7 The Acoustics of Speech Production

Introduction

This chapter provides a thorough account of the acoustics of speech. The first eight sections are a basic introduction to the nature of sound and sound waves, laying a foundation for the understanding of speech as sound:

- the nature of sound (7.1)
- the propagation of sound (7.2)
- simple harmonic motion (7.3)
- complex vibrations (7.4)
- resonance (7.5)
- amplitude (7.6)
- duration in sound waves (7.7)
- frequency components in sound waves (7.8).

The chapter then addresses the relevance of these basic acoustic insights to the analysis of speech:

- perceptual properties of sound waves (7.9)
- acoustic modelling of speech production (7.10)
- phonation considered as a source of sound (7.11)
- frication considered as a source of sound (7.12)
- the vocal tract considered as a filter in vowel production, and the significance of formants (7.13).

The next sections explain how and why spectrographic analysis has played a major role in modern phonetics:

- spectrographic analysis (7.14)
- acoustic properties of vowel quality (7.15)
- the vocal tract as a filter in consonant production (7.16)
- the acoustic properties of consonants in syllables (7.17).

The chapter ends with comments on the relationship between articulation and acoustics (7.18) and the acoustic analysis of prosody, with particular attention to fundamental frequency as a measure of pitch (7.19).

7.1 The nature of sound

As sensory beings we see, touch, taste, smell and hear. What we hear, we call sound. More technically, the scientific study of sound and how we hear it is ACOUSTICS. In this chapter we examine some of the acoustic properties of speech sounds, to complement the physiological, articulatory and phonemic accounts of previous chapters.

All sound results from vibration of one kind or another. In turn, vibration depends on some source of energy to generate it. Fry (1979) takes the example of a symphony orchestra: the players perform such actions as moving their arms or blowing, and their work (under skilful control) generates various kinds of vibration which we hear as sound.

Vibration alone is not enough to produce audible sound, and three accompanying criteria must be satisfied as well. In the first place, there must be a PROPAGATING MEDIUM, something the sound can travel through. Most commonly this medium is air, but any other physical substance, including wood, metal, liquid, or living tissue such as bone, can, with varying degrees of efficiency, serve as the propagating medium. If there is no medium – in a vacuum, that is – no sound can be heard. A classic experiment is used to demonstrate this. If an electric bell or buzzer is placed inside a bell-jar and the air is pumped out of the jar, the sound of the bell fades away. (Some faint sound usually remains because the mounting of the electric bell in the jar still provides a connection to the outside air.)

The two other criteria that must be satisfied concern properties of sound relative to the sensitivity of the ear. Much more will be said about these properties later in this chapter, but we must note here that vibrations vary in their rate or FREQUENCY from very rapid to very slow. The ear detects only a certain range of these frequencies, commonly down to about 20 vibrations per second and up to about 20,000 vibrations per second, although this varies among individuals and is certainly affected by ageing. Thus the second criterion is that a sound must be within the normal audible frequency range.

Thirdly, a vibration has not only a frequency, but also an AMPLITUDE – a measure of the size of vibration or the extent of movement in the vibration. Amplitude relates to what we normally call loudness, and as the amplitude of a vibration diminishes, it becomes less audible. Thus the third criterion is that a vibration must have an amplitude great enough to be detectable. This is not just a matter of the level of vibration at the sound source itself, for audibility also falls rapidly as the distance between the sound source and the listener increases. In addition, the general level of sound in the surroundings can have a masking effect, and the connection or coupling between the source of vibration and the propagating medium may also be inefficient. The effect of inefficient coupling is easily demonstrated with a tuning fork. If the fork is struck and held between finger and thumb in the air, it is scarcely audible because its vibrating prongs are not coupled efficiently to the air. If the same fork is placed on a wooden table top, the sound can be clearly heard at greater

volume because the fork's vibrations are transmitted to the air far more efficiently by means of the larger surface area of the table.

Sounds are not all perceived as identical in quality. For a broad categorization, we can make two basic distinctions. The first of these distinguishes between continuous and impulse-like sounds. A jet plane and an electric power drill are examples of essentially continuous sounds, whereas a door slamming shut or a gunshot are examples of impulse-like sounds. Continuous sounds involve vibrations which last for some time, from seconds to hours. In impulse-like sounds, the vibrations start very suddenly and build up to their maximum amplitude very rapidly (usually in a fraction of a second). The vibrations die away relatively quickly, but mostly not as rapidly as they build up.

The second distinction separates what are often called musical sounds from noise-like sounds. Almost all musical instruments produce PERIODIC sounds, so called because their vibration follows a certain pattern which is repeated regularly. (We shall see below what kinds of patterns vibrations may show.) The number of times the vibration pattern is repeated per second will determine whether we perceive them as high- or low-pitched sounds. Thus a tuba, a foghorn and a bass guitar all generate low-pitched sounds, with a small number of repeated vibration patterns per second, while a violin, a kettle whistle and a piccolo all generate high-pitched sounds with a large number of repeated vibration patterns per second. By contrast, noise-like sounds are APERIODIC and result from vibrations which are much more random and do not repeat their pattern regularly. The hiss of a steam pipe and the steady roar of a large waterfall are good examples of sound sources which have continuous yet quite random patterns of vibration.

Although these distinctions are useful, they are not quite as tidy as they may seem. A single explosion, for example, will have the character of an impulse-like sound, but if a series of explosions is rapid and sustained (as in a fast-running internal combustion engine) the effect may be that of a continuous sound. In fact many sounds have a quite complex nature. For instance, when compressed air or steam is suddenly released from a valve (as in an espresso coffee machine) the initial impulse sound decays into a continuous sound. Furthermore, there are both musical and noise-like elements in the sound, because some of the vibration which produces it is repeated regularly and some is quite random in pattern.

A further important way in which sounds differ is in their quality or TIMBRE. Consider a violin and a flute each playing the same note. Both instruments produce a continuous periodic set of sound vibrations repeated at the same rate – if this were not so, they would not be heard as playing the same note. Yet there is a distinct difference in the quality of their sound, which enables us to hear that different instruments are being played. This difference, commonly described as a difference in timbre, can be judged in impressionistic terms – the violin is perhaps 'sharp edged' whereas the flute is 'rounded and smooth' – but, as we shall see below, it is possible to analyse and explain the difference in terms of patterns of vibration.

These comments on the nature of sound are relevant to speech, which is a very complex form of sound. Speech includes impulse sounds (as in stops, such as [t] or [k]) and continuous sequences (as in vowels, such as [i] or [a]). It has periodic components (again in vowels), aperiodic components (as in fricatives, such as [f] or [s]), and mixes of both (in voiced fricatives, such as [v] or [z]). Differences among the vowels (for instance [i] versus [a] versus [u]) are heard in much the same way that we discern a violin from a flute. It is therefore important to understand the nature of sound itself if we are to have some grasp of the acoustics of speech.

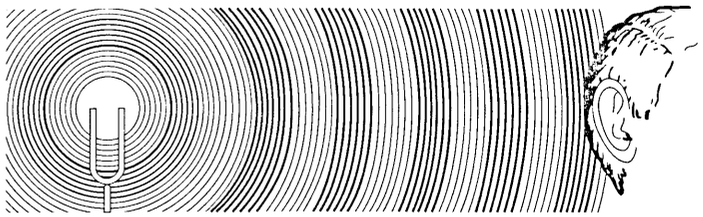
7.2 The propagation of sound

When sound travels from its source to a hearer, vibration is transmitted or propagated, through some medium. This transmission is of course invisible, but we do see something comparable (though by no means identical) when vibration is propagated through water. If a small stone is thrown into the centre of a pool of water, it will start the water vibrating by temporarily displacing water at this point; the vibration is then propagated outwards as a series of ripples of displacement, moving in ever increasing concentric circles until they reach the edge of the pool. Because a single stone thrown into a pool cannot sustain vibration, these ripples will die away, but if a stick or paddle is used to produce repeated displacement of the water at the centre of the pool, the vibration will keep on spreading outwards from this point.

While rippling water offers a simple and easily observed illustration, it is important to realize that the propagation of sound is rather more complicated. Unlike water, air is an elastic medium. Hence, when a sound is produced, the air immediately around the source is compressed. Being elastic, the air will tend to expand again after being compressed, and as it does so, it compresses the air next to it, which will in turn expand again and propagate the compression outwards. Thus when water ripples, the displacement is at right angles to the direction of the wave – the water ripples upwards and downwards from its normal surface plane. But a sound wave travelling through air varies the local air pressure in the same plane as the direction of the wave (figure 7.2.1). In both cases the wave motion amounts to a succession of local displacements. In the case of water, however, the wave is transverse (the plane of displacement is at right angles to the plane of propagation) while in air, the wave is generally longitudinal (displacement and propagation are in the same plane). The velocity of propagation of sound in air at normal temperature and pressure is around 345 metres per second.

7.3 Simple harmonic motion

Sound consists of mechanical vibrations transmitted to the ear through a physical medium, usually air. The simplest form of mechanical vibration is found



(a)

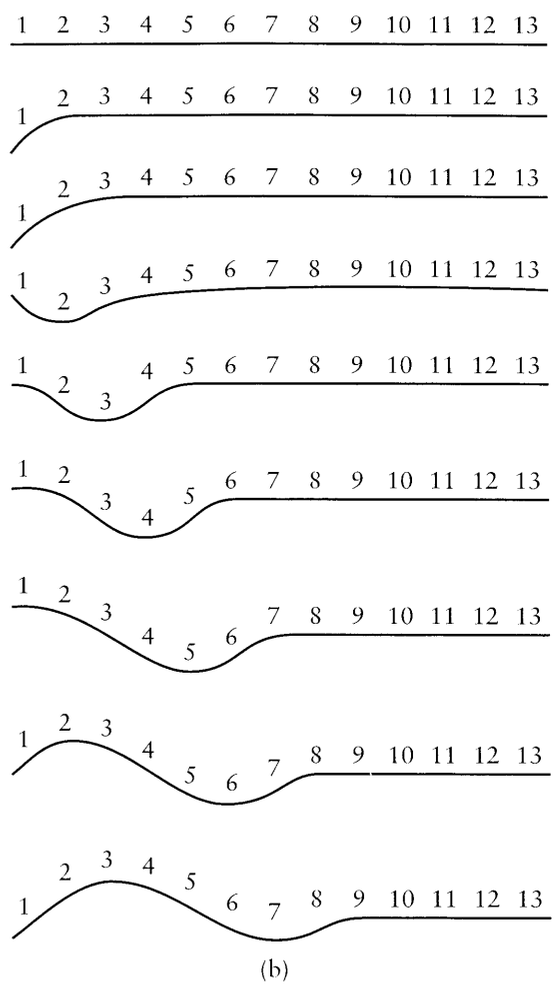
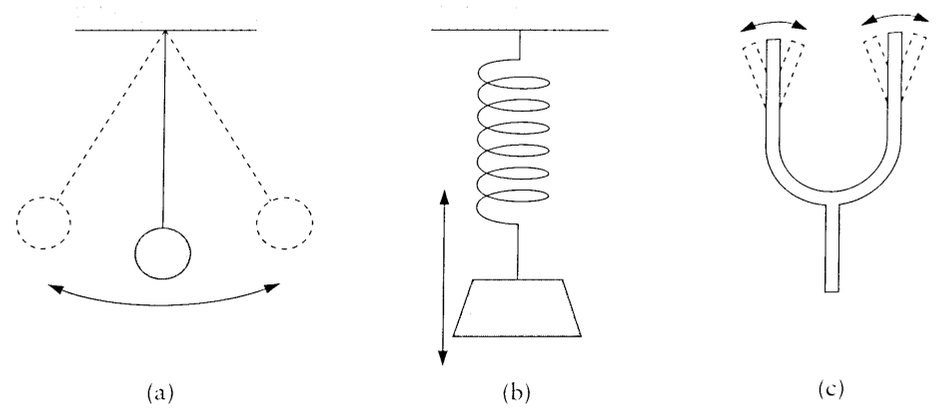


Figure 7.2.1 Propagation of a sound wave: (a) longitudinal; (b) transverse

in systems of the kind shown in figure 7.3.1. When the pendulum (a) is set in motion, or the spring-mass (b) is pulled into oscillation, or the tuning fork (c) is struck, each system will vibrate in a similar fashion. We can show the nature of the vibration by plotting the displacement of the vibrating object from its



(a) (b) (c)

Figure 7.3.1 Simple vibrating systems: (a) pendulum; (b) spring-mass; (c) tuning fork

rest position, measured in relation to time. In the case of the pendulum, this means measuring the distance that the pendulum weight moves from right to left relative to its rest position. For the spring-mass, it is the distance the mass moves up and down. For the tuning fork, it is the movement of the fork prongs either side of their rest position, in this case a very slight movement requiring extremely delicate measurement. When plotted against time, the vibration (displacement) of each of these systems will have the pattern shown in figure 7.3.2. Vibration represented as a graph of this kind is known as a WAVEFORM.

Vibration with a pattern like the one shown is the simplest kind found and is known as SINUSOIDAL vibration, or simple harmonic motion. The term 'sinusoidal wave' is generally abbreviated as SINE WAVE. If idealized, these simple mechanical systems would keep vibrating indefinitely once set in motion. In practice, energy is lost because of factors such as friction and air resistance. As a result, the amplitude of displacement in the vibrations will decrease over time. The vibration is therefore said to be 'damped'. Damped vibration is normal (unless the energy is replenished) and common in speech.

The waveform of figure 7.3.2 is characteristic of undamped vibration. Though idealized, this simple harmonic motion can usefully be taken to be the basic building block of most other more complex forms of vibration.

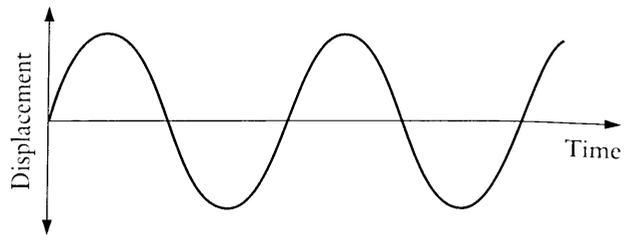


Figure 7.3.2 Simple vibration waveform

For purposes of measurement and calculation, it is helpful to rethink the graph. Let us first be clear about the wave pattern of the graph itself. If we trace the waveform of figure 7.3.2 from the left, the line of displacement curves up and over, returning to the rest point, then curving down to register the displacement in the opposite direction before returning to the axis. At this point, when we meet the axis for the second time, the wave has completed one CYCLE; the rest of this particular idealized wave is simply a repetition of the same pattern for an indefinite number of cycles. Now in acoustic phonetics several values are of significance:

- *A* the maximum amplitude of vibration: the distance between the axis and the highest (or lowest) point on the wave;
- *D* the instantaneous amplitude of vibration at some point of time: the distance between the axis and some selected point on the wave;
- *T* the period of vibration: the time taken by one complete cycle;
- *f* the frequency of vibration: the number of cycles per second, usually expressed as Hertz (Hz); hence 5 Hz is five cycles per second, 10 kHz is 10,000 cycles per second, and so on.

