do not distinguish the vowels in *hood* and *who'd* (or *look* and *Luke*). Some Northern English speakers do not distinguish between *look* and *luck*. Speakers of Canadian English are often distinguishable by the quality of the diphthong in words such as *out* and *about*.

There are also differences in the qualities of the vowels that occur in different consonantal contexts. Everybody distinguishes the vowels in *seen* and *sin*, but only one of these two vowels can occur before 'ng'. For speakers of BBC English the vowel in *sing* is very like that in *sin*; but for many speakers of American English, notably those from California, it is almost identical with the vowel in *seen*. British speakers have three different vowels in the words *merry*, *Mary*, and *marry*, but most Americans do not distinguish all these vowels before 'r'. Californian English speakers typically pronounce all three of these words in exactly the same way.

### 3.3 Summary

Many languages, like Spanish and Japanese, have only five vowels. General American English as spoken by national newscasters has 14 or 15 distinct vowels that can contrast in monosyllabic words. The corresponding form of British English, BBC English, has 20 distinct vowels.

### 4 The Sounds of Vowels

#### 4.1 Acoustic Structure of Vowels

We all know what a vowel sounds like, but there are no popular terms for describing how the quality of one vowel differs from that of another. We need to consider more than the acoustic properties we have been considering so far, the frequency (pitch) and intensity (loudness) of different sounds. Vowels can be produced on any pitch within the range of a speaker's voice. I can say the vowels in *head*, *hid*, *head*, *had* on a low pitch, when the vocal folds are vibrating about 80 times a second (a low E), and then I can say them again with vocal folds vibrating 160 times a second (the E an octave above). The pitch of my voice will have changed, but the vowels will still have the same quality. I can also say any vowel loudly or softly. The quality, the factor that distinguishes one vowel from another, remains the same when I shout or talk quietly.

Different vowels are like different instruments. One can play concert pitch A on a piano, a clarinet, or a violin. In each case it will be the same note because the rate of repetition of the sound wave as a whole – the fundamental frequency - is the same. The quality will be different because the smaller variations within each repetition of the sound wave – the overtones – will be different. Similarly, vowels will retain their individual qualities irrespective of the pitch produced by the vocal folds.

When we listen to a vowel or a musical note, we can tell which vowel it is, or which instrument produced it, by the overtones that occur. The reed of the clarinet or the vocal folds may be vibrating 100 times a second, but the sound that is produced at the mouth of the clarinet or at the lips will contain characteristic groups of overtones at higher frequencies. We can see an indication of
To see how these higher frequencies arise, we can liken the air in the mouth and throat to the air in a bottle. When you blow across the top of a bottle the air inside it will be set in vibration. The resonance of the bottle – the note that it produces when the air vibrates – will depend on the size and shape of the body of air inside it. If the bottle is full of air, with nothing else in it, it will have a low-pitched resonance. Pouring water into it so that the body of air becomes smaller makes the pitch go up. Smaller bodies of air vibrate more quickly. They have a higher resonant frequency.

Producing different vowels is like altering the size and shape of the bottle. For a vowel the relevant shape is the tube formed by the mouth and throat, which is known as the vocal tract. The air in this tube is set in vibration by the pulses of air from the vocal folds. Every time they open and close, the air in the vocal tract above them will be set in vibration. Because the vocal tract has a complex shape, the air within it will vibrate in more than one way. Often we can consider the body of air behind the raised tongue (i.e. in the throat) to be vibrating in one way, and the body of air in front of it (i.e. in the mouth) to be vibrating in another. In the vowel in *heed* the air behind the tongue will vibrate at about 250 Hz, and the air in front of it at about 2,100 Hz.

The resonances of the vocal tract are called formants. Trying to hear the separate formants in a vowel is difficult. We are so used to a vowel being a single meaningful entity that it is difficult to consider it as a sound with separable bits. But it is possible to say vowels so that some of their component parts are more obvious. One possibility is to whisper a series of vowels, as I have done in recording 4.1. When whispering, the vocal folds do not vibrate; they are simply drawn together so that they produce a random noise like that of the wind blowing around a corner. Because this noise is in the pitch range of one of the resonances of the vocal tract, we can hear that resonance more plainly. If you whisper you will not hear a note with a specific pitch; but you will be able to hear the changes in the vowel resonances. Try whispering *heed*, *hid*, *lud*, *had*, *haved*, as in recording 4.1; there will be a general impression of a descending pitch.

Another way of making one of the resonances more obvious is to say a series of words on a very low pitch. Say the vowel in *had* as low a pitch as you can, and then try to go even lower so that you produce a kind of creaky voice. Now say the words *had*, *head*, *hid*, *heed* in this kind of voice, as I have done in recording 4.2. You may be able to hear not only the constant low buzzing sort of pitch associated with the vocal folds, but also a changing pitch in one of the overtones. When saying the words *had*, *head*, *hid*, *heed*, this pitch goes down.

The sound that you hear when whispering is mainly that of the vibrations of the air in the front of the mouth. Conversely, the pitch changes associated with saying *had*, *head*, *hid*, *heed* in a creaky voice are due to the vibrations of the air in the back of the vocal tract. This resonance is the lower in pitch of the two, and is called the first formant. The height of the bars in figure 4.2 shows the

---

**Figure 4.1** The vowel a as in *father* (from figure 1.1) and the vowel i as in *see* (from figure 1.4), showing the different overtones that occur when these vowels are said on the same pitch.

these higher frequencies in the waveforms of the vowels a as in *father* and i as in *see*, which were reproduced in figures 1.1 and 1.4 and are shown here together in figure 4.1. In both cases the wave as a whole repeats itself one hundred times a second. Within one-hundredth of a second in the upper wave there are seven peaks in air pressure. Seven peaks in one-hundredth of a second corresponds to 700 peaks in a second. It is as if there is a 700 Hz wave that starts up and then diminishes every one-hundredth of a second. But within one-hundredth of a second in the lower wave, there is a wave that starts up and repeats itself a little over twice in one-hundredth of a second, and a number of other, smaller, waves that repeat themselves over 20 times in this interval. For my voice (all these figures show sound waves that I produced), the vowel a is characterized by a group of overtones around 700 Hz; and the vowel i is characterized by two groups of overtones, one around 250 Hz, and the other around 2,100 Hz. As we will see later, the situation is actually a little more complicated than this, as the vowel a has two groups of overtones in the 700 range, and the vowel i also has additional overtones that we must take into account.

