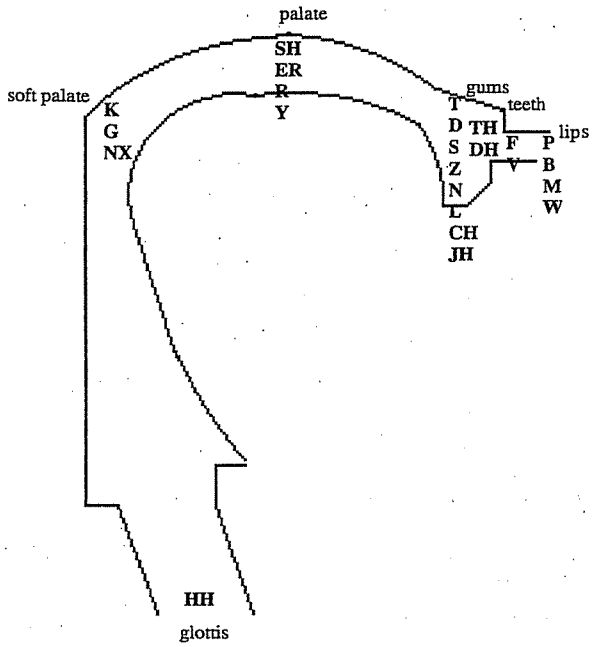


FIGURE 3.16 Place of articulation for plosives, fricatives, nasals, and semivowels.

3.18 NONCONTINUANTS

These phonemes have a changing vocal tract, for example, diphthongs, stops, and affricates.

3.19 SIBILANTS

Sibilants are high-pitched sounds with an obvious hiss as in sigh and shy.

3.20 NONSIBILANTS

Nonsibilants are the other fricatives, not classified as sibilants.

3.21 SONORANTS

These include the nasals and semivowels (liquids and glides) and are produced with less constricted airflow than the obsruents.

3.22 OBSTRUENTS

These include the stops and fricatives. An alternate phoneme classification to Figure 3.14 is shown in Figure 3.17.

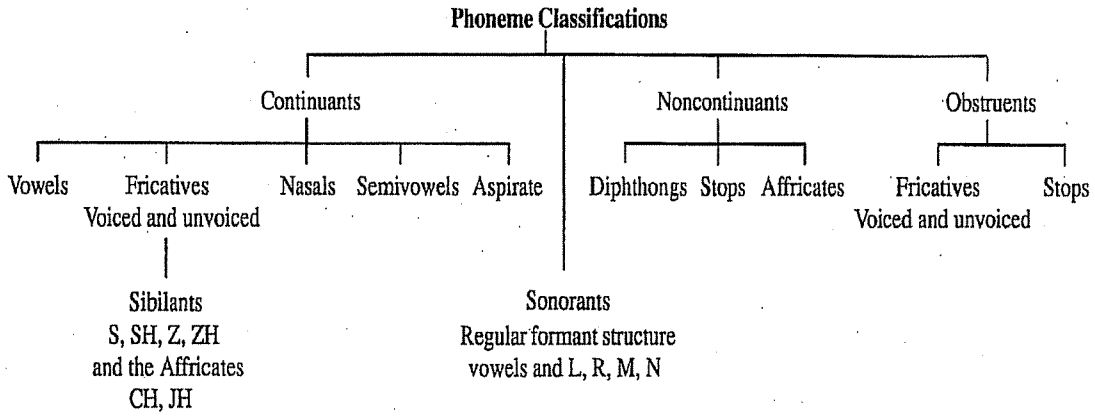


FIGURE 3.17 Summary of phoneme classifications.

3.23 VOICE BAR

A voice bar is a very low-frequency formant, typically near 150 Hz. A voice bar occurs when the vocal folds are vibrating and exciting an occluded nasal and oral tract. Voiced fricatives often contain a voice bar, e.g., /Z/.

3.24 ALLOPHONES

Allophones are a variant of a phoneme, as for /K/, as in keel and cool.

