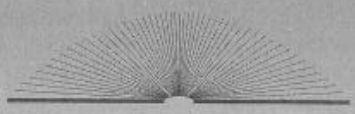
12



SPEECH PERCEPTION[1]

CHAPTER CONTENTS

SOME QUESTIONS WE WILL CONSIDER

- Can computers perceive speech as well as humans? (410)
- Why does an unfamiliar foreign language often sound like a continuous stream of sound, with no breaks between words? (413)
- Does each word that we hear have a characteristic pattern of air pressure changes associated with it? (414)

Speech sounds are, like other sounds, a disturbance of the air. But speech sounds, unlike other sounds, are produced as air is pushed up from the lungs and is shaped into patterns of air pressure changes by actions of the various structures in the vocal tract (Figure 12.1).

Although we perceive speech easily under most conditions, beneath this ease lurks processes as complex as those involved in perceiving the most complicated visual scenes. One way to appreciate this complexity is to consider attempts to use computers to recognize speech. After decades of research into

[1] This chapter is dedicated to the memory of Kerry Green, speech researcher and friend.

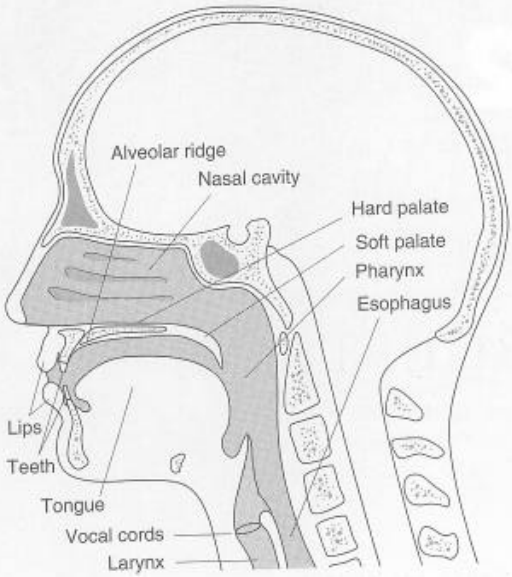


Figure 12.1
The vocal tract includes the nasal and oral cavities and the pharynx, as well as components that move, such as the tongue, lips, and vocal cords.

computer speech recognition, useful computer speech recognition systems are only now becoming available. The phone company uses computer speech recognition to identify simple messages such as telephone numbers or phrases such as "I want to make a credit card call," and there are software programs that make it possible for personal computers to respond to spoken commands (Markowitz, 1996).

There are dictation machines that can translate speech that is spoken clearly and in a quiet environment into printed text, but these programs must be "trained" to recognize the voices and pronunciations of individual speakers, and even after this training they are still prone to error. For example, the 2000 version of a popular speech recognition program translated "Friends, Romans, countrymen, lend me your ears" into "Friends, Romans, countryman, linear years."

In contrast to computers, humans can perceive speech under a wide variety of conditions, including the presence of various background noises, sloppy pronunciation, speakers with different accents, and the often chaotic give-and-take that routinely occurs when people talk with one another.

This chapter will help you appreciate the complex perceptual problems posed by speech and will describe research that has helped us begin to understand how the human speech perception system has solved some of these problems. We begin by describing the nature of the vibrations produced by our vocal apparatus.

The Speech Stimulus

Think about what you hear when someone speaks to you. You perceive a series of sounds called syllables, which create words, and these syllables and words appear strung together one after another like beads on a string. For example, we perceive the phrase "perception is easy" as the sequence of units: "per-sep-shon-iz-ee-z." But although our perception of speech may be easy and the sounds we hear may appear to be discrete sounds that are lined up one after another, the actual situation is quite different.

Rather than following each other with one unit of sound ending and then the other beginning, neighboring sounds overlap one another. In addition, the pattern of air pressure changes for a particular word can vary greatly depending on whether the speaker is male or female, is young or old, speaks rapidly or slowly, or has an accent. To understand the way speech sounds overlap and why the same sound can be represented by many different patterns of pressure changes, we need to describe the speech stimulus and how it is produced. We will do this in two ways: (1) in terms of short segments of sound, called phonemes; and (2) in terms of the patterns of frequencies and intensities of the pressure changes in the air, called the acoustic signal.

Phonemes: Sounds and Meanings

Our first task in studying speech perception is to separate speech sounds into manageable units. What are these units? The flow of a sentence? A part

A syllable? The sound of a letter? A sentence is ...rge a unit for easy analysis, and some letters ...no sounds at all. Although there are arguments ...the idea that the syllable is the basic unit of speech (Mehler, 1981; Segui, 1984), most speech research has been based on a unit called the **phoneme**. The phoneme is the shortest segment of speech that, if changed, changes the meaning of a word. Consider the word *bit*, which contains the phonemes /b/, /i/, and /t/. (Phonemes and other speech sounds are indicated by setting them off with slashes.) We know that /b/, /i/, and /t/ are phonemes, because we can change the meaning of the word by changing each phoneme individually. Thus, *bit* becomes *pit* if /b/ is changed to /p/, it becomes *bat* if /i/ is changed to /a/, and it becomes *bid* if /t/ is changed to /d/.

The phonemes of English, listed in Table 12.1, are represented by phonetic symbols that stand for speech sounds: 13 phonemes have vowel sounds, and 24 phonemes have consonant sounds. Your first reaction to this table may be that there are more vowels than the standard set you learned in grade school (a, e, i, o, and u). The reason is that some vowels can have more than one pronunciation, so there are more vowel sounds than vowel letters. For example, the vowel o sounds different in *boat* and *hot*, and the vowel e sounds different in *head* and *heed*. Phonemes, therefore, refer not to letters but to speech sounds that serve to distinguish meaning.

Table 12.1
Major consonants and vowels of English and their phonetic symbols

Consonants				Vowels	
p	*p*ull	s	*s*ip	i	h*ee*d
b	*b*ull	z	*z*ip	I	h*i*d
m	*m*an	r	*r*ip	e	b*a*it
w	*w*ill	š	*sh*ould	ɛ	h*ea*d
f	*f*ill	ž	plea*s*ure	æ	h*a*d
v	*v*et	č	*ch*op	u	wh*o*'d
θ	*th*igh	ǰ	*g*yp	U	p*u*t
ð	*th*y	y	*y*ip	ʌ	b*u*t
t	*t*ie	k	*k*ale	o	b*oa*t
d	*d*ie	g	*g*ale	ɔ	b*ough*t
n	*n*ear	h	*h*ail	a	h*o*t
l	*l*ear	ŋ	si*ng*	ə	sof*a*
				ɨ	*m*any

Because different languages use different sounds, the number of phonemes varies in different languages. While there are only 11 phonemes in Hawaiian, there are 47 in English, and as many as 60 in some African dialects. Thus, phonemes are defined in terms of the sounds that create meaning in a specific language. Each phoneme is produced by the position or the movement of structures within the vocal apparatus, which produce patterns of pressure changes in the air which are called the **acoustic stimulus** or the **acoustic signal**.

The Acoustic Signal: Patterns of Pressure Changes

The acoustic signal for speech is created by air that is pushed up from the lungs past the vocal cords and into the vocal tract. The sound that is produced depends on the shape of the vocal tract as air is pushed through it. The shape of the vocal tract is altered by moving the **articulators**, which include structures such as the tongue, lips, teeth, jaw, and soft palate (Figure 12.1).

Let's first consider the production of vowels. Vowels are produced by vibration of the vocal cords, and the specific sounds of each vowel are created by changing the overall shape of the vocal tract. This change in shape changes the resonant frequency of the vocal tract and produces peaks of pressure change at a number of different frequencies (Figure 12.2). The frequencies at which these peaks occur are called **formants**.

