Chapter 5 (of *Acoustic and Auditory Phonetics, 3rd Edition - in press)*

Speech perception

When you listen to someone speaking you generally focus on understanding their meaning. One famous (in linguistics) way of saying this is that "we speak in order to be heard, in order to be understood" (Jakobson, Fant & Halle, 1952). Our drive, as listeners, to understand the talker leads us to focus on getting the words being said, and not so much on exactly how they are pronounced. But sometimes a pronunciation will jump out at you - somebody says a familiar word in an unfamiliar way and you just have to ask - "is that how you say that?" When we listen to the phonetics of speech - to how the words sound and not just what they mean - we as listeners are engaged in speech perception.

In speech perception, listeners focus attention on the sounds of speech and notice phonetic details about pronunciation that are often not noticed at all in normal speech communication. For example, listeners will often not hear, or not seem to hear, a speech error or deliberate mispronunciation in ordinary conversation, but will notice those same errors when instructed to listen for mispronunciations (see Cole, 1973).

--------begin sidebar----------

Testing mispronunciation detection

As you go about your daily routine, try mispronouncing a word every now and then to see if the people you are talking to will notice. For instance, if the conversation is about a biology class you could pronounce it "biolochi". After saying it this way a time or two you could tell your friend about your little experiment and ask if they noticed any mispronounced words. Do people notice mispronunciation more in word initial position or in medial position? With vowels more than consonants? In nouns and verbs more than in grammatical words? How do people look up words in their mental dictionary if they don't notice when a sound has been mispronounced? Evidently, looking up words in the mental lexicon is a little different from looking up words in a printed dictionary (try entering "biolochi" in google). Do you find that your friends think you are strange when you persist in mispronouncing words on purpose?

----------end sidebar ----------

So, in this chapter we're going to discuss speech perception as a phonetic mode of listening, in which we focus on the sounds of speech rather than the words. An interesting problem in phonetics and psycholinguistics is to find a way of measuring how much phonetic information listeners take in during normal conversation, but in this book we can limit our focus to the phonetic mode of listening.

5.1 Auditory ability shapes speech perception

As we saw in the last chapter, speech perception is shaped by general properties of the auditory system that determine what can and cannot be heard, what cues will be recoverable in particular segmental contexts, and how adjacent sounds will influence each other. For example, we saw that the cochlea's nonlinear frequency scale probably underlies the fact that no language distinguishes fricatives on the basis of frequency components above 6000 Hz.

Two other examples illustrate how the auditory system constrains speech perception. The first example has to do with the difference between aspirated and unaspirated stops. This contrast is signalled by a timing cue that is called the "voice onset time" (abbreviated VOT). VOT is a measure (in milliseconds) of the delay of voicing onset following a stop release burst. There is a longer delay in aspirated stops than in unaspirated stops - so in aspirated stops the vocal folds are held open for a short time after the oral closure of the stop has been released. That's how the short puff of air in voiceless aspirated stops is produced. It has been observed that many languages have a boundary between aspirated and unaspirated stops at about 30 msec VOT. What is so special about a 30 millisecond delay between stop release and onset of voicing?

Here's where the auditory system comes into play. Our ability as hearers to detect the non-simultaneous onsets of tones at different frequencies probably underlies the fact that the most common voice onset time boundary across languages is at about ±30 milliseconds. Consider two pure tones, one at 500 Hz and the other at 1000 Hz. In a percetion test (see, for example, the research studies by Pisoni, 1977 and Pastore & Farrington, 1996), we combine these tones with a small onset asynchrony - the 500 Hz tone starts 20 milliseconds before the 1000 Hz tones. When

we ask listeners to judge whether the two tones were simultaneous or whether one started a little before the other, we discover that listeners think that tones separated by a 20 msec onset asynchrony start at the same time. Listeners don't begin to notice the onset asynchrony until the separation is about 30 msec. This parallelism between non-speech auditory perception and a cross-linguistic phonetic universal leads to the idea that the auditory system's ability to detect onset asynchrony is probably a key factor in this cross-linguistic phonetic property.

Example number two: Another general property of the auditory system is probably at work in the perceptual phenomenon known as "compensation for coarticulation". This effect occurs in the perception of place of articulation in CV syllables. The basic tool in this study is a continuum of syllables that ranges in equal acoustic steps from [dɑ] to [gɑ] (see figure 5.1). This figure needs a little discussion. At the end of chapter 3 I introduced spectrograms, and in that section I mentioned that the dark bands in a spectrogram show the spectral peaks that are due to the vocal tract resonances (the formant frequencies). So in figure 5.1a we see a sequence of five syllables with syllable number 1 labeled [da] and syllable number 5 labeled [ga]. In each syllable the vowel is the same, it has a first formant frequency (F1) of about 900 Hz, a second formant frequency (F2) of about 1100 Hz, an F3 at 2500 Hz, and an F4 at 3700 Hz. The difference between [da] and [ga] has to do with the brief formant movements (called formant transitions) at the start of each syllable. For [da] the F2 starts at 1500 Hz and the F3 starts at 2900 Hz, while for [ga] the F2 starts at 1900 Hz and the F3 starts at 2000 Hz. You'll notice that the main difference between [al] and [ar] in figure 5.1b is the F3 pattern at the end of the syllable.

Virginia Mann (1980) found that the perception of this [da]-[ga] continuum depends on the preceding context. Listeners report that the ambiguous syllables in the middle of the continuum sound like "ga" when preceded by the VC syllable [ɑl], and sound like "da" when preceeded by [ɑr].

