CHAPTER 3

Basic Audition

The human auditory system is not a high-fidelity system. Amplitude is compressed; frequency is warped and smeared; and adjacent sounds may be smeared together. Because listeners experience auditory objects, not acoustic records like waveforms or spectrograms, it is useful to consider the basic properties of auditory perception as they relate to speech acoustics. This chapter starts with a brief discussion of the anatomy and function of the peripheral auditory system, then discusses two important differences between the acoustic and the auditory representation of sound, and concludes with a brief demonstration of the difference between acoustic analysis and auditory analysis using a computer simulation of auditory response. Later chapters will return to the topics introduced here as they relate to the perception of specific classes of speech sounds.

3.1 Anatomy of the peripheral auditory system

The peripheral auditory system (that part of the auditory system not in the brain) translates acoustic signals into neural signals; and in the course of the translation, it also performs amplitude compression and a kind of Fourier analysis of the signal.

Figure 3.1 illustrates the main anatomical features of the peripheral auditory system (see Pickles, 1988). Sound waves impinge upon the outer ear, and travel down the ear canal to the eardrum. The eardrum is a thin membrane of skin which is stretched like the head of a drum at the end of the ear canal. Like the membrane of a microphone, the eardrum moves in response to air pressure fluctuations.

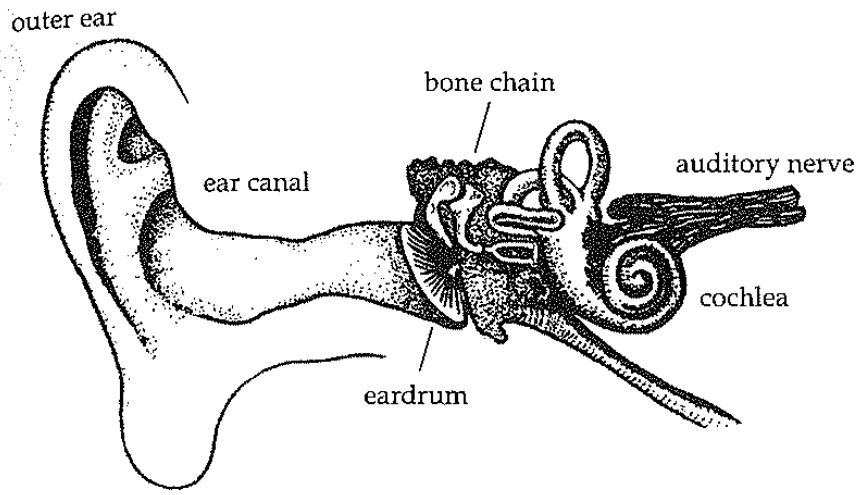

Figure 3.1 The main components of the human peripheral auditory system. Adapted from Brödel, 1946, fig. 1.

These movements are conducted by a chain of three tiny bones in the middle ear to the fluid-filled inner ear. There is a membrane (the basilar membrane) that runs down the middle of the conch-shaped inner ear (the cochlea). This membrane is thicker at one end than the other. The thin end, which is closest to the bone chain, responds to high-frequency components in the acoustic signal, while the thick end responds to low-frequency components. Each auditory nerve fiber innervates a particular section of the basilar membrane, and thus carries information about a specific frequency component in the acoustic signal. In this way, the inner ear performs a kind of Fourier analysis of the acoustic signal, breaking it down into separate frequency components.

3.2 The auditory sensation of loudness

The auditory system imposes a type of automatic volume control via amplitude compression, and as a result, it is responsive to a remarkable range of sound intensities (see Moore, 1982). For instance, the air pressure fluctuations produced by thunder are about 100,000 times larger than those produced by a whisper (see table 3.1).

Look at the values listed in the pressure column in the table. For most people, a typical conversation is not subjectively ten times louder than a quiet office, even though the magnitudes of their sound pressure fluctuations are. In general, subjective auditory impressions of loudness differences do not match sound pressure differences. The mismatch between differences in sound pressure and loudness has been noted for many years. For example, Stevens (1957) asked listeners to adjust the loudness of one sound until it was twice as loud as another or, in

How the inner ear is like a piano

For an example of what I mean by "responds to," consider the way in which piano strings respond to tones. Here's the experiment: go to your school's music department and find a practice room with a piano in it. Open the piano, so that you can see the strings. This works best with a grand or baby grand, but can be done with an upright. Now hold down the pedal that lifts the felt dampers from the strings and sing a steady note very loudly. Can you hear any of the strings vibrating after you stop singing? This experiment usually works better if you are a trained opera singer, but an enthusiastic novice can also produce the effect. Because the loudest sine wave components of the note you are singing match the natural resonant frequencies of one or more strings in the piano, the strings can be induced to vibrate sympathetically with the note you sing. The notion "natural resonant frequency" applies to the basilar membrane in the inner ear. The thick part naturally vibrates sympathetically with the low-frequency components of an incoming signal, while the thin part naturally vibrates sympathetically with the high-frequency components.

Table 3.1 A comparison of the typical maximum pressure fluctuations (given in both micro-Pascals (μPa), and decibels SPL) of some common sounds.

Typical experience	*Pressure (μPa)*	*Decibel level (SPL)*
Absolute threshold	20	0
Faint whisper	200	20
Quiet office	2,000	40
Conversation	20,000	60
City bus	200,000	80
Subway train	2,000,000	100
Loud thunder	20,000,000	120
Pain and damage	200,000,000	140

another task, until the first was half as loud as the second. Listeners' responses were converted into a scale of subjective loudness, the units of which are called "sones." The sone scale is plotted with open squares in figure 3.2. The left vertical axis of the graph shows listeners' judgments of relative loudness, scaled so that a sound about as loud as a quiet office (2,000 μPa) has a value of 1, a sound that is subjectively half as loud has a value of 0.5, and one that is twice as loud has a value of 2. As is clear in the figure, the relationship between sound pressure and

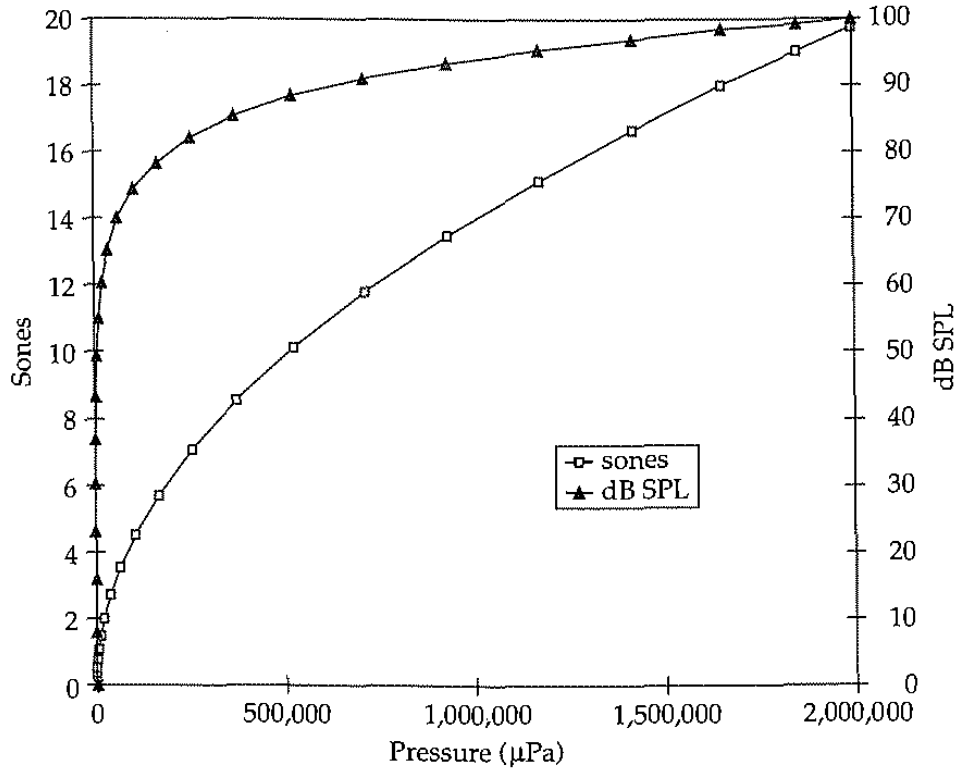

Figure 3.2 The nonlinearity of loudness perception is illustrated in this figure by plotting several sound pressure levels on both the sone scale and the dB scale.

loudness is not linear. For soft sounds, large changes in perceived loudness result from relatively small changes in sound pressure (the left end of the function is quite steep), while for loud sounds, relatively large pressure changes produce only small changes in perceived loudness. For example, if peak amplitude changes from 100,000 μPa to 200,000 μPa, the change in sones is greater than 2, but a change of the same pressure magnitude from 2,000,000 μPa to 2,100,000 μPa produces less than a 1 sone change in loudness.

Figure 3.2 also shows an older relative loudness scale that is named after Alexander Graham Bell. This unit of loudness, the bel, is too big for most purposes, and it is more common to use tenths of a bel, or decibels (abbreviated dB). This easily calculated scale is widely used in auditory phonetics and psychoacoustics, because it provides an approximation to the nonlinearity of human loudness sensation.

As the difference between dB SPL and dB SL implies, perceived loudness varies as a function of frequency. Figure 3.3 illustrates the relationship between subjective loudness and dB SPL. The curve in the figure represents the intensities of a set of tones that have the same subjective loudness as a 1,000 Hz tone presented at 60 dB SPL. The curve is like the settings of a graphic equalizer on a stereo. The lever on the left side of the equalizer controls the relative amplitude

Decibels

Although it is common to express the amplitude of a sound wave in terms of pressure, or, once we have converted acoustic energy into electrical energy, in volts, the decibel scale is a way of expressing sound amplitude that is better correlated with perceived loudness. On this scale the relative loudness of a sound is measured in terms of sound intensity (which is proportional to the square of the amplitude) on a logarithmic scale.

Consider a sound with average pressure amplitude x. Because sound intensity is proportional to the square of amplitude, the intensity of x relative to a reference sound with pressure amplitude r is x^2/r^2. A bel is the base 10 logarithm of this power ratio: $\log_{10}(x^2/r^2)$, and a decibel is 10 times this: $10 \log_{10}(x^2/r^2)$. This formula can be simplified to $20 \log_{10}(x/r) = \text{dB}$.

There are two common choices for the reference level r in dB measurements. One is 20 μPa, the typical absolute auditory threshold (lowest audible pressure fluctuation) of a 1,000 Hz tone. When this reference value is used, the values are labeled dB SPL (for **S**ound **P**ressure **L**evel). The other common choice for the reference level has different reference pressure levels for each frequency. In this method, rather than use the absolute threshold for a 1,000 Hz tone as the reference for all frequencies, the loudness of a tone is measured relative to the typical absolute threshold level for a tone at that frequency. When this method is used, the values are labeled dB SL (for **S**ensation **L**evel).

