

MIT OpenCourseWare
<http://ocw.mit.edu>

MAS.632 Conversational Computer Systems
Fall 2008

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.

2

Speech Production and Perception

A basic knowledge of the physiology of speech production and its perception is necessary to understand both speech synthesis and speech recognition as well as compression techniques. This chapter provides an overview of the production and perception of speech. It briefly treats the organs of speech, i.e., the vocal tract and the auditory system. An articulatory model of speech is presented; this explains the different speech sounds of a language in terms of how they are produced. Finally, it introduces some basic findings of psychoacoustics that relate to the manner in which sound is perceived by the listener.

To distinguish the various types of sounds in English and the mechanisms whereby they are produced some signal processing terms are introduced. In this text, very simple and intuitive descriptions of such terms are given; more formal definitions may be found in texts on signal processing.

VOCAL TRACT

The vocal tract is the set of organs that produce speech sounds; these organs are also used for eating, drinking, and breathing. As seen in Figure 2.1, the vocal tract includes portions of the throat, mouth, and nasal cavities. These organs, the **articulators**, are moved to various configurations to produce the different sounds that constitute speech. The primary topic of this section is the production of speech sounds in general; the next section classifies the sounds specific to English.

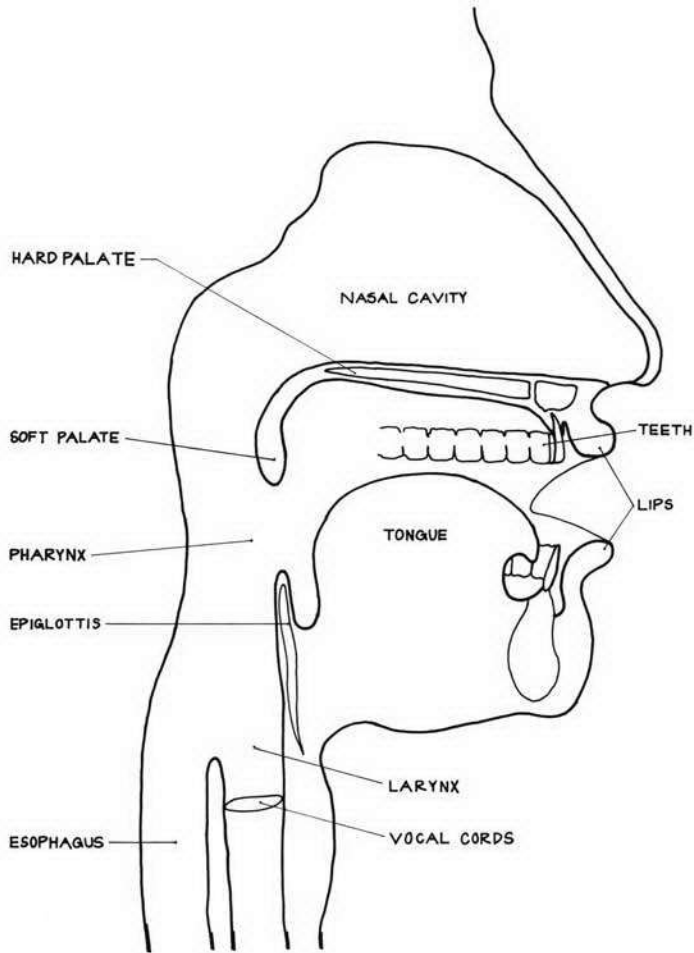


Figure 2.1. The major components of the vocal tract.

Sound is vibration in a medium, usually air, at frequencies that we can hear. In air, sound is carried as variations in pressure over time transmitted by the motion of air molecules. The air molecules create pressure variations by moving back and forth while not traveling a substantial distance. This variation is similar to the wave pattern generated when one end of a Slinky toy is moved back and forth. Some force must be created to cause this motion of molecules. In the vocal tract, this force is provided by air being expelled from the lungs by muscles in the diaphragm.

Without further modification, the air coming out of the lungs does not create a significant sound. If you open your mouth wide and exhale, you hear a little noise created as the air rushes out of the mouth, but this is usually not audible across a room. To become speech this air flow must be focused or concentrated in some

way. Focusing is accomplished by one of two mechanisms, either vibration of the vocal cords to produce a periodic sound or turbulent air flow through a constriction to produce an aperiodic, or “noisy,” sound.¹

If the vocal cords are vibrating, the speech produced is termed **voiced**. The vocal cords are folds of tissue capable of opening and closing in a regular fashion as controlled by several sets of muscles and driven by the pressure of air being expelled from the lungs. In operation the vocal cords remain closed until the pressure behind them is strong enough to force them open. As the air rushes by and pressure is released, a partial vacuum (Bernoulli effect) helps pull the cords back together again to shut off the flow of air.

This opening and closing pattern occurs at a regular rate and results in a series of pulses at that frequency. This frequency, called the **fundamental frequency** of voicing, or **F₀**, corresponds to the pitch that we perceive in speech. F₀ is typically higher for female and young speakers and may be varied within a limited range by muscle tension. The waveform of this glottal pulse is roughly triangular (see Figure 2.2), which results in a spectrum rich in harmonics, i.e., higher frequency components. The energy of speech drops at about 12 dB² per octave from the fundamental. Figure 2.3 illustrates the spectrum of the sound produced by the vocal chords. The **spectrum** is a measure of the magnitude of all the frequency components that constitute a signal. Since the source is of a single frequency F₀, energy is found at that frequency and at integral multiples of that frequency. For example, the author’s typical fundamental frequency is at 110 Hz,³ which means that energy will be found at 220 Hz, 330 Hz, 440 Hz, etc.