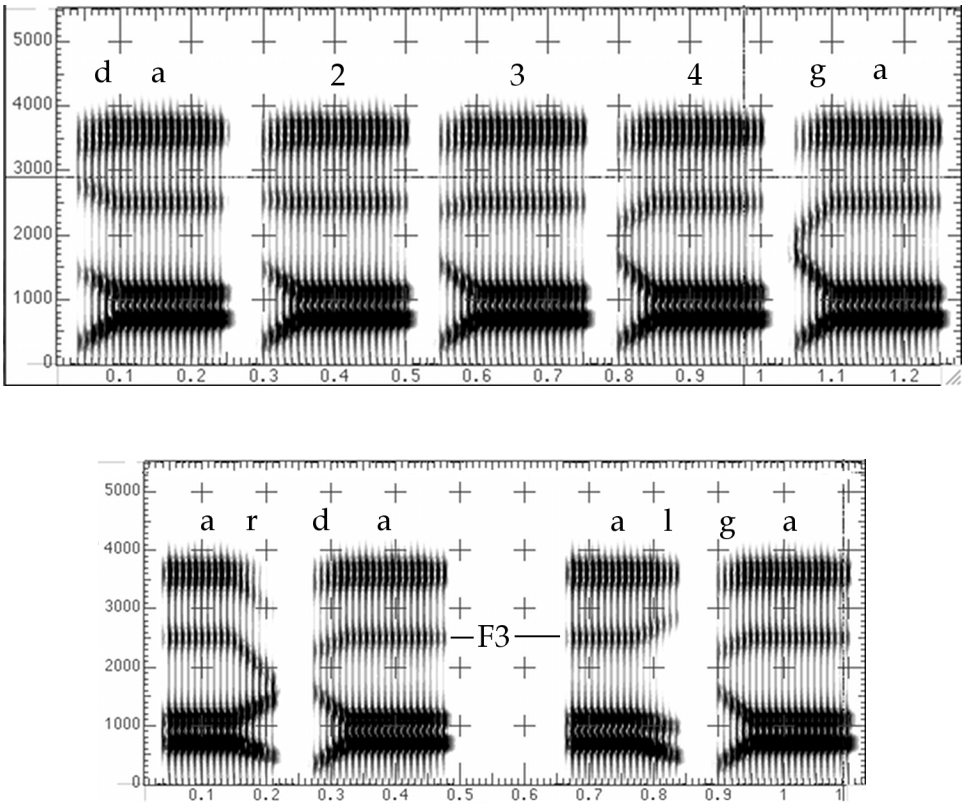

Figure 5.1 Panel (a): A continuum of synthetic consonant-vowel syllables ranging from "da" to "ga" in five acoustically equal steps. Panel (b): Token number 3 from the "da/ga" continnum sounds like "da" when preceeded by "ar" and like "ga" when preceeded by "al".

As the name implies, this "compensation for coarticulation" perceptual effect can be related to coarticulation between the final consonant in the VC context token ([ɑl] or [ɑr]) and the initial consonant in the CV test token ([dɑ]-[gɑ]). However, an auditory frequency contrast effect probably also plays a role. The way this explanation works is illustrated in figure 5.1 (b). The relative frequency of F3 distinguishes [dɑ] from [gɑ], F3 is higher in [dɑ] than it is in [gɑ]. Interestingly, though, the perceived frequency of F3 may also be influenced by the frequency of the F3 just prior to [dɑ/gɑ]. When F3 just prior to [dɑ/gɑ] is low (as in [ɑr]), the [dɑ/gɑ] F3 sounds contrastively higher, and when the F3 just prior is high the [dɑ/gɑ] F3 sounds lower. Lotto and Kluender (1998) tested this idea by replacing the precursor syllable with a simple sinewave that matched the ending frequency of the F3 of [ɑr], in one condition, or matche the

ending F3 frequency of [ɑl] in another condition. They found that these nonspeech isolated tones shifted the perception of the [dɑ]/[gɑ] continuum in the same direction that the [ɑr] and [ɑl] syllables did. So evidently, at least a part of the compensation for coarticulation phenomenon is due to a simple auditory contrast effect having nothing to do with the phonetic mode of perception.

---------- begin side bar ----------

Two explanations for one effect

Compensation for coarticulation is controversial. For researchers who like to think of speech perception in terms of phonetic perception - i.e. "hearing" people talk - compensation for coarticulation is explained in terms of coarticulation. Tongue retraction in [r] leads listeners to expect tongue retraction in the following segment and thus a backish stop (more like "g") can still sound basically like a "d" in the [r] context because of this context-dependent expectation. Researchers who think that one should first and foremost look for explanations of perceptual effects in the sensory input system (before positing more abstract cognitive parsing explanations), are quite impressed by the auditory contrast account.

It seems to me that the evidence shows that both of these explanations are right. Auditory contrast does seem to occur with pure tone context tokens, in place of [ɑr] or [ɑl], but the size of the effect is smaller than it is with a phonetic precursor syllable. The smaller size of the effect suggests that auditory contrast is not the only factor. I've also done research with stimuli like this where I present a continuum between [ɑl] and [ɑr] as context for the [dɑ]-[gɑ] continuum. When both the precursor and the target syllable are ambiguous, the identity of the target syllable (as "da" or "ga") depends on the perceived identity of the precursor. That is, for the same acoustic token, if the listener thinks that the context is "ar" he/she is more likely to identify the ambiguous target as "da". This is clearly not an auditory contrast effect.

So, both auditory perception and phonetic perception seem to push listeners in the same direction.

---------- end sidebar ----------

5. 2 Phonetic knowledge shapes speech perception

Of course, the fact that the auditory system shapes our perception of speech, does not mean that all speech perception phenomena are determined by our auditory abilities. As speakers, not just hearers, of language, we are also guided by our knowledge of speech production. There are main two classes of perceptual effects that emerge from phonetic knowledge: categorical perception and phonetic coherence.

5.2.1 Categorical perception

Take a look back at figure 5.1a. Here we have a sequence of syllables that shifts gradually (and in equal acoustic steps) from a syllable that sounds like "da" at one end to a syllable that sounds like "ga" at the other (see Table 5.1). When we play these synthesized syllables to people and ask them to identify the sounds - with an instruction like "please write down what you hear" - people usually call the first three syllables "da" and the last two "ga". Their response seems very categorical - a syllable is either "da" or "ga". But, of course, this could be so simply because we only have two labels for the sounds in the continuum, so by definition people have to say either "da" or "ga". Interestingly though, and this is why we say that speech perception tends to be categorical, the ability to hear the differences between the stimuli on the continuum is predictable from the labels we use to identify the members of the continuum.

To illustrate this, suppose I play you the first two syllables in the continuum shown in figure 5.1a - tokens number 1 and 2. These are both labelled "da", but they are slightly different from each other. Number one has a third formant onset of 2750 Hz while the F3 in token number two starts at 2562 Hz. People don't notice this contrast - the two syllables really do sound as if they are identical. The same thing goes for the comparisons of token two with token three and of token four with token five. But when you hear token three (a syllable that you would ordinarily label as "da") compared with token four (a syllable that you would ordinarily label "ga") the difference between them leaps out at you. The point is that in the discrimination task - when you are asked to detect small differences - you don't have to use the labels "da" or "ga". You should

be able to hear the differences at pretty much the same level of accuracy no matter what label you would have put on the tokens because the difference is the same (188 Hz for F3 onset) for token one versus two as it is for token three versus four. The curious fact is that even when you don't have to use the labels "da" and "ga" in your listening responses your perception is in accordance to the labels - you can notice a 188 Hz difference when the tokens have different labels and not so much when the tokens have the same label.

Table 5.1 The main acoustic parameters and idenfication results for the syllables show in Figure 5.1a.