Each vowel sound has a characteristic series of formants. The first formant has the lowest frequency. The second formant is the next highest, and so on. The formants for the vowel /ae/ are shown on a display called a **sound spectrogram** in Figure 12.3. The sound spectrogram indicates the pattern of

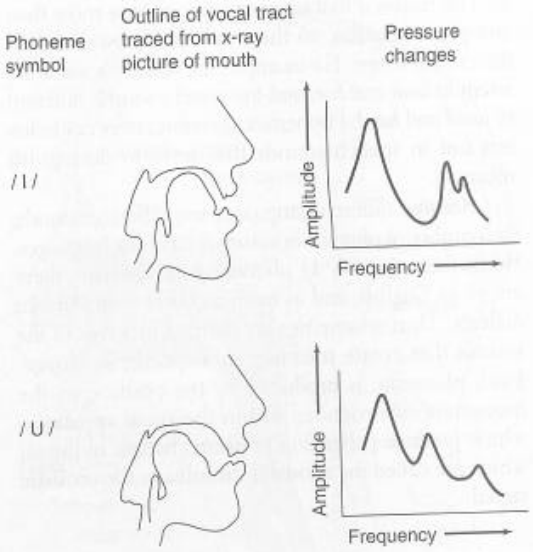


Figure 12.2
Left: The shape of the vocal tract for the vowels /I/ and /oo/. Right: The amplitude of the pressure changes produced for each vowel. The peaks in the pressure changes are the formants. Each vowel sound has a characteristic pattern of formants which is determined by the shape of the vocal tract for that vowel. (Adapted from Denes & Pinson, 1992.)

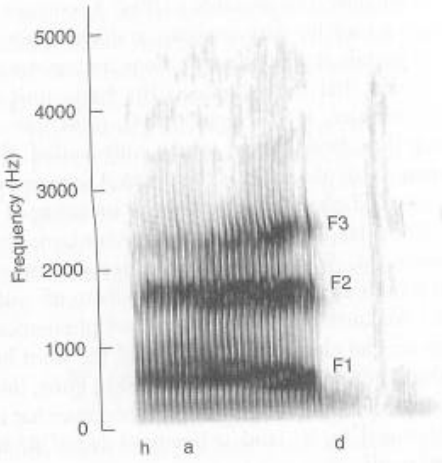


Figure 12.3
Spectrogram of the word had showing the first (F1), second (F2), and third (F3) formants for the vowel /ae/. (Spectrogram courtesy of Kerry Green.)

frequencies and intensities over time that make up the acoustic signal. Frequency is indicated on the vertical axis, time on the horizontal axis, and intensity is indicated by the darkness, with more darkness indicating greater intensity. From Figure 12.3 we can see that /ae/ has formants at 500, 1,700, and 2,500 Hz. The vertical lines in the spectrogram are pressure oscillations caused by vibrations of the vocal cord.

Consonants are produced by a constriction or closing of the vocal tract. To illustrate how different consonants are produced, let's focus on the sounds /d/ and /f/. Make these sounds and notice what your tongue, lips, and teeth are doing. As you produce the sound /d/, you place your tongue against the ridge above your upper teeth (the alveolar ridge of Figure 12.1) and then release a slight rush of air as you move your tongue away from the alveolar ridge (try it). As you produce the sound /f/, you place your bottom lip against your upper front teeth and then push air between the lip and the teeth.

These movements of the tongue, lips, and other articulators create patterns of energy in the acoustic signal that we can observe on the sound spectrogram. For example, the spectrogram for the sentence "Rob read the will" shown in Figure 12.4 shows aspects of the signal associated with vowels and consonants. For example, the three horizontal bands marked F1, F2, and F3 are the three formants associated with the sound of *read*. Rapid shifts in frequency preceding or following formants are called **formant transitions** and are associated with consonants. For example, T2 and T3 are formant transitions associated with the /d/ of *read*.

Now that we know how speech is produced and how it is represented on a speech spectrogram, we are ready to look at some of the problems that we must solve in order to understand speech perception.

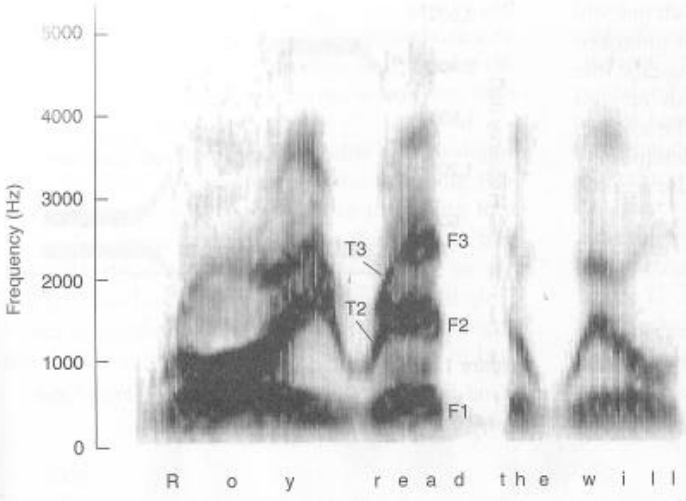


Figure 12.4
Spectgrogram of the sentence "Roy read the will," showing formants such as F1, F2, and F3, and formant transitions such as T2 and T3. (Spectrogram courtesy of Kerry Green.)

Problems Posed by the Speech Stimulus

The main problem facing researchers who are trying to understand speech perception is that the relationships between the acoustic signal and the sounds we hear are extremely complex. One reason for this complexity is that the acoustic signal is not neatly separated into individual words. This lack of separation between the signal for each word creates the **segmentation problem**: How do we perceptually segregate this continuous stream into individual words?

The Segmentation Problem

In Chapter 5 we saw that one task of the visual system is segmentation—separating a visual scene into individual objects. The auditory system faces a similar problem for speech—separating speech stimuli into individual words.

Just as we effortlessly see objects when we look at a visual scene, we usually have little trouble perceiving individual words as we have a conversation with another person. But when we look at the speech spectrogram, we see that the acoustic signal is continuous, with either no physical breaks in the signal or breaks that don't necessarily correspond to the breaks we perceive between words (Figure 12.5). The fact that there are usually no spaces between words becomes obvious when you listen to someone speaking a foreign

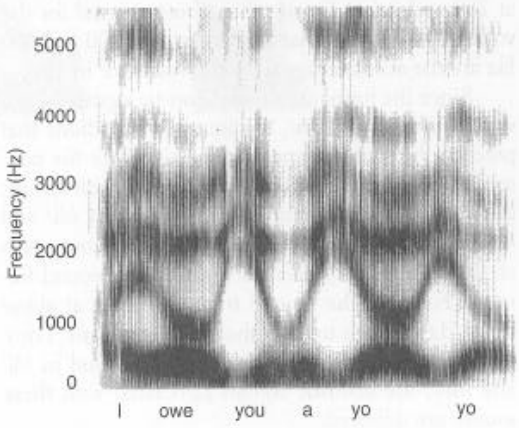


Figure 12.5
Spectrogram of "I owe you a yo-yo." This spectrogram does not contain pauses or breaks that correspond to the words that we hear. The absence of breaks in the acoustic signal creates the segmentation problem. (Spectrogram courtesy of David Pisoni.)

language. To someone who is unfamiliar with that language, the words seem to speed by in an unbroken string. However, to a speaker of that language, the words seem separated, just as the words of your native language seem separated to you. The solution to the segmentation problem involves determining how we divide the continuous stream of the acoustic signal into a series of individual words.

The Variability Problem

Another problem posed by the speech stimulus is that the acoustic signal is so variable that there is no simple correspondence between the acoustic signal and individual phonemes. This variability comes from the following sources.

Variability from a Phoneme's Context The acoustic signal associated with a phoneme changes depending on its context. For example, look at Figure 12.6, which shows spectrograms for the sounds /di/ and /du/. These are smoothed hand-drawn spectrograms that show the two most important characteristics of the sounds: the formants and the formant transitions. Since formants are associated with vowels, we know that the formants at 200 and 2,600 Hz are the acoustic signal for the vowel /i/ in /di/ and that the formants at 200 and 600 Hz are the acoustic signal for the vowel /u/ in /du/.

Since the formants are the acoustic signals for the vowels in /di/ and /du/, the formant transitions that precede the formants must be the signal for the consonant /d/. But notice that the formant transitions for the second (higher frequency) formants of /di/ and /du/ are different. For /di/, the formant transition starts at about 2,200 Hz and rises to meet the second formant. For /du/, the second transition starts at about 1,100 Hz and falls to meet the second formant. Thus, even though we perceive the same /d/ sound in /di/ and /du/, the acoustic signals associated with these sounds are different.