If we looked at some other vowels, we would see that for the vowel in *heed* the most prominent overtones will be at about 550 and 1,600 Hz. For the vowel in *lud* they will be around 750 and 1,200 Hz. These overtones will occur although the vocal folds may be vibrating at any rate from about 80 to around 250 Hz for a male speaker.
The Sounds of Vowels

4.2 The Acoustic Vowel Space

We will now return to the question of why so many languages have the five vowels a, e, i, o, u, pronounced as in the Spanish words masa, mesa, misa, mesa, musa. Figure 4.3 shows the five Spanish vowels in relation to the boundaries of the formant space that can be produced by an average male speaker. The scale used for the second formant frequency has the frequency values unequal distances apart. They have been arranged this way because the ear hears the higher frequencies as if they were closer together on the scale.

In the case of Spanish (and many other languages with five vowels) the vowel u has a low second formant. It thus occupies the space on the left of the diagram, where the English vowel in who might have been, and indeed was until comparatively recent times. In most English dialects the vowel in who now has a higher second formant than the corresponding Spanish vowel.

The Spanish vowels are fairly evenly distributed near the perimeter of the vowel space. The three vowels i, a, u are near the corners of this space, and thus as far apart from each other as possible. The other two vowels, e and o, are spaced at intermediate distances. Because the three vowels i, a, u are as auditorily distinct as possible, they are very effective ways of distinguishing words, and many languages make use of them. If we think of languages from an evolutionary point of view, it is hardly surprising that by far the majority have evolved so that they have vowels similar to the three Spanish vowels i, a, u, which provide such efficient means of communication. If there are two further vowels, as in Spanish, they are likely to be placed so that the resulting set of five vowels is distributed in the possible vowel space in the most efficient way. This will involve arranging the vowels as shown in figure 4.3 for Spanish. These vowels are as auditorily distinct as any five vowels can be.

When we consider the development of vowel systems we must note that the vowel systems of the world’s languages also show evidence of another constraint discussed in chapter 1, the pressure to form patterns. Given that the auditory space for possible vowels is somewhat triangular, the selection of the three most distant vowels i, a, u is obviously beneficial. It would be possible for languages to add just one vowel to these basic three, and, indeed, some languages do have only four vowels. But it turns out that far more languages have five or seven vowels than have four or six. With five or seven vowels it is possible to have a nicely symmetrical triangular vowel space.

If a language uses a still larger number of vowels, then they may be distinguished not only by differences in their formant frequencies but also by other differences such as length. English has several vowels that are kept distinct in this way. For example, the vowel in heed is different from that in hid not only by having a lower first formant, but also by being longer. In the next chapter

Figure 4.2 The values of the first two formants in some English vowels. The solid curve marks the limits of the possible vowel space.

The mean pitch of this formant in the vowels in hawed, hod, had, head, hid, heed. The words are listed from left to right mainly in order of increasing frequency of the overtone that is heard when whispering (the second formant). The highest first formants will be when the second formant is in the middle of its range. The lowest first formants will be when the second formant is very high (as you can hear it is when you whisper heed). The solid curve in the figure shows the limits on the first and second formant frequencies that can occur, given an average-size male vocal tract.

There is a gap on the left of the curve in figure 4.2 where English could have had an additional vowel or two. As we will see, this is because in most dialects of English the vowels in hooed and who’d at one time occupied this space. They are now pronounced with formants that would place them more in the middle of the diagram. If you say moo, the sound a cow is conventionally supposed to make, you may produce something more like the vowel that would be in the lower left of the figure.

You can hear a descending pitch when you say the words hod, hawed, hooed, who’d in a creaky voice, just as you did when you said hooed, head, hid, heed in that way. This is because the first formant goes down in pitch in both these sets of words. It is the second formant that has shifted upwards in the last two words so that the vowels in these words are more in the center of the space.
we will see how we can represent the English vowels on a graph showing the formant frequencies.

### 4.3 Sound Spectrograms

We have been describing vowels as if they were distinguished by only two formants, but actually the situation is more complicated. There is a third formant that is important for distinguishing some sounds, notably the r-colored vowel that occurs in American English pronunciations of words such as _bird_, and the French vowel that occurs in _tu_ (you). There are also formants with even higher pitches that add to the overall vowel quality. We can see the more complete set of formants that occur by making a computer analysis of the sound waves in a set of words. The top part of figure 4.4 shows the sound wave of the words _bead, bid, bed, bad_. The lower part of the figure is a computer analysis showing the component frequencies (in the form of a sound spectrogram). Time runs from left to right, as for the sound wave. The frequency scale (shown on the left) goes up to 4,000 Hz. The dark bands with white lines running through their centers are the formants, with the degree of loudness (the amplitude) of each formant being shown by the darkness of the band. The white lines are not part of the computer analysis but were added so as to make the locations of the formants more obvious. Sound spectrograms of this kind (with or without the white lines) are powerful tools for describing speech sounds, and we will be using them extensively in this book.

In this spectrogram, the formants are far from straight lines. But in general you can see that the first formant frequency in the first word is lower than it is in the second word, and gets steadily higher in each succeeding word. You can hear this change in pitch when you say these words with a creaky voice. The second formant frequency goes steadily down, as it does when you whisper them. The third formant also moves down slightly.
4.4 Summary

The principal characteristics of the sounds of vowels are the groups of overtone pitches, known as formants, which are produced by the vibrations of the body of air in the vocal tract. The acoustic vowel space can be considered to be an area bounded by the possible ranges for the frequencies of the first two formants. Computer analysis of speech can be used to produce spectrograms, graphical displays in which the formants appear as dark bars.