3.25 STATISTICS

The eleven most frequently occurring phonemes of English are (Denes, 1963):

/AX/, as in about, 9.04%
/T/, as in ten, 8.40%
/IH/, as in bit, 8.25%
/N/, as in net, 7.08%
/S/, as in sat, 5.09%
/D/, as in debt, 4.18%
/L/, as in let, 3.69%
/M/, as in met, 3.29%
/DH/, as in that, 2.99%
/K/, as in kit, 2.90%
/AY/, as in buy, 2%

3.26 WHISPER

A whisper is produced like an aspirate, by a steady flow of air through the glottis.

3.27 SPECTRAL CHARACTERISTICS

Each phoneme has certain spectral characteristics. One example of the waveform, spectrum, and spectrogram for the phoneme /IY/ is shown in Figure 3.18. Additional examples for the vowels and some consonants appear in Appendix 5. Figure 3.19 shows a typical vocal tract shape, waveform, cross-sectional vocal tract area, and spectrum for the vowel /IY/

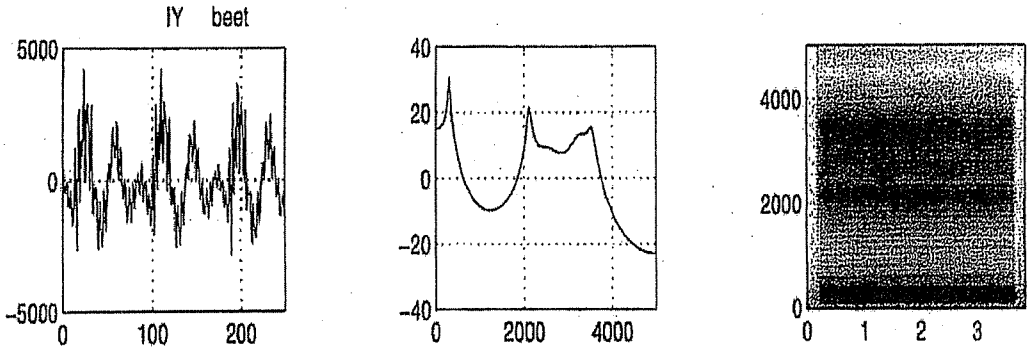


FIGURE 3.18 The waveform, spectrum, and spectrogram for the vowel /IY/.

calculated using the articulatory speech synthesizer described in a later chapter. Additional examples for the vowels appear in Appendix 5.

The formant features for the vowel /IY/ for both a male and female speaker appear in Figure 3.20. The formant frequencies are shifted upward by about 20% for the female speaker relative to the male speaker.

Figure 3.21 illustrates the "phoneme" labeling of the all voiced sentence "We were away a year ago," spoken by a male speaker. In this figure, the labeling is done by indicating the spelling of the words. This is not the conventional procedure, but it serves as an aid to mark the location of the various phonemes and to indicate how the phonemes vary in duration for this sentence. Figure 3.22 shows the pitch contours for both a male and female speaker for the sentence, "We were away a year ago."

It is often useful to know the formant frequencies and bandwidths for the various vowels. Such data were collected some years ago by Peterson and Barney (1952), and recalculated by Childers and Wu (1991) using digitized data and more modern analysis techniques. The latter data are presented in Figures 3.23 and 3.24 for both male and female speakers. These data can be used for speech and speaker recognition tasks. The vowel