This effect of context occurs because of the way speech is produced. The articulators are constantly moving as we talk, so the shape of the vocal tract for a particular phoneme is influenced by the shapes for the phonemes that both precede it and follow it. This overlap between the articulation of neighboring

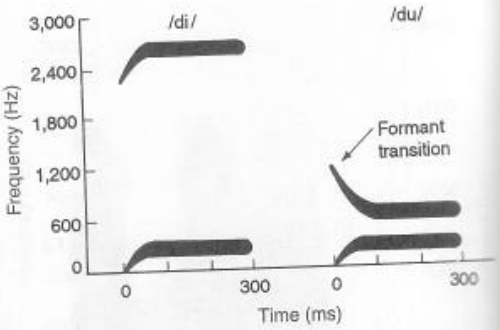


Figure 12.6
Hand-drawn spectrograms for /di/ and /du/. (From Liberman et al., 1967.)

phonemes is called **coarticulation**. You can demonstrate coarticulation to yourself by noting how you produce phonemes in different contexts. For example, say *bat* and *boot*. When you say *bat*, your lips are unrounded, but when you say *boot*, your lips are rounded, even during the initial /b/ sound. Thus, even though the /b/ is the same in both words, you articulate them differently. In this example, the articulation of /oo/ in *boot* overlaps the articulation of /b/, causing the lips to be rounded even before the /oo/ sound is actually produced.

The fact that we perceive the sound of a phoneme as the same, even though the acoustic signal is changed by coarticulation, is an example of perceptual constancy. This term may be familiar to you from our observations of constancy phenomena in the sense of vision, such as color constancy (we perceive an object's chromatic color as constant even when the wavelength distribution of the illumination changes) and size constancy (we perceive an object's size as constant even when the size of its image changes on our retina). Perceptual constancy in speech perception is similar. We perceive the sound of a particular phoneme as constant even when the phoneme appears in different contexts that change its acoustic signal.

Variability from Different Speakers People say the same words in a variety of different ways. Some people's voices are high pitched, and some are low pitched,

k with accents; some talk extremely rapidly and
eak e-x-t-r-e-m-e-l-y s-l-o-w-l-y. These wide vari-
speech in different speakers mean that for dif-
peakers a particular phoneme or word can have
different acoustic signals.

Speakers also introduce variability by their sloppy
nunciation. For example, say the following sen-
nce at the speed you would use in talking to a
iend: "This was a best buy." How did you say "best
buy"? Did you pronounce the /t/ of best, or did
you say "bes buy"? What about "She is a bad girl"?
While saying this rapidly, notice whether your tongue
hits the top of your mouth as you say the /d/ in bad.
Many people omit the /d/ and say "ba girl." Finally, what about "Did you go to the store?" Did you say "did you" or "dijoo"? You have your own ways of producing various words and phonemes, and other people have theirs. Analysis of how people actually speak has determined that there are 50 different ways to produce the word "the" (Waldrop, 1988).

That people do not usually articulate each word individually in conversational speech is reflected in the spectrograms in Figure 12.7. The spectrogram in Figure 12.7a is for the question "What are you doing?" spoken slowly and distinctly, whereas the spectrogram in Figure 12.7b is for the same question taken from

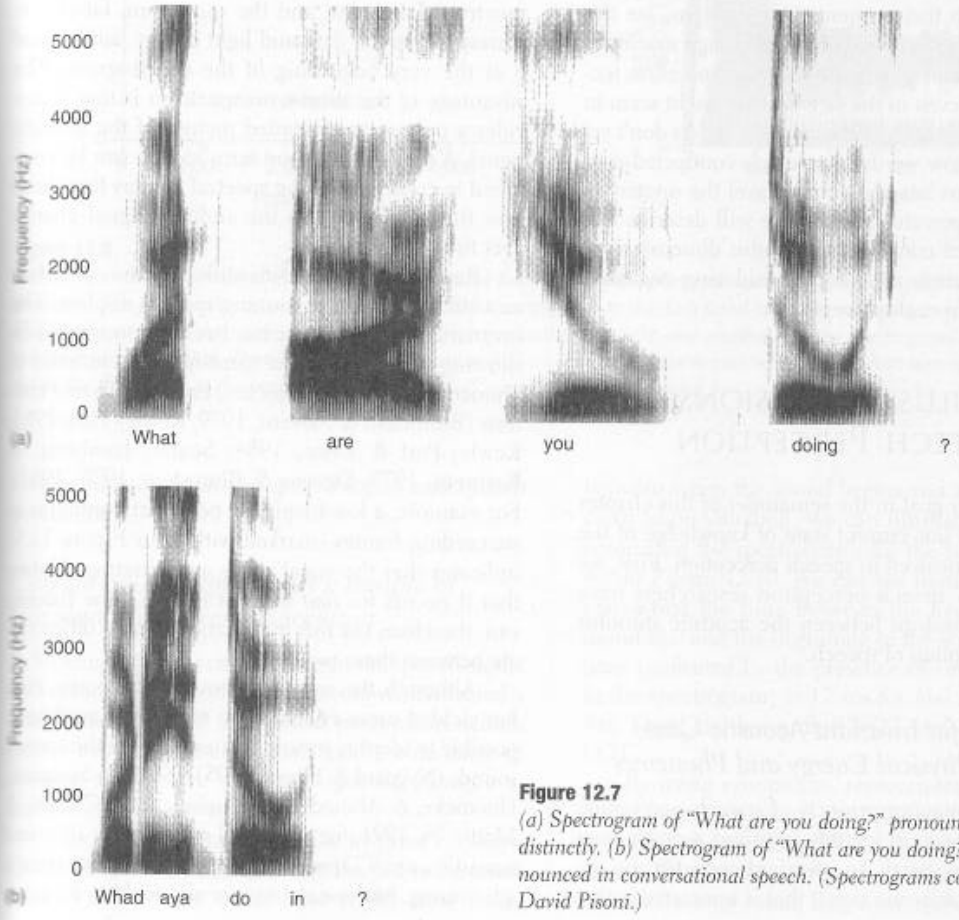


Figure 12.7
(a) Spectrogram of "What are you doing?" pronounced slowly and distinctly. (b) Spectrogram of "What are you doing?" as pronounced in conversational speech. (Spectrograms courtesy of David Pisoni.)

conversational speech, in which "What are you doing?" becomes "Whad'aya doin'?" This difference shows up clearly in the spectrogram, which indicates that, although the first and last words (*what* and *doing*) create similar patterns in the two spectrograms, the pauses between words are absent or are much less obvious in the spectrogram of Figure 12.7b, and the middle of this spectrogram is completely changed, with a number of speech sounds missing.

The variability in the acoustic signal that is caused by coarticulation, by different speakers, and by sloppy pronunciation creates a problem for the listener: He or she must somehow transform the information contained in this highly variable acoustic signal into familiar words. This variability problem, combined with the segmentation problem, are the reasons that it has been so difficult to design machines that can recognize speech. But humans somehow recognize speech even in the face of what might seem to be extreme difficulties. Although researchers don't yet know exactly how we do it, research conducted over the past 50 years has begun to unravel the mystery of how humans perceive speech. We will describe this research by first considering stimulus dimensions of speech perception and then considering cognitive dimensions of speech perception.

Stimulus Dimensions of Speech Perception

WebTUTOR Our goal in the remainder of this chapter is to highlight our current state of knowledge of the mechanisms involved in speech perception. First, we focus on how speech perception researchers have studied relationships between the acoustic stimulus and the perception of speech.

The Search for Invariant Acoustic Cues: Matching Physical Energy and Phonemes

One of the ongoing projects of speech perception research has been to identify invariant acoustic cues in the acoustic signal. An **invariant acoustic cue** is a feature of the acoustic signal that is associated with a particular phoneme and that remains constant even when phonemes appear in different contexts or are spoken by different speakers. Since invariant acoustic cues are not that obvious from the normal speech spectrogram, researchers searching for these invariant cues have devised new ways of displaying and analyzing the acoustic signal.

