6
Acoustics of speech sounds

When we speak to each other the sounds we make have to travel from the mouth of the speaker to the ear of the listener. This is true whether we are speaking face to face, or by telephone over thousands of miles. What is important for us in our study of speech is that this acoustic signal is completely observable: we can capture everything that the listener hears in the form of a recording, and then measure whichever aspect of the signal that we want to know about. There is an interesting observation to make here: for each of the phonetic classes of sound that we have identified, we can find corresponding acoustic patterns. However, if we had started by studying the types of acoustic pattern without knowing anything about how they were made by a human speaker, we would probably have set up a quite different way of classifying them. We will begin by setting out a classification of acoustic patterns, and then see how this fits with the traditional phonetic classification of speech sounds.

Acoustic waveforms

All audible sound is the result of variations in air pressure that produce vibration. In vibration, the pressure in a particular place (for example, inside the ear) becomes alternately higher and lower. This is usually described in terms of wave motion, using diagrams like Figure 6.1 that suggest up-and-down movement, though sound waves do not really move up and down like waves on the sea. They are more like the shock waves that travel outwards from an explosion. We can show the pattern of a particular sort of vibration by displaying its **waveform**. If the vibration

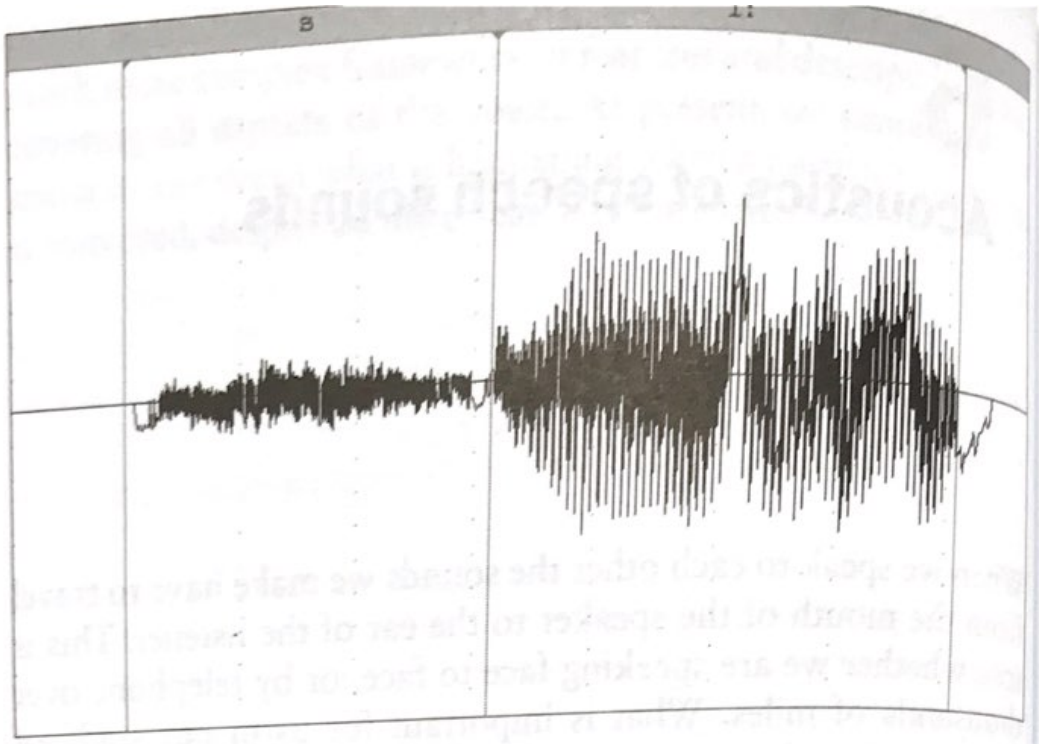

FIGURE 6.1 *Acoustic waveform of 'see' /si:/*

happens rapidly, we say it has a high **frequency**, and if it happens less rapidly, we say it has a lower frequency. If the vibration is regular, repeating its pattern over and over, we call the sound **periodic**, while a pattern of vibration which does not have such a pattern of regular vibration is called aperiodic. If the sound contains a large amount of energy, we say that it has high **amplitude**. Figure 6.1 shows the waveform for the word 'see'; the first part, /s/, is **aperiodic**, having an irregular, rather messy pattern, while the vowel /i:/ is periodic, and we can see a more regular pattern in its vibration.