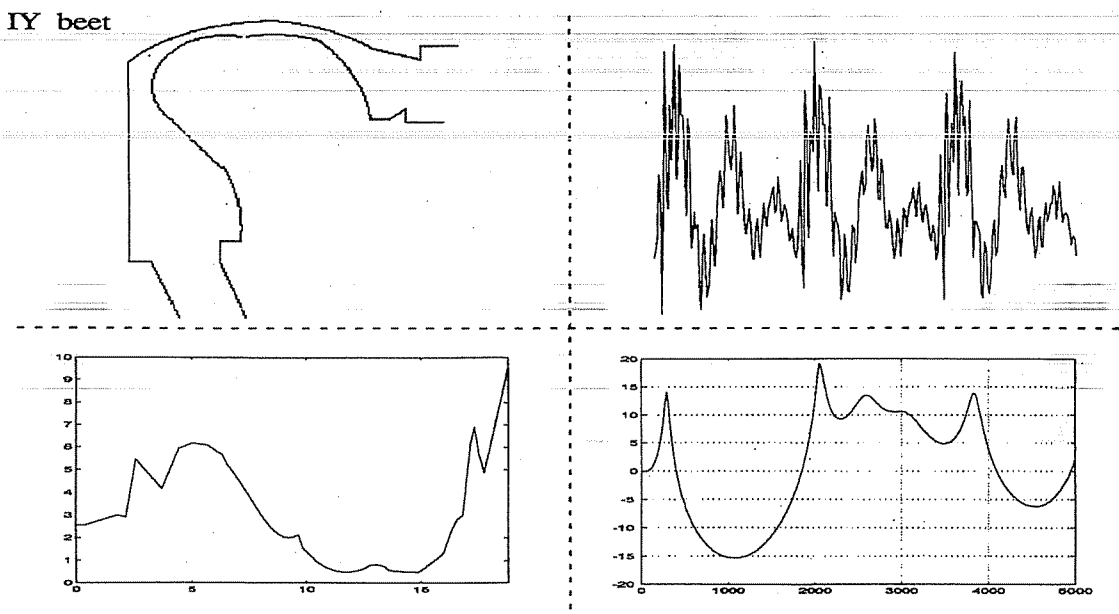


FIGURE 3.19 The vocal tract shape, waveform, cross-sectional area, and spectrum for the vowel /IY/.

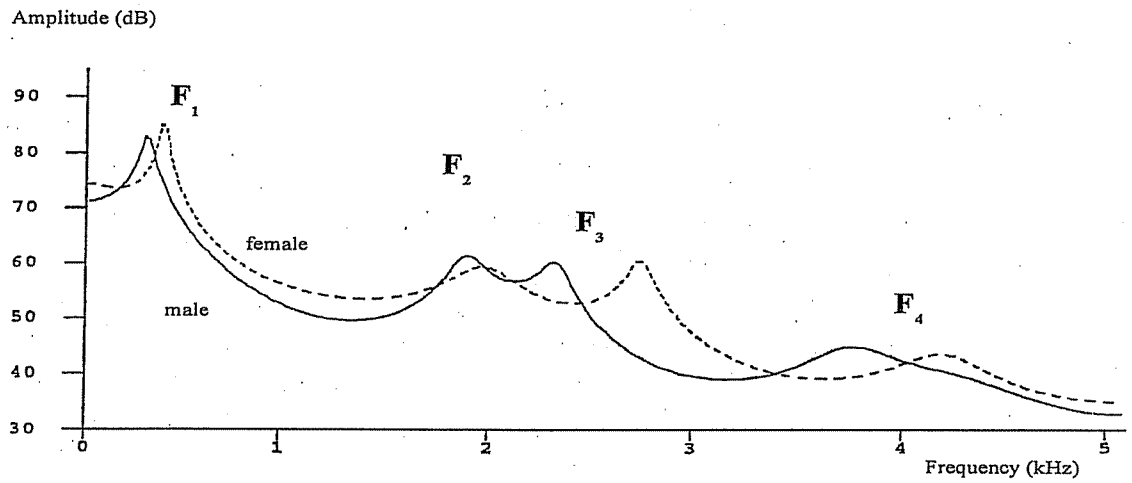


FIGURE 3.20 The formants for the vowel /IY/ for a male and female speaker.