One way of displaying the acoustic signal is called the **short-term spectrum**. A short-term spectrum creates a detailed picture of the frequencies that occur within a short segment of time. For example, the short-term spectrum on the left of Figure 12.8 shows the frequencies that occur during the first 26 ms of the sound /ga/ along with the regular spectrogram for /ga/, on the right. The peak in the short-term spectrum labeled a, and the minimum, labeled b, correspond to the dark and light energy bands a and b at the very beginning of the spectrogram. The advantage of the short-term spectrum is that it provides a precise and detailed picture of the acoustic signal. A sequence of short-term spectra can be combined to create a **running spectral display** that shows how the frequencies in the auditory signal change over time (Figure 12.9).

Researchers have identified some invariant acoustic cues in these running spectral displays. The invariance of these cues has been demonstrated by showing that people can identify characteristics of phonemes based on these cues, even in different contexts (Blumstein & Stevens, 1979; Kewley-Port, 1983; Kewley-Port & Luce, 1984; Searle, Jacobson, & Rayment, 1979; Stevens & Blumstein, 1978, 1981). For example, a low-frequency peak that continues in succeeding frames (marked with V in Figure 12.9) indicates that the vocal cords are vibrating. Notice that it occurs for /da/ but not for /pi/. The listener can, therefore, use this information to help differentiate between these two sounds.

Although the search for invariant acoustic cues has yielded some encouraging results, it hasn't been possible to identify invariant cues for all of the speech sounds (Nygaard & Pisoni, 1995). (See also Sussman, Hoemeke, & Ahmed, 1993; Sussman, McCaffrey, & Matthews, 1991, for additional evidence for invariant acoustic cues.) Thus, the search for invariant cues continuing, but researchers are also looking for

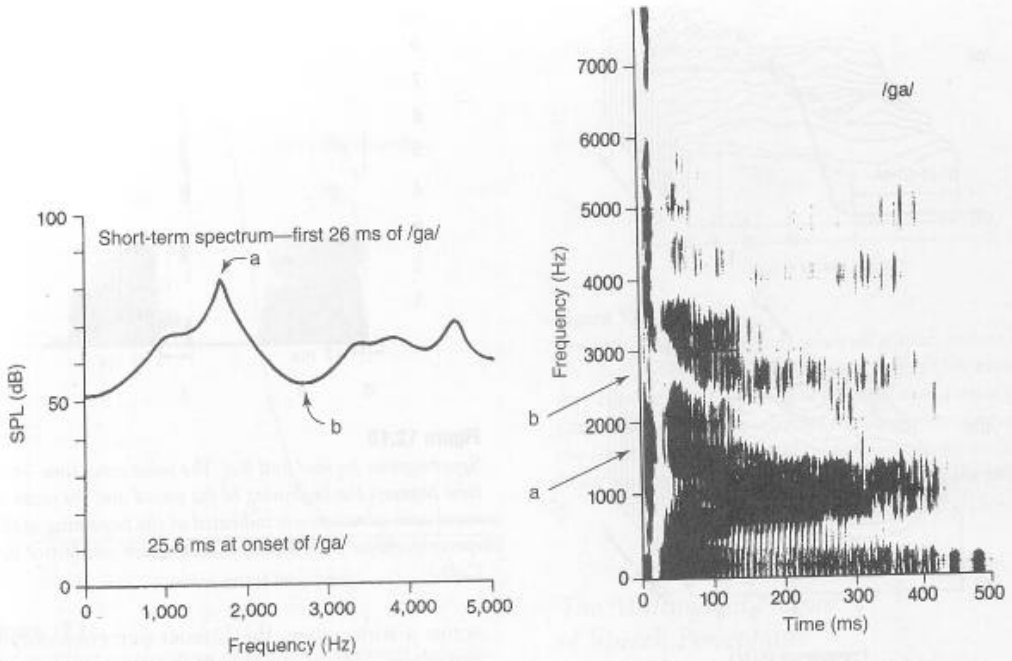


Figure 12.8

Left: A short-term spectrum of the acoustic energy in the first 26 ms of the phoneme /ga/. Right: Sound spectrogram of the same phoneme. The peak in the short-term spectrum marked a *corresponds to the dark band of energy marked* a *in the spectrogram. The minimum in the short-term spectrum marked* b *corresponds to the light area marked* b *in the spectrogram. Note that the spectrogram on the right shows the energy for the entire 500-ms duration of the sound, whereas the short-term spectrum only shows the first 26 ms at the beginning of this signal. (Courtesy of James Sawusch.)*

connections between the speech signal and speech perception.

Categorical Perception: An Example of Constancy in Speech Perception

In looking for connections between the speech signal and speech perception, researchers have discovered a phenomenon called **categorical perception**, which creates two categories of sounds from a wide range of acoustic signals. We will use a specific example to explain what this means.

The example we will describe involves varying a characteristic of the acoustic signal called **voice onset time (VOT)**. Voice onset time is the time delay between when the sound begins and when the vocal cords begin vibrating. We can illustrate this delay by comparing the spectrograms for the sounds /da/ and /ta/ in Figure 12.10. We can see from these spectrograms that the time between the beginning of the sound /da/ and the beginning of the vocal cord vibrations (indicated by the presence of vertical striations in the spectrogram) is 17 ms for /da/ and 91 ms for /ta/. Thus, /da/ has a short VOT and /ta/ has a long VOT.

By using computers, researchers have created sound stimuli in which the VOT is varied in small steps from short to long. When they vary VOT, using stimuli like the ones in Figure 12.10, and ask subjects to indicate what sound they hear, the subjects report

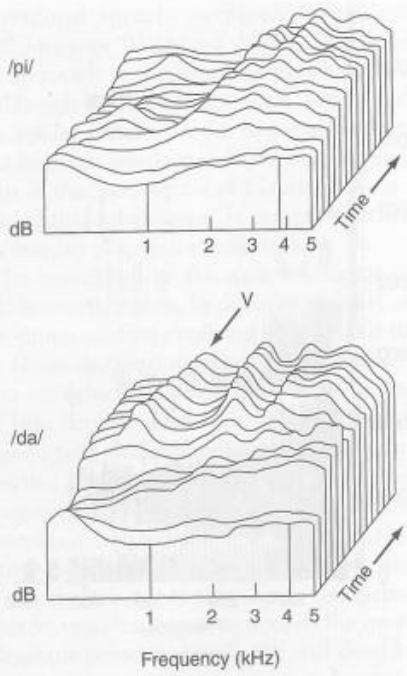


Figure 12.9

Running spectral displays for /pi/ and /da/. These displays are made up of a sequence of short-term spectra like the one in Figure 12.8. Each of these spectra is displaced 5 ms on the time axis, so that each step we move along this axis indicates the frequencies present in the next 5 ms. The low-frequency peak (V) in the /da/ display is a cue for vibration of the vocal cords. (From Kewley-Port & Luce, 1984.)

hearing only two sounds, /da/ or /ta/, even though a large number of stimuli with different VOTs are presented.

This result is shown in Figure 12.11 (Eimas & Corbit, 1973). At short VOTs, subjects report that they hear /da/, and they continue reporting this even when the VOT is increased. But when the VOT reaches about 35 ms, their perception abruptly changes, so at VOTs above 40 ms, they report hearing /ta/. The VOT when the perception changes from /da/ to /ta/ is called the **phonetic boundary**. The key result of the categorical perception experiment is that, even though the VOT is changed continuously

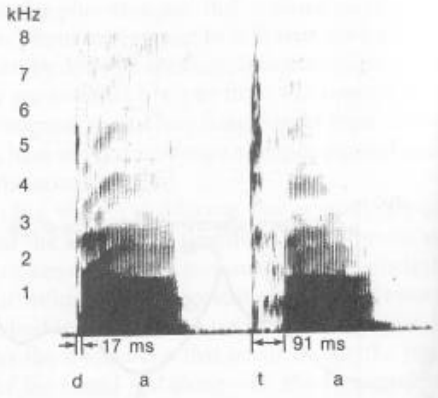


Figure 12.10

Spectrograms for /da/ and /ta/. The voice onset time—the time between the beginning of the sound and the onset of vocal cord vibration—is indicated at the beginning of the spectrogram for each sound. (Spectrogram courtesy of Ron Cole.)

across a wide range, the listener perceives only two categories: /da/ on one side of the phonetic boundary and /ta/ on the other side.