For many calculations involving these values, it is convenient to think in terms of the rotation of a wheel rather than the wave motion of figure 7.3.2. Any point *X* on the wave is now a point on the rim of a wheel rotating at uniform speed, and *D* is still the distance of *X* from a base line drawn horizontally through the wheel. But by thinking in terms of rotation, we can now also express the position of *X* as an angle (theta, θ) relative to the base line. Figure 7.3.3 provides a diagram of rotation alongside a wave. With this in mind we can note the following methods of deriving values.

The value of *D* (expressed as a fraction of *A*) at any instant of time *t* is given by:

(7.3.1) $D = A \sin 2\pi t/T.$

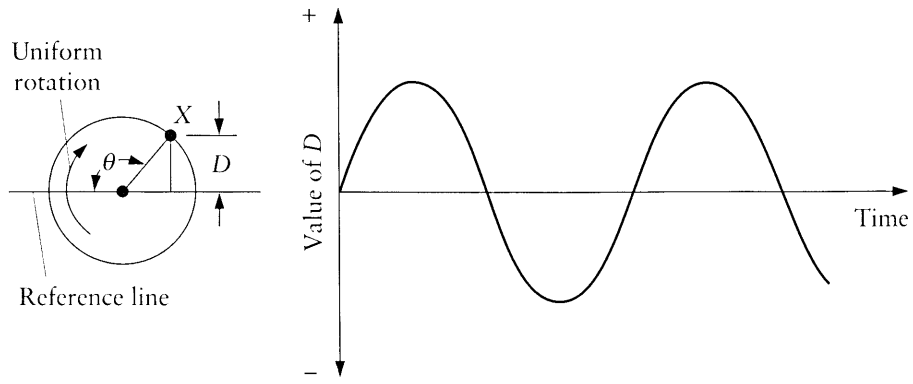


Figure 7.3.3 Simple harmonic motion

The common form of this equation is given in physics texts as:

(7.3.2) $D = A \sin \omega t$

where

(7.3.3) $\omega = 2\pi f.$

Since *f* is the frequency of vibration (or, equivalently, the frequency of rotation of the wheel) it can be derived from *T*, the time it takes to complete a single rotation:

(7.3.4) $f = 1/T$ (expressed in Hz).

Making use of the angle θ (as shown in figure 7.3.3) we can also calculate displacement as ANGULAR DISPLACEMENT:

(7.3.5) $D = A \sin \theta$

To take simple examples, when point *X* is on the axis, $\theta = 0$, $\sin \theta = 0$ and $D = 0$. When point *X* is farthest from the axis, $\theta = 90°$, $\sin \theta = 1$, and $D = A$.

Most waveforms – including those studied in acoustic phonetics – are not sinusoidal but can be analysed as the sum of two or more sine waves. To approach this analysis, we need to understand the time relationship between two or more waveforms, known as their PHASE relationship. Consider the two sinusoidal waveforms in figure 7.3.4. Each has the same frequency (or period), and each has the same maximum amplitude, but they are displaced from each other such that they pass through their maximum and minimum values at different points of time. The phase relationship between two or more waveforms is always relative: we have to take one of the waveforms as the point of reference for measurement. The phase relationship can be expressed as the time displacement between two waveforms, as in figure 7.3.4, but this has the disadvantage of giving an absolute measurement, a value that will vary according to the frequency of the vibration involved. What is more significant is

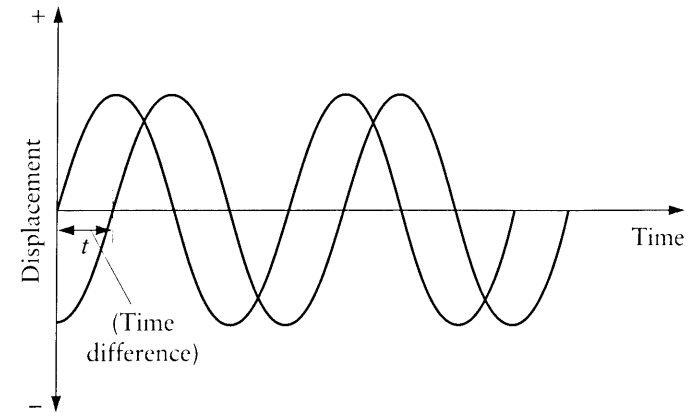


Figure 7.3.4 Time displacement between two waveforms

the relative relationship between the cycles of vibration of the two (or more) waveforms. This can then be used in defining the properties of complex vibrations made up of several sine wave components. This relative relationship is based on angular displacement, instead of time. Figure 7.3.5 shows the PHASE ANGLE between two sine waves, in this case 90°. The phase angle between two waveforms is normally expressed as a value between 0° and 360° relative to the waveform used as the reference. This range of values represents one full vibration cycle of the reference wave.

Consider now two sine waves which have the same frequency and amplitude and which are also perfectly in phase (their phase difference is 0°). Add the two waves together, and the amplitude of the combined wave will be double that of the two constituent waves at every point in the cycle, as shown in figure 7.3.6. The addition shows how one wave can be composed of two waves. But

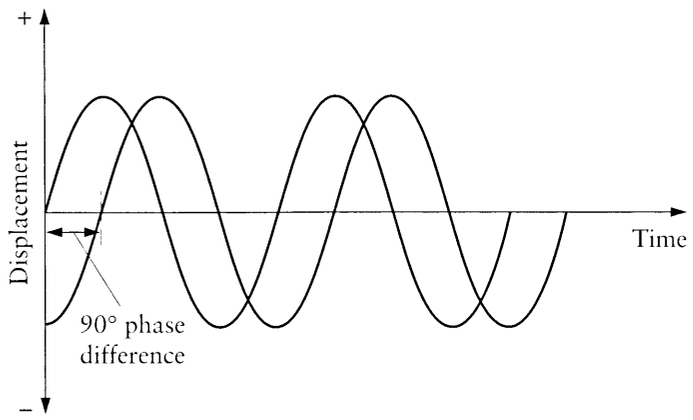


Figure 7.3.5 Phase relationship between two waveforms

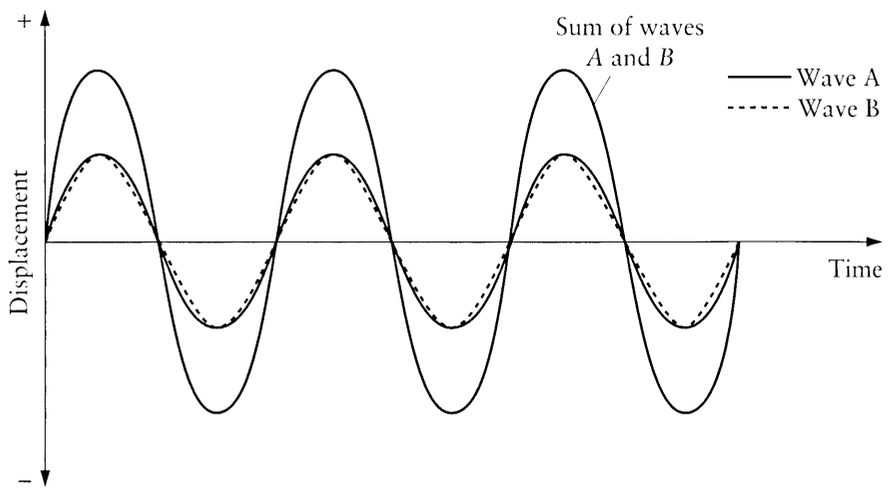


Figure 7.3.6 Effect of combining two waves in phase

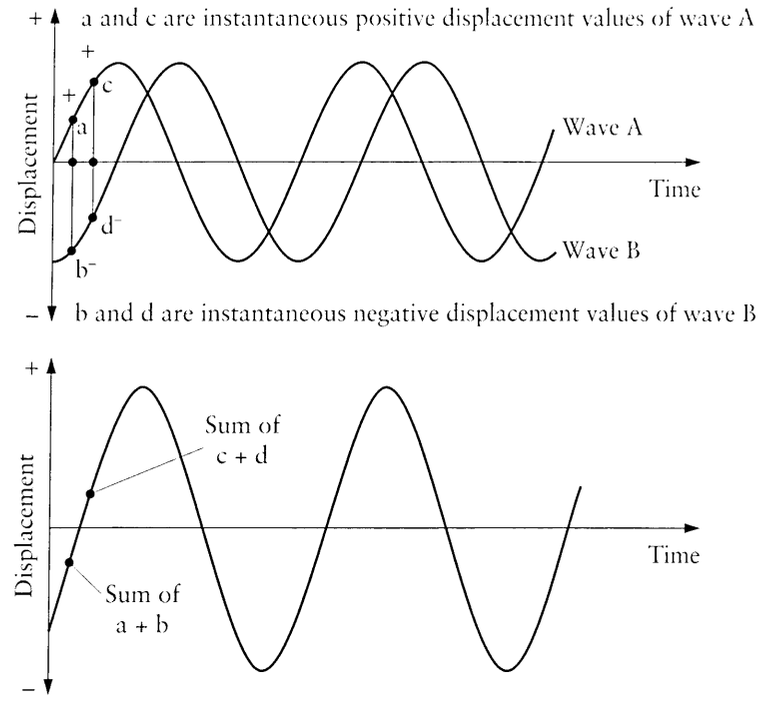


Figure 7.3.7 Effect of combining two waves displaced by 90°

when these same two sine waves are out of phase, say by 90°, the situation is not as simple. In some parts of the vibration cycle the values of the waves have to be added to each other, in others subtracted from each other. Figure 7.3.7 shows the result. In this case, the resultant wave has greater amplitude and is shifted in phase relative to the two component waves. (In practice, calculations are done by trigonometrical methods based on the phase angle between the waveforms, rather than by the time-consuming addition implied by figure 7.3.7.)

Readers wanting a more detailed account of simple harmonic motion may find Small (1973) helpful, and, for a more rigorous mathematical approach, should consult any standard work on acoustics, such as Wood (1964 or 1966).

7.4 Complex vibrations

Sine waves are the building blocks of all forms of vibration, and figure 7.4.1 shows how a complex vibration may consist of the combined effect of three simple sinusoidal vibrations of 100 Hz, 200 Hz and 300 Hz. The result is a complex wave which is not sinusoidal, and which will have a timbre different from that of any simple sine wave. Its frequency of vibration is defined as that of the lowest frequency of the sine waves which compose it. This frequency

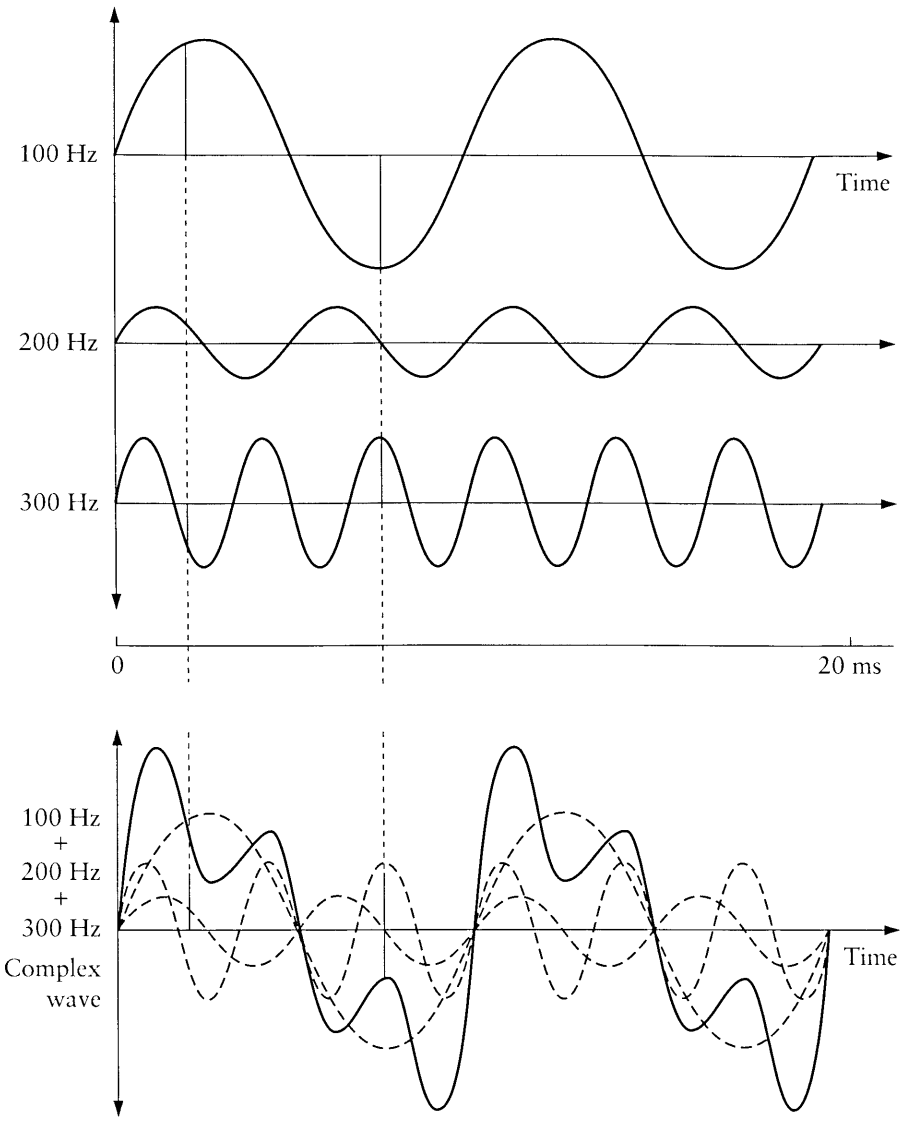


Figure 7.4.1 Complex wave with three sinusoidal components (100 Hz, 200 Hz, 300 Hz)
Adapted from: Ladefoged 1962, p. 35.

(100 Hz in the case of figure 7.4.1) is known as the FUNDAMENTAL FREQUENCY (often just as the FUNDAMENTAL). The three sine waves are the COMPONENTS of the complex wave.