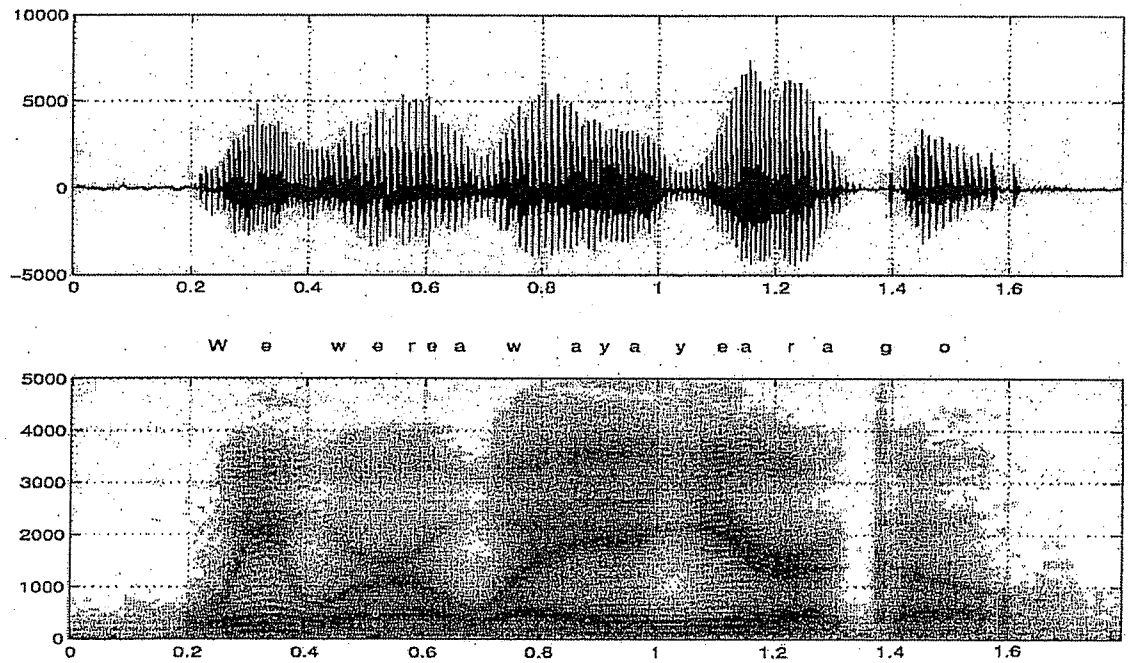


FIGURE 3.21 The labeled waveform and spectrogram for the sentence, "We were away a year ago."

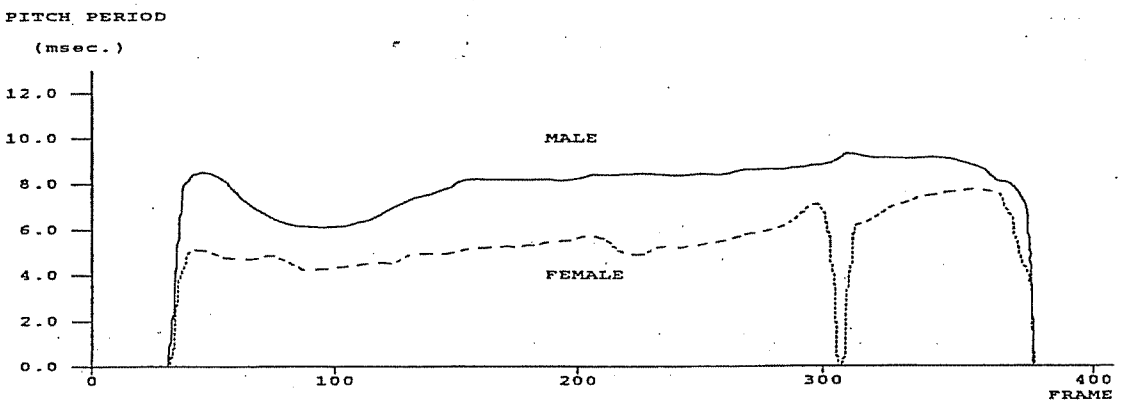


FIGURE 3.22 The pitch contours for both a male and female speaker for the sentence, "We were away a year ago."