Once we have demonstrated categorical perception using the procedure above, we can further confirm the existence of just two categories across the range of VOTs by running a discrimination test, in which we present two stimuli with different VOTs and ask the subject whether they sound the same or different. When we present two stimuli that are on the same side of the phonetic boundary, such as 10 and 30 ms VOTs, the listener says they sound the same (Figure 12.12). However, when we present two stimuli that are on opposite sides of the phonetic boundary, such as 30 and 50 ms VOTs, the listener says they sound different. The fact that all stimuli on the same side of the phonetic boundary are perceived as identical is an example of perceptual constancy (Figure 12.13). If this constancy did not exist, we would perceive different sounds every time we changed the VOT. Instead, we experience one sound on each side of the phonetic boundary. This simplifies our perception of phonemes and helps us to easily perceive the wide variety of sounds in the environment.

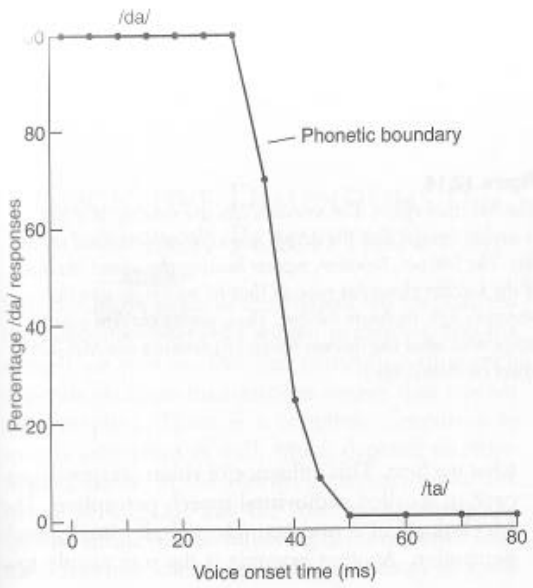


Figure 12.11
The results of a categorical perception experiment indicate that /da/ is perceived for VOTs to the left of the phonetic boundary, and that /ta/ is perceived at VOTs to the right of the phonetic boundary. (From Eimas & Corbit, 1973.) (Listen to WebTutor, Categorical Perception.)

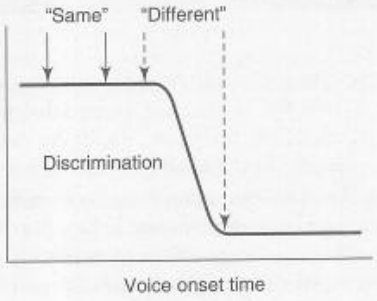


Figure 12.12
In the discrimination part of a categorical perception experiment, two stimuli are presented, and the subject is asked to indicate whether they are the same or different. The typical result is that two stimuli with VOTs on the same side of the phonetic boundary (solid arrows) are judged to be the same, and that two stimuli on different sides of the phonetic boundary (dashed arrows) are judged to be different.

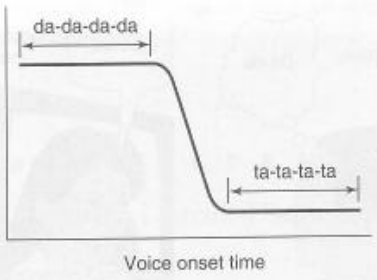


Figure 12.13
Perceptual constancy occurs when all stimuli on one side of the phonetic boundary are perceived to be in the same category even though their VOT is changed over a substantial range. This diagram symbolizes the constancy observed in the Eimas and Corbit (1973) experiment, in which /da/ was heard on one side of the boundary and /ta/ on the other side.

The Multimodal Nature of Speech Perception: Information from Hearing and Vision

Another property of speech perception is that it is **multimodal**. That is, our perception of speech can be influenced by information from a number of different senses. The "Across the Senses" section at the end of the chapter describes how speech can be perceived through touch using a method called Tadoma. Speech perception can also be influenced by visual information, as demonstrated by an effect called the **McGurk effect**, after the man who first described it (McGurk & MacDonald, 1976).

The procedure for achieving the McGurk effect is illustrated in Figure 12.14. A subject observes a videotape showing a person making the lip movements for the sounds /ga-ga/. But as he receives this visual information, he simultaneously receives an auditory sound track, which is the acoustic signal that is usually heard as /ba-ba/. Despite the fact that the subject is receiving the acoustic signal for /ba-ba/, he actually hears the sounds /da-da/. This misperception is a striking perceptual effect in which subjects are convinced that the woman in the videotape is saying /da-da/ even though that stimulus is never actually present. If they close

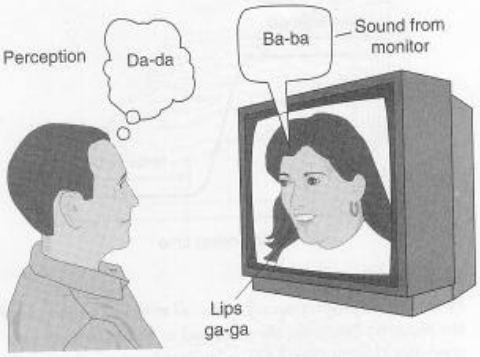


Figure 12.14
The McGurk effect. The woman's lips are moving as if she is saying /ga-ga/, but the actual sound being presented is /ba-ba/. The listener, however, reports hearing the sound /da-da/. If the listener closes his eyes, so that he no longer sees the woman's lips, he hears /ba-ba/. Thus, seeing the lips moving influences what the listener hears. (Experience the McGurk effect on WebTutor.)

their eyes, they hear /ba-ba/. If they open them, they hear /da-da/. Thus, although auditory energy is the major source of information for speech perception, visual information can also exert a strong influence on what we hear. This influence of vision on speech perception is called **audiovisual speech perception.** The McGurk effect is one example of audiovisual speech perception. Another example is the way people rou-

BRAIN SCAN

Activation of Auditory Cortex During Silent Lipreading

Two examples of interactions between speech perception and visual perception are the McGurk effect (above and pages 325–326) and the fact that speech perception is improved by watching a speaker's lips move, particularly in noisy surroundings (Dodd & Campbell, 1987).

A brain imaging study by Gemma Calvert and coworkers (1997) has investigated which brain areas are activated by this lipreading. In the "silent lipreading" condition, subjects watched a silent videotape of a person saying numbers and silently repeated, in their minds, each number they perceived from the person's mouth movements. In the "static" condition, subjects observed a videotape of a static face and repeated the number "one" silently to themselves. When the brain activity, recorded by fMRI, was compared from these two conditions, more activity was observed in the silent lipreading condition than in the static condition in areas of visual cortex associated with movement (since the lips of the face in the videotape were moving) and with areas in temporal cortex that are normally activated by actually hearing speech. The fact that visual cues to speech can activate auditory areas, even though no auditory stimuli are present, may be related to the physiological mechanism responsible for our ability to use perceptions of lip movements to help us perceive speech.

nely use information provided by the speaker's lip movements to help understand speech in a noisy environment. See Summary Table 12.1 for an overview of the material we have covered so far.

Cognitive Dimensions of Speech Perception

Our discussion so far in this chapter has emphasized the relationship between speech perception and the stimuli we receive. But our perception of speech depends on more than just the energy that reaches our receptors. There is a cognitive dimension to speech perception as well, which depends on information stored in the listener's memory about the nature of language and the voice characteristics of specific speakers. The following demonstration illustrates how our knowledge of the meanings of words enables us to perceive these words even when the stimulus is incomplete.

DEMONSTRATION

Perceiving Degraded Sentences

Read the following sentences:

(1) M*R* H*D * L*TTL* L*MB I*S FL**C* W*S WH*T* *S SN*W
(2) TH* S*N *S N*T SH*N*NG T*D**
(3) S*M* W**DS *R* EA*I*R T* U*D*R*T*N* T*A* *T*E*S ●

Your ability to read the sentences, even though up to half of the letters have been eliminated, was aided by your knowledge of English (Denes & Pinson, 1993). This is an example of top-down processing that

SUMMARY TABLE 12.1

The Speech Stimulus

Speech is specified in terms of units called phonemes, which are the basic building blocks of words. The stimulus can also be specified in terms of the patterns of pressure changes in sound spectrograms, which display formants associated with vowels and other patterns associated with consonants.