If the same three waves are combined with different phase relationships among them (figure 7.4.2), the resultant complex vibration can be seen to have a different shape. Although the waveshape is different, the fundamental frequency is the same, and, perhaps surprisingly, the timbre or quality of the sound will strike a listener as almost the same, if not identical. Thus waveshape alone does not reflect the quality of perceived sound, because the human ear is not

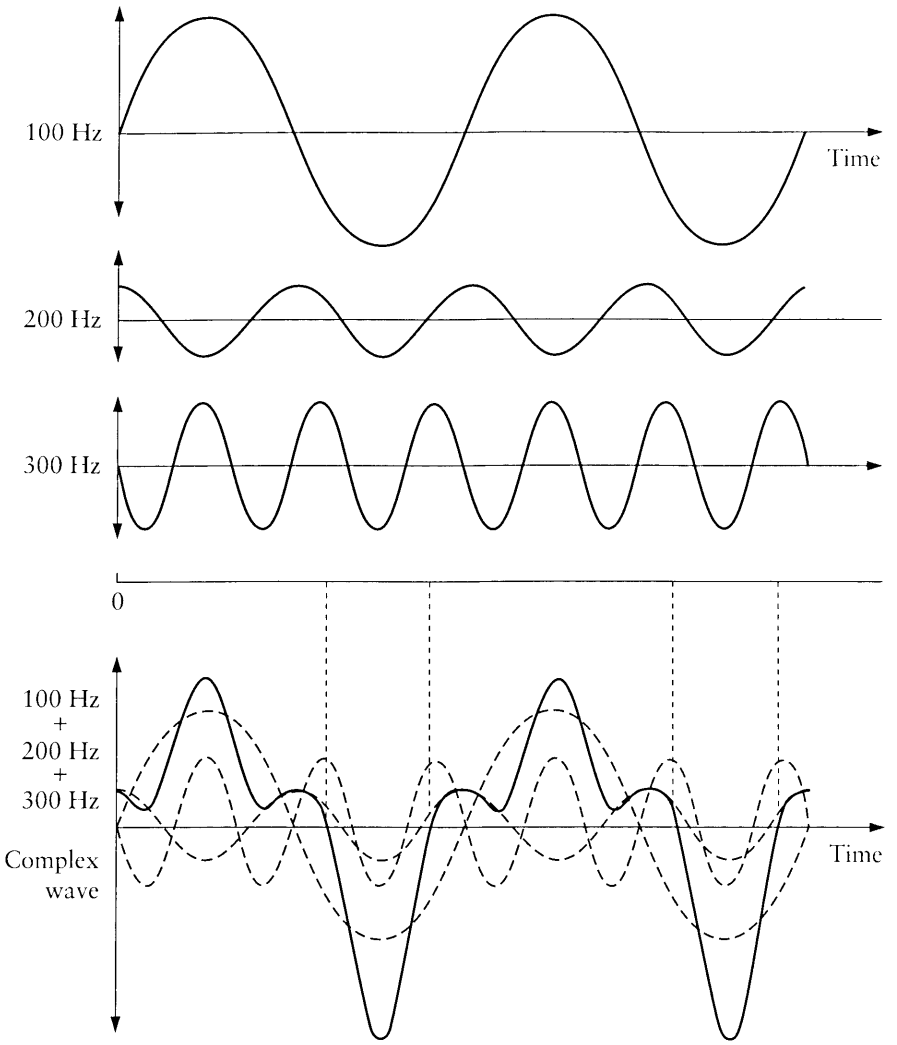


Figure 7.4.2 Complex wave illustrating the effects of phase on waveform pattern
Adapted from: Ladefoged 1962, p. 40.

particularly sensitive to phase. It was once thought that the ear was completely deaf to phase; this is not strictly correct, but it is true that we will hear an appreciable difference in sound quality only if there are changes in the frequency and amplitude of the component waves.

So far we have assumed that the component sine waves are at frequencies which are integral multiples of, and have a fixed phase relationship to, a fundamental. As we have seen, if these component sine waves are shifted statically in their phase relationships, the shape of the complex wave will vary, but the sound quality will generally not change appreciably. If, however, the component sine waves are not integral multiples of the fundamental, their phase relationships will keep changing, as will the resultant complex waveshape. More

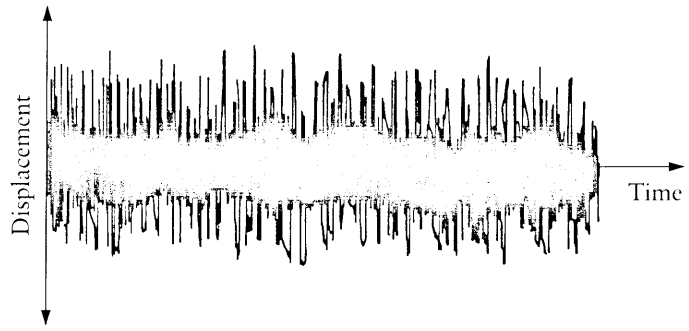


Figure 7.4.3 Noise waveshape

importantly, the timbre or sound quality will be dissonant and unmusical, even though the hearer can perceive a note or pitch. The larger bells in a carillon are a reasonably good example of the sound of this sort of waveform. This sort of complex vibration reaches its limit when all of the components have randomly varying frequencies and randomly varying amplitudes. The result is a complex waveshape that is constantly and rapidly changing, with no general pattern. The sound produced will no longer be periodic in nature, but noise-like. The more truly random the variations in frequency and – to a lesser extent – in amplitude, the more truly noise-like the sound will be. A typical example of a noise waveshape is shown in figure 7.4.3.

The most completely noise-like sound is one in which all possible frequencies in the range of hearing are randomly present, at random amplitudes and in random phase relationships. This is known as WHITE NOISE.

7.5 Resonance

If we were to take the pendulum of figure 7.3.1(a) above and give it a single push, it would swing for a time and the displacement on each swing would get smaller until the pendulum came back to rest. The graph of displacement against time would be as shown in figure 7.5.1. The period of each complete vibration is easily measured, and would be found to be approximately the same, for the pendulum moves more slowly as the distance it travels decreases. In fact the pendulum has a natural frequency at which it vibrates, known as its RESONANT FREQUENCY. If the string on the pendulum were lengthened, the period of vibration would be longer and the resonant frequency lower. Thus the resonant frequency is a function of the length of the pendulum string. All resonant mechanical systems behave much in the same way, and the waveshapes they produce are known as damped vibrations. For reasons that will become apparent later, these damped vibrations are in fact complex vibrations.

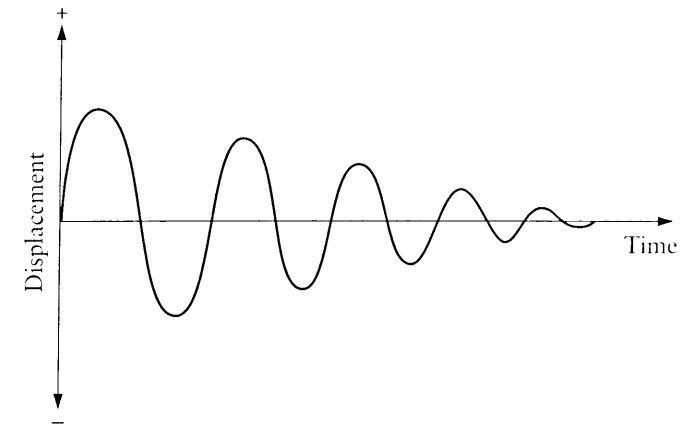


Figure 7.5.1 Pendulum vibration pattern

An important property of mechanical systems is that they respond selectively to vibrations of different frequencies. A simple illustration of this is the spring-mass system of figure 7.3.1(b) above. Imagine that we wish to transmit periodic vibrations through this system, by vibrating the anchoring point. There will be input vibration at the top of the spring and output vibration at the bottom of the mass (figure 7.5.2). If the input vibration is at a frequency very much higher or lower than that of the natural resonance frequency of the system, the input vibrations will be transmitted to the output with very weak displacement amplitude. Assume now that the input vibrations start much lower than the resonant frequency of the system but are gradually increased. Then, as the input frequency approaches the natural resonance frequency of the system, the output vibration amplitude will steadily increase and reach a maximum when the input frequency is equal to the resonance frequency. At this point, the output vibration amplitude may actually exceed the input vibration amplitude.

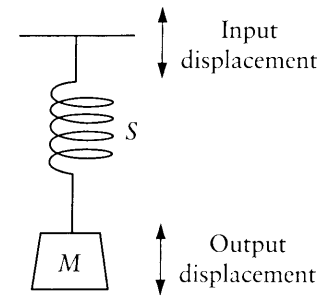


Figure 7.5.2 Transmission of vibration through a spring-mass system

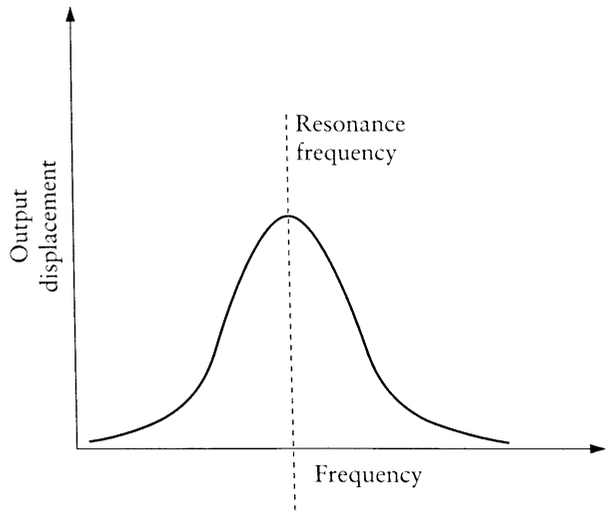


Figure 7.5.3 Resonance curve

If we keep the amplitude of the input vibrations constant while varying their frequency from well below to well above the resonant frequency of the spring-mass system, we can plot the response of the system. Figure 7.5.3 represents this response as a graph of output vibration amplitude against frequency. The display is known as a RESONANCE CURVE. The resonance curve illustrates the extremely important principle that a resonant system transmits the energy of input vibration with selective efficiency, reaching its peak at the resonant frequency of the system. Resonance and its selectivity are important characteristics of the vocal tract.

The degree of selectivity exhibited by a resonant system is determined by its degree of damping. Recall that when a pendulum or spring-mass is given a single impulse of input energy, it will vibrate at its natural (resonant) frequency, with the amplitude of vibration gradually dying away. The duration of this decay in amplitude, relative to the period of the resonance, reflects the effect of losses in the resonant system, and hence its degree of damping. Figure 7.5.4 shows the resonance curves, or frequency responses, of (a) lightly and (b) heavily damped systems.

It is not always convenient to define the selectivity of a resonant system in terms of its damping; a common alternative is to express it in terms of its BANDWIDTH. This is defined as the range of frequencies either side of the centre frequency of the system's resonance curve which have an amplitude of 70.7 per cent or greater of the resonant frequency amplitude (figure 7.5.5).

The selectivity implications of a given bandwidth figure make it necessary to know the resonance frequency as well. Selectivity as an independent property may also be defined by the Q factor of the resonant system, given by:

$$Q \text{ factor} = F_{\text{resonance}}/\text{bandwidth}.$$

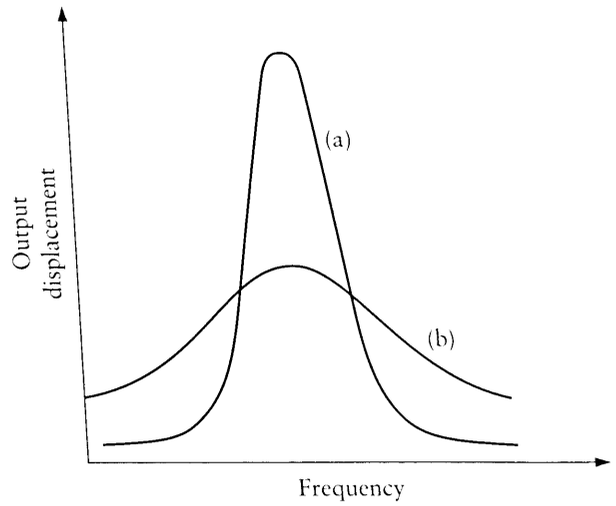


Figure 7.5.4 Resonance curves of damped vibrating systems: (a) lightly damped; (b) heavily damped

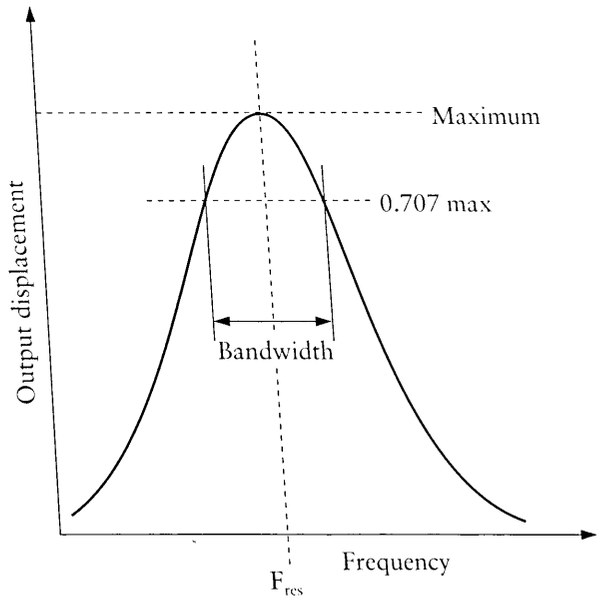


Figure 7.5.5 Method for determining bandwidth

7.6 Basic amplitude properties of sound waves

Amplitude is the term normally used to refer to the magnitude of displacement in a sound vibration. Most commonly, it is air pressure that is varied by this

displacement. Pressure is defined as force per unit area and is measured in Pascals (Pa). Static air pressure at sea level is about 100,000 Pa, but the pressure variations which result in audible sound at normal listening levels are very much smaller than this. For example, the sound pressure variations of conversational speech at a distance of about one metre from the speaker's lips will be in the region of 0.1 Pa.

When a sound is picked up by a conventional microphone, such as the electret type provided with many tape recorders, the pressure variations of the sound propagated in air are transformed into a corresponding electrical voltage. This gives us an electrical representation of sound pressure waves, which then raises the question of what is the most appropriate way to measure and portray this representation.