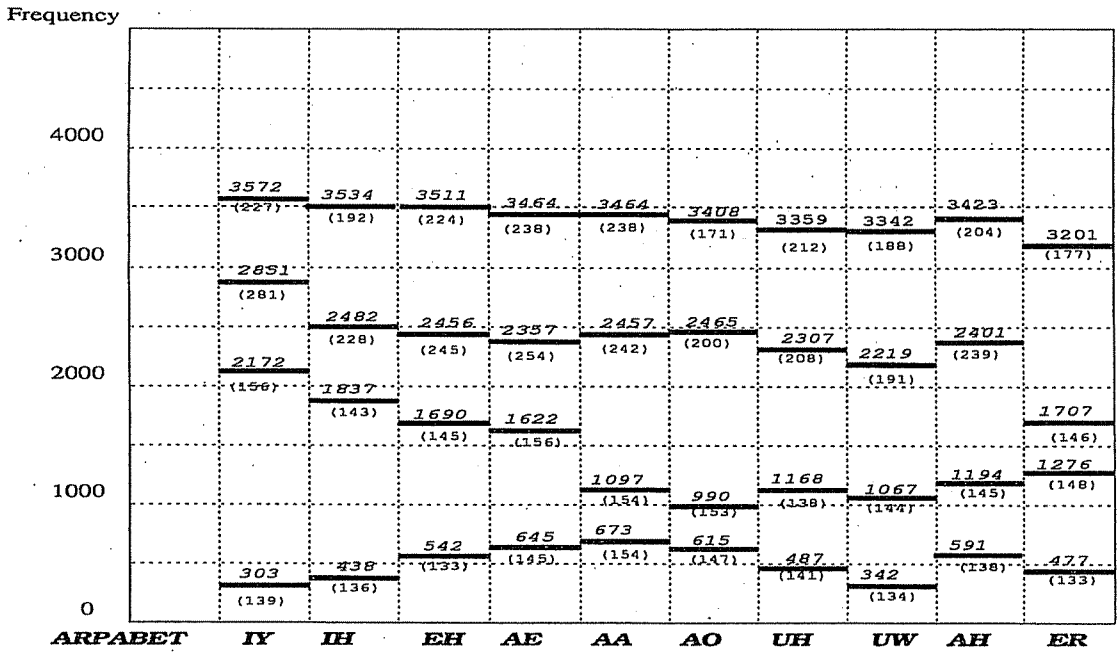


FIGURE 3.23 Average formant frequencies and bandwidths for ten vowels for male speakers.

triangle is obtained by plotting the second formant, F_2, versus the first formant, F_1, as shown in Figure 3.25 for male speakers. This data gives one an idea of how the first two formants vary for the vowels.

Table 3.2 contains a summary of minimum and inherent durations of some of the phonemes. This data was compiled by Klatt and appears in Allen, Hunnicutt, and Klatt (1987). The data is intended for use in speech synthesis. However, it is also useful as a phoneme feature and similar data has been used in isolated word recognition tasks (Gupta