5 Charting Vowels

5.1 Formants One and Two

Earlier in this book, when we were considering the information conveyed by the pitch of the voice, we noted that men’s and women’s voices differed substantially in pitch, but they conveyed the same information. What mattered was the relative pitch within a sentence, whether it went up or down at the end, or which word stood out from the others because of a difference in pitch. It is the pattern of pitch changes that counts, not the exact frequencies involved. We must keep the same considerations in mind when discussing vowels.

The vowels of a particular speaker can be described precisely by stating their formant frequencies. But some speakers with big heads will have large resonating cavities, producing formants with comparatively low frequencies; and others will have higher formant frequencies because they have smaller vocal tracts. In order to represent the vowels of a language we need to show the average values of the formants.

The most useful representation of the vowels of a language is a plot showing the average values of formant one and formant two for each vowel as spoken by a group of speakers. We can also get this plot to reflect the approximate tongue positions in vowels by arranging the scales appropriately. When you say the vowel i as in heed, you pull the tongue up so that it is high and in the front of the mouth. Breathe in while holding the position for i, and you will be able to feel the cold air rushing through the narrow gap between the tongue and the roof of the mouth. Now say the vowel u as in who’d. If you hold this position while breathing in you will find that the tongue is still high in the mouth, but you can feel the rush of cold air further back in the mouth.
The vowel \textit{a} as in \textit{had} has a very low tongue position. Look in a mirror and you will see that the mouth is wide open and the body of the tongue is very far back in the mouth. We should also note that when you say \textit{i} as in \textit{heed}, the corners of your lips are spread apart, but for the vowel \textit{u} as in \textit{who'd} the corners of the lips are more together. We can characterize these two lip positions as being unrounded in \textit{i} as in \textit{heed}, as opposed to rounded in \textit{u} as in \textit{who'd}.

Now let us think about these facts in relation to the formant frequencies. Figure 5.1 is a modified version of part of figure 4.2, showing the first two formants of the vowels in the words \textit{had}, \textit{heed}, and \textit{who'd}. The principal modification of the earlier figure is that the bar for \textit{who'd} has been added to the vowels in \textit{had} and \textit{heed}. Its first formant is similar to that of the corresponding Spanish vowel, so the bar is about the same height. But it has a higher second formant than the corresponding Spanish vowel, so the bar is placed more to the right on the horizontal scale.

As we saw when we first discussed formants, the first formant has a low pitch for \textit{i} (\textit{heed}), a high pitch for \textit{a} (\textit{had}), and a low pitch again for \textit{u} (\textit{who'd}). This is the opposite of what the tongue does (high in \textit{heed}, low in \textit{had}, and high again in \textit{who'd}). So if we want to make a chart that shows both the vowel formants and an approximation of the tongue positions, we have to make the scale for the first formant go downwards, with low values at the top and high values at the bottom. By turning figure 5.1 upside down, as in figure 5.2, we can plot the formants in a way that will match the articulations in this respect. An indication of the tongue height has been added on the right of the figure.

We also saw that the second formant is high for \textit{i} (\textit{heed}) and lower for \textit{u} (\textit{who'd}). The tongue is in the front of the mouth for \textit{i} (\textit{heed}) and in the back in \textit{u} (\textit{who'd}). Traditional phonetic diagrams have the front of the mouth on the left and the back on the right. We can achieve this pattern by reversing figure 5.2, making the scale go from right to left, so that a low value suitable for \textit{u} (\textit{who'd}) will be on the right and high values like those in \textit{i} (\textit{heed}) will be on the left, as in figure 5.3. The only problem with this figure is that it is not a simple representation of the front–back position of the tongue. The tongue is certainly in the front of the mouth for \textit{i} (\textit{heed}) and towards the back for \textit{u} (\textit{who'd}), but the vowel in which the tongue is furthest back in the mouth is \textit{a} (\textit{had}). So it is apparent that the frequency of the second formant is related to something else as well as tongue position. The answer is that a major part of the lowering of the second formant is due to the increase in lip rounding. The lips are much closer together for \textit{u} (\textit{who'd}) than they are for \textit{a} (\textit{had}). Accordingly, the scale should be taken as reflecting lip rounding as much as the backness of the tongue, as indicated at the bottom of the picture.

The five vowels of Spanish, which we first illustrated in figure 4.3, can be represented in this way, as shown in figure 5.4, a reversed and inverted version of figure 4.3. In this case the bars for each vowel have been replaced by points. This kind of representation of vowels is known as a formant chart.
Figure 5.5  Formant charts of a conservative style of General American English. The vowels in bay and bow (rhyming with bee, not hoo) are represented by lines rather than points, as these vowels are diphthongs.

5.2 Comparing English Vowels

Figure 5.5 is a formant chart of the vowels of one form of General American English. Most of the vowels are represented by points, but the vowels in bay and bow (as in bow tie) are indicated by lines, to show that these vowels are diphthongs, starting with one vowel quality and ending with another. The scales in this figure (and in all the formant charts in this book) have been arranged to reflect how people hear the distances between vowels. The points representing the vowels in bee and bid are about as far apart as those for bid and bed, which is how listeners judge the distances between these pairs of vowels. The vowels in bid and bed are much further apart, in accordance with the fact that listeners hear these two vowels as being very different. Remember that these charts are for one form of General American English, so the vowels may not be the same as those with which you are familiar. They are based on data from speakers in the eastern part of the United States, who were recorded in the 1950s. Accordingly they represent a more old-fashioned dialect than that of most readers of this book.

The first thing to note about these charts is the difference between the women on the left of the figure, and the men on the right. The men's vowels have lower formant frequencies, resulting in their chart being more compressed, with all the points being moved upward and to the right. This is because men have larger vocal tracts, containing bigger bodies of air. These larger bodies of air vibrate more slowly, so that the formants have lower frequencies. But the pattern of vowels is much the same for both men and women.