You can experience voicing directly for yourself. Place a finger at the front bottom of your throat, slightly above the level of the shoulders. Speak out loud, alternating making “s” and “z” sounds. During the voiced “z” you can feel the vocal cords vibrate with your finger.

Another source of sound in the vocal tract is turbulence as air rushes through a constriction. Turbulence is not periodic; it is caused by the continuous stream of air molecules bouncing around randomly. It is similar in nature to the sound made by a babbling brook or the wind blowing through trees or surf moving onto shore. Such sounds contain a relatively large amount of high frequency energy, or “hiss,” because the acoustic waveform varies a great deal from moment to moment. Turbulence is characteristic of both continuous **frication** sounds, such as the phoneme “s,” and sudden **plosive** ones associated with a quick opening of a closed vocal tract, as in the phoneme “p.” This turbulence can take place at a number of locations in the vocal tract. In the fricative “s,” it is at a constriction formed between the tongue and the roof of the mouth. With the plosive “p,” it is

¹A periodic signal repeats a pattern at a regular interval or frequency. Aperiodic signals do not exhibit such patterns. Wind instruments produce periodic sounds, while percussion instruments are generally aperiodic.

²1 dB, short for decibel, corresponds to a just noticeable difference in sound energy.

³Hz, or Hertz, means cycles (or periods) per second. 1 kHz, or one kiloHertz, is one thousand cycles per second. 1 MHz, or MegaHertz, is one million cycles per second.

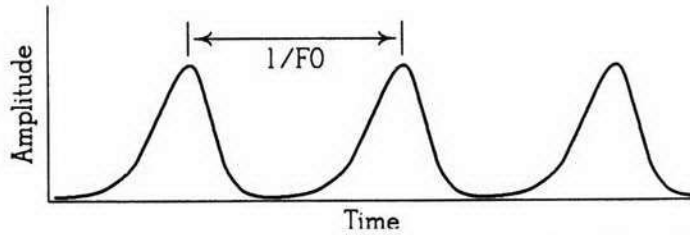


Figure 2.2. The glottal pulse is a roughly triangular-shaped waveform.

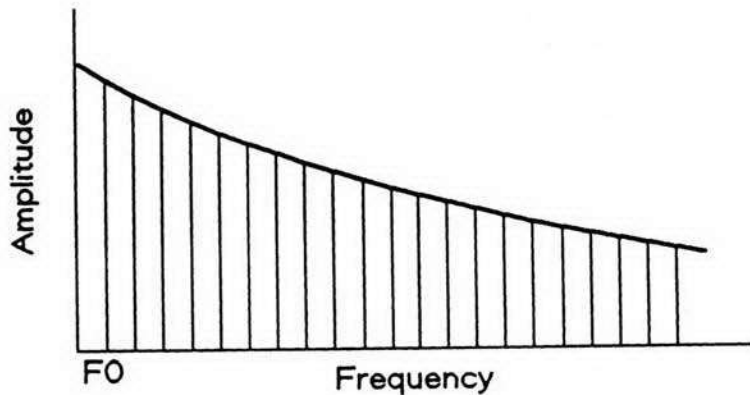


Figure 2.3. The spectrum of the sound produced by the vocal chords shows gradually decreasing magnitude with increasing frequency. This is properly a **line spectrum**, as energy is found only at multiples of F_0 , shown by the vertical line. The curve connecting the tops of these lines shows the **envelope**, or overall shape of the spectrum.

formed at the lips. The friction in “f” occurs around the lower lip making contact with the upper teeth.

There is more to classifying speech than identifying the source of the noise. The sound created by the noise source must then pass through the rest of the vocal tract that modifies it. These modifications are responsible for many of the distinguishing features of the different phonemes, or speech sounds. Even if the sound source is at the lips such that it does not have to pass through any further portions of the vocal tract, the size and shape of the oral cavity behind the lips affects the sound quality.

The vocal tract is usually analyzed as a series of **resonances**. A resonator is a physical system that enhances certain frequencies or range of frequencies. Resonators that amplify a very narrow range of frequencies are referred to as “high

Q",⁴ and those which have a very broad band are "low Q" (see Figure 2.4). Because the vocal tract is composed of soft tissue, its resonators are rather low Q. In contrast, a very "clear" sound, such as a flute note, is produced by a high Q resonator.

The vocal tract is about 17 cm long and can be approximated as a series of joined tubes of varying diameters and lengths that add up to 17. One tube extends from the vocal cords up to the top of the mouth, the other, at approximately a right angle to the first, extends from there to the lips. The relative lengths of these tubes can be varied by the positioning of the tongue. If the tongue is low and flat (as in "ah"), then the front tube formed by the mouth is pronounced and the back tube is shorter. When the tongue is high in the mouth (as when saying the letter "e"), the front tube in front of the tongue is quite short, while the back tube is elongated.

If a column of air in a tube is excited by a source, it resonates at a characteristic frequency dependent upon the length of the tube. This is the principle upon which the flute is designed; opening the finger holes allows the air to escape at various points along the tube, thereby changing the length of the air column and thus its frequency. In the same way, the sound produced from the vocal cord vibration is enhanced at the frequencies corresponding to the length of the tubes in the vocal tract (see Figure 2.5). The frequencies at which these resonances occur are called **formants**. The lowest-frequency formant is labeled **F1** and called the first formant; the second formant is labeled **F2**, and so on. Most analysis stops at five or six formants. In the multitube model, each formant corresponds to a tube section.