It is a fundamental principle in acoustic analysis that any waveform, however complex it might be, can be broken down into simple waveforms of different frequencies. The operation of doing this is called **spectral analysis**, and in some ways is rather like breaking down white light into the rainbow pattern of colours that make up its spectrum. In carrying out the acoustic analysis of speech sounds, we can discover much more by looking at the result of a spectral analysis than by looking at the original waveform that was captured by the microphone. Figure 6.2 shows the picture resulting from the spectral analysis of the word 'see' that we have already looked at in Figure 6.1. This type of picture is called a **spectrogram**. At one time there was a fashion

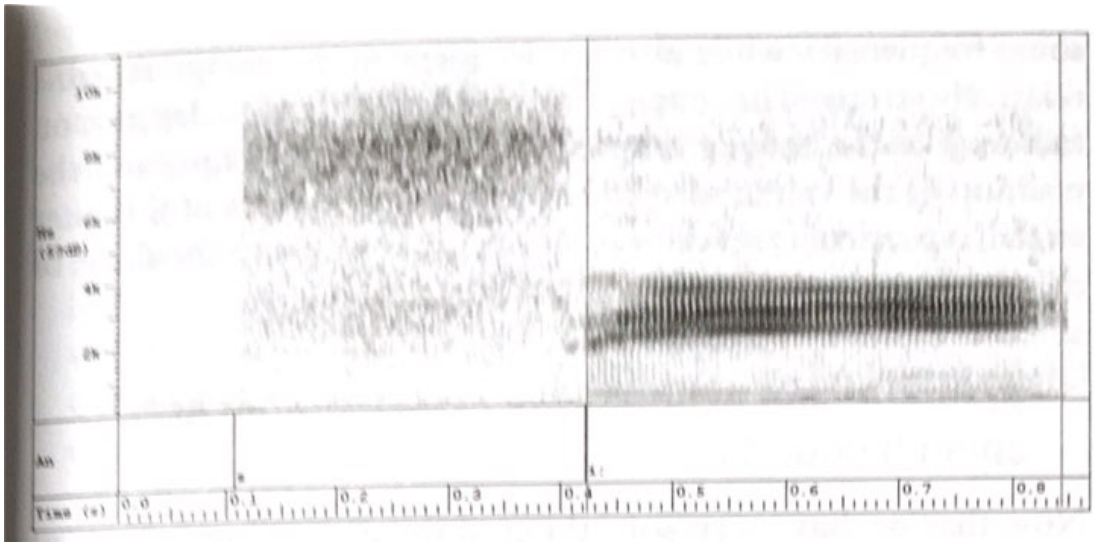

FIGURE 6.2 *Spectrogram of 'see' /siː/*

for calling such pictures 'voice-prints', but this led to some very dubious claims being made about identifying people by their voices for legal purposes, and the name is not now used except (sometimes) by gullible journalists.

In a spectrogram, the vertical axis of the picture represents the frequency scale: the lowest frequencies are shown at the bottom. From left to right is the time axis, with the beginning of the sound analysed shown at the left. The degree of blackness of the markings shows the amplitude at different frequencies in the signal at a particular point in time. You can see that in /s/ the energy is spread widely on the frequency range, but it is greater towards the higher frequencies and almost non-existent at the lowest frequencies. In /iː/, the energy is concentrated in three or four narrow bands (**formants**) in the lower part of the spectrum. Some spectrographic displays now show levels of energy with different colours instead, but although these look pretty and are nice to pin on your wall, most people find they are harder to interpret than the grey-scale spectrograms that have been around since the 1940s.