Vowel charts provide an excellent way of comparing different dialects of a language. Figure 5.6 shows the average vowels of Southern Californian women and men. There are several differences from the more conservative General
American dialect represented in figure 5.5. First, in contemporary Californian English the vowels in bay and bow have only a small change in quality from beginning to end, so it is possible to represent them by points rather than lines. The quality of the vowel in bay is in between that of the vowels in bee and bid. What is not shown on the chart is that the vowels in bay and bid (or bait and bit, to give a minimally contrasting pair) also differ in length. The vowel in bait is longer than that in bit. Another difference between the two dialects is that the contemporary Californian vowels in bee and hood have a much higher second formant frequency — they are further to the left in the diagram — than their conservative General American counterparts. Finally, as we noted in chapter 4, Californians do not distinguish between the vowels in bod and bawd (or cot and caught), making all these words with a vowel sound having a comparatively high first formant frequency.

There are several other distinct dialects of American English, but we will consider only one of them which is notable because of an interesting sound change that has occurred in comparatively recent years. In many of the northern metropolitan areas of the United States, such as Detroit and Rochester, a form of speech has developed that has come to be known as Northern Cities dialect. Formant charts of this form of speech are shown in figure 5.7. The most striking feature of this dialect is the relative position of the vowels in bod and bed, which is the reverse of their positions in General American English. The vowel in bod has come to have a lower first formant, so it has moved to a higher position on the chart. Although they are close together on the chart, the vowels in bod and bed are still quite distinct because bod is longer and has a slightly diphthongal quality.

The vowels of BBC English are shown in figure 5.8. Many of them are in fairly similar positions to the corresponding vowels shown in figures 5.5 and 5.6 for American English. This is hardly surprising as most of us can readily understand both accents. The major differences are due to the additional vowels of BBC English that we discussed in chapter 4, and the large change in quality that occurs in the diphthong in words such as no, dough, and bow.

In the middle of the chart in figure 5.8 is the vowel in bird, which has the same quality as the first vowel in about or the last vowel in sofa, and like them has no r-coloring. As a result of this vowel in bird occupying the space in the center, the BBC English vowel in bud is pushed down so that it is somewhat lower on the chart, with a higher first formant frequency than its General American counterpart. Another additional vowel is that in bard, which also has no r-coloring, but is nevertheless distinct from the vowels in bod and bawd. In BBC English cart, cot, caught are all pronounced differently, all of them without any suspicion of an r sound. The presence of this new vowel in bard and caught pushes the vowel in bod and cot higher on the chart. It has a lower first formant frequency than its General American counterpart. The vowel in bawd and caught is also pushed higher, and in addition has a considerably lower
second formant frequency than the corresponding American vowel, almost pushing it off the chart on the right-hand side. This is largely because the British English vowel has more lip rounding.

As with American English, there are many varieties of British English. London Cockney English differs from BBC English in having large diphthongs in words such as buy and mute, which sound superficially like those in buy and might. They are actually somewhat different, and Cockney speakers do not confuse these words. Buy has changed so that it begins with a vowel more like that in BBC English bud, but buy is also different from BBC English in that it begins with a vowel more similar to that in BBC English father.

5.3 Formant Three

Formant charts such as those we have been discussing do not take the frequency of the third formant into account. This formant has very little function in distinguishing the vowels shown. For all English vowels, with one notable exception, the third formant frequency can be predicted fairly accurately from the frequencies of the first two formants. The exception is the vowel in most forms of American English bird. As with many other sounds associated with r, this vowel has a very low third formant frequency. Its first and second formants are often very similar to those in loud, but it has a very different quality. It would be misleading to plot it on a chart that shows only the first and second formant frequencies.

The frequency of the third formant is very much affected by the position of the lips. This makes it more important in some languages. It so happens that neither General American English nor BBC English have vowels with the same tongue positions but different lip positions. In English, when the tongue is high in the front of the mouth (as for i as in head), the lips are fairly far apart. But in French, German, and many other languages, the position of the lips (and hence the frequency of the third formant) is not so predictable. We will discuss these languages in chapter 14.

5.4 Summary

Vowels can always be accurately described in terms of the frequencies of the first three formants. It is often sufficient to plot the frequencies of the first two formants on a formant chart. Given proper scales that reflect how the differences between vowels are perceived we can give good descriptions of the dialects of English. The only English vowel in which the third formant plays a significant role is the vowel in bird as pronounced in General American English.

6

The Sounds of Consonants

6.1 Consonant Contrasts

There are only minor differences in the consonants of the principal dialects of English, so we need not discuss British and American English separately. Table 6.1 shows all the consonants before e as in bet and a as in buy. Recording 6.1 illustrates my pronunciation of these words. The first column in the table lists the appropriate symbols of the International Phonetic Alphabet. Nearly all of them are familiar letters of the alphabet. We will note the special symbols when we refer to them. The final column of this table groups the consonants in accordance with certain features that they have in common. We will discuss the sounds of consonants with reference to these features.

6.2 Stop Consonants

Many consonants are just ways of beginning or ending vowels. This is particularly true of consonants such as b, d, g, each of which has a rapid movement of the lips or tongue before or after another sound such as a vowel. They are called stop consonants because the air in the vocal tract is completely stopped at some point. When forming a b in a word such as bib, the lips are firmly closed at the beginning and end of the word. In the case of d as in did it is the tip of the tongue that blocks the vocal tract by forming a closure just behind the upper front teeth. For g as in gag, the back of the tongue is raised to make a closure against the roof of the mouth.
Table 6.1  English consonant sounds, illustrated as far as possible before the vowel e as in bet, and the vowel aɪ ('AI') as in buy. The first column shows a set of IPA symbols that can be used to transcribe them, and the final column names a feature that groups sets of consonants together.