Whichever the noise source, either voiced or turbulent, its spectrum is enhanced around the formant frequencies, which depend on the physical configuration of the vocal tract. Another part of the vocal tract that influences the noise source is the **nasal cavity**, a fairly soft region above the roof of the mouth and behind the nose. The passage from the back of the mouth to the nasal cavity can be closed by the **velum**, a small flap of tissue toward the top back of the mouth. When the velum is opened, the nasal cavity is physically connected and acoustically coupled with the vocal tract. The nasal cavity absorbs a significant amount of sound energy in the lower portion of the energy spectrum.

The radiation of the speech signal from the vocal tract into the surrounding air must also be considered. In addition to the mouth, sound is radiated from the cheeks, throat, and bony structure of the head and jaw, resulting in a signal that is much less directional than it would be otherwise with radiational patterns that are heavily frequency dependent. Flanagan [Flanagan 1960] measured sound energies of various frequencies at multiple locations about the head and found that at approximately 45 degrees (horizontally) away from the axis of the mouth

⁴Q, or Quality factor, is a measure of the difference between the higher and lower bounds of the frequencies amplified by the resonator, as measured at the point of half-power relative to the peak power at the center frequency. This number is divided by the center frequency; i.e., it is a factor of frequencies rather than a difference.

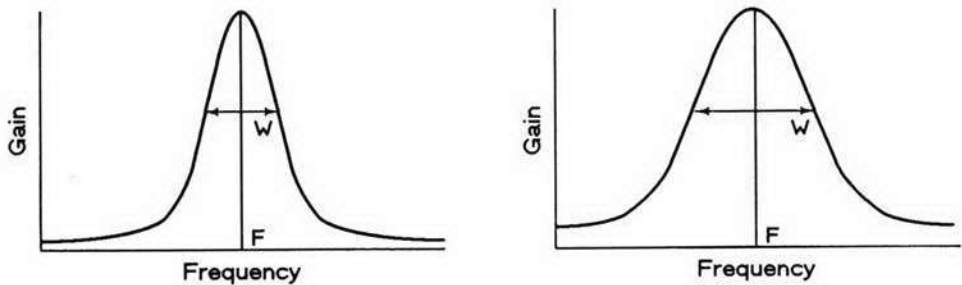


Figure 2.4. A high Q filter, left, and a low Q filter, right. F is the center frequency, and W is its bandwidth.

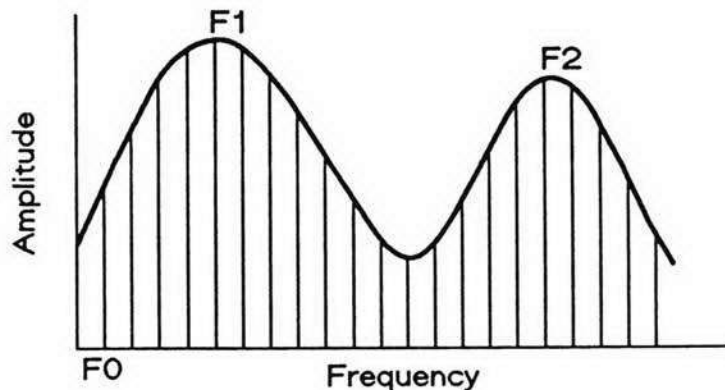


Figure 2.5. Resonances or formants of the vocal tract. The formants, labelled F1 and F2, are peaks in the vocal tract's filter function.

the signal was one half as strong, while directly behind the head the signal decreased by one order of magnitude less than in front of the head.

The noise source is referred to as the **source**, or **excitation**, signal. The manner in which the vocal tract enhances or diminishes the various frequencies of the signal as a function of its physical configuration is called its **transfer function**. A well-defined mathematical relationship allows computation of the signal that results from a source and a transfer function, but a detailed analysis of this relationship is beyond the scope of this book.

THE SPEECH SOUNDS

Each language employs a characteristic set of phonemes or allophones to convey words. These speech sounds are generated by manipulating the components of the vocal tract into specific physical configurations. The set of sounds comprising

the phonemes of a language must be distinctive because the listener perceives only the sound, not the vocal tract configuration that produces it. The classification of phonemes by the manner in which they are produced is referred to as an **articulatory model** of speech.

The three fundamental articulatory classifications are whether or not the phoneme is **voiced**, the **place** where the sound of the phoneme is made, and the **manner** in which it is made. The remainder of this section describes the various sounds of English according to their articulation.

Vowels

Vowels are sounds produced by vocal cord vibration (voicing) and a relatively open vocal tract, i.e., the lips, tongue, or teeth do not close the passageway. In the English alphabet there are only five letters we call vowels, but in spoken English there are closer to 17 vowel phonemes. Vowels are steady state; after we begin speaking a vowel, we can continue it until we run out of breath. Because vowels are voiced, they are periodic as can be seen in Figure 2.6.

The vowels can be differentiated by the acoustical effect of the position of the tongue and lips during their pronunciation. We can distinguish three degrees of constriction of the vocal tract that indicate how close to the roof of the mouth the hump of the tongue gets (high, medium, and low). We can also differentiate three positions laterally where the hump of the tongue makes the greatest constriction (front, central, and back). In addition, vowels can be *nasalized* by moving the velum to open the air passage leading from the throat to the nose. English does not employ nasalization as a phonemic distinction; however, certain other languages such as French use nasalization as a distinguishing feature to contrast phonemes.