There is a general theory of how the acoustic signal is produced by the human vocal tract, based on the principle that we have some way of producing sound (a **source**), and for most sounds also a way of modifying that sound (a **filter**). This source–filter theory is widely accepted as a fundamental concept in speech acoustics. To take vowels as an example, the source for a vowel is the vibration of the vocal folds; as the vibrating flow of air passes through the vocal tract, the vocal tract acts as a filter, weakening the energy at

some frequencies while at other frequencies the energy remains relatively strong. The shape of the vocal tract (which depends on factors like the tongue-shape, the position of the lips, and the position of the velum) determines the characteristics of this filter so that a particular vowel is produced; if you change the shape of the vocal tract, you change the resulting vowel.

Acoustic and articulatory classification of speech sounds

Now that we have seen something of the physical properties of sound, we can see how these properties correspond to our more familiar and traditional phonetic categories. We can say that all speech sounds are made up of just four possible types of acoustic pattern:

1 Periodic sound
2 Aperiodic sound
3 A mixture of periodic and aperiodic sound
4 Silence.

1 Vowels. These are periodic sounds with a regular pattern of vibration. When their spectrum is analysed, it is possible to see peaks of energy at different frequency, rather like the notes in a musical chord. These peaks of energy (which we call **formants**) are different for every vowel, and acoustic phoneticians have analysed the frequencies of many different vowels so that we know a lot about how formants are related to vowel quality. Formants are seen on spectrograms as dark horizontal bars, as in the /i:/ vowel that you can see in Figure 6.2. Although the relationship is certainly not exact, it has been found that the formant with the lowest frequency (Formant 1) corresponds roughly to the traditional open/close dimension of vowels: a low Formant 1 corresponds to a **close** vowel like [i] or [u]. Formant 2, which is higher than Formant 1, corresponds roughly to the front/back dimension of vowels: a vowel with a high Formant 2 is likely to be a **front** vowel like [e] or [a], while a vowel with a low Formant 2 is more likely to be a **back** vowel like [o] or [ɑ]. It is not possible to give exact frequency values

for the different formants, because these vary from speaker to speaker, but the graph shown in Figure 6.3 places some English vowels spoken by an adult female speaker, with the axes arranged so that the positions of the vowels are roughly where they would be placed on a traditional vowel diagram. If you look carefully at textbooks which describe the acoustics of vowels, you will notice how depressingly often they assume that an adult male voice is 'normal' and give little or no detail of women's and children's voices.

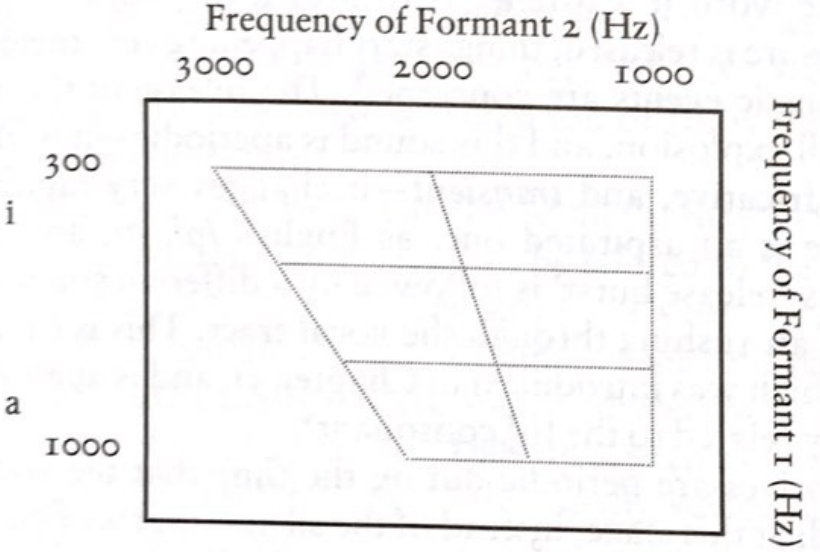

FIGURE 6.3 *Traditional (dotted line) and acoustic (solid line) representation of vowel positions*