In speech analysis programs, amplitude may be expressed in dB relative to the largest amplitude value that can be taken by a sample in the digital speech waveform, in which case the amplitude values are negative numbers; or it may be expressed in dB relative to the smallest amplitude value that can be represented in the digital speech waveform, in which case the amplitude values are positive numbers. These choices for the reference level in the dB calculation are used when it is not crucial to know the absolute dB SPL value of the signal. For instance, calibration is not needed for comparative RMS or spectral amplitude measurements.

of the lowest-frequency components in the music, while the lever on the right side controls the relative amplitude of the highest frequencies. This "equal loudness" contour shows that you have to amplify the lowest and highest frequencies if you want them to sound as loud as the middle frequencies (whether this sounds good is another issue). So, as the figure shows, the auditory system is most sensitive to sounds that have frequencies between 2 and 5 kHz. Note also that sensitivity drops off quickly above 10 kHz. This was part of my motivation in chapter 2 for choosing a sampling rate of 22 kHz (11 kHz Nyquist frequency) for acoustic/phonetic analysis.

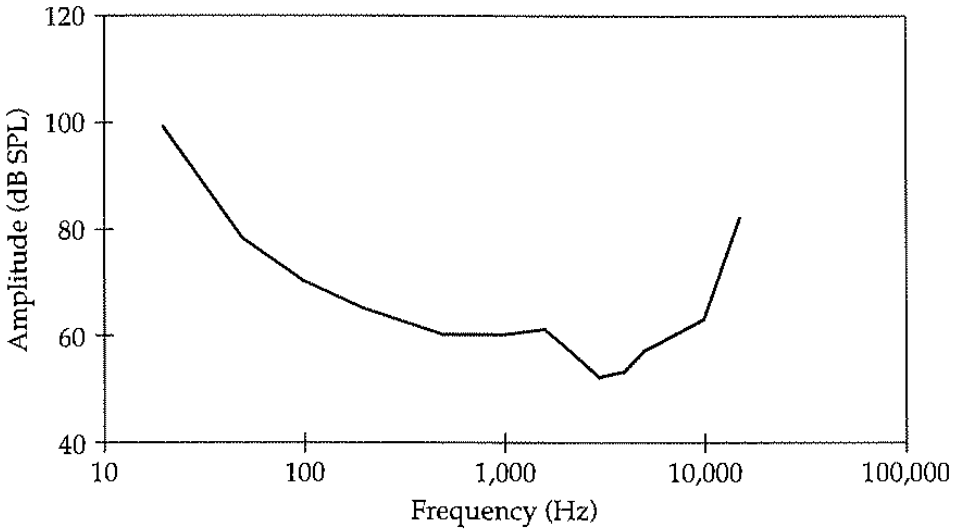

Figure 3.3 Equal loudness curve for pure tones presented over a loudspeaker. The line connects tones that have the same subjective loudness as a 1,000 Hz tone presented at 60 dB SPL.

3.3 Frequency response of the auditory system

As discussed in section 3.1, the auditory system performs a running Fourier analysis of incoming sounds. However, this physiological frequency analysis is not the same as the mathematical Fourier decomposition of signals. The main difference is that the auditory system's frequency response is not linear. Just as a change of 1,000 μPa in a soft sound is not perceptually equivalent to a similar change in a loud sound, so a change from 500 to 1,000 Hz is not perceptually equivalent to a change from 5,000 to 5,500 Hz. This is illustrated in figure 3.4, which shows the relationship between an auditory frequency scale called the Bark scale (Zwicker, 1961; Schroeder et al., 1979), and acoustic frequency in kHz. Zwicker (1975) showed that the Bark scale is proportional to a scale of perceived pitch (the Mel scale) and to distance along the basilar membrane. A tone with a frequency of 500 Hz has an auditory frequency of 4.9 Bark, while a tone of 1,000 Hz is 8.5 Bark, a difference of 3.6 Bark. On the other hand, a tone of 5,000 Hz has an auditory frequency of 19.2 Bark, while one of 5,500 Hz has an auditory frequency of 19.8 Bark, a difference of only 0.6 Bark. The line shown in figure 3.4 represents the fact that the auditory system is more sensitive to frequency changes at the low end of the audible frequency range than at the high end.

This nonlinearity in the sensation of frequency is related to the fact that the listener's experience of the pitch of periodic sounds and of the timbre of complex sounds is largely shaped by the physical structure of the basilar membrane. Figure 3.5 illustrates the relationship between frequency and location along the basilar membrane. As mentioned earlier, the basilar membrane is thin at its base and thick at its apex; as a result, the base of the basilar membrane responds to high-frequency sounds, and the apex to low-frequency sounds. As figure 3.5 shows, a relatively large portion of the basilar membrane responds to sounds

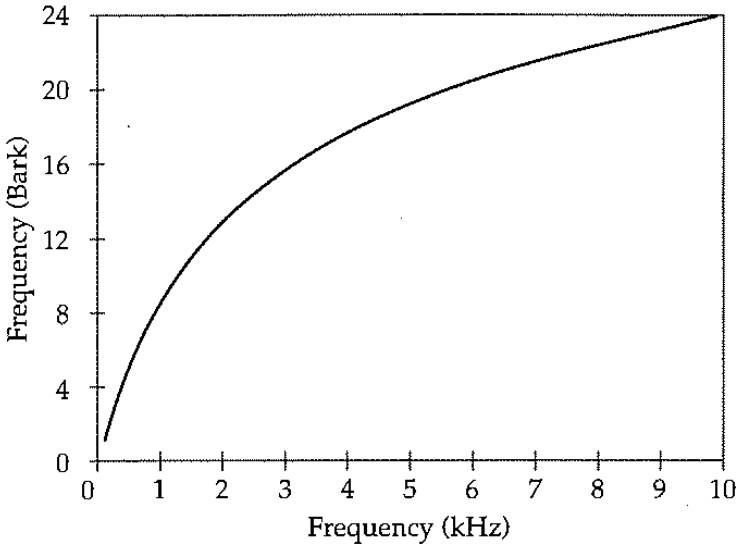

Figure 3.4 Comparison of an auditory frequency scale (the Bark scale) and an acoustic frequency scale (in kHz), showing that the auditory system is more sensitive to small changes in frequency at the low end of the audible range.

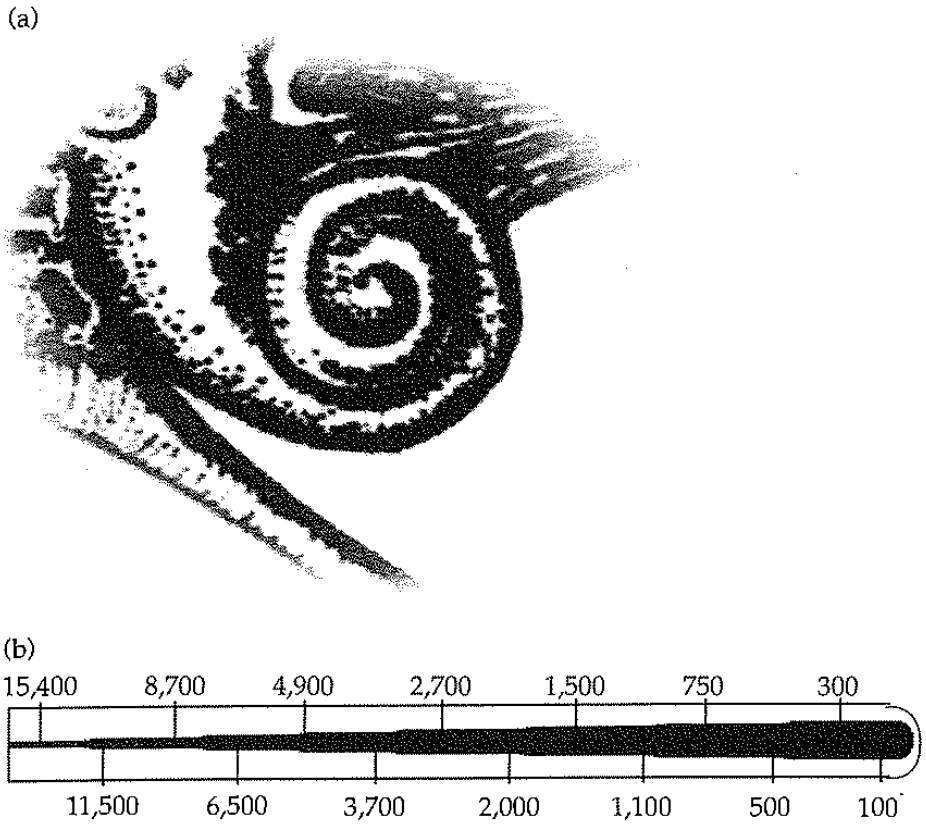

Figure 3.5 (a) shows the inner ear, (b) the approximate locations of sensitivity to representative frequencies along the basilar membrane when the inner ear is "uncoiled." See figure 3.1 to locate the inner ear in relation to other parts of the peripheral auditory system.

below 1,000 Hz, whereas only a small portion responds to sounds between 12,000 and 13,000 Hz, for example. Therefore, small changes in frequency below 1,000 Hz are more easily detected than are small changes in frequency above 12,000 Hz. The relationship between auditory frequency and acoustic frequency shown in figure 3.4 is due to the structure of the basilar membrane in the inner ear.