Consider the sinusoidal and complex waves shown in figure 7.6.1. A simple measure of amplitude is to take the maximum values of displacement in the wave. This is useful if we wish to know the peak or peak-to-peak values of a periodic waveform, or if we need to measure the peak values of impulse sounds. This is the kind of measure we need to know to avoid overload when recording sounds or processing them in some form of computer analysis. Unfortunately, however, this measure tells us little about the rest of the waveform. To account for an entire waveform, we need measurements all the way along a cycle, a series of instantaneous values, as shown in figure 7.6.2. Note that values above the axis line will be positive, those below the axis will be negative. These sample

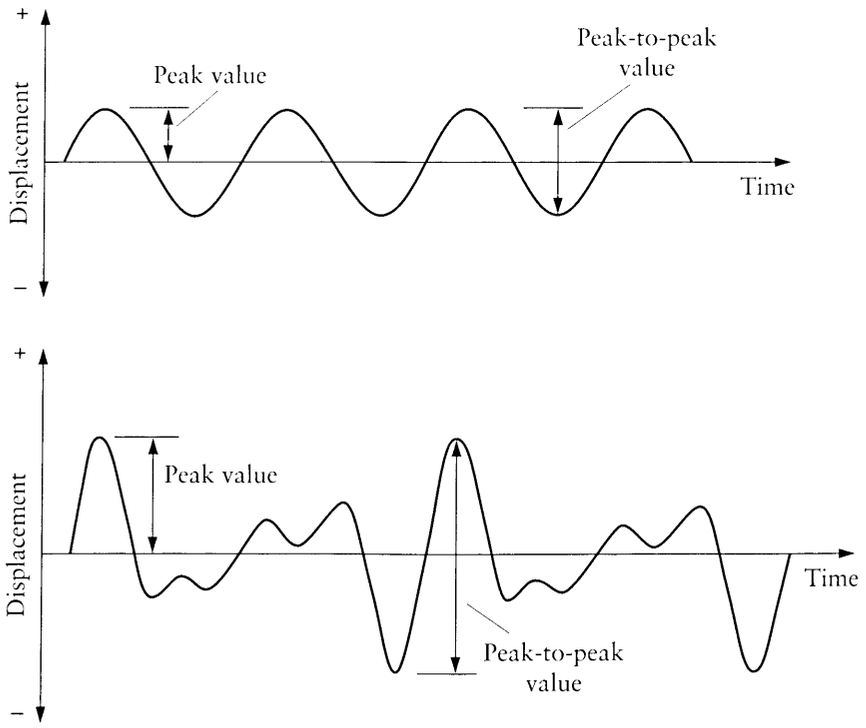


Figure 7.6.1 Waveform amplitude

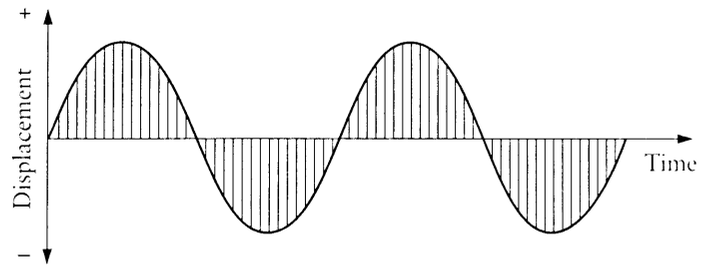


Figure 7.6.2 Instantaneous values of wave amplitude

values can give us a picture of the behaviour of the wave, and it would be useful to summarize or average them in some way. But note what happens if we simply add and average a series of values taken along a sine wave: the positive and negative values cancel each other out when added, leaving a sum of zero. We could avoid this by making all the signs of the samples positive, but it is even more useful to turn to another kind of calculation.

We can derive from amplitude a property called INTENSITY. Intensity is power per unit area, or the way power is distributed in a space. Power itself is a measure of the rate at which energy is being expended – for our purposes, in producing sound. Now it can be shown that intensity is proportional to the square of pressure. Hence, if we take our sample values (as in figure 7.6.2), square them, and then add them and find the average, we have a measure of amplitude over the cycle that relates well to effective intensity. If we then take the square root of this average, we can express pressure rather than pressure squared. This value is known as the ROOT MEAN SQUARE or RMS value. The method of calculation, using samples as shown in figure 7.6.2, is as follows.

Assume we have a set of N samples of the instantaneous values of a waveform defined as X_1, X_2, $X_3 \ldots X_n$. Let Z = the sum of the squared values of the N samples, i.e.

$$Z = (X_1)^2 + (X_2)^2 + (X_3)^2 + (X_n)^2.$$

Let A = the mean (average) of these squared sample values, i.e.

$$A = Z/N.$$

Then the RMS value for the waveform is the square root of A. Suppose, for instance that the sample values are

1, 3, 5, 3, 1, –1, –3, –5, –3, –1.

The squared values are 1, 9, 25, 9, 1, 1, 9, 25, 9, 1 and their sum Z is 90. There are 10 values (N = 10) and A is therefore equal to 90/10. The RMS value is the square root of A, namely 3.

For sine waves, the RMS value of amplitude over the period of the wave is actually 0.707 of the peak amplitude. For any other complex waveshape, the

RMS value must be calculated from samples, as above. With the benefit of any of the modern general-purpose computer-based speech waveform editing and analysis packages, the calculation is quite straightforward.

The RMS value of sound pressure is thus proportional to sound intensity. In fact, when measured under plane wave conditions (i.e. with pressure variations in one plane only) and using 20 microPascals as a reference pressure, the RMS pressure may be equated with intensity. (20 μPa is the usual threshold value of sound pressure which can be detected by a normal adult listener.) It is indeed commonly assumed that intensity and RMS pressure are equivalent, although the assumption is accurate only when a sound is picked up close to the microphone with minimum interference from reflected sound.

It is often useful to know the intensity of a sound over more than one period of vibration. For example, it may be of interest to know the intensity of a whole syllable or word or clause relative to another. The RMS intensity of an entire word may be readily calculated from a stored speech waveform of any defined length using an appropriate number of samples from the speech waveform. Most speech analysis software allows the user to set cursors or markers on the time axis of the speech waveform and to calculate the RMS intensity over the period within the markers. Figure 7.6.3 gives two examples of marked waveforms. The intensity of a speech wave can be expected to vary over time – during a syllable or longer utterance – but can be calculated on a continuous basis. Effectively, the method determines intensity over a defined window and thus deliberately provides no detail of intensity variation within the period of the window.

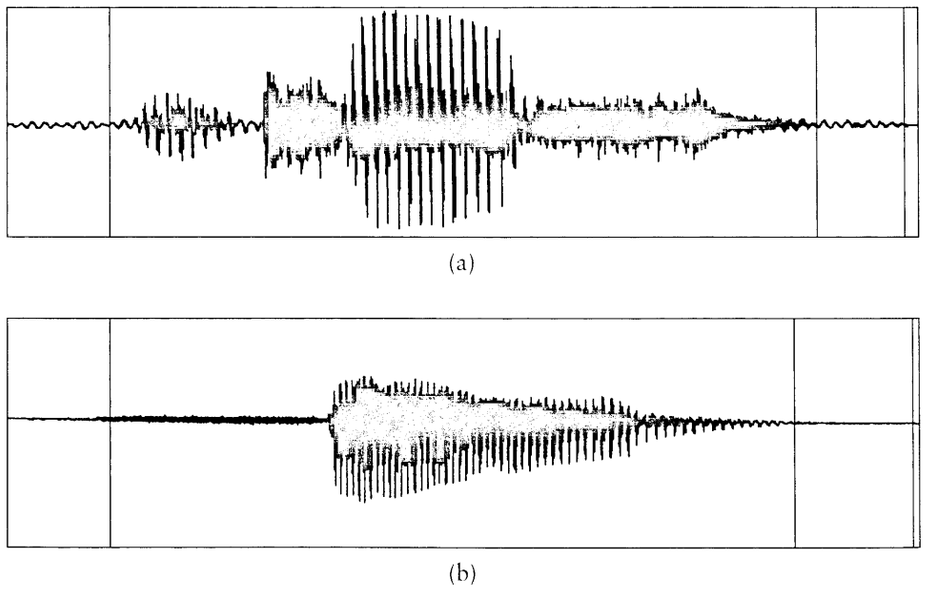


Figure 7.6.3 Waveforms marked for computation of RMS intensity: (a) *juice*; (b) *farm*. The overall intensity of waveform (a) between the markers is approximately 1.5 times that of (b)

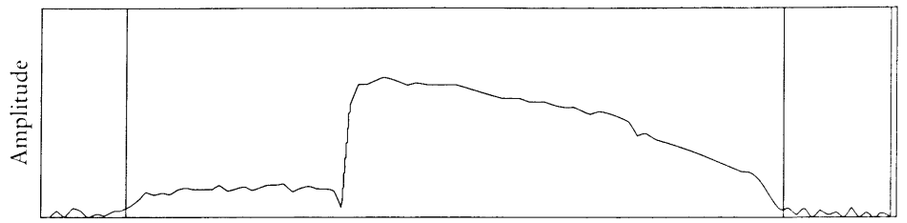


Figure 7.6.4 Intensity envelope for the waveform of figure 7.6.3(b)

Such output can be obtained from a traditional pitch and intensity meter of the stand-alone analog electronic type once found in most phonetics laboratories, although today computer analysis packages are normally used. The period over which intensity is instantaneously and continuously calculated is known as the INTEGRATION TIME of the instrument and is usually adjustable from about 5 ms to about 50 ms. Commonly, integration times of from 10 to 20 ms are used, being short enough to detect any significant fluctuations within a syllable, but long enough to avoid including any effects of the period or fundamental frequency of the speech wave itself. Figure 7.6.4 shows the intensity envelope calculated from the waveform of figure 7.6.3(b) by means of a computer speech-processing package.

7.7 Time domain properties of sound waves

When we analyse sound waves, we can consider them to have various properties, some of which are related to time and some to frequency. All the examples we have so far considered in this chapter are in fact time domain waveforms: they display changes (e.g. in the value of pressure) over time. In this section we will focus on time itself as a property.

The duration of a speech sound or an utterance is often phonetically important (sections 2.8 and 2.15 above). Durations we need to consider may be as small as a fraction of one cycle of a periodic waveform, or may be one complete period of vibration, or may be far longer. In some instances we want to know the duration of a whole word or utterance, or even the duration of a silence such as may occur in the closure phase of a voiceless stop.

To measure duration from a speech waveform we must be able to set reference markers on that waveform which have some meaningful relationship to the phonetic structure of the speech signal being measured. This normally means displaying the waveform on a computer screen (usually using a speech editing and analysis package) which allows the experimenter to replay the section of waveform between the markers. The experimenter can then place the markers accurately and confirm by ear that the phonetically appropriate section of the

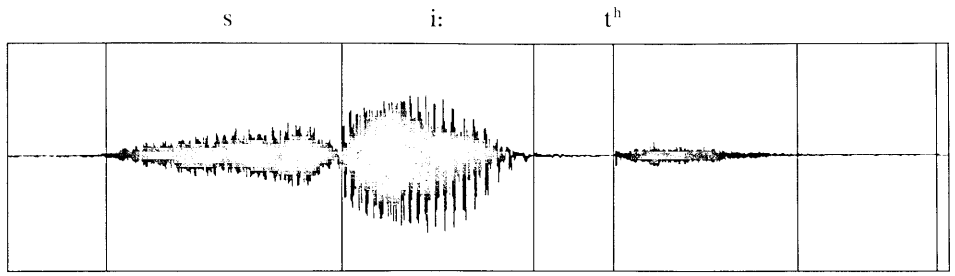


Figure 7.7.1 Segmentation of the waveform for /si:t/ (*seat*)

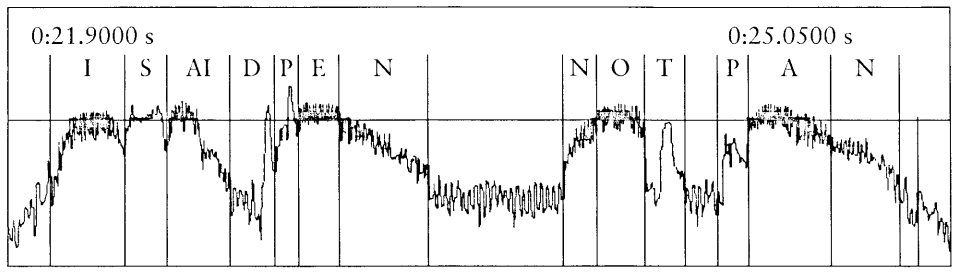


Figure 7.7.2 Segmentation of the intensity envelope of a sentence: *I said 'pen', not 'pan'*

waveform has been marked. Figure 7.7.1 shows the waveform of the word *seat* /si:t/, marked to identify and measure the duration of the vowel nucleus. If larger durations are being measured and very high accuracy is not required, it is more practical to use a time-varying intensity graph to identify the start and end of the word or utterance to be measured. The major problem with this technique lies in establishing a reliable and consistent means of determining the appropriate thresholds of intensity which mark the start and end of the speech to be measured. Figure 7.7.2 shows an example of segmentation based on intensity (applied to the sentence 'I said "pen", not "pan"').

It is possible to become quite skilled at reading time domain waveforms and relating these to the phonetic structure of which they are realizations, but much of speech cannot be segmented and labelled as easily as the straightforward example of figure 7.7.1. In fact, the most accurate method of measuring duration and of segmenting and labelling the time course of the acoustic speech signal involves the combined use of time and frequency domain information. The techniques are discussed in the following sections.

7.8 Frequency domain properties of sound waves

If we are interested in the frequency-related properties of sound waves, it is possible, at least in the simplest cases, to take a fixed section of the waveform

(extracted from the time course) and to analyse it without reference to the time domain.

As we have seen, sinusoidal vibrations are the simplest form of vibration, and they can be taken to be the components which are added together to constitute all other forms of vibration. The mathematical technique of breaking a complex wave down into its sinusoidal components is known as FOURIER ANALYSIS, after the nineteenth-century French scientist who developed its mathematical basis. The example given earlier as figure 7.4.1 illustrates a complex periodic vibration consisting of three sinusoidal components. The lowest frequency sine wave component is the FUNDAMENTAL frequency and the two higher frequency components are the second and third HARMONICS. All three form the harmonic components of the wave, the fundamental frequency being the first harmonic. Note that in periodic waves such as this one, the frequency values of the harmonics are integral multiples of the fundamental. Aperiodic vibrations may also be analysed into sinusoidal components but there will not be any simple arithmetic relationship among the components, which are then referred to not as harmonics but as OVERTONES or simply FREQUENCY COMPONENTS. Among speech sounds, vowels are characteristically periodic, while fricatives are examples of aperiodic sounds.

The frequency distribution and amplitudes of the harmonic components of a complex wave may be represented as its line SPECTRUM. For this display, the horizontal axis represents frequency and the vertical axis amplitude: each harmonic appears as a single vertical line located at the appropriate point along the horizontal axis, and the height of the harmonic line indicates its amplitude. The complex wave of figure 7.4.1 has a line spectrum as shown in figure 7.8.1. Note that this representation does not include any phase information. More examples of common waveform shapes are shown in figure 7.8.2.