<table>
<thead>
<tr>
<th>IPA</th>
<th>Vowel e</th>
<th>Vowel aɪ</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>bet</td>
<td>buy</td>
<td>stop</td>
</tr>
<tr>
<td>d</td>
<td>debt</td>
<td>die</td>
<td>stop</td>
</tr>
<tr>
<td>g</td>
<td>get</td>
<td>guy</td>
<td>stop</td>
</tr>
<tr>
<td>p</td>
<td>pet</td>
<td>pie</td>
<td>stop</td>
</tr>
<tr>
<td>t</td>
<td>ten</td>
<td>tie</td>
<td>stop</td>
</tr>
<tr>
<td>k</td>
<td>ken</td>
<td>kite</td>
<td>stop</td>
</tr>
<tr>
<td>w</td>
<td>wet</td>
<td>why</td>
<td>approximant</td>
</tr>
<tr>
<td>j</td>
<td>yet</td>
<td>lie</td>
<td>approximant</td>
</tr>
<tr>
<td>l</td>
<td>let</td>
<td>rye</td>
<td>approximant</td>
</tr>
<tr>
<td>m</td>
<td>met</td>
<td>my</td>
<td>nasal</td>
</tr>
<tr>
<td>n</td>
<td>net</td>
<td>nigh</td>
<td>nasal</td>
</tr>
<tr>
<td>ŋ</td>
<td></td>
<td></td>
<td>nasal</td>
</tr>
<tr>
<td>f</td>
<td>fed</td>
<td>fie</td>
<td>fricative</td>
</tr>
<tr>
<td>θ</td>
<td>set</td>
<td>sigh</td>
<td>fricative</td>
</tr>
<tr>
<td>s</td>
<td>shed</td>
<td>shy</td>
<td>mission</td>
</tr>
<tr>
<td>j</td>
<td>hen</td>
<td>high</td>
<td>fricative</td>
</tr>
<tr>
<td>h</td>
<td>vet</td>
<td>vie</td>
<td>fricative</td>
</tr>
<tr>
<td>ð</td>
<td>then</td>
<td>thy</td>
<td>fricative</td>
</tr>
<tr>
<td>z</td>
<td>Zen</td>
<td>Zion</td>
<td>mizzen</td>
</tr>
<tr>
<td>ʒ</td>
<td></td>
<td>vision</td>
<td>fricative</td>
</tr>
<tr>
<td>ʃ</td>
<td>Chet</td>
<td>chime</td>
<td>affricate</td>
</tr>
<tr>
<td>ʒ</td>
<td>jet</td>
<td>jive</td>
<td>affricate</td>
</tr>
</tbody>
</table>

Each of these stop consonants forms a quickly changing sound. The resonances of the vocal tract, the formants, are being produced while the stop closure is being formed or is opening. The shape of the vocal tract is changing, and as a result the formant frequencies are moving. Figure 6.1 shows the formant frequency movements associated with the stop consonants b, d, g in the words bab, dad, gag. The location of each of the first three formants at the start of the syllable is marked by an arrow.

At the beginning of the first word, bab, all three formants move rapidly upward. The resonances of the vocal tract when the lips are almost closed are in the low range. The formant frequencies move up from this low starting point to the higher resonant frequencies that occur in the vowel. At the beginning of the word dad the first formant goes up, the second moves very little and the third moves slightly down. The d closure, with the tip of the tongue behind the teeth, makes a vocal tract shape that has resonances as shown by the positions of the three arrows. As the stop is released and the vocal tract moves away from this shape, the formants move away from the position indicated by the arrows. For the final word, gag, the most notable movements are those of the second and third formants. The shape of the vocal tract when the back of the tongue is almost touching the roof of the mouth, as it does in g, makes the second and third formants have very similar frequencies. The resonances of the vocal tract when the stop is being released are as shown by the arrows.

The movements of the second and third formants are the distinguishing characteristics of the stop consonants. The movements of the first formant simply mark them as having a stop closure. For all three of the sounds we have been considering, the frequency of the first formant increases when they are at the beginning of a syllable, and falls when they are at the end. The movements of the other two formants distinguish these sounds from one another. In general, if a word or syllable starts with both the second and third formants increasing in frequency, then the sound is a b. If the third formant falls and the second formant has only a small movement it is a d. If the second and third formants are close together just after a stop has been formed, then the back of the tongue has contacted the roof of the mouth, as in a g.

The stops b, d, and g also affect the formants at the ends of the words in figure 6.1. In each of these words the movements of the formants at the end are the reverse of those at the beginning. The first formant moves down as the stop is formed in all three words. The second and third formants move down
Figure 6.2  Spectrograms of stops in bib, did, gig. The arrows mark the origins of the first three formants.

at the end of bab, reversing the movement at the beginning of the word. There is very little movement of the formants at the end of dad, just as there was little movement at the beginning. With a little imagination one can see that at the very end of dad there is a small downward movement of the second formant and an indication of an upward movement of the third formant. At the end of gig the second and third formants clearly come together, forming a kind of mirror image of the way they came apart at the beginning of the word.

There are similar movements of the formants when these consonants occur with other vowels. Figure 6.2 shows spectrograms of the words bib, did, gig. In these words, too, you can see that the sound b is associated with a lowering of all three formants, for d the second and third formants are level or pointing upwards, and for g the second and third formants are closer together. These formant movements characterize these stop consonants.

English has another set of stop consonants, the sounds p, t, k, as in the words pip, tit, kick. (We use the symbol k for this sound, irrespective of whether it is spelled with 'c', as in cat, or 'k' or 'ck' as in kick, or 'q' as in quick, which can be considered as beginning with kw, or even 'x' as in tax, which ends in -ks.) The sounds p, t, k are made with the same gestures of the lips or tongue as b, d, g. For p and b the lips close, for t and d the tip of the tongue makes contact just behind the teeth, and for k and g the back of the tongue contacts the roof of the mouth. Consequently the movements of the formants are similar for these two sets of sounds. The difference between the two sets is in the action of the vocal folds. When you say a word beginning with b, d, or g the vocal folds are vibrating while the lips or tongue are moving apart. But for p, t, and k the vocal folds are apart at the beginning of the movement. Consequently at the beginning of each of the words pip, tit, kick there is a burst of air that produces a different kind of sound. Instead of the sound produced by the action of the vocal folds, which has a definite pitch and overtones corresponding to a particular vocal tract shape, there is a more noisy sound with a less well-defined pitch. You can hear a difference between the noisy bursts that occur at the beginning and end of each of the words pip, tit, kick, if you say just the p, t, k sounds by themselves. The bursts of noise for p at the beginning and end of pip are very small and with no definite pitch. There is a comparatively high pitch associated with the burst of noise for t in tit, and a somewhat lower pitch for the k noise in kick.