Token number	F2 onset	F3 onset	Identified as:
1	1480	2750	"da"
2	1522	2562	"da"
3	1565	2375	"da"
4	1607	2187	"ga"
5	1650	2000	"ga"

One classic way to present these hypothetical results is shown in figure 5.2 (see Liberman et al., 1957 for the original graph like this). This graph has two "functions" - two lines - one for the proportion of times listeners will identify a token as "da", and one for the proportion of times that listeners will be able to accurately tell whether two tokens (say number 1 and number 2) are different from each other. The first of these two functions is called the identification function and I have plotted it as if we always (probability equals 1) identify tokens 1, 2, and 3 as "da". The second of these functions is called the discrimination function and I have plotted a case where the listener is reduced to guessing when the tokens being compared have the same label (where "guessing" equals probability of correct detection of difference is 0.5), and where he/she can always hear the difference between token 3 (labeled "da") and token 4 (labeled "ga"). The pattern of response in figure 5.2 is what we mean by "categorical perception" - within category discrimination is at chance and between category discrimination is perfect. Speech tends to be perceived categorically, though interestingly, just as with compensation for coarticulation, there is an auditory perception component in this kind of experiment so that speech perception is never perfectly categorical.

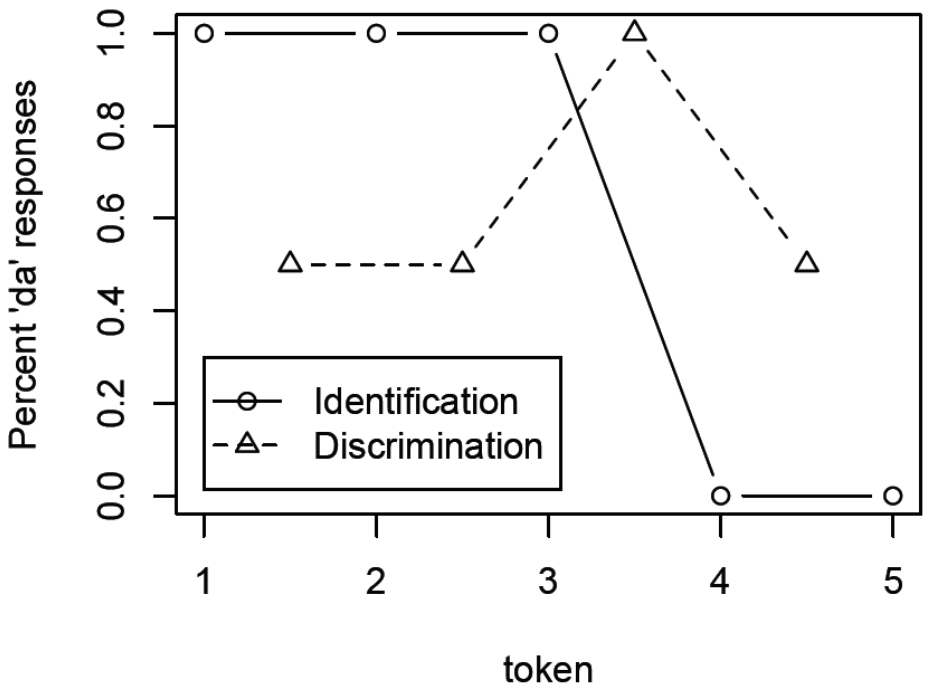

Figure 5.2 The classic categorical perception pattern of results. Identification performance (plotted with open circles) has a sharp transition from one category to the other, and discrimination performance (plotted with open triangles and a dashed line) is no better than chance for within-category discrimination.

Our tendency to perceive speech categorically has been investigated in many different ways. One of the most interesting of these lines of research suggests (to me at least) that categorical perception of speech is a learned phenomenon (see my old article with Jim Ralston on this topic - Johnson & Ralston, 1994). It turns out that perception of "sine wave analogs" of the [dɑ] to [gɑ] continuum is much less categorical than is perception of normal sounding speech. Robert Remez and colleagues (Remez, Rubin, Pisoni & Carrell, 1981) pioneered the use of sine wave analogs of speech to study speech perception. In sine wave analogs, the formants are replaced by time-varying sinusoidal waves (see figure 5.3). These signals, while acoustically comparable to speech do not sound at all like speech. The fact that we have a more categorical response to speech signals than to sine wave analogs of speech suggests that there is something special about hearing formant frequencies as speech versus hearing them as nonspeech video

game noises. One explanation of this was that as humans we have an innate ability to recover phonetic information from speech so that we hear the intended, categorical, gestures of the speaker.

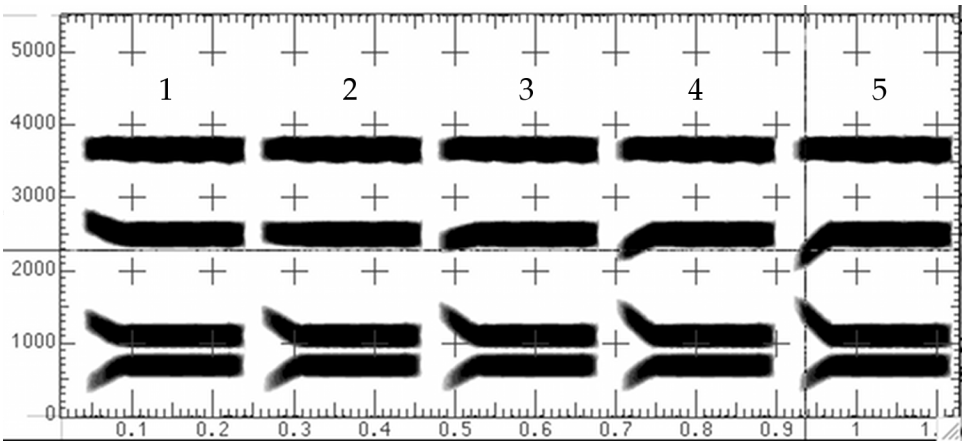

Figure 5.3 A continuum of sinewave analog syllables modeled on the "da/ga" continuum shown in figure 5.1.

A simpler explanation of why speech tends to be heard categorically is that our perceptual systems have been tuned by linguistic experience. As speakers, we have somewhat categorical intensions when we speak - for instance, to say "dot" instead of "got". So as listeners we evaluate speech in terms of the categories that we have learned to use as speakers. Several kinds of evidence support this "acquired categoriality" view of categorical perception.

For example, as you know from trying to learn the sounds of the international phonetic alphabet, foreign speech sounds are often heard in terms of native sounds. For instance, if you are like most beginners, when you were learning the implosive sounds [ɓ], [ɗ], and [ɠ] it was hard to hear the difference between them and plain voiced stops. This simple observation has been confirmed many times and in many ways, and indicates that in speech perception, we hear sounds that we are familiar with as talkers. Our categorical perception boundaries are determined by the language that we speak. [The theories proposed by Best (1995) and Flege (1995) offer explicit ways of conceptualizing this.]