Problems Posed by the Speech Stimulus

The relationship between the speech stimulus and perception is not a simple one. Among the problems the speech system must solve in order to decode the speech stimulus are the segmentation problem, since the speech signal is continuous, and the variability of the speech stimulus, which is caused by coarticulation and across-speaker variability.

Invariant Cues

It has been difficult to find invariant acoustic cues—characteristics of the speech stimulus that are associated with a particular sound and remain constant under a number of different conditions.

Categorical Perception

Categorical perception occurs in speech when a wide range of sound stimuli result in the same perception. This is an example of perceptual constancy.

Multimodal Nature of Speech Perception

Speech perception can be influenced by information from a number of different senses. An example of this is the McGurk effect, in which visual information provided by seeing a speaker's mouth move can change the perception of the sound stimulus.

Brain Scan: Activity in Auditory Cortex During Silent Lipreading

fMRI measurements indicate that watching a person's lips make speech movements activates the auditory areas of the brain even though no sound is present. This provides a physiological mechanism for auditory–visual interactions.

we have discussed in connection with visual perception (see page 8): Knowledge brought to the situation by the perceiver is used to supplement the bottom-up information provided by stimulation of the receptors (Figure 12.15). One example of the effect of top-down processing in speech perception is provided by the process of segmentation, in which the speaker breaks the continuous acoustic signal into individual words.

Meaning and Segmentation

To help you appreciate the role of meaning in achieving segmentation, do the following demonstration.

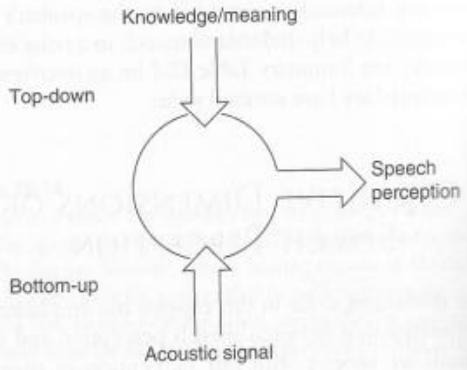


Figure 12.15
Speech perception is the result of top-down processing (based on knowledge and meaning) and bottom-up processing (based on the acoustic signal) working together.

DEMONSTRATION

Segmenting Strings of Sounds

1. **Read the following words: Anna Mary Candy Lights Since Imp Pulp Lay Things. Now that you've read the words, what do they mean? If you think that this is a list of unconnected words beginning with the names of two women, Anna and Mary, you're right; but if you read this series of words a little faster, ignoring the spaces between the words on the page, you may hear a connected sentence that does not begin with the names Anna and Mary. (For the answer see the bottom of page 424—but don't peek until you've tried reading the words rapidly.)**

2. **Read the following phrase fairly rapidly to a few people: "In mud eels are, in clay none are," and ask them to write the phrase.** ●

If you succeeded in creating a new sentence from the series of words in the first demonstration, you did so by changing the segmentation, and this change was achieved by your knowledge of the meaning of the sounds. When Raj Reddy (1976) asked subjects to write their perception of "In mud eels are, in clay none are," he obtained responses like "In muddies, sar, in clay nanar"; "In may deals are, en clainanar"; and "In madel sar, in claynanar." In the absence of any context, some of Reddy's listeners had difficulty figuring out what the phrase meant and therefore forced their own interpretation on the sounds they heard. Had the listeners known that the passage was taken from a book about amphibians, their knowledge about possible meanings of the words would have facilitated segmentation and increased probability of their decoding the sentence correctly.

Pairs of words that flow together in speech also exemplify how segmentation results from meaning. "Big girl" can be interpreted as "big Earl," and the interpretation you pick will probably depend on the overall meaning of the sentence in which these words appear. This example is similar to the familiar "I scream, you scream, we all scream for ice cream" that many people learn as children. The acoustic stimuli for "I scream" and "ice cream" are identical, so the different segmentations must be achieved by the meaning of the sentence in which these words appear.

Top-down processing not only helps us to segment the acoustic signal, but also helps us to recognize phonemes and words. We will now describe some experiments that show how meaningful contexts can enhance a listener's ability to recognize phonemes and words.

Meaning and Phoneme Perception

A large amount of research has shown that it is easier to perceive phonemes that appear in a meaningful context. Philip Rubin, M. T. Turvey, and Peter Van Gelder (1976) showed that meaning enhances a listener's ability to recognize phonemes by presenting a series of short words like SIN, BAT, and LEG, or nonwords like JUM, BAF, and TEG, and asking subjects to respond by pressing a key as rapidly as possible whenever they heard a sound that began with /b/. On the average, subjects took 631 milliseconds to respond to the nonwords and 580 ms to respond to the real words. Thus, when a phoneme is at the beginning of a real word, it is identified about 8 percent faster than if it is at the beginning of a meaningless syllable.

The effect of meaning on the perception of phonemes was demonstrated in another way by Richard Warren (1970), who had subjects listen to a recording of the sentence "The state governors met with their respective legislatures convening in the capital city." Warren replaced the first /s/ in "legislatures" with the sound of a cough and told his subjects that they should indicate where in the sentence the cough occurred. No subject identified the correct position of the cough, and, even more significantly, none of the subjects noticed that the /s/ in "legislatures" was missing. This effect, which Warren called the **phonemic restoration effect**, was experienced even by students and staff in the psychology department who knew that the /s/ was missing.

Warren not only demonstrated the phonemic restoration effect but also showed that it can be influenced by the meaning of words following the missing phoneme. For example, the last word of the phrase "There was time to *ave . . ." (where the * indicates the presence of a cough or some other sound) could be shave, save, wave, or rave, but subjects heard the word wave if the remainder of the sentence had to do with saying goodbye to a departing friend.

The phonemic restoration effect was used by Arthur Samuel (1981) to show that speech perception is determined both by a context that produces expectations in the listener (top-down processing) and also by the nature of the acoustic signal (bottom-up processing). Samuel demonstrated top-down processing by masking various phonemes in sentences with a white-noise masker, like the sound produced by a television set tuned to a nonbroadcasting channel, and showing that longer words increase the likelihood of the phonemic restoration effect. Apparently, subjects use the additional context provided by the long word to help identify the masked phoneme. Further evidence for the importance of context is Samuel's finding that more restoration occurs for a real word like prOgress (where the capital letter indicates the masked phoneme) than for a similar "pseudoword" like crOgress (Samuel, 1990).

Samuel demonstrated bottom-up processing by showing that restoration is better if the masking sound and the masked phoneme sound similar. Thus, phonemic restoration is more likely to occur for a phoneme such as /s/, which is rich in high-frequency acoustic energy, if the mask also contains a large proportion of high-frequency energy.

What's happening in phonemic restoration, according to Samuel, is that we use the context to develop some expectation of what a sound will be. But before we actually perceive the sound, its presence must be confirmed by the presence of a sound that is similar to it. If the white-noise mask contains frequencies that make it sound similar to the phoneme we are expecting, phonemic restoration occurs, and we are likely to hear the phoneme. If the mask does not sound similar, phonemic restoration is less likely to occur (Samuel, 1990).

Meaning and Word Perception

Meaningfulness also makes it easier to perceive whole words. An early demonstration of this effect by George Miller and Steven Isard (1963) showed that words are more intelligible when heard in the context of a grammatical sentence than when presented as items in a list of unconnected words. They demonstrated this by creating three kinds of stimuli: (1) normal grammatical sentences, such as *Gadgets simplify work around the house*; (2) anomalous sentences, that follow the rules of grammar but make no sense, such as *Gadgets kill passengers from the eyes*;

and (3) ungrammatical strings of words, such as *Between gadgets highways passengers the steal.*

Miller and Isard used a technique called **shadowing**, in which they presented these sentences to subjects through earphones and asked them to repeat aloud what they were hearing. The subjects reported normal sentences with an accuracy of 89 percent, but their accuracy fell to 79 percent for the anomalous sentences, and 56 percent for the ungrammatical strings.