3.4 Auditory representations

In practical terms what all this means is that when we calculate an acoustic power spectrum of a speech sound, the frequency and loudness scales of the analyzing device (for instance, a computer or a spectrograph) are not the same as the auditory system's frequency and loudness scales. Consequently, acoustic analyses of speech sounds may not match the listener's experience. The resulting mismatch is especially dramatic for sounds like some stop release bursts and fricatives that have a lot of high-frequency energy. One way to avoid this mismatch between acoustic analysis and the listener's experience is to implement a functional model of the auditory system. Some examples of the use of auditory models in speech analysis are Liljencrants and Lindblom (1972), Bladon and Lindblom (1981), Johnson (1989), Lyons (1982), Patterson (1976), Moore and Glasberg (1983), and Seneff (1988). Figure 3.6 shows the difference between the auditory and

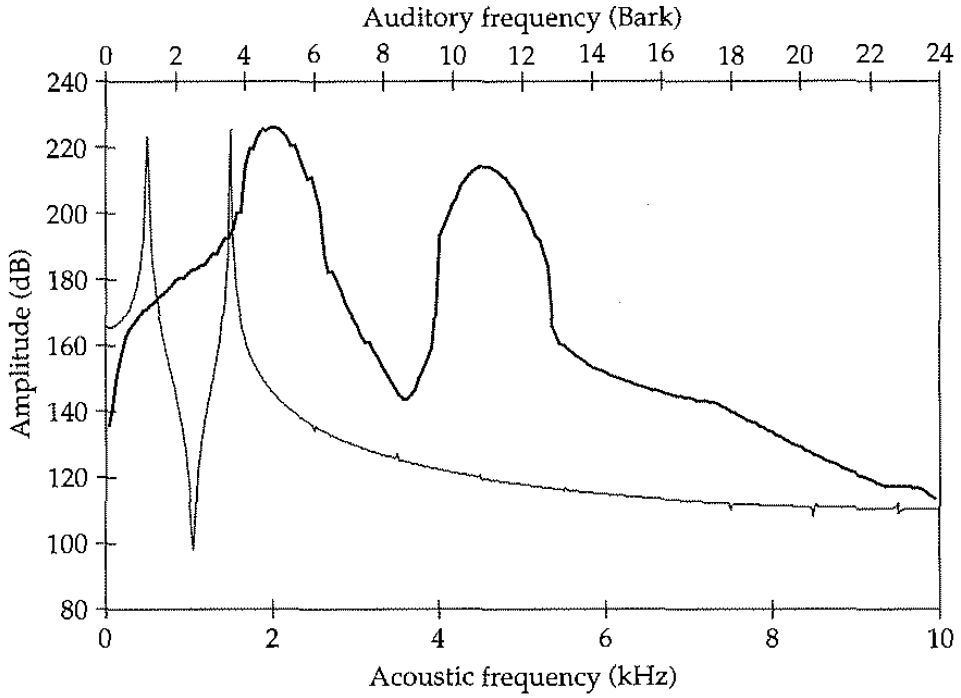

Figure 3.6 A comparison of acoustic (light line) and auditory (heavy line) spectra of a complex wave composed of sine waves at 500 at 1,500 Hz. Both spectra extend from 0 to 10 kHz, although on different frequency scales. The auditory spectrum was calculated from the acoustic spectrum using the model described in Johnson (1989).

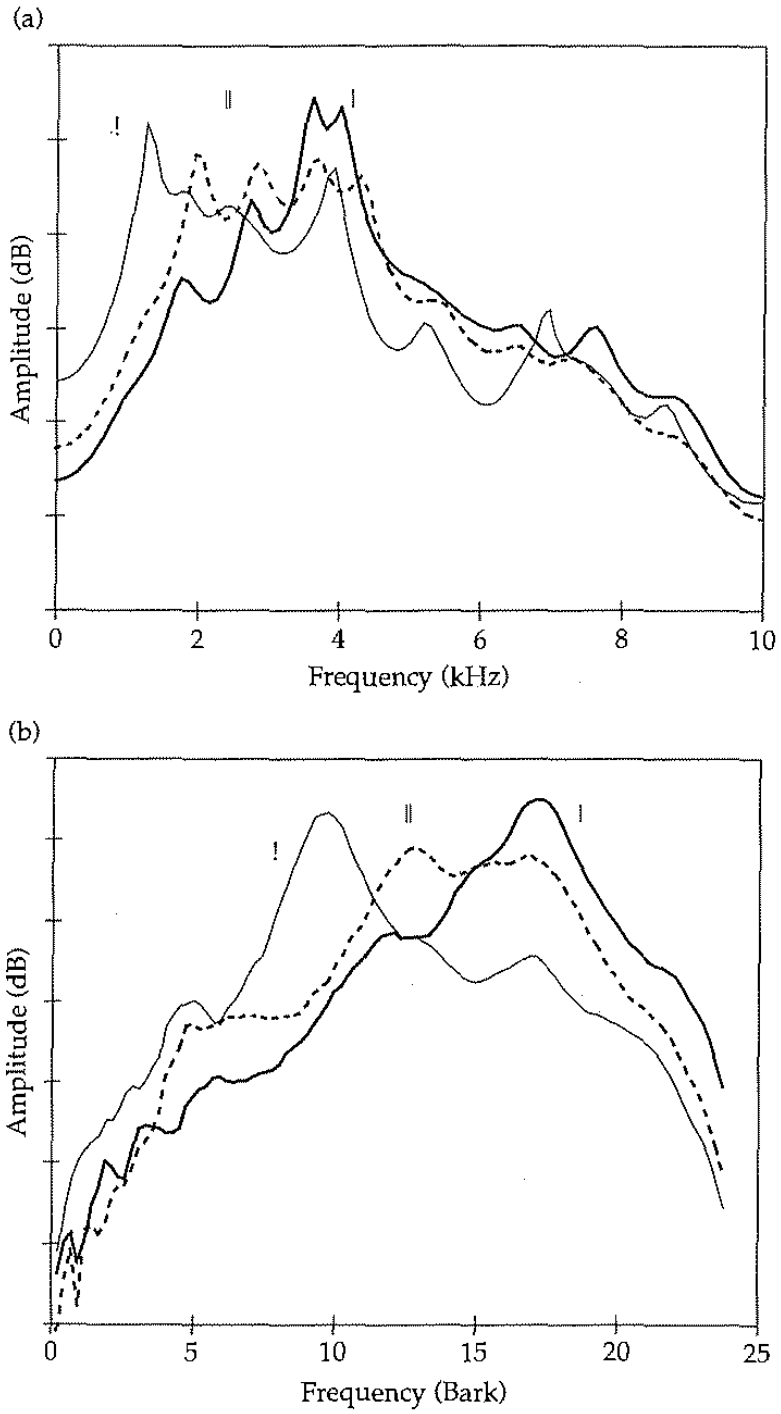

Figure 3.7 A comparison of acoustic power spectra (a) and auditory spectra (b) of three click release bursts in Xhosa. The IPA symbols are [|] dental, [!] post-alveolar, and [||] lateral. In the acoustic spectra the clicks differ from each other below 6 kHz. In the auditory spectra the nondistinctive frequencies above 6 kHz are compressed at the top end of the scale (above) 20 Bark (after Johnson, 1992). The interval between ticks on the vertical axes is 10 dB.

acoustic spectra of a complex wave composed of a 500 Hz and a 1,500 Hz sine wave component. The vertical axis is amplitude in dB, and the horizontal axis shows frequency in Hz, marked on the bottom of the graph, and Bark, marked on the top of the graph. I made this auditory spectrum, and others shown in later figures, with a computer program (Johnson, 1989) that mimics the frequency response characteristics shown in figure 3.4 and the equal loudness contour shown in figure 3.3. Notice that because the acoustic and auditory frequency scales are different, the peaks are located at different places in the two representations, even though both spectra cover the frequency range from 0 to 10,000 Hz. Almost half of the auditory frequency scale covers frequencies below 1,500 Hz, while this same range covers less than two-tenths of the acoustic display. So, low-frequency components tend to dominate the auditory spectrum. Notice too that in the auditory spectrum there is some frequency smearing that causes the peak at 11 Bark (1,500 Hz) to be somewhat broader than that at 5 Bark (500 Hz). This spectral-smearing effect increases as frequency increases.

Figure 3.7 shows an example of the difference between acoustic and auditory spectra of speech. The acoustic spectra of the release bursts of the clicks in Xhosa are shown in (a), while (b) shows the corresponding auditory spectra. Like figure 3.6, this figure shows several differences between acoustic and auditory spectra. First, the region between 6 and 10 kHz (20–4 Z in the auditory spectra), in which the clicks do not differ very much, is not very prominent in the auditory spectra. In the acoustic spectra this insignificant portion takes up two-fifths of the frequency scale, while it takes up only one-fifth of the auditory frequency scale. This serves to visually, and presumably auditory, enhance the differences between the spectra. Second, the auditory spectra show many fewer local peaks than do the acoustic spectra. In this regard it should be noted that the acoustic spectra shown in figure 3.7 were calculated using LPC analysis to smooth them; the FFT spectra which were input to the auditory model were much more complicated than these smooth LPC spectra. The smoothing evident in the auditory spectra on the other hand is due to the increased bandwidths of the auditory filters at high frequencies.

Auditory models are interesting, because they offer a way of looking at the speech signal from the point of view of the listener. The usefulness of auditory models in phonetics depends on the accuracy of the particular simulation of the peripheral auditory system. Therefore, the illustrations in this book were produced by models that implement only well-known, and extensively studied, nonlinearities in auditory loudness and frequency response, and avoid areas of knowledge that are less well understood for complicated signals like speech.

These rather conservative auditory representations suggest that acoustic analyses give only a rough approximation to the auditory representations that listeners use in identifying speech sounds.

Recall from chapter 2 that digital spectrograms are produced by encoding spectral amplitude in a series of FFT spectra as shades of gray in the spectrogram. This same method of presentation can also be used to produce auditory spectrograms from sequences of auditory spectra. Figure 3.8 shows an acoustic