So that you can compare the sounds p, t, k with b, d, g in bab, dad, gag, figure 6.3 shows a spectrogram of pap, tat, kack. The first two are real words in my vocabulary, but the third is a nonsense word beginning in the same way as cat, but ending so that it rhymes with back. The bursts of noise at the beginning of each of these words are enclosed within dashed lines. (The bursts that sometimes occur at the ends of these words were not produced when these words were said. They are more commonly heard when the next word begins with a vowel.) The burst associated with p is fairly faint, and scattered over a wide range. That associated with t is in the higher frequency range, and in fact extends far above the range of frequencies shown in this analysis. The k burst has its greatest intensity between 2,000 and 3,000 Hz. The burst for k is also slightly longer than that for t, and both of them are longer than that for p.

You cannot see the movements of the formants at the beginnings of pat, tat, cat, as the corresponding movements of the lips and tongue take place during the noise bursts. But you can see some of the changes that occur for the stops at the ends of these words. There are small downward movements at the end of pap, just as there were at the end of bab. There is very little movement at the end of tat, which is what we found at the end of dad. At the end of kack there is a tendency for the second and third formants to come together, as happened at the end of gag. The movements of the formants are not so clear at the ends
of *pat*, *tat*, *tak* because in my speech (as for many speakers of both BBC English and General American English) the vocal folds close tightly together just as the final consonants in words of this type are being made, cutting off the last part of the syllable. This is also the reason why there are no bursts at the ends of these words when they are said in a normal unemphatic way.

### 6.3 Approximants

The next set of consonants listed in table 6.1 are those in words such as *wet*, *yet*, *let*, and *retch*. The IPA symbols for them are w, j, l, r, all of them having the same values as in English words except for j, which corresponds to English ‘y’. The International Phonetic Association chose j to represent the sound at the beginning of *yet*, as ‘j’ has this value in several Northern European languages such as German and Danish. In these languages words such as ja (yes) and jung (young) begin with a sound we would write with ‘y’. The sounds w, j (‘y’), l, r are called ‘approximants’. They are the opposite of stop consonants in that they do not involve any kind of closure of the vocal tract. Instead there is simply a narrowing at some point. For w the lips are close together and the back of the tongue is raised, but air still flows freely out of the mouth. For j (‘y’), it is the front of the tongue that is raised, but not far enough to hinder the airflow. The sound l is peculiar in that the tip of the tongue comes up and touches the upper teeth or roof of the mouth much as it did for d and t, but air flows out freely over the sides of the tongue. The sound r is more difficult to describe, partly because different speakers make it in different ways. It usually involves some raising of the tip of the tongue toward a point on the roof of the mouth well behind the upper front teeth.

Each of these gestures is associated with particular formant frequencies. As you can see in figure 6.4, the most conspicuous aspect of the w in *wet* is the rising second formant. The first formant also goes up, but less noticeably (partly because it is very faint at the start of the word), and the third formant has much the same frequency at the beginning and end of the word. The j (‘y’) in *yet* has a falling second formant, a more visible rise in the first formant, and a drop in the third formant.

The l in *let* differs from the first two sounds (and from the last) in that there is a distinct break in the pattern at the moment indicated by the arrow. Before that time there is a faint formant bar at a very low frequency, and another faint bar at about 1,500 Hz. Immediately after the arrow the formants have a much higher intensity (the bars are darker) and are at a distinctly different frequency. The same kind of change in the pattern also occurs in the higher frequencies above 3,000 Hz. These changes occur because there is an abrupt change in the articulation – the tip of the tongue is in contact with the roof of the mouth for the l, and then breaks away from it for the vowel.

#### Figure 6.4

Spectrograms of approximants in *wet*, *yet*, *let*, *retch*. The arrow below the third spectrogram marks the moment when the tip of the tongue, which is raised for l, comes away from the roof of the mouth. The arrow in the fourth spectrogram shows the low beginning of the third formant.

The r at the beginning of *retch* is characterized by the very low frequency of the third formant. All the formants rise at the beginning of this word, but it is the movement of the third formant that is most significant. Whenever there is an r in a word the third formant will be below 2,000 Hz, sometimes, as shown by the arrow in this example, falling to as low as 1,500 Hz.

### 6.4 Nasals

The next set of sounds listed in table 6.1 are called nasals because they involve sound radiated while air comes out through the nose. The nasal sounds m, n, and ng (the phonetic symbol for ‘ng’) occur at the ends of the words *ram*, *ran*, and *rang*. They are like vowels and approximants in that they can be characterized largely in terms of their formant frequencies, but they differ in that the formants are not as loud as they are in vowels. The nasals are made by blocking the sound from coming out of the mouth while allowing it to come out through the nose, and this affects the relative amplitude (the loudness) of the formants.

Spectrograms illustrating the three nasals in the words *ram*, *ran*, *rang* are shown in figure 6.5. (These spectrograms also provide further illustrations of the r with its low third formant at the beginning of each of these words.) Just before the nasals at the end of each of the words in figure 6.5 there is a sharp discontinuity (marked by an arrow) when the lips come together or the tongue comes up to contact the roof of the mouth. After this point there is less
amplitude in the nasal consonant itself. All three nasals have a first formant which has distinctly less energy (is fainter) than in the preceding vowel, and which has a very low frequency, around 200 Hz. For each of them there is also another formant visible in the neighborhood of 2,500 Hz, but there is comparatively little energy in the region normally occupied by the second formant. This pattern typifies the nasal consonants.