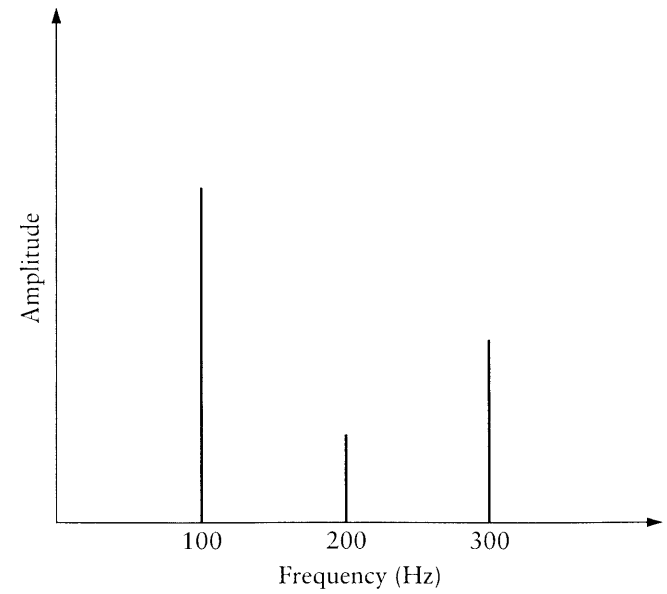


Figure 7.8.1 Line spectrum for figure 7.4.1

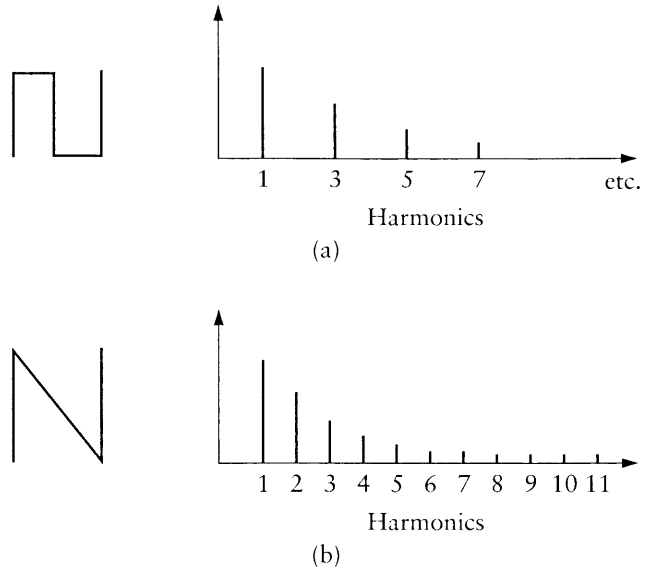


Figure 7.8.2 Line spectra of two common waveforms: (a) square wave; (b) sawtooth wave

The spectral analyses in these examples assume that the complex waves are perfectly periodic and that the same waveform is repeated indefinitely. In practice, the analysis is valid as long as the waveform remains consistent over successive cycles. Many sounds, however, including some of those in speech, do not exhibit this continuity from cycle to cycle of vibration. One commonly occurring type of sound wave is that known as QUASIPERIODIC. Imagine that a pendulum is given a push, then allowed to swing freely for several cycles before being given another push. The cycles following each push will decrease in amplitude at a rate dependent on the degree of damping in the system; and it remains for the next push to restore the amplitude of swing to its initial value. If the pushes occur at uniform intervals, the resultant wave will have an effective period set by this interval, with a damped train of sine waves occurring in between. Two examples are shown in figure 7.8.3, where (a) has only

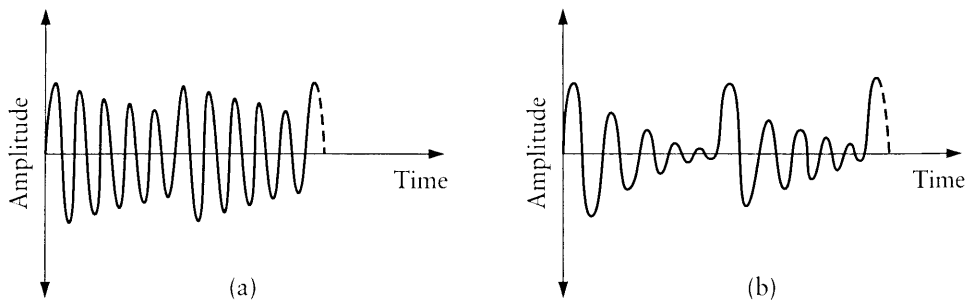


Figure 7.8.3 Quasiperiodic waveforms: (a) lightly damped; (b) heavily damped

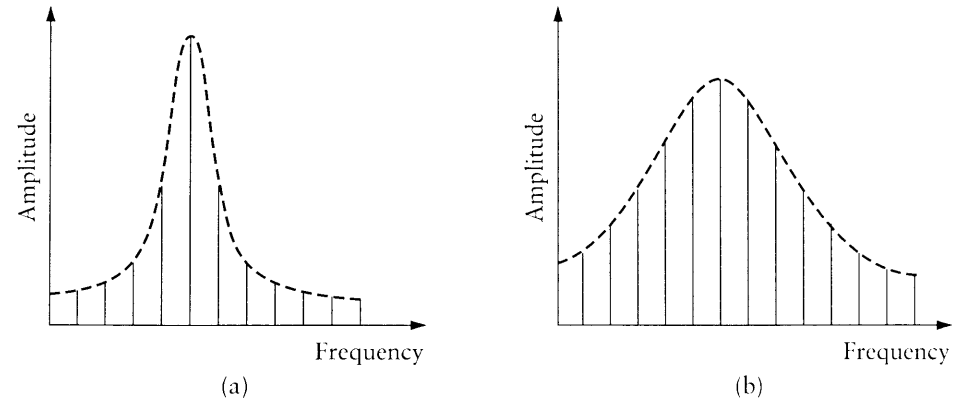


Figure 7.8.4 Line spectra: (a) figure 7.8.3(a); (b) figure 7.8.3(b)

a small amount of damping, and (b) substantial damping. The spectral properties of such waves are determined by three factors:

1 the natural resonant or oscillation frequency of the system;
2 the degree of damping (or losses) in the system;
3 the period/frequency of the external energy input.

Figure 7.8.4 gives line spectra for the waveforms of figure 7.8.3. The spacing of the harmonic lines is determined by the frequency with which the energy is restored (which is the effective fundamental frequency); the frequency of maximum amplitude of harmonic energy is set by the natural resonant frequency of the vibrating system; and the pattern of the amplitude peaks (forming a shape usually referred to as the ENVELOPE) is determined by the degree of damping in the vibrating system. In these examples, the only difference between the two spectra is in their envelope shape, since it is only the degree of damping in the resonant systems that varies.

If, on the other hand, the damping is kept constant, and only the frequency of energy restoration is changed, the resultant spectra have amplitude envelopes with the same shape and same centre frequencies; only the frequency spacing between their harmonic lines differs. Figure 7.8.5 shows examples of such waveforms and their corresponding spectra. Finally, if the damping and energy-restoring frequency remain constant, and only the oscillation frequency is altered, then the only significant change will be the centre frequency of the spectral envelope amplitude peak. Waveforms and spectra to illustrate this are shown in figure 7.8.6.

The shape of the spectral envelopes shown above can be seen to depend on the frequency and damping properties of the resonant system alone, and not on the frequency of energy restoration. A comparison of these spectral envelopes with those of the resonance curve in figure 7.5.3 will show that they correspond in shape quite directly. This is exactly as it should be: such spectra effectively display the frequency-selective properties of resonant vibrating systems.

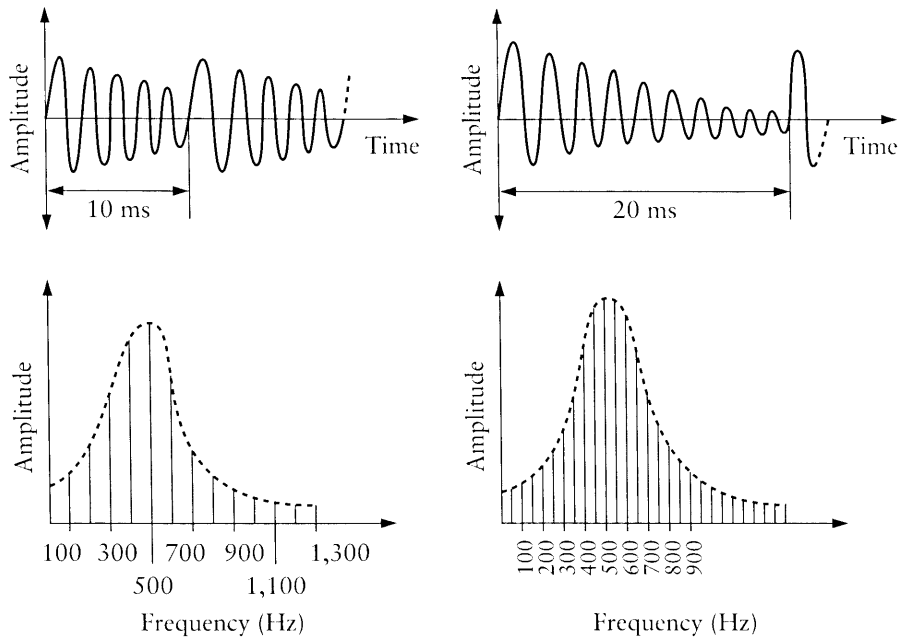


Figure 7.8.5 Waveforms with constant damping and differing energy restoration rates, with their line spectra

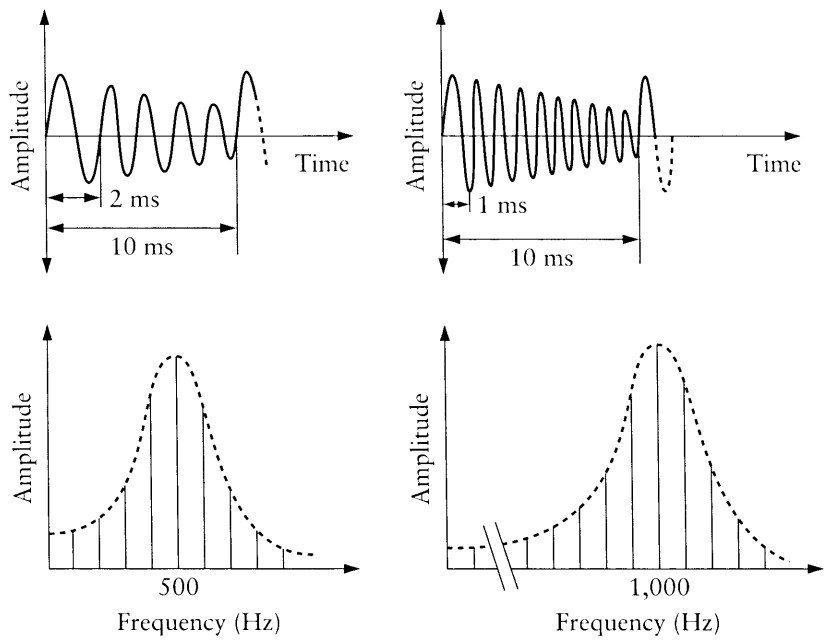


Figure 7.8.6 Waveforms with constant damping and energy restoration rates, and differing oscillation frequencies, with their line spectra (harmonics at 100 Hz intervals)

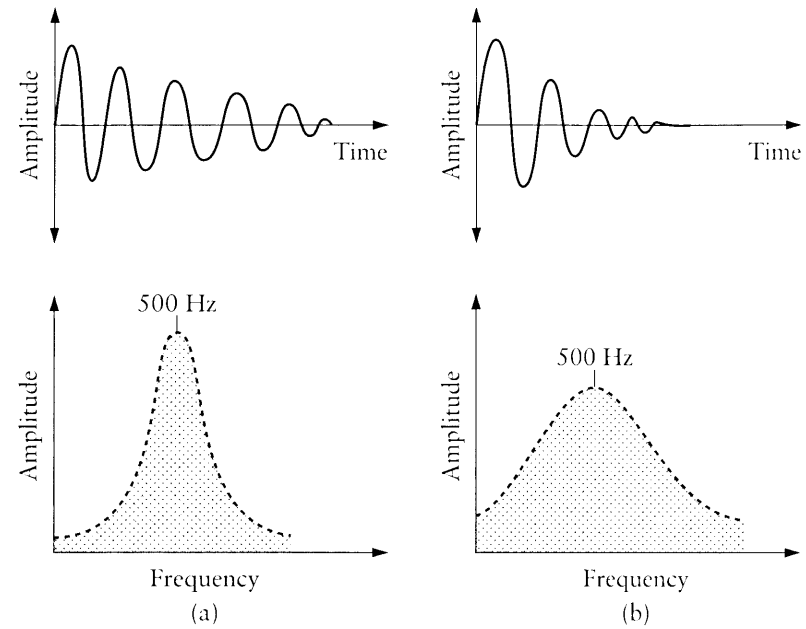


Figure 7.8.7 Aperiodic damped 500 Hz waves and their corresponding spectra: (a) low damping; (b) high damping

Aperiodic sounds have more complex spectra than any of the preceding examples. The most straightforward case is that of a single damped sinusoidal vibration. Examples with high and low damping are shown in figure 7.8.7. The spectra of these two waveforms have no harmonic lines, simply because the waves are not periodic. The vibration (sinusoidal in these examples) converges to zero amplitude over a number of cycles and there is no cyclical repetition of the waveform. Despite this, their spectra still exhibit a peak of energy in the amplitude envelope with a centre frequency equal to that of the damped sinusoidal vibration, and the sharpness of the envelope peak still depends on the degree of damping.

The reason for the evident lack of harmonic structure in these spectra is that there is, in effect, an infinity of harmonics; the notional period of such waveforms is itself infinite. As a result, there is uniform density of spectral energy throughout the frequency range of the spectrum. Sounds of this sort are, so to speak, on the way towards pure noise, in which there is no periodicity in any aspect of the waveform. In the extreme case of white noise, the spectrum has no peak of energy, but simply reveals uniform amplitude energy across the entirety of the relevant frequency spectrum. The component frequencies of the noise vary constantly and randomly in their amplitude and frequency and in their phase relationships to each other.

All the spectral types discussed in this section are found in speech. In particular, the quasiperiodic type – though somewhat more complex than the

examples given here – forms the acoustic basis of vowels, while the various aperiodic spectra are fundamental to stops and fricatives.