The differences among the three types of stimuli became even greater when the subjects heard the stimuli in the presence of a background noise. For example, at a moderately high level of background noise, accuracy was 63 percent for the normal sentences, 22 percent for the anomalous sentences, and only 3 percent for the ungrammatical strings of words.

This result is telling us that when words are arranged in a meaningful pattern, we can perceive them more easily. But most people don't realize that it is their knowledge of the nature of their language that helps them fill in sounds and words that might be difficult to hear. For example, our knowledge of permissible word structures tells us that ANT, TAN, and NAT are all permissible sequences of letters in English, but that TQN or NQT cannot be English words.

At the level of sentences, our knowledge of the rules of grammar tells us that *the cat is weird* is a permissible sentence, but *is weird cat the* is not a permissible sentence. Since most of our everyday experience is with meaningful words and grammatically correct sentences, we are continually using our knowledge of what is permissible in our language to help us understand what is being said. This becomes particularly important when listening under less than ideal conditions such as in noisy environments, or if the speaker's voice quality or accent is difficult to understand (see also Salasoo & Pisoni, 1985).

All of these results support the idea that the knowledge a listener brings to a situation helps the listener decode the acoustic signal into phonemes and meaningful words and sentences. In addition, there is also evidence that a listener's experience in listening to specific speakers can enhance his or her ability to perceive what is being said.

Answer to question on page 422: "An American delights in simple play things."

Speaker Characteristics

When you're having a conversation, hearing a lecture, or listening to dialogue in a movie, you usually focus on determining the meaning of what is being said. But as you are taking in these messages, you are also, perhaps without realizing it, taking in characteristics of the speaker's voice. These characteristics, which are called **indexical characteristics**, carry information about speakers such as their age, gender, where they are from, their emotional state, and whether they are being sarcastic or serious. Consider, for example, the following joke:

> A linguistics professor was lecturing to his class one day. "In English," he said, "A double negative forms a positive. In some languages, though, such as Russian, a double negative is still a negative. However, there is no language wherein a double positive can form a negative."
>
> A voice from the back of the room piped up. "Yeah, right."

This joke is funny because "Yeah, right" contains two positive words that, despite the linguistics professor's statement, produce a negative statement that most people who are aware of contemporary English usage would interpret as "I disagree" (or "No way," to use a more colloquial expression.) The point of this example is not just that "Yeah, right" can mean "I disagree," but that the meaning of this phrase is determined by our knowledge of current English usage and also (if we were actually listening to the student's remark) by the speaker's tone of voice, which in this case would be highly sarcastic.

The speaker's tone of voice is one factor that helps listeners determine the meaning of what is being said. But most research on indexical characteristics has focused on how speech perception is influenced by the speaker's identity. Thomas Palmer

Stephen Goldinger, and David Pisoni (1993) demonstrated the effect of speaker identity by having subjects listen to a sequence of words. After each word, the subject indicated whether the word was a new word (this was the first time it appeared) or an old word (it had appeared previously in the sequence). They found that subjects reacted more rapidly and were more accurate when the same speaker said all of the words than if different speakers said the words. This means that the listeners are taking in two levels of information about the word: (1) its meaning and (2) characteristics of the speaker's voice.

In another experiment that demonstrates the importance of the speaker's voice for speech perception, subjects listened to the voices of 10 different speakers. Following this training, the listeners were given a word intelligibility test to determine how well they could identify words spoken by the speakers. When the results of this test were compared to the performance of a control group who were also trained to recognize the same 10 speakers but who heard unfamiliar speakers for the intelligibility test, it was found that those hearing the familiar speakers performed better on the test (Nygaard, Sommers, & Pisoni, 1994).

In order for performance to be better with the familiar speakers, the voice characteristics of these speakers would have to be stored in the listener's long-term memory. When presented with the word intelligibility test, listeners retrieve this information about the familiar speakers from their memory and use it to help them identify the words.

From the results of this experiment and the others we have discussed, we can conclude that speech perception depends both on the bottom-up information provided by the acoustic signal and on the top-down information provided by the meanings of words and sentences, the listener's knowledge of the rules of grammar, and information that the listener has about characteristics of the speaker's voice.

We can appreciate the interdependence of the acoustical and meaningful units of speech when we realize that, although we use the meaning to help us to understand the acoustic signal, the acoustic signal is the starting point for determining the meaning. Look at it this way: There may be enough information in my sloppy handwriting so that a person using bottom-up processing can decipher it solely on the basis of the squiggles on the page, but my handwriting is much easier to decipher if, by using top-down processing, the person takes the meanings of the words into account. And just as previous experience in hearing a particular person's voice makes it easier to understand that person later, previous experience in reading my handwriting would make it easier to read the squiggles on the page. Speech perception apparently works in a similar way. Although most of the information is contained in the acoustic signal, taking meaning and indexical properties into account makes understanding speech much easier.

The Physiology of Speech Perception

Researchers have investigated the physiology of speech perception in a number of different ways: (1) by recording neural responses to both natural speech stimuli and stimuli resembling parts of the speech signal; (2) by studying patients who, because of brain damage, have difficulty understanding or producing speech; and (3) by recording the change in blood flow in different parts of the brain during speech perception.

Neural Responses to Speech and Complex Sounds

Most of the research explaining the neural response to speech stimuli has focused on how neurons in the auditory nerve respond to speech sounds. For example, Figure 12.16 shows the short-term spectrum for the sound /da/ and the firing pattern for a representative population of cat auditory nerve fibers with low, medium, and high characteristic frequencies. (Remember from Chapter 10 that the neuron's characteristic frequency is the frequency to which this neuron responds best.) The match between the speech stimulus and the firing of a number of neurons in the auditory nerve indicates that information in the speech stimulus is represented by the pattern

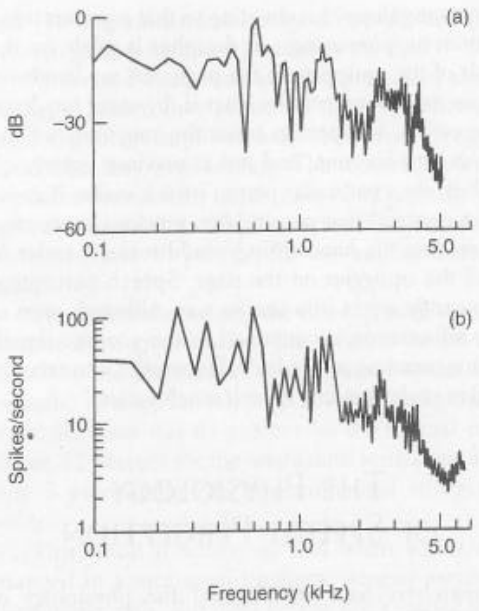


Figure 12.16
(a) Short-term spectrum for /da/. This curve indicates the energy distribution in /da/ between 20 and 40 ms after the beginning of the signal. (b) Nerve firing of a population of cat auditory nerve fibers to the same stimulus. (From Sachs, Young, & Miller, 1981.)

of firing of auditory nerve fibers (also see Delgutte & Kiang, 1984a,b).

Another approach to studying the relation between neural responding and speech perception is to look for neurons that respond to parts of the speech stimulus, such as the formant transitions and formants shown in Figure 12.4. Neurons have been found in a number of animals, including bats, birds, frogs, and cats, that respond best to combinations of tones that, like speech stimuli, have specific timing and frequencies (Fuzessery & Feng, 1983; Margoliash, 1983; Nelson, Erulkar, & Bryan, 1966; Olsen & Suga, 1991a,b). Most significantly, neurons have been found in an area of the monkey cortex that is analogous to human speech areas, which respond to recordings of monkey "calls" (Rauschecker, Tian, & Hauser, 1995). The existence of these neurons in the monkey as well as the other animals opens the possibility that there may be neurons in the human cortex that respond best to complex, speechlike stimuli.

Localization of Function

It has been known for over 150 years that the brain operates according to a principle called **localization of function**—specific functions are localized in specific areas of the brain. One form of localization is lateralization—a particular function is processed more strongly in either the left hemisphere or the right hemisphere. For most people, much of speech is processed in specific areas in the left hemisphere of the brain (Figure 12.17) (although some linguistic information is processed in the right hemisphere; Chairello, 1991).