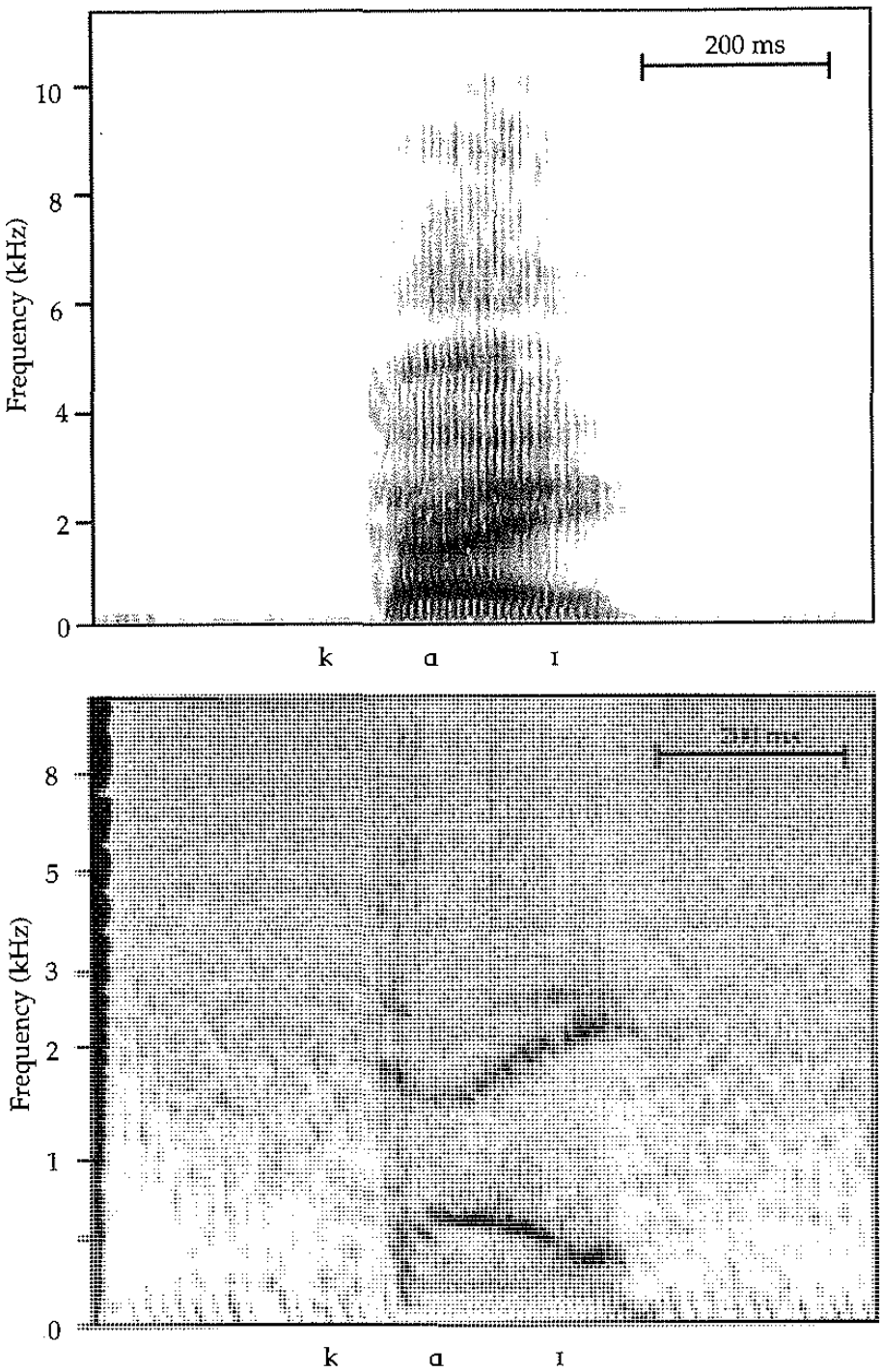

Figure 3.8 Comparison of a normal acoustic spectrogram (top), and an auditory spectrogram, or cochleagram (bottom), of the Cantonese word [kɑ˥] "chicken." The cochleagram was produced by Lyons's (1982) cochlear model.

spectrogram and an auditory spectrogram of the Cantonese word [kɑ˥] "chicken" (see figure 2.17). To produce this figure, I used a publicly available auditory model (Lyons's cochlear model (Lyons, 1982; Slaney, 1988), which can be found at: ftp://ftp.ling.ohio-state.edu/pub/phonetics/ear). The auditory spectrogram, which is also called a "cochleagram," combines features of auditory spectra and spectrograms. As in a spectrogram, the simulated auditory response is represented with spectral amplitude plotted as shades of gray, with time on the horizontal axis and frequency on the vertical axis. Note that although the same frequency range (0–11 kHz) is covered in both displays, the movements of the lowest concentrations of spectral energy in the vowel are much more visually pronounced in the cochleagram because of the auditory frequency scale. The following chapters will make use of cochleagrams to explore the auditory properties of various classes of speech sounds.

Exercises

Sufficient jargon

Define the following terms: peripheral auditory system, eardrum, middle ear, inner ear, cochlea, basilar membrane, natural resonant frequency, sone, decibel, dB SPL, dB SL, equal loudness contour, Bark scale, auditory model, auditory spectrum, cochleagram.

Short-answer questions

1 Refering to figure 3.2, what is the sound pressure level (in μPa) of a sound that is 6 sones loud? What is the sound pressure level of a sound that is subjectively twice as loud as a 6 sone sound?
2 Refering to figure 3.3, which is subjectively louder: a 100 Hz sine wave at 65 dB SPL or a 1 000 Hz sine wave at 60 dB SPL, a 15,000 Hz sine wave at 80 dB SPL or a 2,000 Hz sine wave at 70 dB SPL, a 500 Hz sine wave at 60 dB SL or a 9,000 Hz sine wave at 60 dB SL?
3 Refering to figure 3.4, what is the auditory frequency in Bark of each of the following sine waves: 7 kHz, 8 kHz, 1 kHz, 2 kHz? Which is the bigger difference in auditory frequency, 7 kHz to 8 kHz or 1 kHz to 2 kHz?
4 Plot the relationship between distance traveled by sounds and their subjective loudnesses. Here is a set of sounds that you could use in this exercise: a card (like a business card, or a credit card) tapping on a table or car hood, a pen tapping on a table, a ruler tapping on a table, and a coffee mug tapping on a table. The idea is that the sounds will cover a managable range of loudness. You can use other sounds if you want, but choose minimally four things and choose things that are similar in time (like tapping).

   Do this exercise with a partner. First, both of you should independently rate the loudness of the sounds by filling in the following blanks with numbers (use fractions if you want).

   (a) The pen is————times louder than the card.
   (b) The ruler is————times louder than the pen.
   (c) The mug is————times louder than the ruler.

You can make a subjective loudness scale by setting the loudness of the card to 1, then set the loudness of the pen to 1 times the number given in (a), and the loudness of the ruler as the loudness of the pen times the number given in (b), and the loudness of the mug as the loudness of the ruler times the number given in (c). How different is your scale from your partner's scale?

Now measure loudness a different way. Find a hallway or an open yard or field. As one person makes the noise, the other should walk away and stop walking when he/she can no longer hear the noise. The noise-maker should count the number of steps the listener takes and mark this down for the sound. Repeat this for each sound and for both partners. How different is your scale from your partner's scale?

How do these two loudness scales compare with each other? Try plotting them the way sones are plotted against pressure in figure 3.2. Is the relationship linear? non-linear? Is one scale more like the sones scale and one more like the sound pressure scale? How would you convert the distance scale into sound pressure?

CHAPTER 4

Speech Perception

The perceptual representation of speech is different from its auditory representation. Acoustically, speech sounds are sound waves that are produced by vocal tracts in motion. Auditorily, they are shaped by the nonlinearities of the auditory system. Perceptually, speech sounds are no longer even limited to audition – the visual image of the talker plays a role in perception – and perception takes place in the context of a lifetime of experience with language.

The acoustic descriptions which form the primary topic of this book open the way to some very important insights about language. But as we saw in the previous chapter, the ear is not a high fidelity transmission system. The mental representation of speech has been warped by the auditory system. So, we will pay attention to the auditory mapping of speech. But now, in this chapter, we will consider the ways that the brain codes speech. This chapter has two main points. First, I hope to convince you that it is relatively easy to study speech perception; so easy, in fact, that you can start studying speech perception right away. Second, I hope to show that the perceptual representations of speech are built from auditory input **plus** other information.

4.1 A speech perception experiment

The study of speech perception starts with confusion. Maybe all scientific endeavors do – why else conduct research? But speech perception research literally starts with confusion because in the typical experiment we play sounds to people and ask them what they heard and hope that they will sometimes get the wrong answer. We analyze the pattern of wrong answers, the confusions, to determine the structure of the "perceptual space" used by listeners.

The next few pages will walk you through the steps of a speech perception experiment. Here's a preview of the steps:

1 Decide what to study – select examples to use in an experiment
2 Record multiple repetitions of each test syllable (call these tokens)
3 Digitize the tokens and make a separate sound file for each one
4 Add noise to the tokens
5 Make an answer sheet
6 Put the tokens in random order
7 Play the tokens in random order onto tape
8 Play the test tape to listeners
9 Tabulate the responses into a confusion matrix

Experimental paradigms

There is no end to how clever people can be when it comes to devising listening experiments. Things can get really complicated as we attempt to control our listeners' thought patterns during the course of the experiment. We don't want them to guess the purpose of the experiment so we embed the "test trials" in a large number of "filler trials." Or, we try to give very detailed specific instructions so that listeners will approach the task in a particular way. I have found that listeners tend to be influenced by the structure of the experiment and by the structure of the stimuli much more than by the instructions that you give (see, for example, the failed instruction manipulations in Johnson, Flemming and Wright, 1993). In speech perception experiments this dependence on stimuli and experiment design can be so complete that the listener can know exactly what the experiment is for, what the expected results are, and even have lots of practice listening to the stimuli and still produce results that are practically the same as listeners who have no idea about these things and don't care either. I find that my own listening responses are often very much like those of "naive" subjects.

When it comes to describing the experimental procedures that you might use in a speech perception experiment the options listed below may be helpful. The experiment that we are describing in this section uses naturally produced stimuli (just a recording of someone saying nonsense syllables) that have been modified by adding white noise. The listener's task is to identify the initial phoneme of each test syllable. In a discrimination task we might ask listeners to hear a pair of stimuli and tell whether the two presented syllables are the same or different. Or in a matching task we might present three stimuli in a row and ask if the middle one is more like the first one or the last one. In most experiments we are simply interested in the response choices that people make, but sometimes we also measure how long it takes to make that choice. Finally, experiments can be characterized according to the group of people who

participate as listeners – whether we are studying hearing deficits or cross-linguistic perception.

Stimuli:	Measure:
natural	response choice
modified natural	response time
synthesized	
Listener's task:	Subject population:
identification	hearing ability
discrimination	talker's language
matching	listener's language

The first step is to decide what to study. I'm interested in why English [ð] is sometimes pronounced as [z] and sometimes as [d] by speakers who learn English as a second (or third, or fourth) language. I notice this especially with the function words "this" and "that" [zɪs] and [zæt] or [dɪs] and [dæt]. This variation also occurs across some dialects of English. I'm also personally interested in a similar phenomenon with [θ] and [f] – as for example when my niece Erin decided that my name is [kif].

So, that's the general context for a study of speech sounds. In particular though, let's hypothesize that perceptual confusability underlies patterns of substitution involving fricatives in English. This hypothesis predicts that the consonants that substitute for each other (e.g. [ð]/[z]/[d] and [θ]/[f]) are perceptually similar to each other. So now we need to do an experiment to test this hypothesis.

Here's how it works. First, we record someone saying some test syllables. These can be monosyllabic nonsense words. To start with we can record syllables with initial fricative sounds (and the stop [d] – to test the "dis" and "dat" hypothesis). For example, let's choose [fɑ], [vɑ], [sɑ], [zɑ], [θɑ], [ðɑ], and [dɑ]. We'll have our speaker pronounce each test syllable ten times so that we don't end up relying on one particular, possibly weird, pronunciation of the test syllable. Then we digitize them into a computer and make separate sound files (e.g. fa1.wav, va1.wav, sa1.wav, za1.wav, tha1.wav, dha1.wav, and da1.wav, fa2.wav . . .) so that we end up with 70 sound files – seven test syllables with ten renditions of each.