The main differences between the three nasals are not in the wholly nasal portions, but in the onsets to these portions. The ways the formants move as the nasal is being formed are very reminiscent of the formant movements in /ab/, /dad/, /gog/, which we saw in figure 6.1. As the lips close for m the formants (particularly the second) lower in frequency, just as they did before the lip closure for b. The formant frequencies just before the tongue tip closure for n are very similar to those for m, but the third formant is very slightly higher. The \( \eta \) ('ng') at the end of rang is far more distinct. The second and third formants come together just as they did in gog.

6.5 Fricatives

There are several consonants that are produced without vibrations of the vocal folds. Prominent among them are the consonants in the words /fi/, /thigh/, /sigh/, /shy/, each of which begins with a sound in which the vocal folds are held apart so that they do not vibrate. In these consonants the noise is made by air being forced through a narrow gap. Instead of formants – resonances of the vocal tract – their most prominent acoustic features are higher-pitched, more random, noises, akin to the screech of the wind as it blows around a corner. The sound at the beginning of each of the words /fi/, /thigh/, /sigh/, /shy/ is called a voiceless fricative – voiceless because the vocal folds are not vibrating, and fricative to indicate that the noise is produced by the friction, the resistance to the air as it rushes through a narrow gap.

Figure 6.6 shows a spectrogram of the four voiceless fricatives that occur in English. The pattern of the noise that occurs at the beginning of each of the words /fi/, /thigh/, /sigh/, /shy/ is similar to the bursts that we saw at the beginning of each of the words /pap/, /tat/, /lack/ in figure 6.3. In all these voiceless sounds there are no pulses produced by the vocal folds. Whenever the vocal folds are vibrating there is a series of pulses that produce amplitude peaks in the sound wave. These peaks are visible on spectrograms as vertical striations. Figure 6.7 shows an expanded spectrogram of the first part of the word /fi/, with the waveform showing the vocal fold pulses immediately above. As may be seen, there are about 10 pulses in 100 milliseconds (ms), or one pulse in one-hundredth of a second (10 ms). There are no such regular pulses in the fricative portion before the vowel. The energy is scattered over the higher frequencies. It is not, however, randomly distributed, and each of the four English fricatives has a distinct pattern.

The spectrogram of /f/ as in /fi/ on the left of figure 6.6 has noise spread over a wide range of frequencies. Insofar as there is a region in which there is greater intensity it is between 3,000 and 4,000 Hz. The spectrogram of /b/ (the phonetic symbol for 'th' as in /thigh/) also shows energy over a range of frequencies, but in this case it is centered in the higher frequency range, above 8,000 Hz. There is often very little difference in the fricative noises of these two sounds – neither of them is very loud. There are, however, differences in the formants of the adjacent vowels. The fourth formant is below 4,000 Hz in /fi/ and above it in /thigh/. The second formant in /fi/ also starts at a slightly lower frequency, around 1,200 Hz, and then moves noticeably upwards. At the start
of *thigh* the second formant is fairly level at around 1,250 Hz. In many circumstances, such as over a telephone or in a room full of people, the fricative noise is inaudible, and the formant movements are the only cues distinguishing these two sounds. But in any case the differences in the fricative noise and in the formant movements are both very small, and it is not surprising that English is one of the few languages in the world that uses both these sounds. Greek and Spanish are the only other well-known European languages that contrast them.

The next fricative is *s* as in *sigh*. It has a large amount of energy in the upper part of the figure, extending even above the 10,000 Hz shown in the figure. There is comparatively little energy below 3,500 Hz, and a noticeable intense band above 5,000 Hz. The sound *f* (the phonetic symbol for *sh* as in *shy*) has more energy at a slightly lower frequency, centered at a little above 3,000 Hz. Say a long *f* (‘*sh*’), as if you were saying *shush* to a child, and then a long *s*, as if you were hissing a villain. The *f* (‘*sh*’) sounds lower in pitch than the *s*. Neither of these sounds said in isolation has a real pitch in the way that a vowel does. For that you have to have a repeating waveform of the kind that is produced by regular repetitive movements of the vocal folds. But you get some impression of pitch from the fact that the energy is in specific parts of the frequency range.

The fricatives *s* as in *sigh* and *f* as in *shy* are sometimes called sibilants. They have a greater intensity – they are louder – than the other two voiceless fricatives *f* and *θ* (‘*th*’). As a result there are darker marks in the upper frequency range on the spectrogram. If you make each of the sounds *f*, *θ* (‘*th*’), *s*, *f* (‘*sh*’) one after the other without any vowel in between, you can hear that the first two fricatives are far less loud than the other two.

The next consonant listed in table 6.1 is *h* as in *high*. It is convenient to discuss it at this point, although it is not really a voiceless fricative as the source of the noise is not air being forced through a narrow gap. Instead the origin of the sound is the turbulence – the random variations in air pressure – caused by the movement of the air across the edges of the open vocal folds and other surfaces of the vocal tract. Because the principal origin of the sound is deep within the vocal tract, rather than near the lips or the front of the mouth, the resonances of the whole vocal tract will be more prominent, and the sound is more like that of a noisy vowel.

Figure 6.8 shows spectrograms of *h* in *high* and *behind*. In *high* there is a noisy third formant at a little below 3,000 Hz, and there are even faint traces of the first two formants. In *behind*, in which *h* is between two vowels, the vocal folds draw apart without completely ceasing the vibrations that occur in the vowels, so that they are still flapping in the airstream. There are noisy forms of the first two formants, as well as energy in the higher frequencies.