---------- begin sidebar ----------

Categorical magnets

One really interesting demonstration of the language-specificity of categorical perception is the "perceptual magnet effect", which was demonstrated by Pat Kuhl and her colleagues (Kuhl et al., 1993). In this experiment, you synthesize a vowel that is typical of the sound of [i] and then surround it with vowels that systematically differ from the center vowel. In figure 5.4 this is symbolized by the white star, and the white circles surrounding it. A second set of vowels is synthesized, again in a radial grid around a center vowel. This second set is not centered around a typical [i] but instead around a vowel that is a little closer to the boundary between [i] and [e].

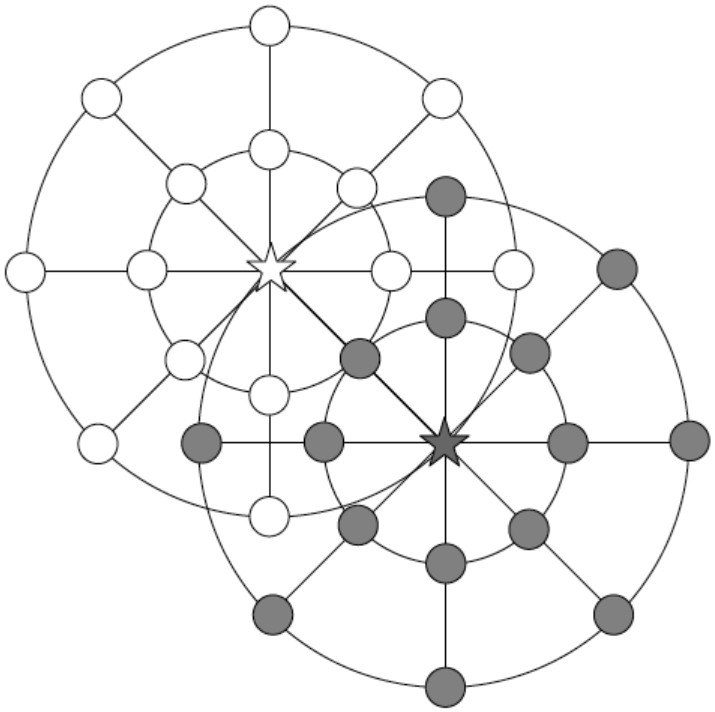

Figure 5.4 A schematic display of the stimuli used to compare perceptual sensitivity around a prototype vowel (the white star) to sensitivity around a non-prototypical example of the same vowel category.

When you ask adults if they can hear the difference between the center vowel (one of the stars) and the first ring of vowels it turns out that they have a harder time distinguishing the white star (a prototypcial [i]) and its neighbors than they do distinguishing the black star (a nonprototypical [i]) from its neighbors. This effect is interesting because it seems to show that

categorical perception is gradient within categories (note that all of the vowels in the experiment sound like variants of [i], even the ones in the black set that are close to the [i]/[e] boundary). However, even more interesting, is the fact that the location of a perceptual magnet differs depending on the native language of the listener - even when those listeners are mere infants!

---------- end sidebar ----------

5.2.2 Phonetic coherence

Auditory sensory experience forms a coherent "picture" of the world by means of a number of gestalt organizing principles that have been called "Auditory Scene Analysis". When we are perceiving speech, however, we can experience phonetic coherence with acoustic components that according to scene analysis principles should be incoherent.

Duplex perception is a good example of this. In this phenomenon, which was discovered by Timothy Rand in 1974, the stimulus has on the left channel a small "chirp" noise - a little 80 ms tone glide that corresponds to the typical frequency of F3 during either a [dɑ] or a [gɑ] syllable and on the right channel a "base" stimulus that is composed of [dɑ] or [gɑ] missing only the F3 chirp component. Interestingly, the base can be identical for [dɑ] and [gɑ] so the only difference between the stimuli is present in the "chirp". Figure 5.5 shows the acoustic wave forms for the left and right ears for a sequence of five syllables - the first one in the series sounds like [dɑ] and the last one sounds like [gɑ] (this is just like the continuum in figure 5.1a). The base signal is presented to the right ear, and the "chirp" noises are presented to the left ear.

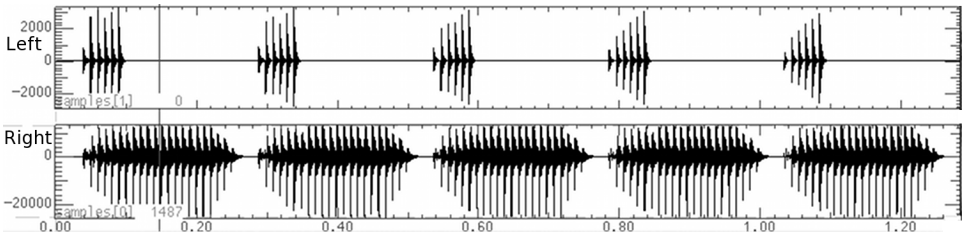

Figure 5.5 Wave forms of stimuli used to test for duplex perception. The top panel shows a trace of the signal that is presented in the left ear and the bottom panel shows a trace of the signal that is presented to the right ear.

The spectrograms in figure 5.6 highlight what is so special about these duplex perception stimuli. As you can see in the spectrograms in figure 5.6b, the base stimuli are identical for each stimulus in the sequence of five syllables, and in each case there is a gap where the third formant should be. The chirps shown in figure 5.6a fill these gaps exactly. The first one in the series has a downward going chirp, and the last one has an upward going chirp. when you add the chirps to the bases you get, almost exactly, the [dɑ]/[gɑ] continuum that was shown in figure 5.1a.