In English the vowels are distinguished solely by the position of the tongue, which changes the lengths of the two primary vocal tract tube sections. In other words, the first two formants form a set of distinctive features for the vowels. Although the typical positions of the formants will vary with gender and among speakers, they can generally be clustered together into regions in a space defined

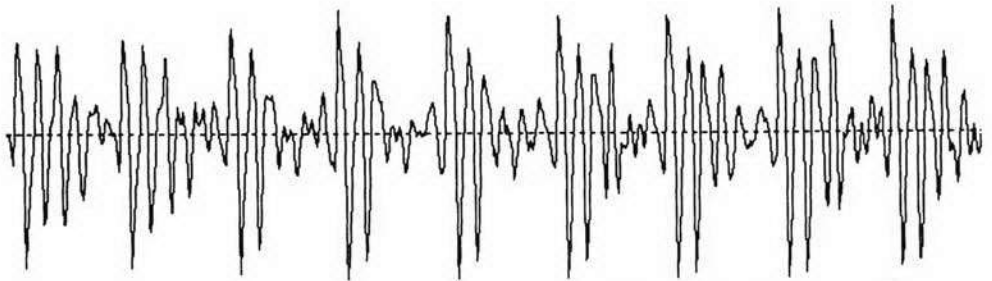


Figure 2.6. A waveform of the vowel in “but.” The display shows amplitude mapped against time, over a 100 millisecond interval. Nine pitch periods can be seen; the frequency of voicing is about 90 Hz.

by F1 and F2 as illustrated in Figure 2.7. Such a diagram is sometimes called the “vowel triangle” because all of the vowels fit in a space that approximates a triangle in the chosen coordinate system of F1 versus F2.

Some vowels are **diphthongs**. Diphthongs are actually combinations of two vowel sounds with the vocal tract moving from one position to another during their production. The vowel sounds in “bay” and “boy” are examples of diphthongs.

Consonants

Consonants are characterized by constrictions in the vocal tract and can be differentiated by place of closure, manner (type or degree) of closure, and whether they are voiced.

Place refers to the location in the vocal tract of the closure associated with a consonant. **Labial** closure uses both lips as in “p” and “b” sounds. **Labial-dental** closure involves the lower lip and upper teeth as in “f” and “v” sounds. **Alveolar** closure involves the tongue and the gum ridge behind the front teeth as in “n,” “d,” “s,” and “z” sounds. **Palatal** closure uses the tongue on the soft palate or roof of the mouth slightly farther back as in “sh” and “zh” sounds.

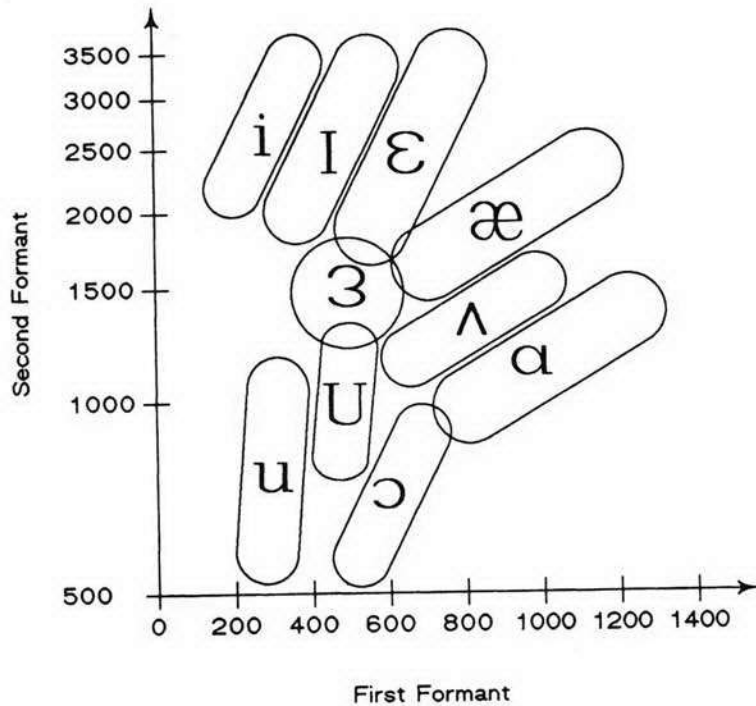


Figure 2.7. The vowel triangle displays English vowels mapped onto a Cartesian space defined by the the first and second formants. The IPA notation from Figure 1.4 is used for labeling.

Velar closure occurs at the back of the mouth, or hard palate, as in “g” and “k” sounds. Each of these places of closure results in a distinct sound as the shape of the vocal tract changes and consequently the frequencies of resonance vary. The closure divides the vocal tract into cavities in front of and behind the closure; these have different resonant frequencies depending on their size. For example, the position of the third formant is often a strong cue to the identity of a stop consonant.

Closure of the vocal tract is achieved in several different manners, each resulting in a different sound. **Stops** (“t,” “b,” “p,” “d”) involve a sudden and total cessation of air flow. Stop consonants are very dynamic. During the closure, no air flows through the vocal tract, resulting in silence. Noise resumes suddenly as the closure is released, and the air pressure from behind the closure may result in **aspiration** noise, such as in the “t” in “top.” Stops are also called **plosives**, which focuses on the release rather than the closure of the consonant.

Fricatives (“s,” “z,” “sh,” “zh”) involve constriction to the point of producing turbulence and hence noise but not total closure. The partial closure for “s” is alveolar and that for “sh” is palatal. Because closure is incomplete the fricatives result in a continuous sound as air flows by. Sound created by air rushing through a small opening is aperiodic (see Figure 2.8) and dominated by high-frequency components. Combining these two traits, the characteristic acoustic feature of the fricatives is a moderate duration of high-frequency energy.

Nasals (“m,” “n,” “ng” as in “sing”) are produced by closing the oral cavity but opening the velum to the nasal cavity. The nasal cavity absorbs a significant amount of low-frequency energy giving a distinctive cue to nasalization. The three English nasals are all voiced. The nasals are differentiated from each other by the place at which the vocal tract is closed, which is either labial (“m”), palatal (“n”), or velar (“ng”).