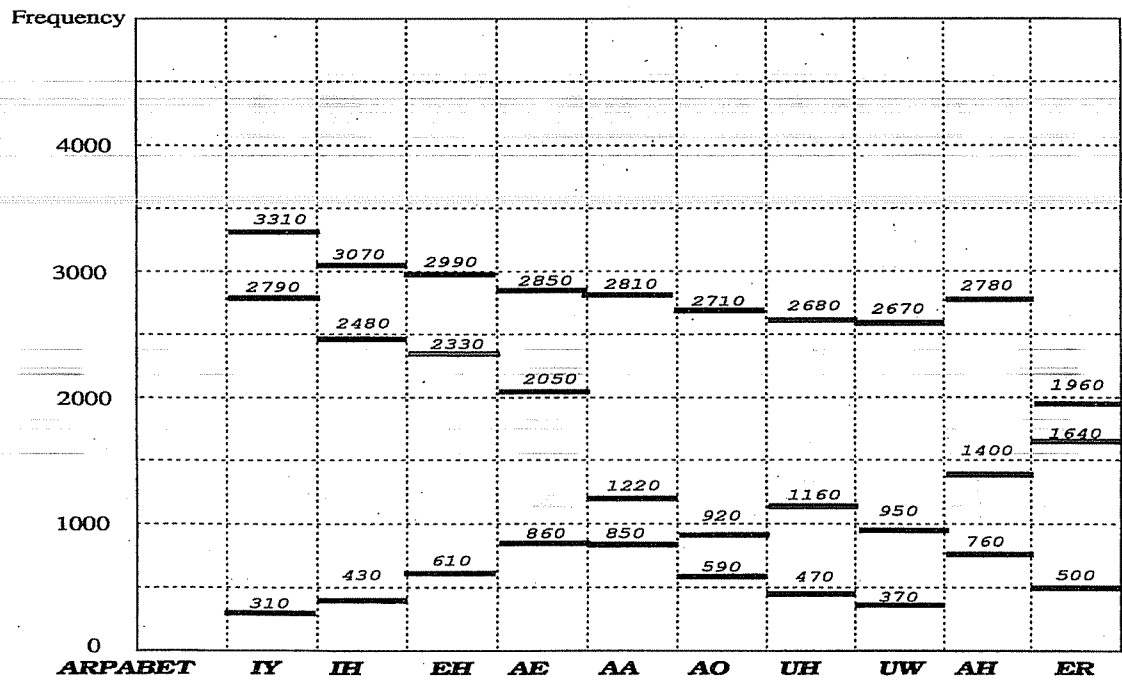


FIGURE 3.24 Average formant frequencies and bandwidths for ten vowels for female speakers.

TABLE 3.2. A Summary of Minimum and Inherent Phoneme Durations (msec)

Vowels								
AA	100	240	AE	80	230	AH	60	140
AO	100	240	AW	100	260	AX	60	120
ER	80	180	AY	150	250	EH	70	150
IH	40	135	IX	60	110	UW	70	210
IY	55	155	OW	80	220			
OY	150	280	UH	60	160			
Sonorant consonants								
L	40	80	HH	20	80	R	30	80
Y	40	80	W	60	70			
Nasals								
M	60	70	N	50	60	NX	60	95
Fricatives								
DH	30	50	F	80	100	S	60	105
SH	80	105	TH	60	90	V	40	60
Z	40	75	ZH	40	70			
Plosives								
B	60	85	D	50	75	T	50	75
G	60	80	P	50	90			
K	40	80						
Affricates								
CH	50	70	JH	50	70			

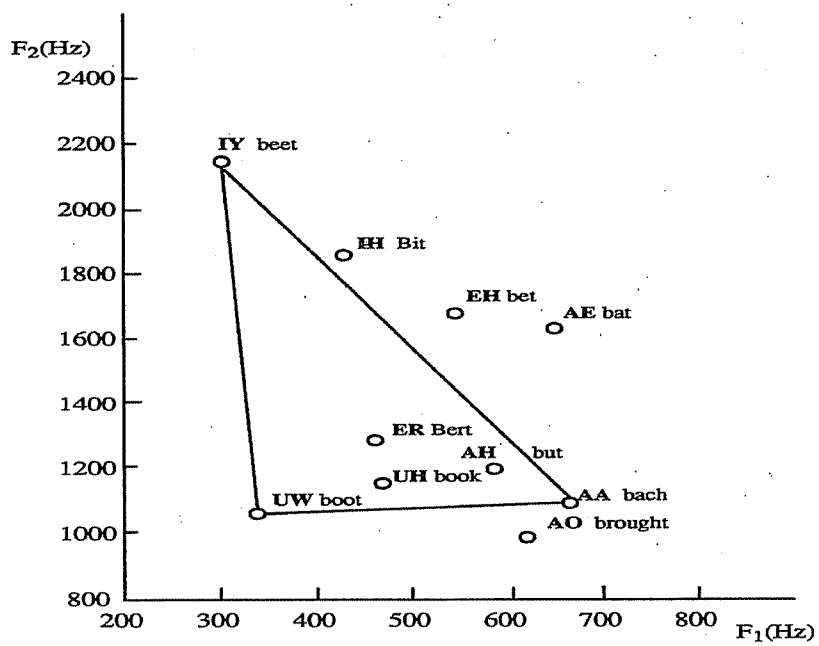


FIGURE 3.25 The vowel triangle for male speakers.

et al., 1992). According to Klatt, the inherent duration has no special significance, other than as a starting point for rule application. It is roughly the duration of the phoneme in non-sense consonant–vowel–consonant or consonant–consonant–consonant context, for example, b–vowel–b or c–consonant–b.

3.28 SUMMARY

Since this text attempts to stress the examination of aspects of speech using toolboxes, we outline certain features of speech data here to assist the reader in understanding characteristics of speech that can be useful in speech synthesis and recognition.

The excitation for speech can be voiced/unvoiced/mixed (V/U/M). Unvoiced excitation tends to be turbulent noise, caused by constrictions in the vocal tract. Unvoiced excitation tends to be high frequency with a high number of zero-crossings in the waveform and low amplitude (low energy). Voiced excitation is generally periodic in nature and high amplitude (high energy).