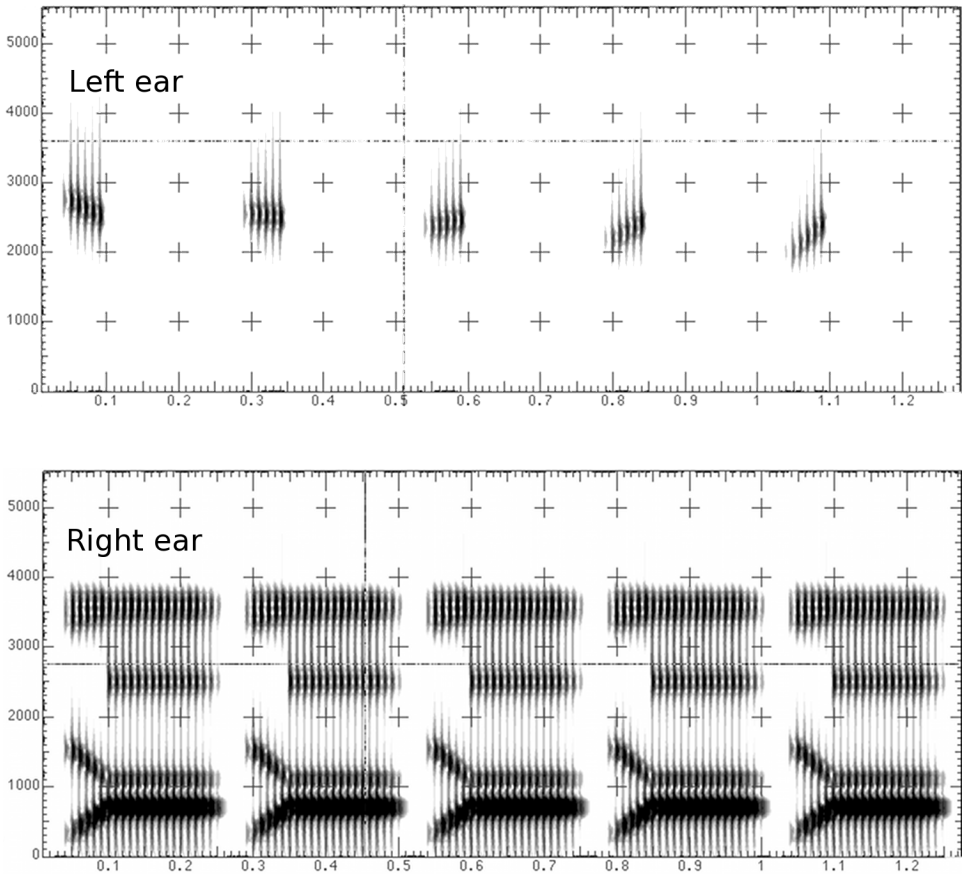

Figure 5.6 Spectrograms of the left and right duplex perception stimuli. The stimuli are practically identical to those shown in figure 5.1, except that the crucial information that distinguishes [dɑ] from [gɑ] is isolated in the left ear.

In normal auditory perception, sounds that are louder in your left ear seem to come from the left side of your body, while sounds that are louder in your right ear seem to come from the right. This is true in the duplex perception phenomenon too. The chirp seems to be on the left and the base seems to be on the right. One additional thing happens though, that we don't see in ordinary

auditory perception. The chirp (even though it is heard as originating from a different location from the base) influences the phonetic perception of the base. If the chirp is like the F3 of [d] listeners will hear the base as "da" and if the chirp is like the F3 of [gɑ] listeners will hear the base as "ga". The reason that this phenomenon is called "duplex" is because the chirp seems to be two places at once - as an isolated nonspeech chirp noise and as a phonetic component determining the place of articulation of the base. This is a neat effect because it indicates a pretty high degree of "phonetic coherence" in speech perception. The perceptual system glues together acoustic bits that would ordinarly not combine with each other.

Here's another phenomenon that illustrates the phonetic coherence of speech perception. Imagine that you make a video of someone saying "ba", "da", and "ga". Now, you dub the audio of each of these syllables onto the video of the others. That is, one copy of the video of [bɑ] now has the audio recording of [dɑ] as its sound track, another has the audio of [gɑ], and so on. There are some interesting confusions among audio/video mismatch tokens such as these and one of them, in particular has become a famous and striking demonstration of the phonetic coherence of speech perception.

Some of the mismatches just don't sound right at all. For example, when you dub audio [dɑ] onto video [bɑ], listeners will report that the token is "ba" (in accordance with the obvious lip closure movement) but that it doesn't sound quite normal.

The really famous audio/video mismatch is the one that occurs when you dub audio [bɑ] onto video [gɑ]. The resulting movie doesn't sound like either of the input syllables, but instead it sounds like “da”! This perceptual illusion is called the “McGurk effect” after Harry McGurk, who first demonstrated it (McGurk & MacDonald, 1986). It is a surprisingly strong illusion that only goes away when you close your eyes. Even if you know that the audio signal is [bɑ], you can only hear “da.”

The McGurk effect is an illustration of how speech perception is a process in which we deploy our phonetic knowledge to generate a phonetically coherent percept. As listeners we combine information from our ears and our eyes, to come to a phonetic judgement about what is being said. This process taps specific phonetic knowledge not just generic knowledge of speech

movements. For instance, Walker, Bruce, & O'Malley's (1995) demonstrated that audio/video integration is blocked when listeners know the talkers, and know that the voice doesn't belong with the face (in a dub of one person's voice onto another person's face). This shows that phonetic coherence is a property of speech perception, and that phonetic coherence is a learned perceptual capacity, based on knowledge we have acquired as listeners.

---------- begin sidebar ----------

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

---------- end sidebar ----------

5.3 Linguistic knowledge shapes speech perception

We have seen so far that our ability to perceive speech is shaped partly by the nonlinearities and other characteristics of the human auditory system, and we have seen that what we hear when we listen to speech is partly shaped by the phonetic knowledge we have gained as speakers. Now we turn to the possibility that speech perception is also shaped by our knowledge of the linguistic structures of our native language.

I have already included in section 5.2 (on phonetic knowledge) the fact that the inventory of speech sounds in your native language shapes speech perception, so in this section I'm not

focussing on phonological knowledge when I say "linguistic structures", but instead I will present some evidence of lexical effects in speech perception - that is, that hearing words is different from hearing speech sounds.

I should mention at the outset that there is controversy about this point. I will suggest that speech perception is influenced by the lexical status of the sound patterns we are hearing, but you should know that some of my dear colleagues will be disappointed that I'm taking this point of view.

---------- begin side bar ----------

Scientific method - on being convinced.

There are a lot of elements to a good solid scientific argument, and I'm not going to go into them here. But, I do want to mention one point about how we make progress. The point is that no one individual gets to declare an argument won or lost. I am usually quite impressed by my own arguments and cleverness when I write a research paper. I think I've figured something out and I would like to announce my conclusion to the world. However, the real conclusion of my work is always written by my audience and it keeps being written by each new person who reads the work. They decide if the result seems justified or valid. This aspect of the scientific method, including the peer review of articles submitted for publication, is part of what leads us to the correct answers.

The question of whether speech perception is influenced by word processing is an interesting one in this regard. The very top researchers - most clever, and most forceful - in our discipline are in disagreement on the question. Some people are convinced by one argument or set of results and others are more swayed by a different set of findings and a different way of thinking about the question. What's interesting to me is that this has been dragging on for a long, long time. And what's even more interesting, is that as the argument drags on, and researchers amass more and more data on the question, the theories start to blur into each other a little. Of course, you didn't read that here!