Most of the consonant types mentioned so far have come in pairs, such as (“d,” “t”) and (“v,” “f”). Both members of each of these pairs is produced with the articulators in the same place but are distinguished by the presence or absence of

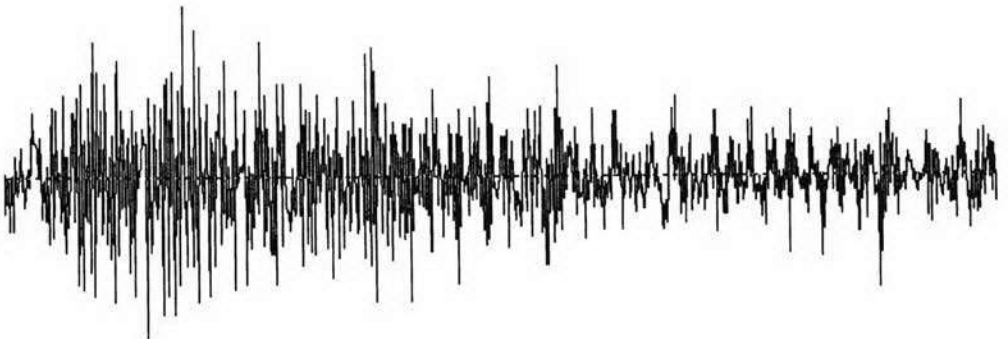


Figure 2.8. The aperiodic waveform of the consonant in “see.” This figure also shows 100 milliseconds of speech. No periodicity is evident.

voicing. Voiced consonants thus have two noise sources: one at the place of articulation and the other at the periodic glottal pulses.

Whether or not a consonant is voiced depends on the type of consonant. In a stop there is no air flow so a stop is considered voiced if voicing resumes soon after the release; otherwise it is unvoiced. Air flow continues during a fricative so voicing is defined by whether the vocal cords are vibrating during the consonant itself.

Liquids and Glides

A small class of special-case phonemes behave much like vowels. The **liquids** ("l," "r") invariably precede a vowel and are dynamic, involving a transition into the vowel. It is sometimes hard to identify "r" as a distinct consonant but rather to observe the effect, or **coloration**, of the vowel caused by the motion of the lips and the rise of the tongue in the middle of the mouth. The **glides** ("w," "y" as in "you") are similar to vowels except that the vocal tract is more constricted than for the vowels. According to another point of view, the glides are simply vowels that do not occur during the energy peak of the syllable.

Acoustic Features of Phonemes

The methods used to articulate the various phonemes of a language must produce sounds with adequate acoustical cues to allow the listener to distinguish between possible phonemes in speech. Some examples of acoustical cues have just been mentioned: the distinctive structure of the first two formants in vowels, the energy at high frequencies that distinguish fricatives, the presence or absence of periodicity due to voicing, and the absorbing effect of the nasal cavity.

The listener can select from a multitude of cues occurring simultaneously to differentiate phonemes. In fluent speech the articulators often move only partially in the direction of target vocal tract configurations described above for each phoneme. The acoustical effect is still sufficiently pronounced to allow the listener to detect at least some of the cues and identify each phoneme.

HEARING

Sound arrives at our ears as variations in air pressure. We can hear vibrations in the range of approximately 20 Hz up to 15 or 20 kHz; this figure varies among individuals, and the range decreases as we age. This sensory stimulus triggers neural activity through a complex but versatile mechanical transformation in the ear. There is a chain of processes whereby the physical sound from its source causes the auditory event of "hearing." In turn, these neural firings stimulate perceptual processing at higher levels of the brain. Although little is known about such processing from direct neural evidence, the domain of **psychoacoustics** studies our perception of sound by observing subjects' responses to acoustic stimuli.

Auditory System

The complement to the vocal tract is the auditory system. Although primarily designed to turn air pressure variations into corresponding neural signals, the ear also contains the **vestibular organs** that are used to maintain physical balance; we are not concerned with them here. The ear can be divided into three distinct sections. The outer ear directs sound toward the eardrum, the middle ear converts the pressure variations of sound into mechanical motion, and the inner ear converts this motion to electrical signals in the auditory neurons. The ear is shown in Figure 2.9.

The **outer ear** includes the **pinna** and the **ear canal**, which leads to the eardrum. The pinna consists of the fleshy protrusion on the side of the head and is what we usually refer to when we use the term “ear” (as in “Van Gogh cut off his ear”). The pinna with its various folds of cartilage around the ear opening serves primarily as a protective mechanism. The pinna provides some amplification of the sound by focusing it into the ear canal in much the same way that we may cup our hands behind our ears to better hear quiet sounds. It is directional at high frequencies and is used as a localization aid to find a sound source because it makes the ear more sensitive to sounds coming from in front of rather than behind the listener. In certain animals such as owls, the outer ear occupies a much larger surface area with respect to the head and is more fundamental to localization.

The ear canal is a tube about 1 cm wide by 2.5 cm long leading to the middle ear. It has a resonance of about 3000 Hz and therefore amplifies sound at this frequency. The length of the ear canal shields the middle ear from physical injury if the side of the head is struck.

The **middle ear** provides the linkage between the outer and inner ear; its function is to effectively convert variations in air pressure to mechanical motion of the liquid inside the inner ear. The **eardrum**, or **tympanic membrane**, covers the interior end of the ear canal. Vibrations in the air cause the eardrum to vibrate; thus it produces physical motion from the sound in the air. The middle ear is filled with air. The **eustachean tube** runs between the middle ear and the throat. When we yawn, the eustachean tube opens, allowing the air pressure to equalize across the eardrum. As experienced when descending rapidly in an airplane, this pressure balance is essential for effective transfer of sound across the eardrum.