2 Fricatives. Voiceless fricatives such as [s] and [ʃ] are aperiodic, and do not have formants in the way that vowels do. They do, however, have concentrations of energy at different frequencies. If you make [s] and [ʃ] alternately, you will hear that [s] sounds more high-pitched than [ʃ]. This is particularly noticeable in English, since English speakers tend to round their lips for /ʃ/, making this fricative sound even lower-pitched. You can check in a mirror whether you round your lips when you say an English /ʃ/. Voiced fricatives combine aperiodic with periodic sound. This is because they have the aperiodic hissing noise typical of voiceless fricatives such as [f], [s], or [ʃ], combined with vibration of the vocal folds. The vibration of the vocal folds happens in a very regular way, which means that the buzzing sound made in the larynx is periodic.

3 Plosives. Plosives occur in several different acoustic forms. We will begin by looking at voiceless plosives such as [p], [t], or [k]. Although most people are not aware of this, the first important component of these sounds is silence. When a voiceless plosive occurs at the beginning of a word (as in 'pin' /pɪn/), the word begins with a complete closure of the mouth, so that no air can escape, and during that time your lips are firmly pressed together for /p/ but no sound is made. If you think of a word like 'upper' /ʌpə/, it is clear that there must be a short silence in the middle of the word if a listener is to hear a /p/ sound there. When the closure is released, things start happening very rapidly as far as acoustic events are concerned. The release of the air causes a small explosion, and this sound is aperiodic—it is like a very brief fricative, and *transient*—it changes very rapidly. If the plosive is an aspirated one, as English /p/, /t/, and /k/ often are, this 'release burst' is followed by a different sound—the sound of air rushing through the vocal tract. This is called **aspiration** (which was introduced in Chapter 3), and is aperiodic (it is closely related to the [h] consonant).

Voiced plosives are periodic during the time that the vocal tract is closed; at this stage, instead of the silence that we find in voiceless plosives, we can hear (but only just) the vibration of the vocal folds coming from the larynx. Although we class English /b/, /d/, and /g/ as voiced plosives, they actually have very little voicing, so to hear a good example of a truly voiced plosive you should listen to some other languages: French, Spanish, and Italian are good examples.

4 Nasals. Nasal sounds such as English /m/, and /n/ are periodic. They are similar to vowels, but they have much less energy at higher frequencies and it is very difficult to identify formants in their spectrum. This is mainly because most of the sound generated in the larynx by the vibration of the vocal folds cannot escape out through the mouth as it does in the case of vowels, but has to pass through the nasal cavity and out through the nostrils. If you put your fingers in your ears and produce a sequence of vowel and nasal sounds like /mɑːmɑːmɑːmɑː/ you will be able to hear the low-frequency humming sound created by the nasal resonance during the time that you are producing /m/.

5 **Affricates.** These are acoustically complex sounds. Voiceless affricates such as [tʃ] begin as plosives, so their initial portion is silence. After this, the closure in the vocal tract is released and we hear a fricative sound, which is aperiodic. Voiced affricates (if they are *really* voiced) are accompanied by vocal fold vibration; as a result, the first part is periodic and the second part is a mixture of periodic and aperiodic.

6 **Approximants.** As explained in Chapter 3, approximants are quite similar in their articulation to vowels. Not surprisingly, then, they are also *acoustically* similar to vowels. Sounds like [l], [r], [w], and [j] are periodic, and have recognizable formants (with the possible exception of [l], where the formants are sometimes hard to identify).