---------- end side bar ----------

The way that "slips of the ear" work suggests that listeners apply their knowledge of

words in speech perception. Zinny Bond (1999) reports perceptual errors like "spun toffee" heard as "fun stocking" and "wrapping service" heard as "wrecking service". In her corpus of slips of the ear, almost all of them are word misperceptions, not phoneme misperceptions. Of course, sometimes we may mishear a speech sound, and perhaps think that the speaker has mispronounced the word, but Bond's research shows that listeners are inexorably drawn into hearing words even when the communication process fails. This makes a great deal of sense, considering that our goal in speech communication is to understand what the other person is saying, and words (or more technically, morphemes) are the units we trade with each other when we talk.

This intuition, that people tend to hear words, has been verified in a very clever extension of the place of articulation experiment we discussed in sections 5.1 and 5.2. The effect, which is named the Ganong effect after the researcher who first found it (Ganong, 1980), involves a continuum like the one in figure 5.1, but with a word at one end and a nonword at the other. For example, if we added a final [g] to our [dɑ]/[gɑ] continuum we would have a continuum between the word "dog" and the non-word [gɑg]. What Ganong found, and what makes me think that speech perception is shaped partly by lexical knowledge, is that in this new continuum we will get more "dog" responses than we will get "da" responses in the [dɑ]/[gɑ] continuum. Remember the idea of a "perceptual magnet" from above? Well, in the Ganong effect words act like perceptual magnets; when one end of the continuum is a word, listeners tend to hear more of the stimuli as a lexical item, and fewer of the stimuli as the nonword alternative at the other end of the continuum.

Ganong applied careful experimental controls using pairs of continua like "tash-dash" and "task-dask" where we have as much similarity between the continuum that has a word on the /t/ end ("task-dask") and the one that has a word on the /d/ end ("tash-dash"). That way there is less possibility that the difference in number of "d" responses is due to small acoustic differences between the continua rather than the difference in lexicality of the endpoints. It has also been observed that the lexical effect is stronger when the sounds to be identified are at the ends of the test words, as in "kiss-kish" versus "fiss-fish". This makes sense if we keep in mind that it takes a little time to activate a word in the mental lexicon.

A third perceptual phenomenon that suggests that linguistic knowledge (in the form of lexical identity) shapes speech perception was called "phoneme restoration" by Warren when he discovered it (Warren, 1970). Figure 5.7 illustrates phoneme restoration. The top panel is a spectrogram of the word "legislation" and the bottom panel shows a spectrogram of the same recording with a burst of broadband noise replacing the [s]. When people hear the noise-replaced version of the sound file in figure 5.7b they "hear" the [s] in [ˌlɛʤɪsˈleˈʃn̩]. Arthur Samuel (1991) reported an important bit of evidence suggesting that the [s] is really perceived in the noise-replaced stimuli. He found that listeners can't really tell the difference between a noise-added version of the word (where the broad band noise is simply added to the already existing [s]) and a noise-replaced version (where the [s] is excised first, before adding noise). What this means is that the [s] is actually perceived - it is restored - and thus that your knowledge of the word "legislation" has shaped your perception of this noise burst.

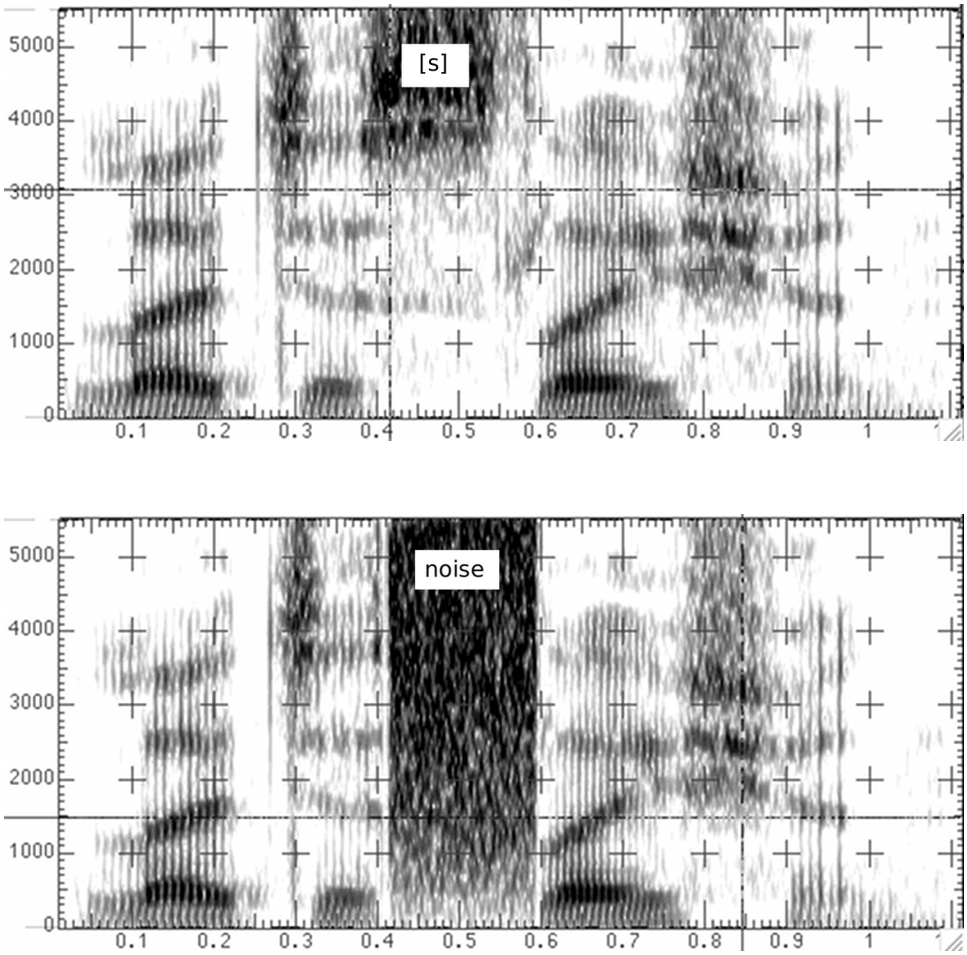

Figure 5.7 Panel (a): A spectrogram of the word "legislation" with the [s] noise marked. Panel (b): The same utterance again, but with the [s] replaced by broadband noise.

Jeff Elman and Jay McClelland (1988) provided another important bit of evidence that linguistic knowledge shapes speech perception. They used the phoneme restoration process to induce the perception of a sound that then participated in a compensation for coarticulation. This two step process is a little complicated, but one of the most clever and influential experiments in the literature.

Step one: compensation for coarticulation. We use a [dɑ]/[gɑ] continuum just like the one in figure 5.1, but instead of context syllables [ɑl] and [ɑr], we use [ɑs] and [ɑʃ]. There is a compensation for coarticulation using these fricative context syllables that is like the effect seen with the liquid contexts. Listeners hear more "ga" syllables when the context is [as] than when it is [ɑʃ].