Now we add some noise to copies of the sound files so that we have fa1_noise.wav, sa1_noise.wav, etc. It is pretty easy to add noise to a sound file because many waveform editors have an editing function called "add noise." You can just play around with the add noise function and find a level so that the noise is subjectively about as loud as the speech signal. When the noise has the same loudness as the speech the signal to noise ratio (SNR) will be at about 0 dB. If you want to be more accurate about signal to noise ratio, you can measure loudness in earphones with a loudness level meter (available at your local stereo

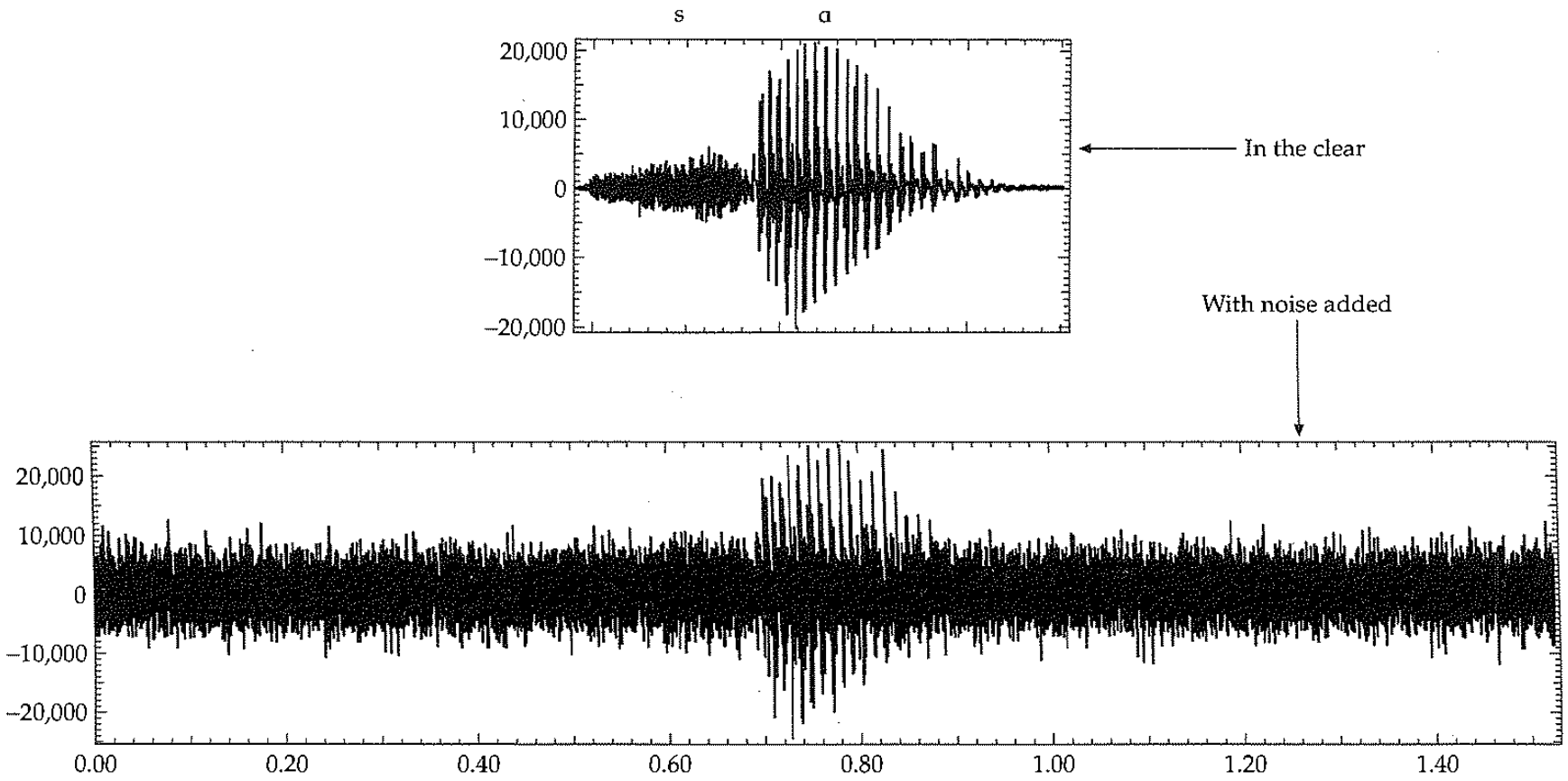

Figure 4.1 The test syllable [sɑ] with no noise added, which is also called presentation "in the clear" (top waveform), and embedded in white noise at a signal to noise ratio of 0 dB (bottom waveform). This figure illustrates that SNR is based on sound power and not on peak amplitude.

shop). By the way, before adding noise you'll want to add a short interval of silence before and after the test syllable so that the test syllable will be embedded in noise. A half a second (500 ms) before and after the test syllable is enough. Figure 4.1 shows one of the test syllables in the clear and with noise added.

Ok, so we have test syllables that differ by only their initial consonant, and we have embedded these in noise. Now it is time to present them to listeners, right? Well, not quite. First, let's make a test tape that has all of the test syllables on it. We'll want to present the test syllables in random order so that listeners have to listen closely to each token. You can do this pretty easily by assigning a random number to each of the 70 test syllables (using a random number function in a spreadsheet program, or by using a random number table from the back of a statistics text book) and then sorting the syllables by their random numbers from smallest random number to largest, or the other way if you want. To make the test "tape" you could construct a long computer sound file that contains the test tokens in random order and silent intervals inserted between the sound files. Then you could just play the sound file from the computer without bothering with a tape recorder. Or you could make the tape by playing sound files out to a tape recorder. This involves a procedure with the following steps – press record, play a file, let the recorder run while you count to three, pause the recorder, load the next file, release the pause, play the file, count to three, pause the tape, and so on. Tedious but low tech. In either case (tape or computer), be sure to leave enough time to complete the task for each token, and leave longer gaps at regular intervals, for example every ten tokens, so that people won't get lost on the answer sheet.

So, of course, we will need an answer sheet for our listeners to use. Here are the first four lines that I would use for the fricative test for American English listeners:

1 fa va tha dha sa za da
2 fa va tha dha sa za da
3 fa va tha dha sa za da
4 fa va tha dha sa za da

The listener's instruction is to listen to the first token and circle the correct answer in row 1 of the answer sheet, then listen to the second token and circle the correct answer in row 2, and so on.

There's an obvious problem isn't there? The test syllables include two sounds that aren't distinguished in English orthography [θ] and [ð]. If I am going to go through with this test with American English listeners I would have to try to teach them about these two sounds, that they will be written with "th" and "dh" on the answer sheet, and practice with them until they get it. We'll come back to this problem later in the chapter.

Let's say that we present each one of our 70 test syllables to ten listeners. This would result in 700 identification responses. To analyze these responses we will tabulate them in a **confusion matrix.** Each row in the matrix corresponds to one

Table 4.1 Fricative (and [d]) confusions from Miller and Nicely (1955).

	"f"	*"v"*	*"th"*	*"dh"*	*"s"*	*"z"*	*"d"*	*Other*	*Total*
[f]	199	0	46	1	4	0	0	14	264
[v]	3	177	1	29	0	4	0	22	236
[θ]	85	2	114	0	10	0	0	21	232
[ð]	0	64	0	105	0	18	0	17	204
[s]	5	0	38	0	170	0	0	15	228
[z]	0	4	0	22	0	132	17	49	224
[d]	0	0	0	4	0	8	189	59	260

of the test syllables (collapsing across all ten tokens of that syllable) and each column in the matrix corresponds to one of the responses available to listeners.

I didn't actually conduct this experiment, but table 4.1 shows the confusion matrix for the 0 dB SNR condition in a larger study conducted by George Miller and Patricia Nicely in 1955. Yep, these data are old, but they're good. Looking at the first row of the confusion matrix we see that [f] was presented 264 times and correctly identified as "f" 199 times and as "th" 46 times. Note that Miller and Nicely have more data for some sounds than for others.

That's it! We designed and constructed a speech perception experiment and now we have the data all nicely laid out in a confusion matrix. We can even get some pretty quick answers out of the confusion matrix. For example, why is it that "Keith" is sometimes pronounced "Keif" by children? Well, according to Miller and Nicely's data, [θ] was called "f" 85 times out of 232 – it was confused with "f" more often than with any other speech sound tested. Cool. But it isn't clear that these data tell us anything at all about the realizations of "this" and "that." To address that question we need to find a way to map the perceptual "space" that underlies the confusions we observe in our experiment. It is to this mapping problem we now turn.

4.2 Maps from distances

So, we're trying to pull information out of a confusion matrix to get a picture of the perceptual system that caused the confusions. The strategy that we will use takes a list of distances and reconstructs them as a map. Consider for example the list of distances below for cities in Ohio.

Columbus to Cincinnati, 107 miles
Columbus to Cleveland, 142 miles
Cincinnati to Cleveland, 249 miles

From these distances we can put these cities on a straight line as in figure 4.2(a), with Columbus located between Cleveland and Cincinnati. A line works to describe

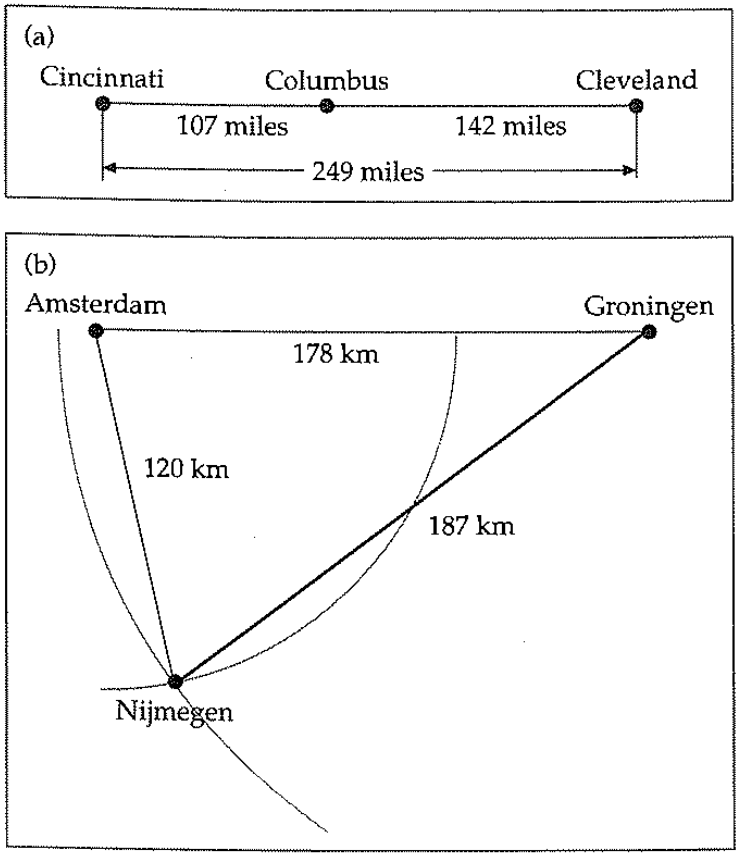

Figure 4.2 (a) A one dimensional map showing the relative locations of three cities in Ohio. (b) A two dimensional map showing the relative locations of three Dutch cities. The gray arcs illustrate how we place Nijmegen using the procedure of triangulation.

these distances because the distance from Cincinnati to Cleveland is simply the sum of the other two distances (107 + 142 = 249). Here's an example that requires a two dimensional plane.