We have now looked at almost all the major classes of speech sound, and seen how each can be related to the major types of acoustic pattern. Only one group of sounds, the trills, flaps, and taps, remains to be examined. **Taps** and **flaps** (such as [ɾ] and [ɽ]) are usually voiced, and are therefore to be seen as being very brief voiced plosives. **Trills**, too, such as the tongue-tip trill [r] and the uvular trill [ʀ], are usually voiced, and we find the strange situation that these sounds are *doubly* periodic: they

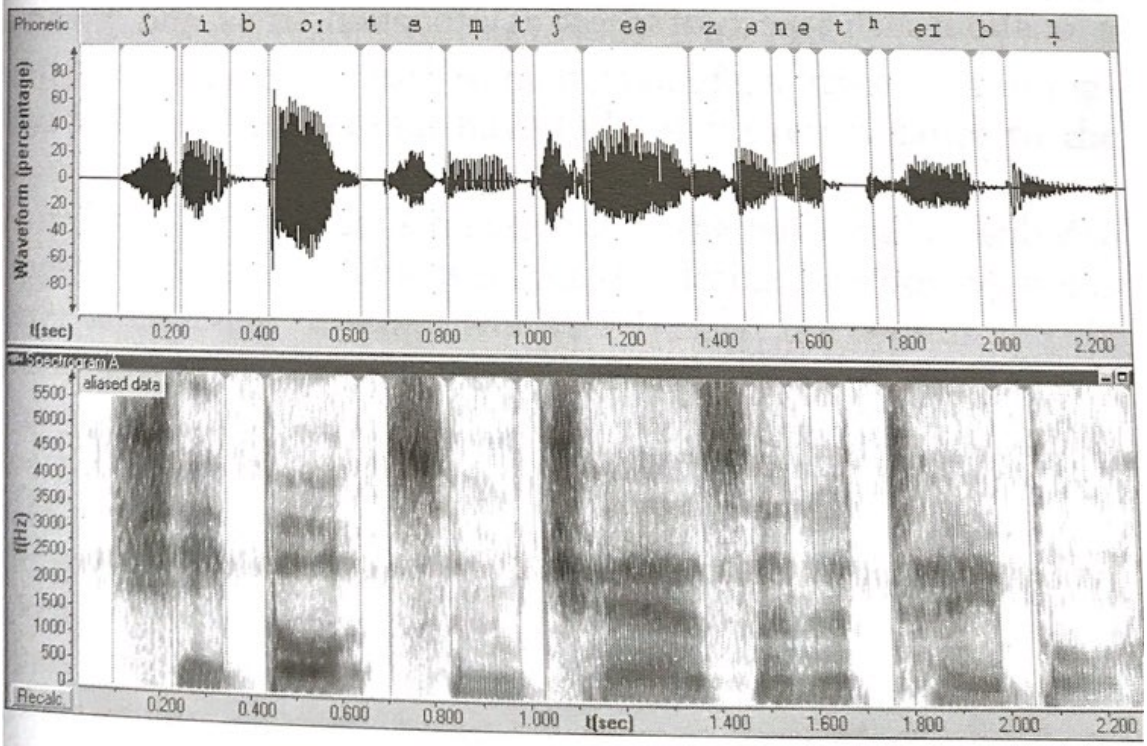

FIGURE 6.4 *Acoustic waveform and spectrogram of the sentence 'She bought some chairs and a table.'*

are periodic as a result of the vocal fold vibration, and also periodic because of the regular (though much slower) vibration of one of the articulators, such as the tongue-tip.

Figure 6.4 presents an acoustic waveform and a spectrogram of an English phrase ('She bought some chairs and a table') which contains examples of many of the above sounds.

Acoustics of suprasegmental features

Another aspect of acoustic phonetics is the analysis of the **suprasegmental features** of speech which were introduced in Chapter 5. When we hear the tones of a tone language, or the intonation of an utterance, we experience the sensation of **pitch**. This only happens in the case of voiced sounds. The sensation of pitch is related to the frequency of vocal fold vibration, which we call **fundamental frequency**. This means that we have one word (pitch) for a subjective sensation and another (fundamental frequency, or FO) for something that we can measure objectively. In a similar way, we can perceive the loudness of a sound or syllable, and we can also use instruments to measure its **intensity**; we perceive the length of a sound, and can measure its **duration**. Using computers to measure fundamental frequency and duration, we can discover a lot about such aspects of speech as intonation, stress, and rhythm.