Each of the four voiceless fricatives *f*, *θ* (‘*th*’), *s*, *f* (‘*sh*’) has a counterpart in which the vocal folds are vibrating while the fricative noise is being formed by forcing air through a gap. The *f* in *fee* contrasts with the *v* in *vie* in this way. A similar contrast occurs in *thigh* and *thy*. In this case English spelling does not distinguish the two forms of ‘*th*’. As we have seen, the phonetic symbol for ‘*th*’ as in *thigh* is *θ*. The symbol for ‘*th*’ as in *thy* is *ð*, a modified version of the old Anglo-Saxon letter for this sound. The voiceless fricative *s* has a voiced counterpart *z*. The nearest word showing the contrast with *s* in *sigh* is *Zion*. A pair of words that differ only in that one has *s* and the other has *z* is *seal* and *zeal*. The counterpart to *f* as in *shy* occurs in relatively few words, usually between vowels as in the middle of *measure*. This sound, for which the phonetic symbol is *ʒ*, cannot appear at the beginning of a word other than in a foreign name such as *Zsa Zsa*, *3a 3a*. The distinction between *f* and *ʒ* is best
exemplified by words such as mission, which has f, versus vision which has ʒ in the middle.

Voiced fricatives have formants produced by pulses from the vocal folds as well as more random energy produced by forcing air through a narrow gap. Figure 6.9 shows spectrograms of the voiced fricatives in vie, thy, Zion. In all three of these words very faint formants can be seen during the initial fricatives. In the first two words, vie and thy, there is only a little random energy in the higher frequencies, but in the third word, Zion, the effects of the turbulent airstream produced by the friction are clearly visible. We saw in the discussion of figure 6.6 that the corresponding voiceless fricatives, f and θ, had a relatively low-intensity noise in the higher frequencies when compared with s.

The difference between the voiced and voiceless fricatives ʒ and f is illustrated in figure 6.10 (which also shows a further example of the voiced fricative v). The fricatives are in the middle of each word, as indicated by the placement of the phonetic symbols. Under the ʒ in the first word, in the area between the dashed lines, there are vertical striations associated with vibrations of the vocal folds. These indications of vocal fold vibrations are rather difficult to see, so I have added lines at the top of the picture that make them a little clearer. Under f there is only the noise due to the turbulent airstream.

The last sounds to be considered in this chapter, the initial consonants in chime and jive, are not really single sounds. The ‘ch’ in words such as chime and chip consists of the t in time or two, followed by the f in shine or ship. You can see this if you compare phrases such as gray chip and great ship. In the first of these phrases the two consonants are at the beginning of the second word, but in the second phrase the t is at the end of the first word and the f is at the beginning of the second. Similarly ‘j’ as in jive begins with the two consonants d and ʒ, although this cannot be illustrated in the same way, as English words do not begin with ʒ. The nearest I can get to showing this combination is the pair of phrases hide Zsa (Zsa) and high jar. The combination of a stop followed by a fricative as in f and θ is called an affricate.

Figure 6.11 illustrates these two sounds. There is, of course, little to see for the initial t in chime, except for the abrupt beginning of the following f. The vertical striations due to the vibrations of the vocal folds are just visible in ʒ in jive. Both the voiceless f and the voiced ʒ are considerably shorter than when they occur on their own.
6.6 Summary

The voiced and voiceless consonants of English can be divided into stops, approximants, nasals, fricatives, and affricates. Spectrograms of stops are marked by the abrupt beginning or ending of the adjacent vowel. The voiced stops b, d, g can be distinguished by the formant movements, b having a low second and third formant, d a second formant in the mid-range, around 1,700 Hz and a high third formant, and g having second and third formants that are very close together. The voiceless stops p, t, k may have formant movements similar to those of b, d, g after a vowel, but at the beginning of a word they are largely distinguished by the frequencies of the bursts of noise that are produced as the stop closure is released. The approximants w, j ('y'), l, r have their own formant patterns, w being similar to a movement away from the vowel u as in boo, and j ('y') being similar to a movement away from the vowel i as in bee. The approximant l has a low-intensity formant at a very low frequency, another low-intensity formant at about 1,500 Hz, and a distinct break in the pattern before a vowel. Whenever there is an r in a word the third formant will be very low, usually below 2,000 Hz. The nasals m, n, ñ ('ng') have first formants with a very low frequency, around 200 Hz, another formant visible in the neighborhood of 2,500 Hz, and comparatively little energy in the region normally occupied by the second formant. They have similar formant movements to the corresponding stops. The voiceless fricatives s and ð ('th') have energy over a wide range of higher frequencies. They are distinguished from each other partly by the formant movements. The other voiceless fricatives, s and ʃ ('sh'), have greater energy, with s being mostly in the high frequency range from about 3,500 Hz upwards, and ʃ having most energy somewhat lower, around 3,000 Hz. The corresponding voiced fricatives v, ð ('th'), z, ð (in vision) have similar energy distributions, but with the addition of formant resonances. The affricates ʧ ('ch') and ʤ ('j') are like the sequences t + ʃ ('ch') and d + ð. The consonant h has noisy forms of the formants in the adjacent vowels.

7

Acoustic Components of Speech

7.1 The Principal Acoustic Components

When I was a graduate student in the early 1950s I had the good fortune to hear a talk by one of the first people to realize how conversational speech could be fully described in a limited set of acoustic terms. He was an engineer named Walter Lawrence, who broke speech down into parts that he could represent by lines painted on a glass slide. Each line represented one of the ingredients - parameters as he called them - of running speech. He used a beam of light to read the lines that he had painted, and was able to turn them back into speech-like sounds. Lawrence's great insight was that it was possible to break speech into a set of components. We will follow (and slightly enlarge) his original system so that we can summarize the acoustic characteristics of speech that we described in chapter 6.

Let's begin with vowels. They can be distinguished by the frequencies of the first three formants. On a spectrogram you can see higher formants, but they do not vary much from one vowel to another. They differ from person to person and make people sound slightly different from one another, but they have little linguistic function. They distinguish people, not words.

Table 7.1 is the start of a list summarizing the acoustic characteristics of speech, showing the things we need to know other than those required for distinguishing one speaker from another. The table gives both the technical acoustic terms and the terms that we can use to describe what we hear. We'll refer to these as the auditory correlates. The vowels of English and most other languages can be described by stating the values of just the three things shown in the table.