The inner ear is filled with water. There is a large impedance mismatch between air- and water-like fluids. Water is much denser than air so air does not cause significant displacement in the liquid that it impinges; even a strong wind on a pond does not penetrate much deeper than to form surface ripples because of this mismatch. A series of three **ossicular bones** (the **malleus**, **incus**, and **stapes**) provide mechanical advantage (leverage) from the ear drum to the **oval window**, a membrane on the surface of the inner ear. As the eardrum is 17 times as large as the oval window, this difference in area provides further amplification. The size difference plus the mechanical advantage of the ossicular bones combine to provide a 22:1 amplification and thereby an impedance match for efficient transfer of acoustical energy to the inner ear.

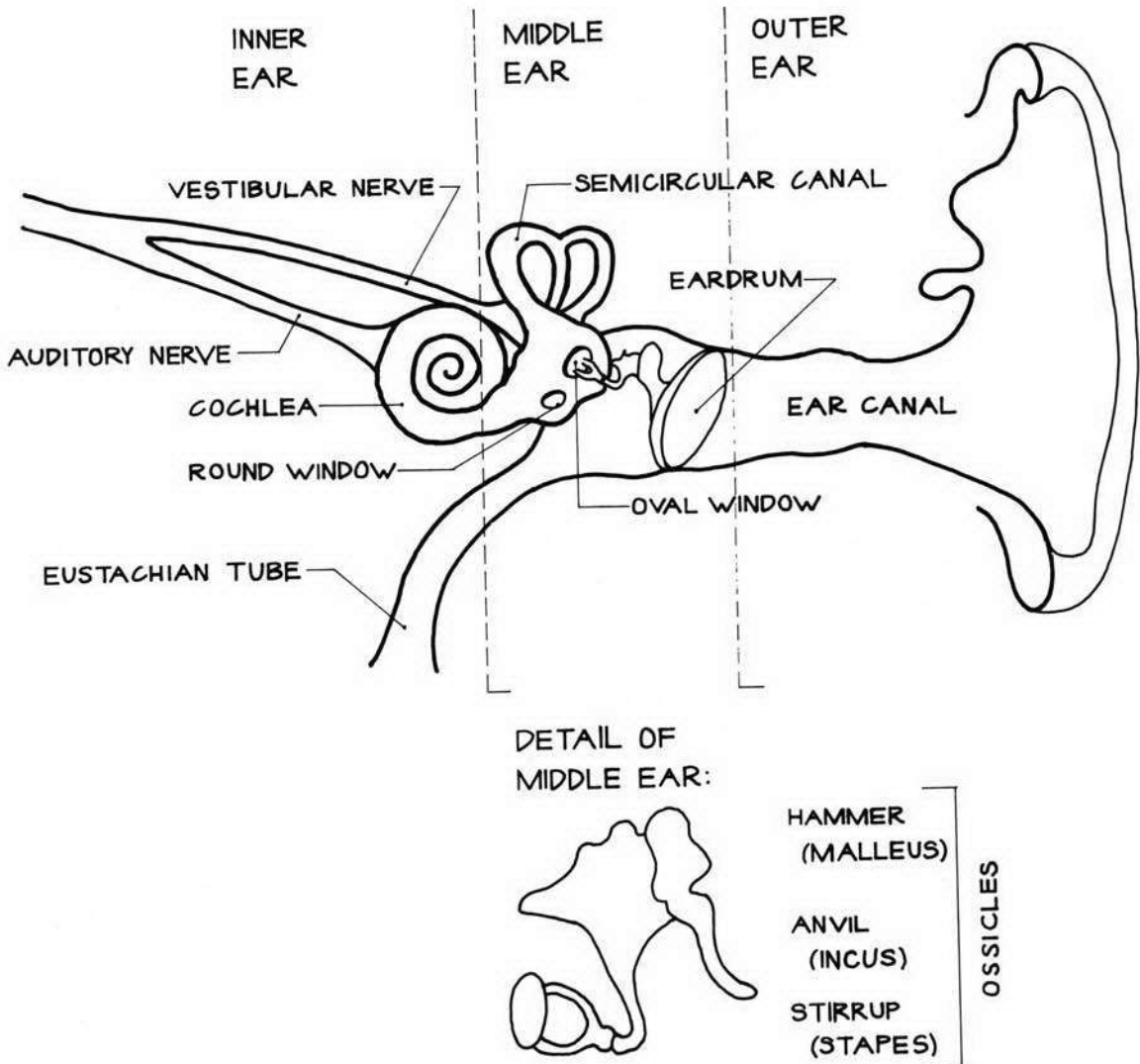


Figure 2.9. The ear.

The middle ear has several different modes of vibration that provide improved dynamic range without distortion (the goal of all good audio amplifiers). For loud sounds, the vibration mode is changed from pumping to a rotational action. In addition, for very loud sounds a reflex occurs in the muscles of the middle ear to damp the vibration and thereby protect the sensitive inner ear.

The **inner ear** is the transducer from mechanical to electrical energy. At its center is the **cochlea**, a spiral-shaped chamber looking somewhat like a snail shell and making $2\frac{1}{2}$ turns. Sound-induced waves in the liquid medium of the cochlea cause vibration on the **basilar membrane**. Different portions of this tapering membrane vibrate in response to sound energy at specific frequencies. At the basal end it is thin and stiff, while at its apex it is flexible and massive.