Amsterdam to Groningen, 178 km
Amsterdam to Nijmegen, 120 km
Groningen to Nijmegen, 187 km

The two-dimensional map that plots the distances between these cities in The Netherlands is shown in figure 4.2(b). To produce this figure I put Amsterdam and Groningen on a line and called the distance between them 178 km. Then I drew an arc 120 km from Amsterdam, knowing that Nijmegen has to be somewhere on this arc. Then I drew an arc 187 km from Groningen, knowing that Nijmegen also has to be somewhere on this arc. So, Nijmegen has to be at the intersection of the two arcs – 120 km from Amsterdam and 187 km from Groningen. This method of locating a third point based on its distance from two known points is called triangulation. The triangle shown in figure 4.2(b) is an

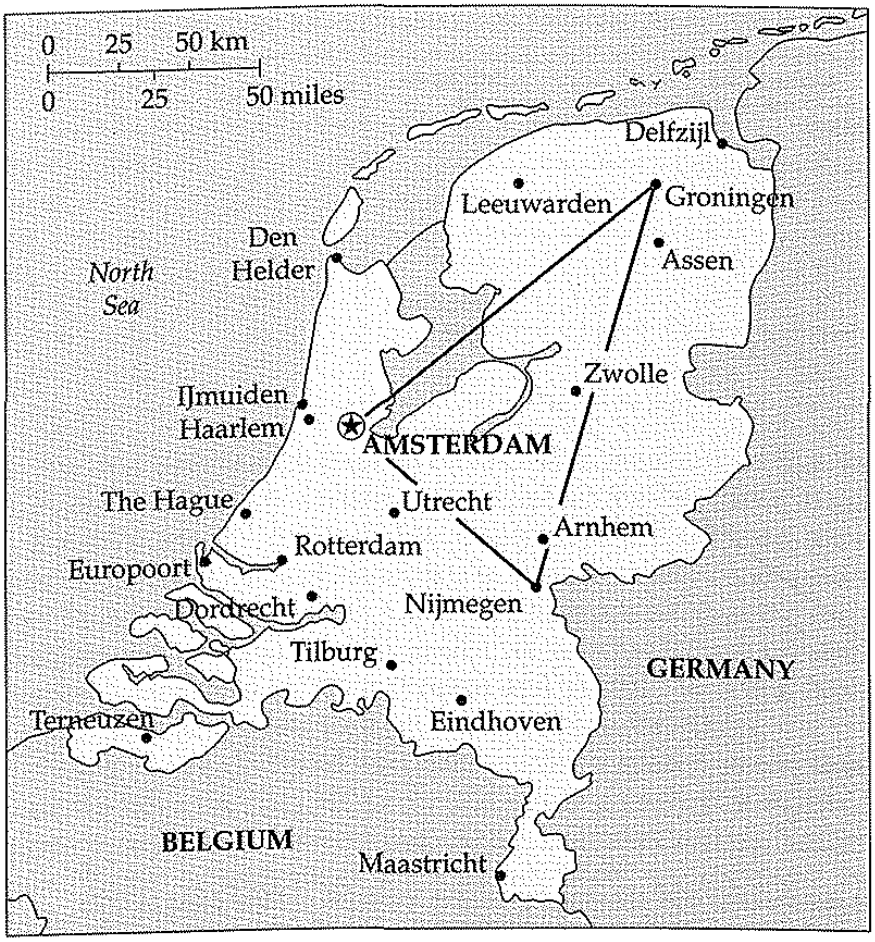

Figure 4.3 A map of the Netherlands showing the orientation of the Amsterdam–Groningen–Nijmegen triangle derived in figure 4.2(b).

accurate depiction of the relative locations of these three cities as you can see in the map in figure 4.3.

You might be thinking to yourself, "Well, this is all very nice, but what does it have to do with speech perception?" Good question. It turns out that we can compute perceptual distances from a confusion matrix. And by using an extension of triangulation called multi-dimensional scaling, we can produce a perceptual map from a confusion matrix.

4.3 The perceptual map of fricatives

In this section we will use multidimensional scaling to map the perceptual space that caused the confusion pattern in table 4.1.

The first step in this analysis process is to convert confusions into distances. We believe that this is a reasonable thing to try to do because we assume that when things are close to each other in perceptual space they will get confused

with each other in the identification task. So the errors in the matrix in table 4.1 tell us what gets confused with what. Notice, for example, that the voiced consonants [v], [ð], [z] and [d] are very rarely confused with the voiceless consonants [f], [θ] and [s]. This suggests that voiced consonants are close to each other in perceptual space while voiceless consonants occupy some other region. Generalized statements like this are all well and good, but we need to compute some specific estimates of perceptual distance from the confusion matrix.

Here's one way to do it (I'm using the method suggested by the mathematical psychologist Roger Shepard in his important 1972 paper "Psychological representation of speech sounds"). There are two steps. First, calculate similarity and then from the similarities we can derive distances.

Similarity is easy. The number of times that you think [f] sounds like "θ" is a reflection of the similarity of "f" and "θ" in your perceptual space. Also, "f"–"θ" similarity is reflected by the number of times you say that [θ] sounds like "f", so we will combine these two cells in the confusion matrix – [f] heard as "θ" and [θ] heard as "f." Actually, since there may be a different number of [f] and [θ] tokens presented, we will take proportions rather than raw counts.

Notice that for any two items in the matrix we have a submatrix of four cells: (a) is the submatrix of response proportions for the "f"/"θ" contrast from Miller and Nicely's data. Note for example that the value 0.75 in this table is the proportion of [f] tokens that were recognized as "f" (199/264 = 0.754). Listed with the submatrix are two abstractions from it.

(a)

	"f"	"θ"
[f]	0.75	0.17
[θ]	0.37	0.49

(b)

	"f"	"f"
[f]	p_{ff}	$p_{f\theta}$
[θ]	$p_{\theta f}$	$p_{\theta\theta}$

(c)

	"i"	"j"
[i]	p_{ii}	p_{ij}
[j]	p_{ji}	p_{jj}

The variables in submatrix (b) code the proportions so that "p" stands for proportion, the first subscript letter stands for the row label and the second subscript letter stands for the column label. So $p_{\theta f}$ is a variable that refers to the proportion of times that [θ] tokens were called "f". In these data $p_{\theta f}$ is equal to 0.37. Submatrix (c) abstracts this a little further to say that for any two sounds i and j, we have a submatrix with confusions (subscripts don't match) and correct answers (subscripts match).

Asymmetry in confusion matrices

Is there some deep significance in the fact that [θ] is called "f" more often than [f] is called "th"? It may be that listeners had a bias against calling things "th" – perhaps because it was confusing to have to distinguish between "th" and "dh" on the answer sheet. This would seem to be the case in table 4.1 because there are many more "f" responses than "th" responses overall. However, the

relative infrequency of "s" responses suggests that we may not want to rely too heavily on a response bias explanation because the "s" to [s] mapping is common and unambiguous in English. Shepard's method for calculating similarity from a confusion matrix assumes that $p_{f\theta}$ and $p_{\theta f}$ are two imperfect measures of the same thing – the confusability of "f" and "θ." These two estimates are thus combined to form one estimate of "f"–"θ" similarity. This is not to deny that there might be something interesting to look at in the asymmetry, but only that for the purpose of making perceptual maps the sources of asymmetry in the confusion matrix are ignored.

Here is Shepard's method for calculating similarity from a confusion matrix. We take the confusions between the two sounds and scale them by the correct responses. In math, that's:

$$S_{ij} = \frac{p_{ij} + p_{ji}}{p_{ii} + p_{jj}} \tag{4.1}$$

In this formula, S_{ij} is the similarity between category i and category j. In the case of "f" and "θ" in Miller and Nicely's data (table 4.1) the calculation is:

$$S_{ij} = 0.43 = \frac{0.17 + 0.37}{0.75 + 0.49}$$

I should say that regarding this formula Shepard simply says that it "has been found serviceable." Sometimes you can get about the same results by simply taking the average of the two confusion proportions p_{ij} and p_{ji} as your measure of similarity, but Shepard's formula does a better job with a confusion matrix in which one category has confusions concentrated between two particular responses, while another category has confusions fairly widely distributed among possible responses – as might happen, for example, when there is a bias against using one particular response alternative.

OK, so that's how to get a similarity estimate from a confusion matrix. To get perceptual distance from similarity you simply take the negative of the natural log of the similarity. Ahhem, in math that is:

$$d_{ij} = -\ln(S_{ij}) \tag{4.2}$$

This is based on Shepard's Law, which states that the relationship between perceptual distance and similarity is exponential. There may be a deep truth about mental processing in this law – it comes up in all sorts of unrelated contexts (Shannon and Weaver, 1949; Parzen, 1962), but that's a different topic.

Table 4.2 Similarities among American English fricatives (and [d]), based on the 0 dB SNR confusion matrix from Miller and Nicely (1955).

	"f"	*"v"*	*"th"*	*"dh"*	*"s"*	*"z"*	*"d"*
[f]	1.0						
[v]	.008	1.0					
[θ]	.434	.010	1.0				
[ð]	.003	.345	.000	1.0			
[s]	.025	.000	.170	.000	1.0		
[z]	.000	.026	.000	.169	.000	1.0	
[d]	.000	.000	.000	.012	.000	.081	1.0

Anyway, now we're back to map-making, except instead of mapping the relative locations of Dutch cities in geographic space, we're ready to map the perceptual space of English fricatives and "d." Table 4.2 shows the similarities calculated from the Miller and Nicely confusion matrix (table 4.1) using equation (4.1).

The perceptual map based on these similarities is shown in figure 4.4. One of the first things to notice about this map is that the voiced consonants are on one side and the voiceless consonants are on the other. This captures the observation that we made earlier, looking at the raw confusions, that voiceless sounds were rarely called voiced, and vice versa. It is also interesting that the voiced and voiceless fricatives are ordered in the same way on the vertical axis. This might be a front/back dimension, or there might be an interesting correlation with some acoustic aspect of the sounds.

In figure 4.4, I drew ovals around some clusters of sounds. These show two levels of similarity among the sounds as revealed by a hierarchical cluster ana-

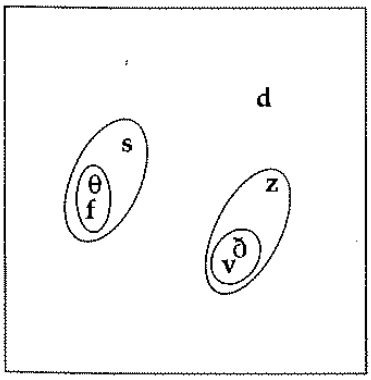

Figure 4.4 The perceptual map of fricatives and [d] in English. The location of the points was determined by multidimensional scaling of the confusion data from Miller and Nicely (1955). The circled groups of sounds are clusters that were found in a hierarchical cluster analysis of the same data.

MDS and acoustic phonetics

In acoustic phonetics one of our fundamental puzzles has been how to decide which aspects of the acoustic speech signal are important and which things don't matter. You look at a spectrogram and see a blob – the question is, do listeners care whether that part of the sound is there? Does that blob matter? Phoneticians have approached the "does it matter?" problem in a number of ways.