Step two: phoneme restoration. We replace the fricative noises in the words "abolish" and "progress" with broadband noise, as was done to the [s] of "legislature" in figure 5.7. Now we have a perceived [s] in "progress" and a perceived [ʃ] in "abolish" but the signal has only noise at the ends of these words in our tokens.

The question is whether the restoration of [s] and [ʃ] in "progress" and "abolish" is truly a perceptual phenomenon, or if it is just something more like a decision bias in how listeners will guess the identity of a word. Does the existance of a word "progress" and the nonexistance of any word "progresh" actually influence speech perception? Elman and McClelland's excellent test of this question was to use "abolish" and "progress" as contexts for the compensation for coarticulation experiment. The reasoning is that if the "restored" [s] produces a compensation for coarticulation effect, such that listeners hear more "ga" syllables when preceeded by a restored [s] than they do when preceeded by a restored [ʃ], then we would have to conclude that the [s] and [ʃ] were actually perceived by listeners - they were actually perceptually there and able to interact with the perception of the [dɑ]/[gɑ] continuum. Guess what Elman and McClelland found? That's right, the phantom, not-actually-there, [s] and [ʃ] caused

compensation for coarticulation. Pretty impressive evidence that speech perception is shaped by our linguistic knowledge.

5.4 Perceptual similarity

Now to conclude the chapter, I'd like to discuss a procedure for measuring perceptual similarity spaces of speech sounds. This method will be useful in later chapters as we discuss different types of sounds, their acoustic characteristics, and then their percepual similarities. Perceptual similarity is also a key parameter in relating phonetic characteristics to language sound change and the phonological patterns in language that arise from sound change.

The method involves presenting test syllables to listeners and asking them to identify the sounds in the syllables. Ordinarily, with carefully produced "lab speech" (that is, speech produced by reading a list of syllables into a microphone in the phonetics lab) listeners will make very few misidentifications in this task, so we usually add some noise to the test syllables to force some mistakes. The noise level is measured as a ratio of the intensity of the noise compared with the peak intensity of the syllable. This is called the signal to noise ratio (SNR) and is measured in decibels. To analyze listeners' responses we tabulate them in a **confusion matrix.** Each row in the matrix corresponds to one 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.

Table 5.2 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

Table 5.2 shows the confusion matrix for the 0 dB SNR condition in George Miller and Patricia Nicely's (1955) large study of consonant perception. 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.

Even before doing any sophisticated data analysis, we can 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 other possible points of interest - for example, why "this" and "that" are sometimes said with a [d] sound. 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.

5.4.1 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 5.8(a), with Columbus located between Cleveland and Cincinnati. A line works to describe these distances because the distance from Cincinnati to Cleveland is simply the sum of the other two distances (107 + 142 = 249).

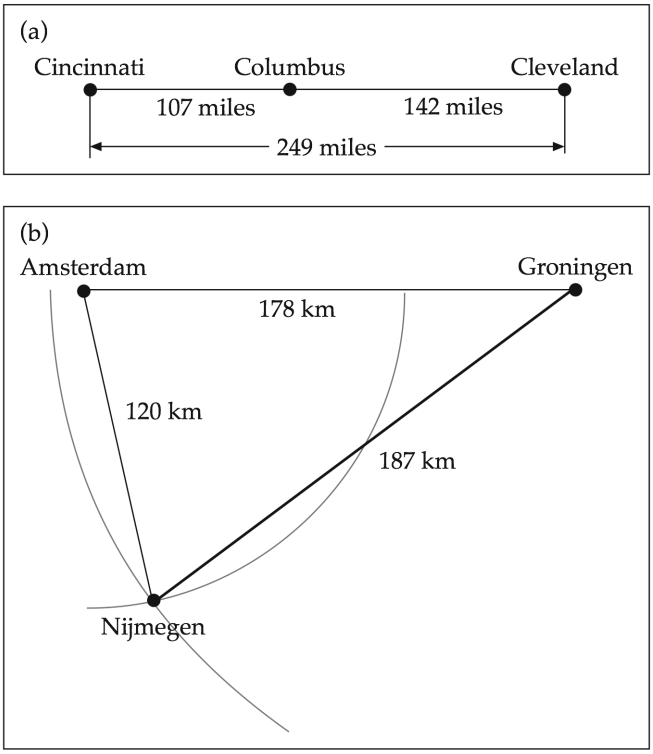

Figure 5.8 Panel (a): A one-dimensional map of three cities in Ohio. One dimension can adequately express the distances between them. Panel (b): A two dimensional map of three cities in The Netherlands. The arcs show how Nijmegen was placed on the map.

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 5.8(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 5.8(b) is an accurate depiction of the relative locations of these three cities as you can see in the map in figure 5.9.

Figure 5.9 A map of the Netherlands showing the orientation of the Amsterdam–-Groningen–Nijmegen triangle derived in figure 5.8(b).

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.

5.4.2 The perceptual map of fricatives

In this section we will use multidimensional scaling (MDS) to map the perceptual space that caused the confusion pattern in table 5.2.

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 5.2 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"	"θ"
[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).

---------- begin side bar ----------

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 5.2 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. One interesting point about the asymmetry of [f] and [θ] confusions is that the perceptual confusion matches the cross-linguistic tendency for sound change (that is, [θ] is more likely to change in to [f] than vice versa). Mere coincidence, or is there a causal relationship? Shepard's method for calculating similarity from a confusion matrix glosses over this interesting point and 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.

---------- end side bar ----------

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:

$$S_{ij} = \frac{p_{ij} + p_{ji}}{p_{ii} + p_{jj}} \qquad (5.1)$$

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:

$$d_{ij} = -\ln(S_{ij}) \qquad (5.2)$$

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.

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 5.3 shows the similarities calculated from the Miller and Nicely confusion matrix (table 5.2) using equation (5.1).

Table 5.3 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		
[zl	.000	.026	.000	.169	.000	1.0	
[d]	.000	.000	.000	.012	.000	.081	1.0

The perceptual map based on these similarities is shown in figure 5.10. 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.

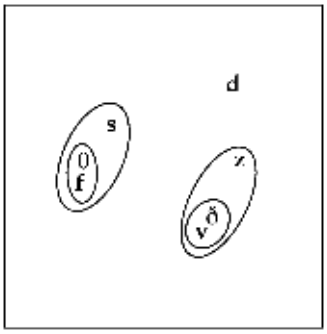

Figure 5.10 The perceptual map of fricatives and [d] in English. The location of the points was determined by multidimensional scaling of the confusion data from Miller and Nicely (1955). The circled groups of sounds are clusters that were found in a hierarchical cluster analysis of the same data.