7.9 Some basic perceptual properties of sound waves

It is characteristic of human perception that the sensations we experience in response to stimuli rarely correspond directly with the values we derive from measurement of those stimuli. Our perception of light and dark, for instance, is not like the operation of a light meter but is highly sensitive to context: grey looks lighter against a dark background, sunshine seems even brighter when you emerge from a dark building, some colours under certain lighting are much easier to distinguish than others, and so on. Similar considerations apply to the perception of sound, studied under the heading of PSYCHOACOUSTICS. Here we are concerned particularly with the perception of loudness, pitch and what we have previously called timbre (section 7.1 above).

The human auditory system is capable of responding to an enormous range of sound intensities, and the upper end of this range is more than a million times greater than the lowest perceivable intensity. Not only does this lead to some very inconvenient numerical values, but, given the nature of perception, the figures do not relate very well to the perceptual effects of differences in intensity. If the intensity of a sound is doubled in numerical value on a simple linear scale, it does not necessarily mean a doubling in the sensation of loudness. The relation between perceived loudness and acoustic intensity is more nearly logarithmic. In a logarithmic scale increments are powers of ten, i.e. 2 corresponds to 10^2 (= 100), 3 to 10^3 (= 1,000), and so on. Hence the most convenient way to express intensity so that it relates to perceived loudness is as a logarithmic ratio, comparing the sound to a reference intensity. In honour of Alexander Graham Bell (the inventor of the telephone) the term BEL was given to a unit of this logarithmic scale: one Bel represents a ratio of 10:1, two Bels a ratio of 100:1. It turned out that this unit was too large for practical purposes, and one tenth of it, the DECIBEL (dB), was adopted as the usual measure. Thus one decibel is ten times the logarithm ($\log_{10}$) of the measured intensity (I_a) divided by the reference intensity (I_b):

$$1 \text{ dB} = 10 \log_{10}(I_a/I_b).$$

Note that any sound intensity expressed in dB is always relative to some reference level of intensity. When dB values are expressed without explicit indication of this reference level (as they often are), it can usually be assumed that the reference level is the threshold of hearing. This threshold can be taken to be an intensity of 10^{-16} Watts per cm^2, which corresponds to a sound pressure of 20 μPa, the statistically normal threshold of absolute hearing for a 1 kHz sinusoidal tone (section 7.6 above).

Since intensity is proportional to the square of sound pressure, a decibel is also equivalent to 20 times the logarithm ($\log_{10}$) of measured pressure (P_x) divided by the reference sound pressure level (P_y):

$$1 \text{ dB} = 20 \log_{10}(P_x/P_y).$$

Intensity calculated in terms of sound pressure in this way, using the threshold of hearing as a reference level, is known as SOUND PRESSURE LEVEL or SPL. Some typical sound pressure levels are:

130 dB	very loud sounds, such as the note of a trumpet at the bell of the instrument, or heavy metal music;
100 dB	a brass band or ambulance or police siren;
80 dB	noise in the cabin of a jet aircraft;
70 dB	normal speech;
60 dB	the background noise of a quiet office;
40 dB	very quiet speech;
20 dB	residual noise in a sound-treated room such as a recording studio;
0 dB	threshold of hearing.

Although the logarithmic dB scale relates intensity to perceived loudness far better than a simple linear scale, there are substantial differences in the perceived loudness of sounds at different frequencies. The auditory system of a young healthy adult will respond to sounds at frequencies ranging from about 20 Hz to about 20,000 Hz, but the system is by no means equally sensitive to sounds at all frequencies within this range. Even simple sinusoidal tones of different frequencies may vary by 40 dB or more in their intensity to yield the same perceived loudness. This is particularly true of low-frequency sounds below 200 Hz; it also applies, to a lesser extent, to sounds above 5,000 Hz. (But most of the useful information in speech lies within the 200–5,000 Hz range.) The loudness of complex sounds is a more difficult matter, which will not be considered in detail here. In general terms, loudness is a function of the range and energy distribution of the frequency components in the sound concerned.

Pitch is the perceived period or frequency of a sound wave. Perceived pitch is largely determined by the fundamental frequency of the sound, and to a minor extent by the intensity of the sound, but the relationship between pitch and fundamental frequency is again nonlinear and varies with the frequency involved.

Our sensitivity to changes in the frequency of a sinusoidal tone – in other words our pitch discrimination – varies enormously as we move up the audible frequency scale. Below 1,000 Hz, listeners can readily hear frequency changes of 4 or 5 Hz; above this frequency our ability to perceive small absolute changes in frequency decreases progressively and substantially. By about 8,000 Hz listeners may have difficulty in discriminating changes that are below 40 or 50 Hz. Figure 7.9.1 shows just noticeable differences (JND) in pitch plotted against test frequency, illustrating this characteristic.

Since the relationship between frequency and pitch is not linear, a perceptual unit called the MEL has been devised to represent equal increments of pitch

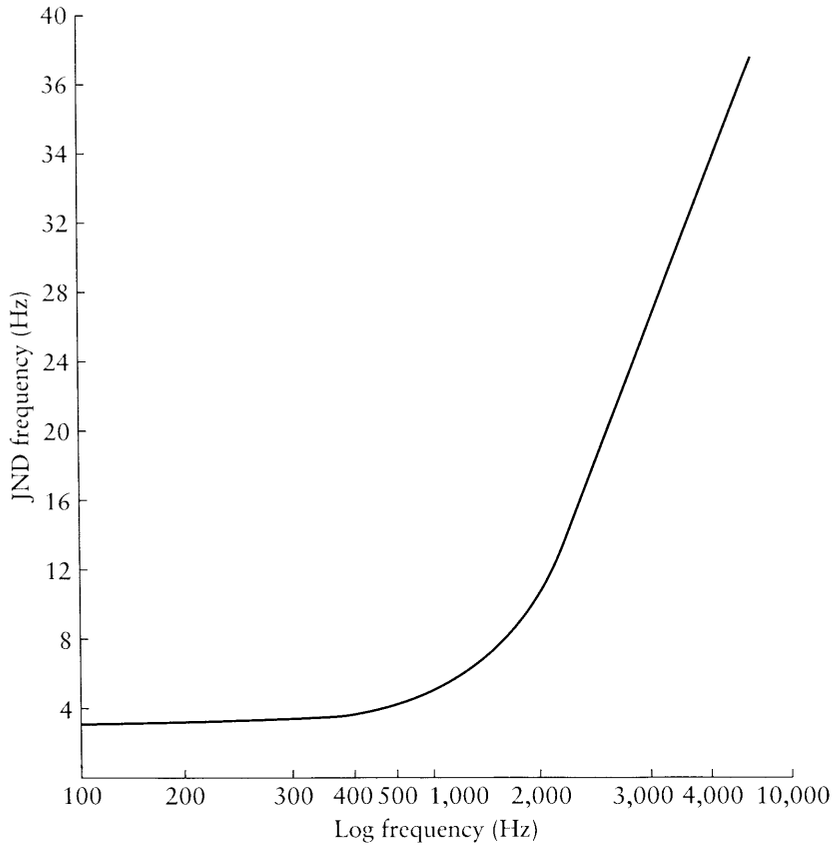


Figure 7.9.1 Just noticeable differences in frequency (JND) against frequency

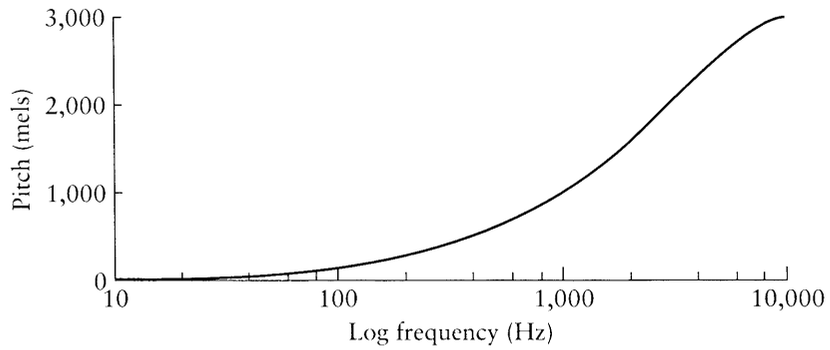


Figure 7.9.2 Mels against frequency

and relate them to frequency. Figure 7.9.2 shows mel values plotted against frequency: below 1,000 Hz there is a fairly direct correspondence between perceived pitch and frequency, and above this point the relationship becomes essentially logarithmic. Readers interested in the calculation may like to note the formula given by Fant (1968):

$P = (1{,}000/\log_{10} 2)\ (\log_{10}(1 + f/1{,}000))$
where P = pitch in mels and f = frequency in hertz.

Alternative transformations from frequency to mels are given by Beranek (1949), Lindsay and Norman (1977) and O'Shaughnessy (1987).

Complex sounds pose a curious problem. The pitch perceived depends on the fundamental, or lowest frequency component in the spectral composition of the sound. It does not matter what the amplitude of that fundamental is in relation to the other harmonic components of the sound. Indeed, even if the fundamental is removed by some form of electronic processing such as filtering, a pitch corresponding to the fundamental, known as the 'phantom fundamental', will still be perceived. (Recall that it is the fundamental frequency which determines the harmonic spacings.) It appears that essential pitch information can be decoded by listeners from the harmonic structure of the complex sound, at least for frequencies up to around 5,000 Hz.

Finally, we return to timbre and the example of a violin and a flute, playing the same note but sounding very different (section 7.1 above). In essence, timbre is a quality perceived in complex sounds: we hear differences of timbre in complex sounds, even when their perceived loudness and pitch are the same, because of the difference in the energy distribution of their spectra.

Speech sounds may be distinguished in just this way, the simplest example being that of pure vowel sounds. A speaker may produce different vowels of the same pitch and loudness, but the vowels are perceived as different sounds for the very reason that the violin and flute sound different, namely that the distribution of their spectral energies is different. As we shall see below, there are additional complexities in speech sounds, but the perceptual processes rest on these basic principles. Further information on the perception of sound can be found in Lindsay and Norman (1977), Moore (2003) and Warren (1982).

7.10 The acoustic model of speech production

The acoustic behaviour and properties of the human vocal tract in speech production are traditionally considered in terms of a source and filter model of the general type shown in figure 7.10.1. In the light of this model, the speech

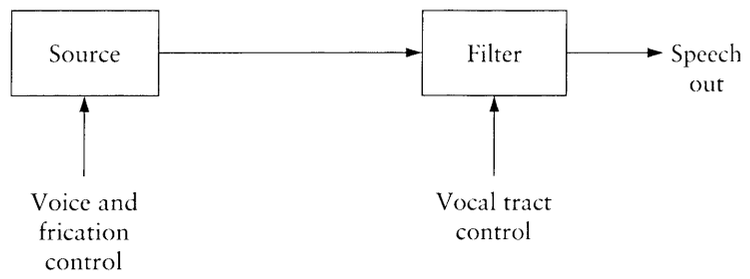


Figure 7.10.1 Source and filter model of speech production

signal can be viewed acoustically as the result of the properties of the sound source, modified by the properties of the vocal tract functioning as a frequency-selective filter. Both the properties of the source and those of the vocal tract can be varied – and are varied continuously – during speech.

A description of speech should of course relate the acoustic properties of the speech signal to the phonological information which the signal conveys. This is by no means a simple task, but the source and filter model provides a convenient functional division of the mechanisms that are active in the process of generating speech sounds.

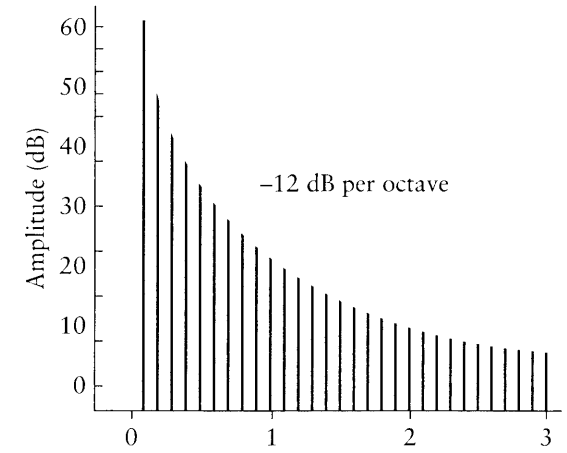
7.11 Phonation as a sound source

The periodic vibration of the vocal folds known as PHONATION (sections 2.6 and 6.6 above) provides the most important and acoustically efficient sound source in the vocal tract. The expiratory airflow from the lungs is interrupted or modulated in a periodic vibratory cycle, and muscular tension settings and aerodynamic forces regulate the frequency and intensity of the output. An idealized form of the phonation airflow waveform is shown in figure 7.11.1(a), corresponding to the waveform of figure 6.6.2. The waveform displays the amount of air flowing through the glottis, plotted against time, and can be described as a volume velocity waveform. Being a form of periodic vibration, the waveform has a harmonic spectrum, as shown in figure 7.11.1(b). The slope of the energy profile of the spectrum for this idealized waveform is –12 dB per octave, which means that the intensity of the harmonics falls away quite rapidly at high frequencies. In normal speech, the slope of the spectrum varies considerably, depending on the phonatory setting being used. To some extent, this setting will be a matter of the individual's choice of speaking style; to some extent it will reflect the speaker's personal voice quality and habitual long-term phonatory setting.

Figure 7.11.2 shows volume velocity waveforms for two varieties of phonatory setting, with their harmonic spectra. Example (a) is breathy voice: the waveform results from relatively slow closure of the folds for quite short periods during the total cycle, such that there is almost continuous airflow. Although not shown here, there is usually some accompanying turbulence (especially in the region of vocal fold closure and minimum airflow). By contrast, example (b) represents quite forceful phonation, usually in the context of high overall articulatory effort, which results in very 'bright' voice quality. In this case, the folds remain closed for more than half the total cycle and the closure action is quite rapid, causing a very sharp fall in airflow rate before the airflow stops altogether in the closed phase of the cycle.

Several aspects of phonation waveforms contribute to their spectral shape. The property which affects voice quality as much as any other is the slope of the spectrum, as described above. This slope is controlled largely by the rate of change of airflow during the phonatory cycle, usually its fall from peak to closure in the pitch pulse. The faster the rate of change, the smaller the spectral

(a)



(b)

Figure 7.11.1 Idealized phonation waveform and spectrum
Adapted from: (b) Pickett 1980, p. 64.

slope and the greater the amount of high-frequency energy available as input to the vocal tract filter. Thus phonation which results in a rapid rate of change in the airflow is generally more efficient as a vocal tract sound source, because of its overall greater distribution of acoustic energy. The number of phonatory settings is of course virtually infinite, and the examples above merely illustrate one or two possibilities and their acoustic consequences.