The frequency dependent vibration of the basilar membrane induces motion in the microscopic hairs that penetrate the membrane. These hairs are linked to neurons to the brain that produce neural firings in response to the stimulus of the hairs' bending. Each neuron is connected to a small number of hairs in a particular location on the basilar membrane, and the basilar membrane responds to different frequencies along its length so the firing of a neuron corresponds to the presence of sound of the appropriate frequency. Because of this relationship between neural firings and basilar membrane vibration, an individual neuron is not very responsive to frequencies much higher or lower than its preferred one. Measurements of neural responses indicate that each acts as a bandpass filter and that all the neurons and associated hair cells exhibit a fairly constant Q.

Neural activity is highest at the onset of a sound and decays rapidly at first and then more slowly. This is known as **adaptation**; we are more sensitive to changing sounds than continual tones. Neurons fire more rapidly for louder sounds, reaching peak rates of up to 1000 firings per second for short periods. At frequencies below about 1 kHz, neurons tend to fire in phase with the vibration of the basilar membrane, i.e., every cycle of membrane motion induces one neural spike.

The ear's response to a sound is a series of neural spikes from many neurons simultaneously. The firing pattern of any particular neuron is a function of amplitude of the acoustic energy (at the basilar membrane) within the frequency range to which that neuron is sensitive. The pattern adapts over a short period of time and may be at least partially in phase with the acoustic signal. The neural firings are transmitted to a number of stages of processing in the central auditory system. At several points along the way are sites that are particularly sensitive to time or amplitude differences between signals arriving from each ear. Such differences are central to our ability to locate a sound spatially.

Localization of Sounds

When we hear a sound, we perceive it as coming from some location in space outside of our head; this is known as **localization**. Much of our ability to localize sound depends on the differences in the sound that arrives at each of our two ears from a single source. These differences are due to the positions of the ears and hence their different distances from the sound source as well as the tendency of the head to partially block sound coming to an ear from the opposite side of the head.

We can localize a sound source within about 4 degrees in front of the head, 6 degrees behind the head, and only within 10 degrees at the sides. Localization is very frequency dependent and we are most sensitive to sounds at around 800 Hz.

Our hearing is about 100 times less sensitive to position than our visual system, but we can hear sounds behind our heads. We have a strong reflex to turn and look at a sound that occurs behind our heads. This moves the source of the sound into our field of view, which could aid survival, and places the source where we can better localize it. The motion of the head also provides cues to enhance localization as we seem to be especially sensitive to differences in location.

If a sound source is off to either side of our head, it takes the sound longer to reach the ear on the other side as it is further away (see Figure 2.10). In other words, when a sound begins one ear hears it before the other. But more importantly, while the sound continues the air vibrations will arrive at each ear out of phase with the other ear; this is defined as the **interaural phase difference**. The human head has a thickness of 21 or 22 centimeters, which means that it can take sound up to about 600 microseconds longer to reach the further ear. This phase difference becomes confused when it is greater than the wavelength of the sound so phase is most effective for localizing sound below 1500 Hz.

When a sound is off to one side, the mass of the head also blocks the direct path from the sound source to the opposite ear. This results in an **interaural intensity difference**; the sound is louder at the ear that has an unobscured path. This effect is most pronounced at frequencies above 10 kHz, as the head more effectively blocks small wavelengths.

The differences in phase and intensity of the sound arriving at each ear can help localize it to the left or right side but leave confusion as to whether the sound is in front of or behind the head. A sound in front of and to the left of the listener creates the same interaural phase and intensity differences as a sound the same distance behind and also to the left. The shape of the pinna interacts with incom-

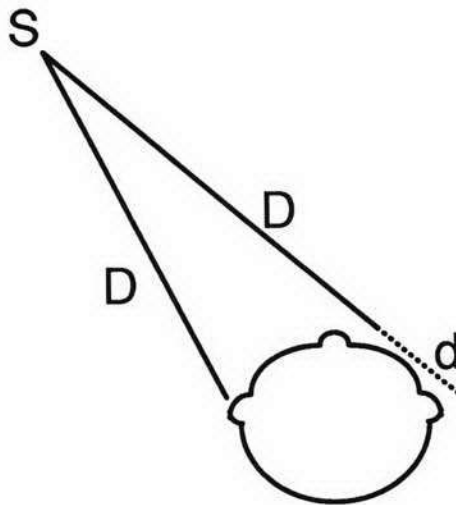


Figure 2.10. A sound on the left side of the head arrives at the right ear later than the left ear because it must travel further.

ing sound and provides front-back cues. Its amplification effect provides directionality through increased gain for sounds from the front. In addition, the folds of the pinna act as a complex filter enhancing some frequencies; this filtering is different for sounds from the front than from the back providing further front/back cues.

Psychoacoustics

So far we have described the process whereby sound energy is converted to neural signals; we have not yet described how the listener perceives the sound characterized by such signals. Although more is known about higher-level processing of sound in the brain than we have mentioned in the brief overview above, there is still not sufficient knowledge to describe a model of such higher auditory processing. However, researchers can build an understanding of how we perceive sounds by presenting acoustical stimuli to subjects and observing their responses. The field of psychoacoustics attempts to quantify perception of pitch, loudness, and position of a sound source.

The ear responds to a wide but limited range of sound loudness. Below a certain level, sound cannot be perceived; above a much higher level, sound causes pain and eventually damage to the ear. Loudness is measured as **sound pressure level** (SPL) in units of **decibels** (dB). The decibel is a logarithmic scale; for power or energy, a signal that is X times greater than another is represented as $10 \log_{10} X$ dB. A sound twice as loud is 3 dB higher, a sound 10 times as loud is 10 dB higher, and a sound 100 times as loud is 20 dB higher than the other. The reference level for SPL is 0.0002 dyne/cm^2 ; this corresponds to 0 dB.