Damage to **Broca's area**, in the frontal lobe, causes difficulty in speaking, and damage to Wernicke's area, in the temporal lobe, causes difficulty in understanding speech (Geschwind, 1979). These difficulties in speaking and understanding are forms of **aphasia** (Kolb & Whishaw, 1985).

There are numerous forms of aphasia, with the specific symptoms depending on the area and extent of the damage. The form that involves speech perception is called **Wernicke's aphasia**—an inability to

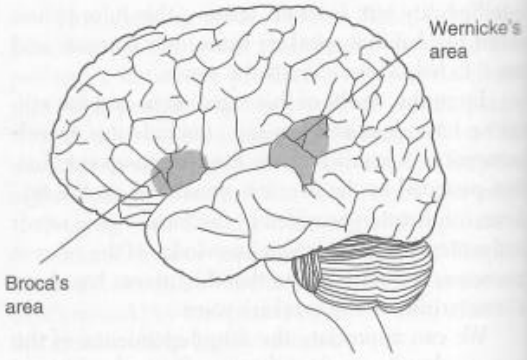


Figure 12.17
Broca's and Wernicke's areas, which are specialized for language production and comprehension, are located in the left hemisphere of the brain in most people.

comprehend words or arrange sounds into coherent speech. Wernicke's aphasia is a form of fluent aphasia, so called because people with this disorder can produce fluent speech, although their production often consists of meaningless strings of words and words in which phonemes are confused, which has been described as "word salad."

We can understand the basis of the problem in Wernicke's aphasia by considering the problem native Japanese speakers have in discriminating between /l/ and /r/. Native Japanese speakers are not able to distinguish /l/ from /r/ when they hear English, because the Japanese they heard spoken when they were infants did not distinguish between the sounds /l/ and /r/. Thus, the necessary templates for these two sounds did not develop in their brains, and as adults they lack the physiological mechanisms to distinguish between these two sounds (see "Plasticity of Perception," page 429). It has been suggested that people with Wernicke's aphasia have a similar problem—because of their brain damage, they cannot isolate phonemes or classify them into known phonemic systems (Kolb & Whishaw, 1985).

Research that used positron emission tomography (PET) on humans to compare brain activation during the perception of pitches to activation during the perception of speech stimuli has shown that pitch stimuli activate areas in the right hemisphere but that speech stimuli activate areas in the left hemisphere (Zatorre et al., 1992). In addition to physiological evidence of the lateralization of speech perception, psychophysical experiments indicate that speech stimuli are more easily processed when presented through earphones to the right ear than when they are presented through earphones to the left ear. Since signals from the right ear are sent preferentially to the left hemisphere, this right-ear preference indicates that most aspects of speech stimuli are processed in the left hemisphere (Kimura, 1961).

Is Speech "Special"?

Some researchers working in the field of speech perception think that there is something special about speech perception that sets it apart from the perception of other auditory stimuli. The idea of a mechanism that is specialized for the perception of speech has its appeal, especially when we remember how the visual system operates. We know that there are nuclei in the visual system that are specialized to process information about different qualities, such as color, depth, and movement (Casagrande, 1994). In addition, there are neurons that are specialized to respond to complex visual stimuli such as faces (Perrett et al., 1992). This evidence from vision, along with the fact that language is processed in specific areas of the cortex, makes it seem reasonable that there could be a specialized mechanism that exists especially to process speech stimuli.

The idea that there is a specialized mechanism for speech is, however, not universally accepted by all speech perception researchers. Some researchers feel that while the speech signal may be extremely complex, the perception of this signal can be explained by regular auditory mechanisms.

The question of whether speech is "special" or if it is served by the same auditory mechanism that serves other auditory stimuli has generated a voluminous amount of research over the more than 30 years since the idea of a special speech mechanism was proposed (Liberman et al., 1967). To illustrate the kinds of evidence that have been presented on both sides of this question, we will consider a phenomenon called duplex perception that was originally introduced to support the idea that speech is special.

Duplex perception is created by splitting the acoustic signal for a sound into two parts and presenting one part to each ear. For example, Figure 12.18a shows the speech spectrogram for the sound /da/, which consists of three formant transitions and their formants. Alvin Liberman and Ignatius Mattingly (1989) created duplex perception by presenting just the transition for the third formant to the left ear (Figure 12.18b), and the rest of the signal, which is called the base, to the right ear (Figure 12.18c). What the listener hears is the speech sound /da/, from the combined acoustic signal from the left and right ears, and a nonspeech "chirp," from the third formant transition from the left ear.

According to proponents of the special speech mechanism, this experiment illustrates that there are

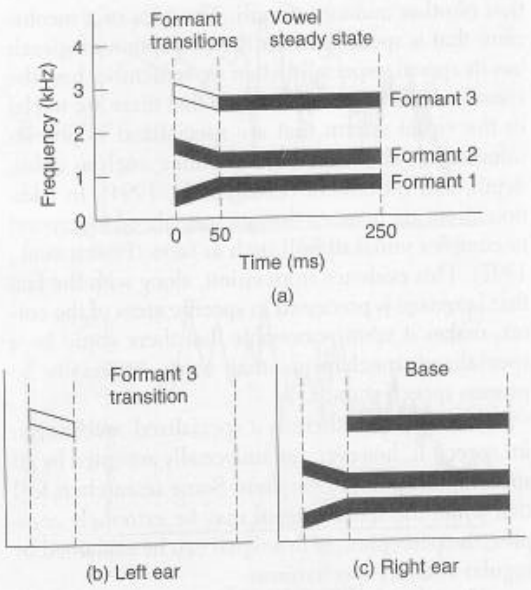


Figure 12.18

(a) The spectrogram for the sound /da/. Duplex perception occurs when the transition for the third formant (b) is presented to the left ear and the rest of the signal, called the base (c), is presented to the right ear. The listener perceives the sound /da/ in the right ear and a "chirp" in the left ear. See text for further details. (From Whalen & Lieberman, 1987.)

two kinds of perception of auditory stimuli: (1) a special speech mode, which combines the formant transition presented to the left ear and the base presented to the right ear to create the speech sound /da/; and (2) the auditory mode, which creates the nonspeech "chirp" sound from the high-frequency formant transition presented to the left ear.

Whalen and Liberman (1987) point out that the third-formant transition has a low intensity; the listener hears only the speech sound. The intensity of the third-formant transition must be increased to a higher level before the listener hears the nonspeech chirp sound as well. They concluded that this result confirms the special nature of speech, since processing the formant transition as speech (which creates the speech sound) takes priority over processing it as a general auditory signal (which creates the chirp).

However, Carol Fowler and Lawrence Rosenblum (1990) challenged this interpretation of duplex perception by creating an effect similar to duplex perception from the sounds of a door slamming. They split a recording of a metal door closing into two parts, a high-frequency component at the beginning of the sound and a lower-frequency component at the end of the sound. These two components are similar to the formant transition and the base of the speech sound used by Liberman and Mattingly. When Fowler and Rosenblum presented the high-frequency component to the left ear and the low-frequency component to the right ear, subjects heard the sound of a metal door slamming (the combination of the left- and right-ear stimuli) plus a "shaking" sound like the sound made by sand or small pellets being shaken in a cup (the high-frequency left-ear stimulus). According to Fowler and Rosenblum, this demonstration of duplex perception with nonspeech stimuli means that duplex perception can occur for environmental stimuli in general, which includes speech, slamming doors, and other stimuli as well (see Hall & Pastore, 1992, for a musical example).

Which interpretation of duplex perception is correct? Is speech perception served by a mechanism that is specialized for the perception of speech or by a mechanism similar to the ones that help us perceive environmental sounds in general? Is our perception of speech analogous to our perception of faces, which involves a specialized cortical area that may have been shaped by our experience with faces? We don't yet know the answers to these questions, and there are many more experiments in addition to our duplex perception example above that present evidence on either side of this issue (Fowler & Dekle, 1991; Mattingly & Studdert-Kennedy, 1991). This debate, which is still unresolved, is a good example of how researchers have approached the same problem from different perspectives.