Apart from its long-term variability, phonation also shows minor inconsistencies from cycle to cycle, which may have some effect on voice quality. All speakers seem to exhibit some inconsistency in duration from cycle to cycle of phonation. This constitutes variation in frequency, known as pitch JITTER. The greatest degree of jitter is usually evident at the start of phonation following a voiceless consonant, after which it reduces greatly in the syllable peak. If jitter is more pervasive, it is likely to be perceived as 'roughness' or 'harshness' in voice quality. Inconsistencies in the amplitude of phonation from cycle to cycle, known as SHIMMER, may also contribute to perceived voice quality.

More regular or periodic changes from cycle to cycle are also common. The perceptual effects are generally covered by the labels 'creak', 'creaky voice' and

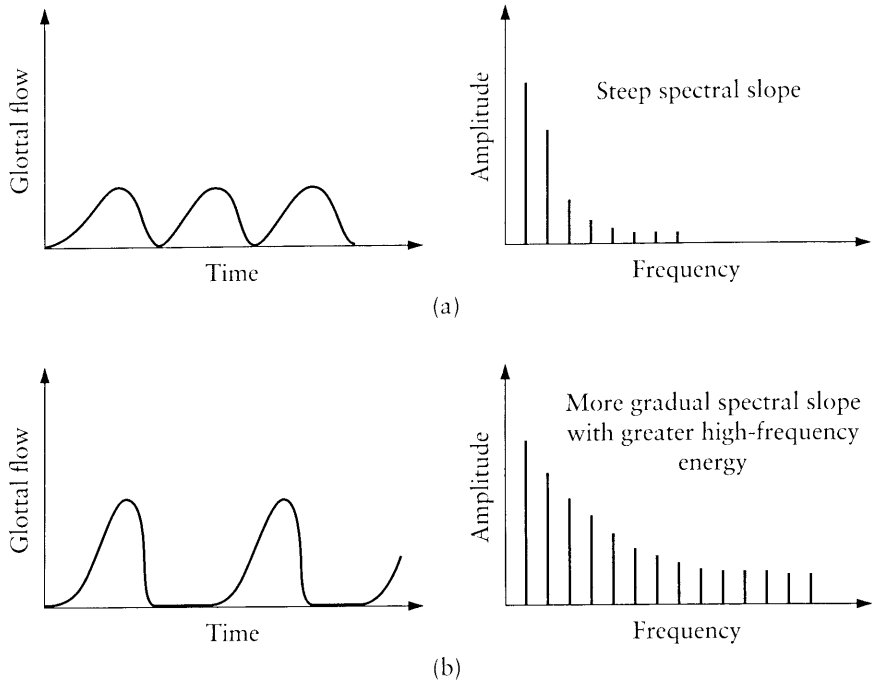


Figure 7.11.2 Phonation airflow (volume velocity) waveforms and their corresponding spectra: (a) breathy voice; (b) bright voice

'vocal fry' (cf. section 2.6 above). These changes occur most often at or below the bottom of a speaker's normal pitch range. The phonation waveform may show either a regular pattern of a pitch pulse of high amplitude followed by one of low amplitude, or pairs of pitch pulses close together with a longer interval between successive pairs (sometimes known as double pulsing).

Vocal creak is pervasive in some speakers, but may also be used deliberately. It is not uncommon for English speakers to switch into creaky voice as they reach the end of an utterance on low pitch, where, functionally, creak may be said to serve as a kind of extension of low pitch into a yet lower range. Flanagan (1958), Miller (1959), Lindqvist (1970), Fant (1979), Sundberg and Gauffin (1979) and Ananthapadmanabha (1984) provide extensive discussion of the acoustic properties of phonation, and Laver (1980) and Nolan (1983) offer useful accounts of how these properties contribute to overall voice quality and its linguistic functions. Ladefoged's review of the phonation process (1971, ch. 2) includes examples of contrasts from a variety of languages, including the breathy (or 'murmured') voicing of south Asian languages such as Gujarati, and the creaky voicing of west African Chadic languages such as Margi.

The phonation waveforms and spectra shown so far in this section have all been idealizations of natural speech, for the sake of simple illustration. But, in a sense, all phonation waveforms are idealizations: they cannot be derived directly from the acoustic output of the vocal tract at the lips, for the vocal tract filter (section 7.10 above) alters the phonation waveform spectrum and

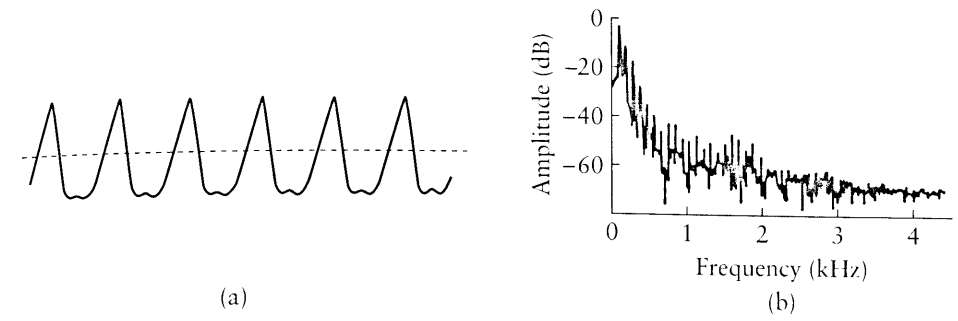


Figure 7.11.3 Waveform and spectrum of typical phonation in natural speech: (a) waveform; (b) line spectrum
Adapted from: Sondhi 1975, p. 230.

thus modifies the output waveform. Hence 'actual' waveforms can be obtained only by a complex measurement technique which largely cancels out the effects of the vocal tract filter. Figure 7.11.3(a) shows such a derived waveform, based on Sondhi (1975), who describes one of the ways of obtaining a phonation waveform which is relatively uncontaminated by vocal tract filter effects. The corresponding spectrum in figure 7.11.3(b) reveals discontinuities of energy and does not have a constant spectral slope, but these 'irregularities' do contribute in some measure to overall voice quality, often in a quite idiosyncratic way.

Two properties of phonation stand out as important in our understanding of the voice source: firstly, FUNDAMENTAL FREQUENCY (F_0), i.e. the frequency of vibration of the larynx in phonation, which can be measured directly from the speech waveform; and secondly, INTENSITY, as the primary determinant of overall speech intensity. Phonation modes cannot be included in quite the same way, for although they can be readily categorized auditorily (section 2.6 above), they stand in a far more complex, more indirect and less consistent relationship to various acoustic values.

The range of fundamental frequency employed by speakers reflects physical differences in the larynx, particularly in the length and muscular settings of the vocal folds in males, females and children (see section 6.5 above). There is wide individual variability, but the general ranges of F_0 for English speakers are:

Adult males	80–200 Hz
Adult females	150–300 Hz
Children	200–500 Hz.

Average values suggested by Peterson and Barney (1952) are around 130 Hz (males), 220 Hz (females) and 270 Hz (children). We should, however, be wary of generalizing about the characteristics of adult male, adult female and children's voices. Peterson and Barney's figures derive from linguistically restricted material, and do not reflect the dynamics of intonation. Moreover, it is especially difficult to generalize from the data on children because of the additional

variation due to the process of maturation. It is also clear that there is appreciable overlap in the ranges used by the three groups, and that frequency range alone does not distinguish among them. There is much to be explored in this area, a point which is underlined by the fact that most research in speech acoustics has used male voices, partly for reasons of convenience in spectrographic analysis (section 7.14 below).

7.12 Sources of frication

Frictional sounds depend on air turbulence (sections 2.6 and 2.12 above) which creates aperiodic acoustic energy (sections 7.1 and 7.8 above). Unlike phonation, frictional sound may be generated at any location in the vocal tract, from the larynx to the lips, provided that it is possible to satisfy the minimum aerodynamic conditions for turbulent airflow: a constriction must be formed between two articulators, and sufficient airflow initiated to meet the aerodynamic conditions required to change laminar into turbulent airflow. These conditions will depend on the cross-sectional area and geometry of the constriction in question, and the acoustic properties of frictional sound are less predictable than those of phonation.

The intensity of frictional sound sources is essentially determined by the aerodynamic conditions and the relevant constriction geometry. Catford (1977) presents some evidence that intensity increases with increasing airstream velocity (which is not to be confused with volume velocity). Arkebauer et al. (1967) have shown that intensity is a function of the differential air pressure across the constriction. Stevens (1972b) and Scully (1979) provide quantitative treatments based on model studies which show that the turbulent noise sources of frication are determined by both the differential pressure across the constriction and its cross-section. The relationship between the two is shown in figure 7.12.1.

For some constrictions, where the fricative constriction area is much smaller than the glottal area, the differential pressure is effectively the subglottal pressure (Psg); but if the two areas are of comparable magnitude, the differential pressure will be defined by the actual pressure drop across the constriction. In general, it appears that the intensity of the sound source is essentially controlled by Psg during the dynamics of articulation, just as with phonation. The constriction area itself has much less influence, and cannot readily be varied systematically for a given fricative sound. As with phonation, we have no direct way of measuring frication sources, and the overall intensities measured for fricative consonants are in many instances strongly influenced by the vocal tract configuration.

Little quantitative information is available on the spectral properties of frictional sound sources. Both Fant (1960) and Stevens (1972b) suggest that the energy distribution is relatively uniform over the frequency range 500–3,000 Hz (within which fricative consonants are distinguished), with a falling slope

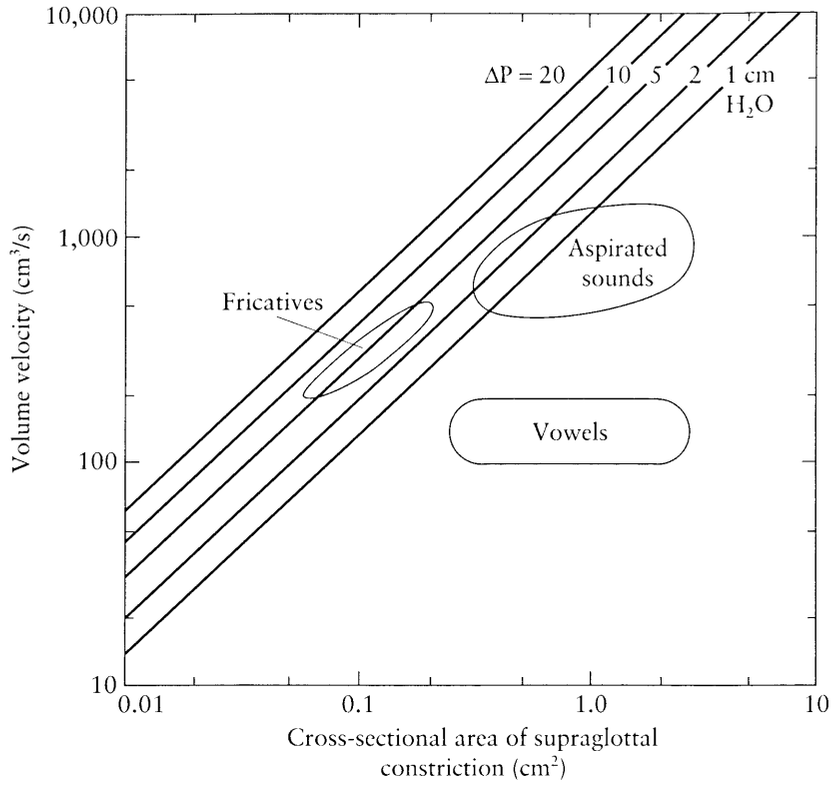


Figure 7.12.1 Relationship between constriction area, differential pressure (ΔP) and airflow
Source: Stevens 1972b, p. 1188.

of 6 dB per octave at both the high and low frequency ends. Figure 7.12.2 is an approximation to the typical frictional spectrum, based on the general characteristics of analogous turbulent airflow.

Voiced frictional sound sources introduce a further complexity. Since phonation occurs simultaneously, the pulses of phonation affect the differential pressure across the frictional constriction. This effect, known as MODULATION, causes regular variation in the pattern of airflow. According to Stevens (1972b), a Psg of 8 cm H_2O will result in a variation of 2–6 cm H_2O across the constriction over each phonatory cycle, causing modulation of around 15 dB at the fundamental frequency. Overall intensity will again be largely controlled by Psg, and the spectrum of this sound source will contain both periodic and aperiodic components.

For descriptive purposes, using categories or parameters of description, two properties of frictional sound sources are significant: their intensity, as reflected in the overall intensity of the speech sounds produced with this source; and their categorization as either voiced or voiceless, determined by the presence or absence of any periodic (harmonic) structure in their spectra.

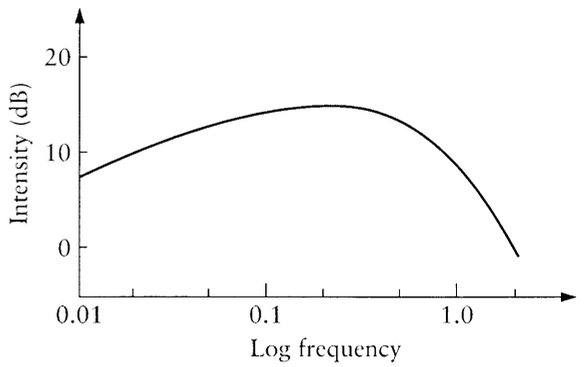


Figure 7.12.2 Fricative source spectrum (based on model studies)
Adapted from: Stevens 1972b, p. 1183.

7.13 The vocal tract filter in vowel production

The source and filter model (section 7.10 above) treats the vocal tract as a filter: the filter is frequency-selective and constantly modifies the spectral characteristics of sound sources during articulation. The properties of the filter vary from moment to moment, for they are determined by the geometry of the vocal tract, which is itself varied as the speaker moves and positions articulatory organs such as the tongue and lips.

We begin with the acoustics of vowel production – consonants are rather more complicated – and take one of the simplest instances, a long central vowel [ɜ] such as is heard in RP *bird* or *fur*. This vowel is formed with the tongue, lips and jaw in a relatively neutral open position, and the cross-section of the supraglottal vocal tract is more or less uniform along its length, as seen in figure 7.13.1(a). In this vowel setting, the configuration of the vocal tract approximates a parallel-sided tube which is closed at one end (the larynx) and open at the other (the lips), as shown in figure 7.13.1(b). With one end closed, the tube acts as a resonator (section 7.5 above); in this case, the resonance is in the air column within the tube, resulting from the reflection of air pressure from one end of the tube to the other in what is known as a STANDING WAVE or stationary wave. Provided that the length of the tube is much greater than its diameter, air pressure will be reflected when a minimum of pressure occurs at the open end of the tube and a maximum of pressure at the closed end. This condition will be met at one quarter of a complete cycle of a sinusoidal vibration and at every half cycle thereafter, as shown in figure 7.13.2. As a result, the air column in the tube will resonate at a basic frequency corresponding to four times the length of the tube (because of the quarter cycle of vibration) and at frequencies corresponding to every successive half cycle – in other words, at frequencies which are three, five, seven (and so on indefinitely) times the basic resonant frequency.