The ear is sensitive to frequencies of less than 20 Hz to approximately 18 kHz with variations in individuals and decreased sensitivity with age. We are most sensitive to frequencies in the range of 1 kHz to 5 kHz. At these frequencies, the threshold of hearing is 0 dB and the threshold of physical feeling in the ear is about 120 dB with a comfortable hearing range of about 100 dB. In other words, the loudest sounds that we hear may have ten billion times the energy as the quietest sounds. Most speech that we listen to is in the range of 30 to 80 dB SPL.

Perceived loudness is rather nonlinear with respect to frequency. A relationship known as the Fletcher-Munson curves [Fletcher and Munson 1933] maps equal input energy against frequency in equal loudness contours. These curves are rather flat around 1 kHz (which is usually used as a reference frequency); in this range perceived loudness is independent of frequency. Our sensitivity drops quickly below 500 Hz and above about 3000 Hz; thus it is natural for most speech information contributing to intelligibility to be within this frequency range.

The temporal resolution of our hearing is crucial to understanding speech perception. Brief sounds must be separated by several milliseconds in order to be distinguished, but in order to discern their order about 17 milliseconds difference is necessary. Due to the firing patterns of neurons at initial excitation, detecting details of multiple acoustic events requires about 20 milliseconds of averaging. To note the order in a series of short sounds, each sound must be between 100 and 200 milliseconds long, although this number decreases when the sounds have

gradual onsets and offsets.⁵ The prevalence of such transitions in speech may explain how we can perceive 10 to 15 phonemes per second in ordinary speech.

Pitch is the perceived frequency of a sound, which usually corresponds closely to the fundamental frequency, or F_0 , of a sound with a simple periodic excitation. We are sensitive to *relative* differences of pitch rather than absolute differences. A 5 Hz difference is much more significant with respect to a 100 Hz signal, than to a 1 kHz signal. An **octave** is a doubling of frequency. We are equally sensitive to octave differences in frequencies regardless of the absolute frequency; this means that the perceptual difference between 100 Hz and 200 Hz is the same as between 200 Hz and 400 Hz or between 400 Hz and 800 Hz.

The occurrence of a second sound interfering with an initial sound is known as **masking**. When listening to a steady sound at a particular frequency, the listener is not able to perceive the addition of a lower energy second sound at nearly the same frequency. These frequencies need not be identical; a sound of a given frequency interferes with sounds of similar frequencies over a range called a **critical band**. Measuring critical bands reveals how our sensitivity to pitch varies over the range of perceptible frequencies. A pitch scale of **barks** measures frequencies in units of critical bands. A similar scale, the **mel** scale, is approximately linear below 1 kHz and logarithmic above this.

Sounds need not be presented simultaneously to exhibit masking behavior. **Temporal masking** occurs when a sound masks the sound succeeding it (**forward** masking) or the sound preceding it (**backward** masking). Masking influences the perception of phonemes in a sequence as a large amount of energy in a particular frequency band of one phoneme will mask perception of energy in that band for the neighboring phonemes.

SUMMARY

This chapter has introduced the basic mechanism for production and perception of speech sounds. Speech is produced by sound originating at various locations in the vocal tract and is modified by the configuration of the remainder of the vocal tract through which the sound is transmitted. The articulators, the organs of speech, move so as to make the range of sounds reflected by the various classes of phonemes.

Vowels are steady-state sounds produced by an unobstructed vocal tract and vibrating vocal cords. They can be distinguished by their first two resonances, or formants, the frequencies of which are determined primarily by the position of the tongue. Consonants are usually more dynamic; they can be characterized according to the place and manner of articulation. The positions of the articula-

⁵The onset is the time during which the sound is just starting up from silence to its steady state form. The offset is the converse phenomenon as the sound trails off to silence. Many sounds, especially those of speech, start up and later decay over at least several pitch periods.

tors during consonant production create spectra characteristic of each phoneme; each phoneme has multiple cues as to its identity.

Sound consists of vibrations in a medium (usually air) at a frequency we can hear. The variations in sound pressure are focused by the outer ear and cause mechanical motion in the middle ear. This mechanical motion results in waves in the liquid in the inner ear, which causes neurons to fire in response. Which neurons fire and the pattern of their firings is dependent on the spectral and temporal characteristics of the sound. We localize a sound based on interaural phase and intensity differences. Each ear receives a slightly different sound signal due to the position of the ears on either side of the head. The pinna enables us to distinguish between sounds in front and behind.

This chapter also discussed several topics in psychoacoustics and the frequency and temporal responses of the auditory system; what we perceive with our sense of hearing is as much due to the characteristics of our auditory system as to the qualities of the sounds themselves. These factors influence our perception of speech in particular as well as all sounds.

The next chapter explores methods to digitally capture, encode, and compress the speech signal for computer storage, analysis, or transmission. Data compression techniques for speech take advantage of temporal and frequency characteristics of the speech signal as well as the sensitivity of the auditory system to reduce the data rate without loss of intelligibility or perceived quality of the reconstructed signal.

FURTHER READING

A concise overview of the human speech system can be found in Denes and Pinson and a more analytic and rigorous one in Flanagan (1972). Ladefoged offers an excellent treatment of phonetics. O'Shaughnessy covers many issues in speech production and hearing as a prelude to describing speech technologies. Handel offers an excellent overview of the perception of speech and music, while Yost is a more introductory text with superb illustrations.