For example, we have looked at lots of spectrograms and asked concerning the mysterious blob – "is it always there?" One of the established facts of phonetics is that if an acoustic feature is always, or even usually, present then listeners will expect it in perception. This is even true of the so-called "spit spikes" in the lateral fricatives [ɬ] and [ɮ]. These sounds get a bit juicy, but this somewhat tangential aspect of their production seems to be useful in perception.

Another answer to "does it matter?" has been to identify the origin of the blob in the acoustic theory of speech production. For example, sometimes room reverberation can "add" shadows to a spectrogram. (Actually in the days of reel-to-reel tape recorders we had to be careful of magnetic shadows that crop up when the magnetic sound image transfers across layers of tape on the reel.) If you have a theory of the relationship between speech production and speech acoustics you can answer the question by saying, "it doesn't matter because the talker didn't produce it." We'll be exploring the acoustic theory of speech production in some depth in the remaining chapters of this book.

One of my favorite answers to "does it matter?" is "Cooper's rule." Franklin Cooper, in his 1951 paper with Al Liberman and John Borst, commented on the problem of discovering "the acoustic correlates of perceived speech." They claimed that there are "many questions about the relation between acoustic stimulus and auditory perception which cannot be answered merely by an inspection of spectrograms, no matter how numerous and varied these might be" (an important point for speech technologists to consider). Instead they suggested that "it will often be necessary to make controlled modifications in the spectrogram, and then to evaluate the effects of these modifications on the sound as heard. For these purposes we have constructed an instrument . . ." (one of the first speech synthesizers). This is a pretty beautiful direct answer. Does that blob matter? Well, leave it out when you synthesize the utterance and see if it sounds like something else.

And finally there is the MDS answer. We map the perceptual space and then look for correlations between dimensions of the map and acoustic properties of interest (like the mysterious blob). If an acoustic feature is tightly correlated with a perceptual dimension then we can say that that feature probably does matter. This approach has the advantages of being based on naturally produced speech, and of allowing the simultaneous exploration of many acoustic parameters.

lysis (another neat data analysis method available in most statistics software packages). At the first level of clustering "θ" and "f" cluster with each other and "v" and "ð" cluster together in the perceptual map. At a somewhat more inclusive level the sibilants are included with their non-sibilant neighbors ("s" joins the voiceless cluster and "z" joins the voiced cluster). The next level of clustering, not shown in the figure, puts [d] with the voiced fricatives.

Combining cluster analysis with multi-dimensional scaling gives us a pretty clear view of the perceptual map. Note that these techniques are largely just data visualization techniques, we did not add any information to what was already in the confusion matrix (though we did decide that a two dimensional space adequately describes the pattern of confusions for these sounds).

Concerning the realizations of "this" and "that" we would have to say that these results indicate that the alternations [ð]–[d] and [ð]–[z] are not driven by auditory/perceptual similarity alone – there are evidently other factors at work – otherwise we would find "vis" and "vat" as realizations of "this" and "that."

4.4 The perceptual map of [place]

Now I'd like to talk about how perceptual maps are different from auditory maps. To make this point let's look at some data reported by Braida, Sekiyama and Dix (1998). They presented audio recordings of test syllables that differed by place of articulation (as in [bɑ], [dɑ], and [gɑ]) to a group of Japanese listeners. The tokens were presented in noise, and as table 4.3 shows, listeners made a substantial number of mistakes.

Using Shepard's formulas (equations 4.1 and 4.2) for similarity and distance, we have the following estimated similarities and distances for "b," "d," and "g" in the Braida et al. data:

	S_{ij}	d_{ij}
b–d	.57	.56
b–g	.41	.89
d–g	.62	.47

Table 4.3 [Place] confusions for auditory tokens presented in noise, from Braida, Sekiyama and Dix (1998). The table shows response proportions. So, for example, only 56 percent of the [b] tokens were correctly identified as "b."

	"b"	*"d"*	*"g"*
[b]	0.56	0.28	0.15
[d]	0.30	0.46	0.24
[g]	0.25	0.31	0.43

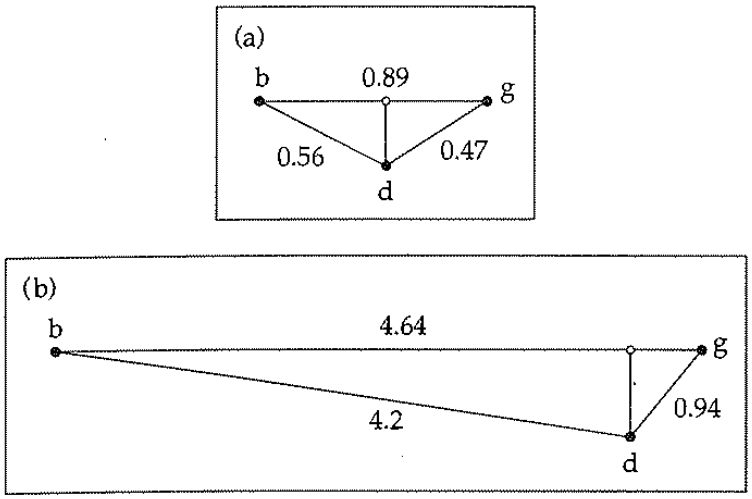

Figure 4.5 Two perceptual maps of place of articulation in Japanese. The data come from Braida, Sekiyama and Dix (1998). (a) A perceptual map derived from the auditory confusion matrix in table 4.3. (b) A perceptual map derived from the visual confusion matrix in table 4.4. In both maps, the position of [d] on the [b]–[g] axis (for a one dimension map) is shown by an open circle.

From these estimates of distance we can draw a perceptual map using triangulation, just like we did when we made part of the geographic map of Holland above. This map of the auditory perceptual space for [b], [d] and [g] is shown in figure 4.5(a). One thing that I've done in this map is to show what it would look like if we "projected" the two dimensional space onto one dimension. This would give us the line from "b" to "g" with "d" located where the open circle is drawn. This one-dimensional projection is not as accurate as the two dimensional map, but below we will find that it is a useful approximation to the auditory perceptual space.

What's interesting about the Braida, Sekiyama and Dix paper is that they also collected visual confusion data for [place]. Here's how to collect visual confusion data. Instead of just making an audio recording of someone saying your test syllables you make a video. Then you can digitize the video and use a digital video editor to chop out little movie clips of the person saying your test words. When you present the video tokens to people for identification responses you turn the volume of the TV or computer all the way down. So they just see a face talking, but can't hear anything.

The Braida et al. visual [place] confusion matrix is shown in table 4.4.

As you can see in the confusion matrix, people are very good at the visual identification of labials, but tend to confuse coronals and dorsals. This pattern of confusions leads to a perceptual map (figure 4.5(b)) in which "b" is quite distant from "d" and "g." It is interesting that "d," the more front of the two non-labials, is a little closer to "b."

As with figure 4.5(a) I have indicated in figure 4.5(b) how "d" projects onto the "b"–"g" line for a one-dimensional version of the visual perceptual space.

Table 4.4 The visual confusion matrix from Braida, Sekiyama and Dix (1998). As in table 4.2, the data are presented as proportions.

	"b"	*"d"*	*"g"*
[b]	0.98	0.02	0.004
[d]	0.01	0.85	0.13
[g]	0.01	0.43	0.56

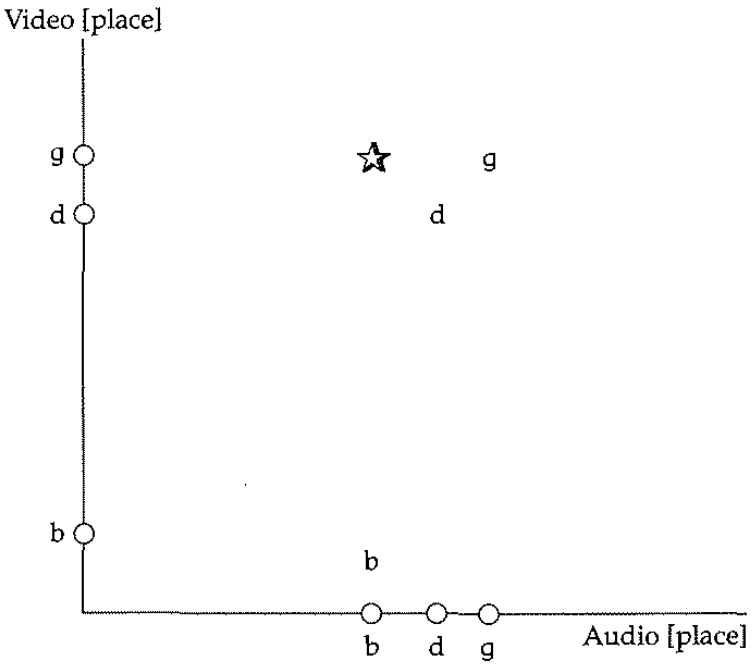

Figure 4.6 The audio-visual perceptual space for consonant place of articulation in Japanese. The horizontal axis shows audio [place] which is the one dimensional projection of the map in figure 4.5(a). The vertical axis shows video [place] which is the one dimensional projection of the map in figure 4.5(b).

Figure 4.6 shows why I wanted to represent these data with one audio dimension and one video dimension (this figure is very much like Braida et al.'s figure 1). In figure 4.6 the audio [place] dimension is on the horizontal axis and the video [place] dimension is on the vertical axis. The phonetic symbols plotted in this two-dimensional map thus show a hypothetical audio-visual perceptual map of [place] in Japanese. It turns out that the actual AV perceptual map (based on perceptual confusions of movies with the sound turned up) is very similar to this predicted AV space.

The star symbol in figure 4.6 shows the location in the perceptual [place] map of a token composed of video [gɑ] and the mismatched audio soundtrack [bɑ].

Why, you may ask, would anyone present such an audio/video mismatch to listeners? I would love to say that McGurk and McDonald (1976) did this to test the validity of the AV [place] map, but actually they were just interested in what would happen. Braida et al. (1998) formulated the hypothesis that the star token tests – namely that listeners have an audio-visual perceptual map for [place] that is a combination of audio and visual maps for [place]. This hypothesis predicts that the mismatch of audio [bɑ] with video [gɑ] will result in the perception "da." Interestingly enough, the $[bɑ]_A/[gɑ]_V$ stimulus does indeed sound like "da." This perceptual illusion is called the "McGurk effect" after Harry McGurk, who first demonstrated it. It is a really surprising effect that only goes away if you close your eyes. Otherwise, even if you know that the audio signal is [bɑ], you can only hear "da."