The vocal and nasal tracts introduce resonances causing the formants in the speech signal. The resonances can be modeled with an all-pole transfer function, as we will see in Chapter 5. When the velum is open, coupling the nasal tract to the oral tract, we have two all-pole models in parallel; thereby introducing zeros in the overall vocal tract model. So nasal sounds tend to introduce antiresonances caused by zeros. Nasal sounds tend to have formants with larger bandwidths than vowels because of the antiresonances.

Some features of the phonemes are summarized in Table 3.3. This summary is only indicative of the features and is not necessarily valid for all situations.

Additional features become apparent in certain applications. For example in word boundary identification tasks, it is difficult to determine when a word begins or ends with

- Weak fricatives (/F/, /TH/, /HH/).
- Weak plosives (/P/, /T/, /K/).
- Nasals (at the end of words).
- Voiced fricatives that die out (or become devoiced) at end of word.
- Trailing off of vowels at the end of a word.

In most speech analysis, we also assume we know the background noise level. This knowledge is useful for setting threshold levels for algorithm decision making.

Be sure that you read Appendices 4 and 6.

Additional references for the material presented in this chapter include Denes and Pinson (1993); Edwards (1992); Fant (1973); Olive, Greenwood, and Coleman (1993); Potter, Kopp, and Kopp (1966); Stevens (1998).

TABLE 3.3. A Summary of Some Phoneme Features

Sound	Energy	Excitation	Zero-Crossing	Formants	Duration
Vowels	High	V	Low	Yes	Long
Stops	Low	U and M	High	No	Short
Fricatives	Low	U and M	High	No	Short
Nasals	Medium	V	Low	Yes	Long

PROBLEMS

3.1 Read Appendix 4, which describes the data set provided with this book. Note that the data includes speech and electroglottographic files for speakers with normal larynges, a set of data for patients with vocal disorders, a set of data that mimics various voice types, an other data folder, and an additional data folder. The purpose of this problem is to examine the data for speakers m01 and f06 in the normal data folder. Calculate the first three formants for the twelve vowels in the data set, i.e., files m0103s.dat through m0114s.dat using the spectrogram and the Pitch/Jitter/Formant options. Repeat this task for subject f06. Do the results generally agree with the data in Figures 3.23 through 3.25? Do the results agree with the data in Appendix 5? Does the pitch for these two subjects agree with the data in Table 3.1?

3.2 Repeat Problem 3.1 for the fricatives, files m0115s.dat through m0123s.dat and files f0615s.dat through f0623s.dat. Only the spectrogram can be used, since for fricatives there is no voicing or it is weak so that the Pitch/Jitter/Formant software gives an error. Do the features you determine generally agree with those summarized in Table 3.3 and in Appendix 5 for the fricatives? Is it generally true that the fricatives have no (or very weak) formants?

3.3 Since the vowels and fricatives for files m0103s.dat through m0123s.dat and f0603s.dat through f0623s.dat are steady waveforms with little variation, we can also examine the FFTs. Calculate an FFT for each file using a centrally located section of the file. Use a hamming window of duration 512. Do the results generally agree with the data in Figures 3.23 through 3.25? Do the results agree with the data in Appendix 5? Does the pitch for these two subjects agree with the data in Table 3.1? Compare your results from Problems 3.1 and 3.2. How can you estimate the pitch from the FFTs?

3.4 Use the Cepstrum option to determine the pitch contour for files m0125s.dat and f0625s.dat. You will have to record the pitch for each segment and plot the data. Use a hamming window length of 512 with 10% overlap. What is the average pitch for each speaker? Does this result agree with Table 3.1? Now use the Pitch/Jitter/Formant option and calculate the pitch contour again. Do the two pitch estimates generally agree?

3.5 For the beginning of the file m0125s.dat, i.e., the "We were" segment, use the autocorrelation option to calculate the pitch contour.

3.6 Design a pitch period estimation algorithm using the autocorrelation function. You do not have to implement this algorithm in MATLAB. However, write the algorithm in some detail and discuss those sections that might introduce errors in the computation.