In figure 5.10, I drew ovals around some clusters of sounds. These show two levels of similarity among the sounds as revealed by a hierarchical cluster analysis (another neat data analysis method available in most statistics software packages - see my book *Quantitative Methods in Linguistics* for more on this). 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 (MDS) 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."

---------- begin sidebar ----------

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" seen sometimes in spectograms of the lateral

fricatives [ɬ] and [ɮ]. (A spit spike looks like a stop release burst, see chapter 8, but occurs in the middle of a fricative noise.) 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.

---------- end sidebar ----------

Recommended Reading:

Best, C.T. (1995) A direct realist perspective on cross-language speech perception. In *Speech perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-language Speech Research*, edited by W. Strange (York: Timonium, MD), pp. 167-200. Describes a theory of cross-language speech perception in which listeners map new, unfamiliar, sounds onto their inventory of native-language sounds.

Campbell, R. (1994) Audiovisual speech: Where, what, when, how? *Current Psychology of Cognition*, 13, 76–80. On the perceptual resilience of the McGurk effect.

Cole, R.A. (1973) Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics*, **13**, 153-156. A study showing that people often don't hear mispronunciations in speech communication.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, *27*, 143–165. One of the most clever, and controversial, speech perception experiments ever reported.

Flege, J.E. (1995) Second language speech learning: Theory, findings, and problems. In *Speech perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-language Speech Research*, edited by W. Strange (York: Timonium, MD), pp. 167-200. Describes a theory of cross-language speech perception in which listeners map new, unfamiliar, sounds onto their inventory of native-language sounds.

Ganong, W.F. (1980) Phonetic Categorization in Auditory Word Recognition. *J. of Exp. Psychol. Hum. Percept.Perform.*, **6**, 110-125. A highly influential demonstration of how people are drawn to hear words in speech perception. The basic result is now known as "the Ganong effect".

Green, K. P., Kuhl, P. K., Meltzoff, A. N., and Stevens, E. B. (1991) Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, 50, 524–36. Integrating gender-mismatched voices and faces in the McGurk effect.

Jakobson, Roman, Fant, Gunnar, and Halle, Morris (1952) *Preliminaries to Speech Analysis*, Cambridge, Mass.: MIT Press. A classic in phonetics and phonology in which a set of distinctive phonological features is defined in acoustic terms.

Johnson, Keith & Ralston, James V. (1994) Automaticity in speech perception: some speech/nonspeech comparisons. *Phonetica* **51**(4), 195-209. A set of experiments suggesting that over-learning accounts for some of the "specialness" of speech perception.

Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N. & Lindblom, B. (1992) Linguistic experiences alter phonetic perception in infants by 6 months of age. *Science*, **255**, 606-608. Demonstrated the perceptual magnet effect with infants.

Liberman, A.M.; Harris, K.S.; Hoffman H.S. & Griffith, B.C. (1957) The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358 - 368. The classic demonstration of categorical perception in speech perception.

Lotto, A.J. & Kluender, K.R. (1998) General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception and Psychophysics* **60**, 602-619. A demonstration that at least a part of the compensation for coarticulation effect (Mann, 1980) is due to auditory contrast.

Mann, V.A. (1980) Influence of preceding liquid on stop-consonant perception. *Percept. & Psychophys*. **28**, 407-412. The original demonstration of compensation for coarticulation in sequences like [al da] and [ar ga].

McGurk, H. and MacDonald, J. (1976) Hearing lips and seeing voices. *Nature*, 264, 746–8. The audio-visual speech perception effect that was reported in this paper has been come to be called the "McGurk" effect.

Miller, George A. and Nicely, Patricia E. (1955) An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27, 338–52. A standard reference for the confusability of American English speech sounds.

Parzen, E. (1962) On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33, 1065–76. A method for estimating probability from instances.

Pastore, R.E. & Farrington, S.M. (1996) Measuring the difference limen for identification of order of onset for complex auditory stimuli. *Perception & Psychophysics* **58**(4), 510-26. On the auditory basis of the linguistic use of aspiration as a distinctive feature.

Pisoni, D.B. (1977) Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America* **61**, 1352-1361. More on the auditory basis of the linguistic use of aspiration as a distinctive feature.

Rand T. C. (1974) Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**(3), 678-80. The first demonstration of the duplex perception effect.

Remez, R.E., Rubin, P.E., Pisoni, D.B. & Carrell, T.D. (1981) Speech perception without traditional speech cues. *Science* **212**, 947-950. The first demonstration of how people perceive sentences that have been synthesized using only time-varying sinewaves.

Rosenblum, L. D., Schmuckler, M. A., and Johnson, J. A. (1997) The McGurk effect in infants. *Perception & Psychophysics*, 59, 347–57.

Sekiyama, K. and Tohkura, Y. (1993) Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21, 427–44. These authors found that the McGurk effect is different for people who speak different languages.

Shannon, C. E. and Weaver, W. (1949) *The Mathematical Theory of Communication*. Urbana: University of Illinois. The book that established "information theory".

Shepard, R. N. (1972) Psychological representation of speech sounds. In E. E. David and P. B. Denes (eds.) *Human Communication: A unified view.* New York: McGraw-Hill, 67–113. Measuring perceptual distance from a confusion matrix.

Walker, S., Bruce, V. and O'Malley, C. (1995) Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, 57, 1124–33. A fascinating demonstration of how top-down knowledge may mediate the McGurk effect.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392–393. The first demonstration of the "phoneme restoration effect".

Exercises

Sufficient jargon

stimulus continuum, duplex perception, Ganong effect, sinewave analog of speech, confusion matrix, identification task, signal to noise ratio, reaction time, multi-dimensional scaling, perceptual distance, projection, triangulation, McGurk effect.

Short answer questions

1 Record yourself saying the words "sue" and "see". Look at these recordings in spectrograms. How are the initial /s/ sounds different from each other? Now, splice the /s/ from "sue" onto the /i/ of "see", and the /s/ from "see" onto the /u/ from "sue". There is a perceptual compensation for coarticulation at work here. Describe the coarticulatory gesture that is involved.

2 Point your browser to a "misheard lyrics" web page like http://www.kissthisguy.com/ and pick three misperceptions that you think you could explain in terms of the acoustic phonetic similarities between the intended utterance and the misperception. You get extra credit (and are more likely to get the answer right) if you include spectrograms of you saying the intended and the heard words. By the way, can you find a case of a misperception that isn't words?

3 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 5.1. So listeners responded "same" 15 percent of the time when they heard talker AJ paired with talker CN [1.9 = -ln(O.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

4 Compute the perceptual distance between "ð" and "d" in table 5.2. Is "ð" closer to "d" or "z"?

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?