McGurking *ad nausea*

The McGurk effect is a really popular phenomenon in speech perception and researchers have poked and prodded it quite a bit to see how it works. In fact it is so popular we can make a verb out of the noun "McGurk effect" – to "McGurk" is to have the McGurk effect. Here are some examples of McGurking:

Babies McGurk – Rosenblum, Schmuckler, and Johnson, 1997
You can McGurk even when the TV is upside down – Campbell, 1994
Japanese listeners McGurk less than English listeners – Sekiyama and Tohkura, 1993
Male faces can McGurk with female voices – Green, Kuhl, Meltzoff and Stevens, 1991
A familiar face with the wrong voice doesn't McGurk – Walker, Bruce and O'Malley, 1995

Let's come back to the main point of this section. Perceptual representations are different from auditory representations. I have tried to show that these two are different from each other by showing a perceptual map that has an auditory dimension on one axis and a visual dimension on the other. Then I suggested that the McGurk effect shows that this perceptual space is the one used by listeners. You "hear" [bɑ] differently when you see the face saying [gɑ]. This means that the perceptual representation of speech has both auditory and visual aspects.

4.5 The limits of perceptual universality: A cross-linguistic map of Chinese tones

This section is about another way that perception is different from audition. It is reasonable to believe that the functioning of the auditory system is not affected

by native language. The bones of the middle ear may vary from person to person (particularly if you've had serious ear infections) and the physiology of the inner ear may vary, but there is no reason to believe that language experience causes auditory system changes. The same is not true of speech perception. Instead, we need to be able to distinguish aspects of perception that may be universal, having a firm basis in pure audition or vision, from aspects of perception that show the effects of linguistic tuning.

A substantial body of research leads to the conclusion that people who speak different languages have different perceptual maps. To illustrate the language specificity of speech perception (the limits of perceptual universality) we will consider some data from a recent study of Mandarin Chinese tones (Huang, 2001).

Mandarin Chinese has four distinctive tones – these are particular pitch patterns that appear on stressed syllables. In IPA notation they are: the high tone, [55]; the falling tone, [51]; the rising tone, [35]; and the dipping tone, [214] which is usually realized as [21] (remember this little fact). One of the phonological rules of Mandarin is a tone rule that in a sequence of two dipping tones the first one changes into a rising tone.

214 214 → 35 214

Huang (2001) agreed with researchers who have suggested that this change makes the sequence of tones easier to say, but she raised the question of why the rising tone in particular is chosen to replace the first dipping tone, rather than choosing one of the other two tones [55] or [51]. She hypothesized that the change 214 → 35 is "a case of perceptually tolerated articulatory simplification" because tone 35 is perceptually similar to tone 214.

This Mandarin tone rule is an old one – perhaps first appearing in the language 700 years ago. We know it is old because many related "dialects" of Mandarin have the rule, even though they have been separated from each other for a long time. Therefore, it is a matter of some speculation to say that the tone shapes were 35 and 214 at the time that the rule came into the language. Chen (1976) suggested that these tones both had a rising contour, and on this dimension differed from the other tones of Mandarin. My assumption, based on Chen (1976), is that the tones were shaped at least approximately like 35 and 214 when the tone rule first appeared, and that Huang's study is therefore a direct test of the role of perception in the historical development of this pattern in Mandarin.

At any rate, Huang conducted a study to map the perceptual space of tones in Mandarin and that's our main focus here. Her study goes beyond the methods that we've seen so far in this chapter. First, she determined the perceptual map for both native speakers of Mandarin Chinese and for American English speakers. It was important to test listeners who don't speak Chinese because of the phonological rule that converts 214 into 35 in Chinese. This rule relating the tones might influence their position in a perceptual map, so that the effect is taken to be the cause. In other words, they might be confusable because of the

rule, instead of the rule being caused or permitted by perceptual confusability. To conduct the test in American English Huang needed to refrain from use of the identification task. This is because her English listeners didn't know the names of the Chinese tones and she didn't want to try to teach them Chinese as a part of the experiment. Therefore, and this is her second methodological extension, she used a discrimination task and measured reaction time.

A trial in her experiment went like this. The listener heard two syllables, call them syllable A and syllable X. A and X always matched on their consonant and vowel, they only differed by which of the tones they had. If A and X where simply repetitions of the same syllable – same tone – then the listener was supposed to press a button labeled "same." But if A and X had different tones then the listener was supposed to press the "different" button. The buttons were connected to a computer that was clocking the amount of time it took to respond "same" or "different." Huang used the reaction time for correct "different" responses as a measure of the perceptual similarity of the two tones.

We think it is OK to use reaction time as a measure of similarity because it makes sense to believe that if two things are similar to each other it will take longer to notice that they are actually different. And we know of some prior research that bears this assumption out (e.g. Shepard, Kilpatrick and Cunningham, 1975).

Huang's data are shown in table 4.5. In this table we see for example that for Mandarin listeners it took 563 milliseconds to answer "different" when the high level tone [55] and the rising tone [35] were presented. We see right away that the longest average reaction times for both Mandarin and English listeners occured when [35] and [214] were paired with each other (these times are underlined in the table). Following Shepard et al. (1975) Huang then calculated distance from the reaction times by simply taking the inverse of the time. Thus, the distance between [55] and [35] for Mandarin listeners was assumed to be 1.776 (= 1/0.563).

This gives us a matrix of distances from our matrix of reaction times, so we can use multidimensional scaling to map the perceptual space just as we did in the previous sections of this chapter. The perceptual tone spaces of Mandarin and English listeners are shown in figure 4.7. Notice in figure 4.7 that the American English tone space fits inside the Mandarin Chinese tone space. This aspect of the data is also apparent in the reaction times in table 4.5 (recalling that perceptual map-making is mainly a way of visualizing patterns in data). On average the English listeners took longer (609 ms) than Chinese listeners (584 ms) to choose

Table 4.5 Average reaction times (in milliseconds) for correct "different" responses in Huang's (2001) study of tone perception.

(a) Mandarin listeners					*(b) American English listeners*				
	[55]	[35]	[214]	[51]		[55]	[35]	[214]	[51]
[55]					[55]				
[35]	563				[35]	615			
[214]	579	<u>683</u>			[214]	536	<u>706</u>		
[51]	588	548	545		[51]	600	592	608	

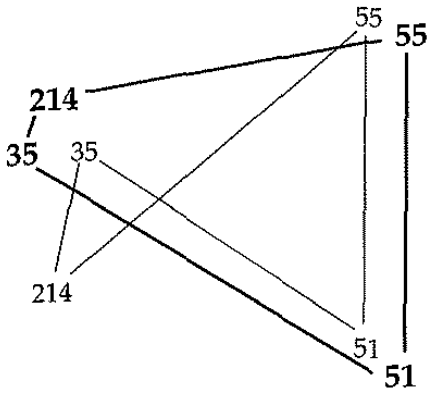

Figure 4.7 Maps of the perceptual tone spaces for Mandarin Chinese listeners (heavy lines, large font) and American English listeners (light lines, small font). These maps were produced by linear multi-dimensional scaling of the data in table 4.5. The MDS dimensions were scaled by the regression coefficent relating actual and predicted distance so that distance in the MDS solution is proportional to distance in the raw data. The lines drawn in this figure show some selected perceptual distances – e.g. the distance between tone [55] and tone [214] is shown as a line for both the Mandarin and American English listeners, while the [55]–[35] distance is not directly drawn in for either group.

"different" when the tones didn't match. This indicates that the tones were less distinct for the English listeners, on average – hence a smaller perceptual space.

Note also that the dipping tone [214] is closer to the rising [35] tone than any other tone for both the Chinese listeners and the English listeners. Huang was right that the rising tone is more confusable with the dipping tone and thus more likely to replace it in a phonological rule. Interestingly, the distance between tone [214] and tone [35] is shorter for Chinese listeners, despite their overall larger tone space. This suggests that the phonological relationship between the tones (that [214] → [35] sometimes) decreased their separation in the Chinese perceptual space.

Finally, what about that odd crossing pattern in the American English tone space? The horizontal dimension is the same for both groups of listeners – tones that start with high pitch are on the right side of the map and tones that start relatively low are on the left side. But the vertical dimension codes different things for the two groups. Now is when it is important to remember that the dipping tone [214] is most often simplified to [21]. For Chinese listeners the vertical dimension distinguishes the basically level tones [55] and [21] from the contour tones [35] and [51]. Perhaps it is no coincidence that Chinese scholars call Mandarin Chinese a "contour tone language." For English listeners, though, the vertical dimension codes a simpler phonetic property – the pitch at the end of the syllable – distinguishing the tones that end low [21] and [51] from the tones that end high [55] and [35]. I think that this is a mighty cool result!

Remember the main point of this section, which was also one of the main points of this chapter, language experience alters your perception of language. Experience changes your perceptual map for speech, tuning it to better perceive the contrasts and structure of your language.

Exercises

Sufficient jargon

test syllable versus test token, confusion matrix, identification task, signal to noise ratio, reaction time, multi-dimensional scaling, perceptual distance, projection, triangulation, McGurk effect.

Short answer questions

1 Randomize the order of this list of words. Describe your method.
"one," "two," "three," "four," "five," "six," "seven," "eight"

2 Use a ruler and a compass to draw the perceptual space that is encoded in the following matrix of distances. Note, this is a matrix of the perceived differences in talker's voices. We played pairs of words to listeners and asked them "does it sound like the same person twice or two different people?" Then we measured similarity as the number of times that different talkers were called the "same." The distance values in this table were then calculated by equation 4.2. So listeners responded "same" 15 percent of the time when they heard talker AJ paired with talker CN [$1.9 = -\ln(0.15)$]. Can you tell which two of the talkers are twins?

	AJ	CN	NJ	RJ
AJ		1.9	0.3	1.9
CN	1.9		2.3	2.5
NJ	0.3	2.3		1.9
RJ	1.9	2.5	1.9	

3 Compute the perceptual distance between "ð" and "d" in table 4.1. Is "ð" closer to "d" or "z"?

4 Now that we think of perceptual maps as having audio and video dimensions, we can reconsider the failure of our perceptual explanation for the variants of "this" and "that". In particular, could the visual component somehow explain why we see no "vis" or "vat" as variants of "this" and "that"? What audio-visual perceptual space (or spaces) could possibly give rise to reduced [v]/ [ð] confusions yet still permit Keif for "Keith"?

5 You may have noticed that in this chapter I used some different notations to refer to speech sounds. Here are the interpretations that were implicitly in the text:

[θ] – phonetic articulatory or acoustic physical aspects of the sound
$[\theta]_A$ – phonetic aspects conveyed acoustically
$[\theta]_V$ – phonetic aspects conveyed visually
"θ" – the perceptual representation of the sound

Some researchers argue that perceptual representations like "θ" are of speech gestures – i.e. that listeners interpret speech in terms of vocal tract activities rather than simply in terms of sensory patterns. What in this chapter is compatible or incompatible with this "gesturalist" view of speech perception?

6 One theory of speech perception claims that the ability to perceive speech is genetically prespecified. In some quarters, this has been taken to mean that humans have innate "feature detectors" used in perceiving speech. How do Huang's (2001) data on tone perception challenge this view?