3.7 The purpose of this problem is to compare the waveforms and spectrograms for the same subject but for different vocal conditions, i.e., normal (modal) voice (file m0125s.dat), vocal fry (creaky voice) (file m0305ms.dat), breathy voice (file m0405ms.dat), hoarse voice (file m0505ms.dat), computer voice (file m0605ms.dat), and rough voice (file m0705ms.dat). The latter five files are in the Mimic folder. Describe the major differences between the features of the waveforms and the spectrograms for these six different voice types. What is the major contributor to the differences between these voice types?

3.8 The purpose of this problem is to compare the jitter for the six voice types considered in Problem 3.7. The files to be compared are as follows: normal (modal) voice (file m0125s.dat), vocal fry (creaky voice) (file m0305ms.dat), breathy voice (file m0405ms.dat), hoarse voice (file m0505ms.dat), computer voice (file m0605ms.dat), and rough voice (file m0705ms.dat). The latter five files are in the Mimic folder. Calculate and compare the pitch/jitter/formant contours for these files. What are the jitter values for all of these voices? Which voice type has the least jitter and which the largest jitter? An average jitter greater than 5% to 7% is often considered a deviant voice by clinicians. Would any of these voices be considered deviant? Note that the jitter is plotted in Hz on the vertical scale just as is the pitch contour. To calculate the percent jitter, calculate the average jitter, divide by the average pitch, and multiply by 100.

3.9 The purpose of this problem is to illustrate the difficulty of identifying and labeling word boundaries for sentences with different phoneme content. The three files to be compared are m0125s.dat,

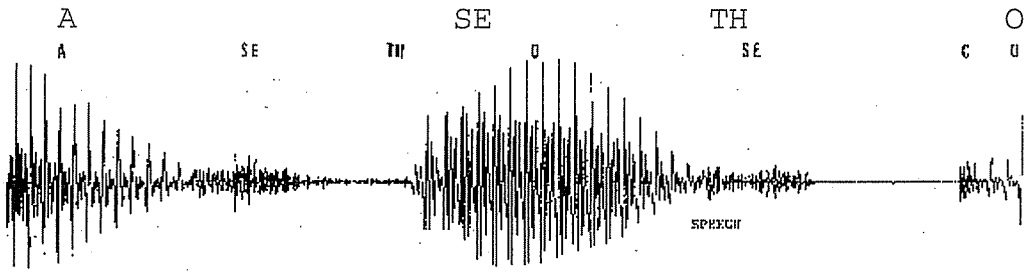


FIGURE P3.10 An illustration of the zoom option for phoneme labeling.

m01zs.dat (in the Other folder), and spe9_1.dat (in the Additional folder). The three sentences are "We were away a year ago" (all voiced), "That zany van is azure" (voiced fricatives), and "We saw the ten pink fish" (unvoiced plosives and fricatives), respectively. Use the Energy and ZCR option with a hamming window of length 256 and 10% overlap to make your first estimate of the word boundaries. Next calculate the spectrogram for each file and label the boundaries by hand. Finally, use the Zoom in and Play options to locate the word boundaries. Compare and discuss your results.

3.10 The purpose of this problem is to practice phoneme labeling for five sentences: "We were away a year ago" (all voiced) (m0125s.dat), "Early one morning a man and a woman ambled along a one mile lane" (many nasals) (m0126s.dat), "Should we chase those cowboys" (fricatives) (m0127s.dat), "That zany van is azure" (voiced fricatives) (m01zs.dat; Other folder), and "We saw the ten pink fish" (unvoiced plosives and fricatives) (spe9_1.dat; Additional folder). Use plots of the waveforms and spectrograms to identify and label the phoneme boundaries. The Zoom option and the Energy and ZCR and the pitch and formant contours may also be helpful. Do not use the Play option. Do this problem without the help of your ears. Make a list of the major features that you found useful for identifying and labeling the phonemes. When you have finished, compare your results with the results in Appendix 5. Discuss the possibility of designing an automatic phoneme recognition system. An illustration of how the zoom option can help phoneme labeling is shown in Figure P3.10. The segment shown is taken from the sentence "Should we chase those cowboys?"