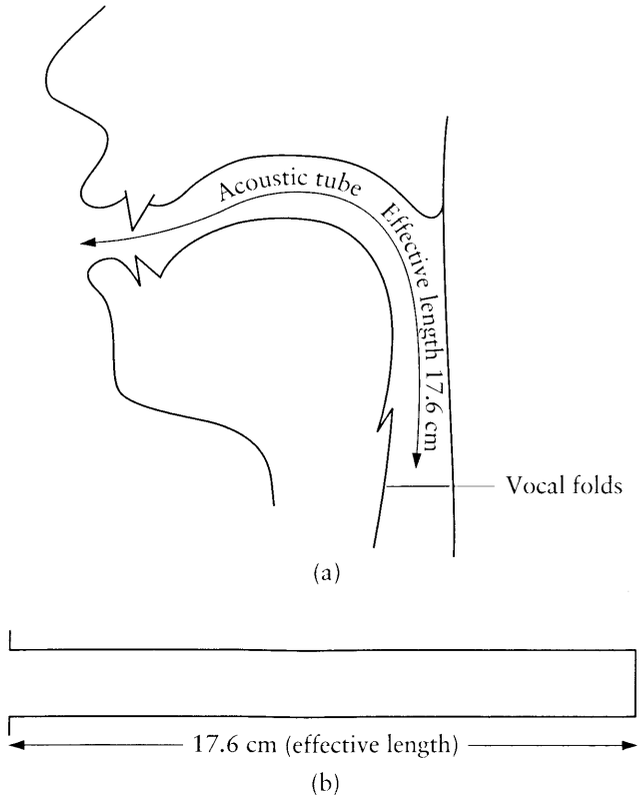


Figure 7.13.1 Resonator configuration for the central vowel [ɜ]: (a) actual vocal tract; (b) simple tube equivalent to (a)

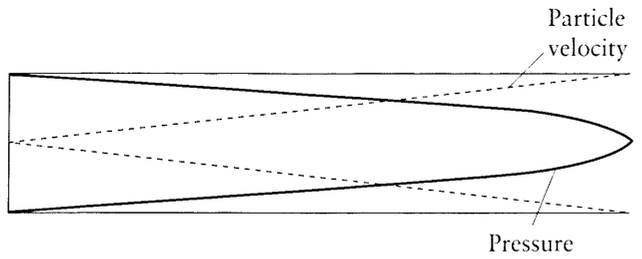


Figure 7.13.2 Quarter wave resonant frequency pressure and particle velocity patterns in a simple tube resonator

If we take *s* to stand for the relevant multiples 1, 3, 5, 7, etc., and *C* to be the speed of sound in air (about 340 m/s at sea level), then *s* multiplied by *C*, divided by 4 times the length of the tube, will give the resonant frequency:

$F_{res} = sC/4l.$

The equation is actually not quite accurate, for the standing wave does not stop precisely at the end of the open tube. In other words, the 'acoustic end' of the tube is slightly beyond the physical end of the tube. The equation therefore needs an 'end correction', which is partly related to the diameter of the tube; further details can be found in Wood (1964). It should also be noted that the reflections in the air column are not perfect, and some acoustic energy is radiated. It is of course desirable that this happens – so that sound is propagated – but it does produce losses in the column which are manifested as damping of the oscillations of resonance (section 7.8 above).

Thus the fundamental difference between the closed tube resonator just described and the resonating systems introduced in section 7.3 is that in the former resonance occurs at a succession of frequencies. This more complex mode of multiple resonance is crucially important in speech production.

We now return to the vowel of figure 7.13.1 and its tube resonator equivalent. The length of this resonator is the length of the vocal tract from the lips to the glottis. Human vocal tracts are of course not all of identical length, and there are appreciable differences, depending on whether the person is male or female, physically mature, and so on. With those reservations in mind, we can nevertheless take a typical male vocal tract to be 17.6 cm long (Fant 1960). According to data in Pickett (1980), the length of a woman's vocal tract is about 80–90 per cent of a man's, while a child's, depending on age, may be around 50 per cent of a man's.

Using the tube resonator equation above – including end-effect corrections – we can show that a vocal tract of 17.6 cm will resonate at 500 Hz, 1,500 Hz, 2,500 Hz, and so on to infinity. Figure 7.13.3 shows the frequency response of this tube resonator. The figure reveals that the frequency-selective characteristics are similar to those of the simple resonator shown in figure 7.5.3, except that a series of resonant peaks now appears at the frequencies predicted by the tube resonator equation. Of these peaks, the three lowest play a major part in determining vowel quality. Higher peaks contribute more to personal voice quality, and become progressively less significant above about 5 kHz.

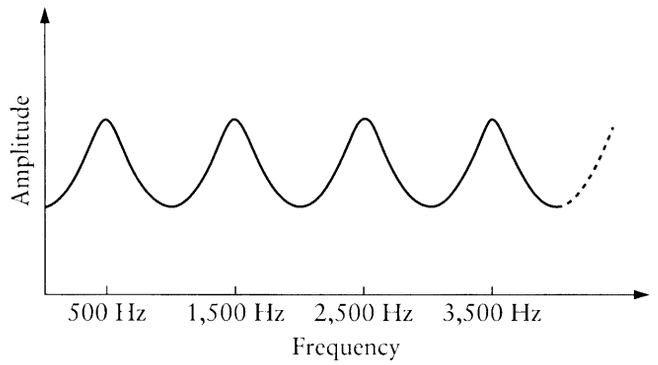


Figure 7.13.3 Frequency response of tube resonator approximating male vocal tract for the vowel [ɜ]

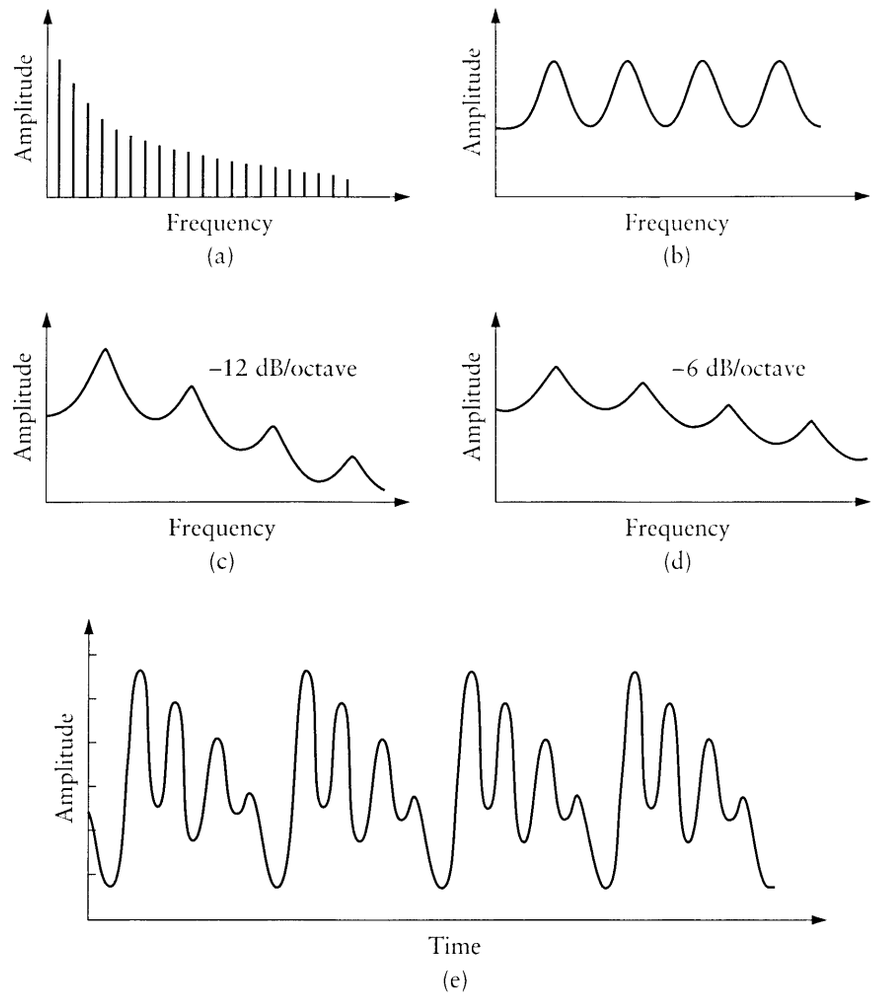


Figure 7.13.4 Acoustic properties of the vowel [ɜ]: (a) harmonic spectrum of phonation source; (b) resonant response of vocal tract; (c) spectral envelope from filtering of source by vocal tract resonance; (d) spectral envelope of radiated sound pressure wave; (e) time domain sound pressure wave

These resonance properties of the vocal tract must of course be considered in conjunction with the sound source, which for vowels is normally phonation at the larynx (section 7.11 above). In fact speech output from the lips actually reflects the combined acoustic properties of phonation, tract resonance and the acoustic radiation properties of the human head. Figure 7.13.4 shows these components for the vowel [ɜ], namely

a the spectrum of laryngeal phonation;
b the resonant frequency response of the tract;

c the phonation spectrum resulting from the effects of tract resonance;
d the spectrum of the final radiated acoustic sound pressure wave;
e the time domain sound pressure wave itself.

In this idealized example, the phonation spectrum (a) is assumed to have a slope of −12 dB per octave, and the resonance peaks in the tract (b) have equal amplitudes. The result is a spectrum (c) with peaks of energy caused by the resonance. These peaks of energy, produced by selective enhancement of the source by tract resonance, are known as FORMANTS. The tract resonances themselves are sometimes referred to as formants, but this is technically imprecise. Formants are a consequence of resonance, not resonance itself. The information-bearing formants of the speech spectrum are conventionally numbered upwards from the lowest in frequency (F_1, F_2, etc.); the three lowest formants are essential parameters in the description of vowel quality.

The final spectrum (d) of the radiated sound pressure wave has a high-frequency slope only half that shown in (c). This is because sound emerges from the lips, and the lips constitute a single point relative to the surface area of the head. The head functions as a kind of reflecting surface, or, more precisely, as a spherical baffle of about 9 cm radius. This favours the propagation of high-frequency sound and causes output to rise by about +6 dB per octave from the region of several hundred Hz upwards. The −12 dB per octave slope of the voice source is thereby reduced to an effective −6 dB per octave, which also enhances information-bearing aspects of the signal.

Our example of the vowel [ɜ] assumes that there is equal damping (or bandwidth) on each of the resonances. It is only if this condition is met that the amplitudes of the resonant peaks are equal. In fact under normal conditions, with modal phonation providing the sound source, there is usually greater damping (wider bandwidth) at the higher resonances, yielding unequal formant amplitudes.

Having looked at frequency characteristics, we can also consider the process of vowel production in the light of the time domain waveform shown in figure 7.13.4(e). This is a damped quasiperiodic wave similar to that shown above in figure 7.8.3 (waveform) and figure 7.8.4 (line spectrum). In both cases, there is a single peak in the line spectrum, and the harmonic energy lines are determined by the repetition rate of the energy restoration. But the vowel waveform is more complex: because of the multiple resonances of the vocal tract, there are several damped sine waves superimposed on each other, and only the one occurring at the lowest frequency is clearly seen. From this perspective, vowel production may be seen as a series of impulses from the larynx which shock the multiple resonator into a series of simultaneous damped sinusoidal vibrations at different frequencies. This gives us an alternative view of the speech production process, a view no less valid than that based on frequency. It is another reminder that reality is multifaceted and cannot be reduced to a single aspect. We must be prepared to think in both ways about the speech signal to understand how phonological structure is encoded in it.

All vowel sounds have spectra which reflect the source, filter and radiation characteristics described above, but other vowels are more complex than our [ɜ] example. Once the articulators are moved from the more or less neutral

posture of a vowel such as [ɜ], the cross-section of the vocal tract is no longer uniform along its length and the tract ceases to approximate a simple tube with parallel sides. The resonance characteristics of the tract are correspondingly more complex. Change of tongue position, vertical movement of the jaw, and protrusion or rounding of the lips can all contribute to variation in cross-sectional area. For the [ɜ] vowel, cross-sectional area is of the order of 6 cm^2; but some vowels, especially high vowels, involve extreme narrowing in part of the tract, and the cross-sectional area at the narrowest point may be as little as one-fifteenth of the area at the widest point in the tract. For the vowel [u:], articulated with lips rounded and tongue fully retracted, Fant (1960) quotes areas as small as 0.32 cm^2 in the region of the lips and as large as 13 cm^2 in the front cavity. The consequence of such variation is that the locations of the resonant peaks on the frequency scale are no longer equally distributed. Their relative amplitudes are also unequal, and are determined by their frequency relationships. A further complication is that there may be absolute differences in the amplitude of individual peaks, caused by differences in their bandwidths. The simple tube resonator calculation cannot be used to find the positions of the vocal tract resonances.

To approximate a vocal tract varying in area along its length, we could imagine two tubes of different size connected to each. And it is possible to estimate the unequally distributed resonance patterns of the vocal tract from such a simple approximation, a compound pair of tube resonators. Fant (1960) provides what are sometimes called nomographs, diagrammatic representations from which values can be calculated. Fant's nomographs allow us to derive the four lowest resonances from the lengths of the resonators and their cross-sectional areas. Figure 7.13.5 shows such compound resonators for various vowels with the approximate positions of their resonant peaks.

It is nevertheless essential to note that the two-tube representation is only a crude approximation of the complex resonant cavity system of the human vocal tract during vowel production. The approximation can be improved by using more than two resonators – indeed, the more tubes the better the approximation – but the calculations also become more complex as the number of tubes increases (cf. section 7.18 below).

Early theories of the acoustics of vowel production did in fact attempt to explain the resonance patterns of vowels in terms of a vocal tract made up of two resonant cavities coupled together. It would indeed be convenient if we could associate resonances (and hence formants) with specific cavities formed by the vocal tract shape characterizing a particular sound. The model does work tolerably well for high vowels such as [i:] where the tongue does divide the tract into a small front cavity and a large back cavity. In this case the first resonance is correspondingly quite low and the second quite high. Unfortunately, the model does not work well either for higher-frequency resonances or for open vowels in general.

Figure 7.13.6 shows the vocal tract configurations and typical spectra for the vowels [i], [a] and [u]. The diagrams of the vocal tract are derived from X-ray data and represent the varying cross-sectional area of the vocal tract from lips to glottis. As already noted, it is the distribution of the three lowest formants

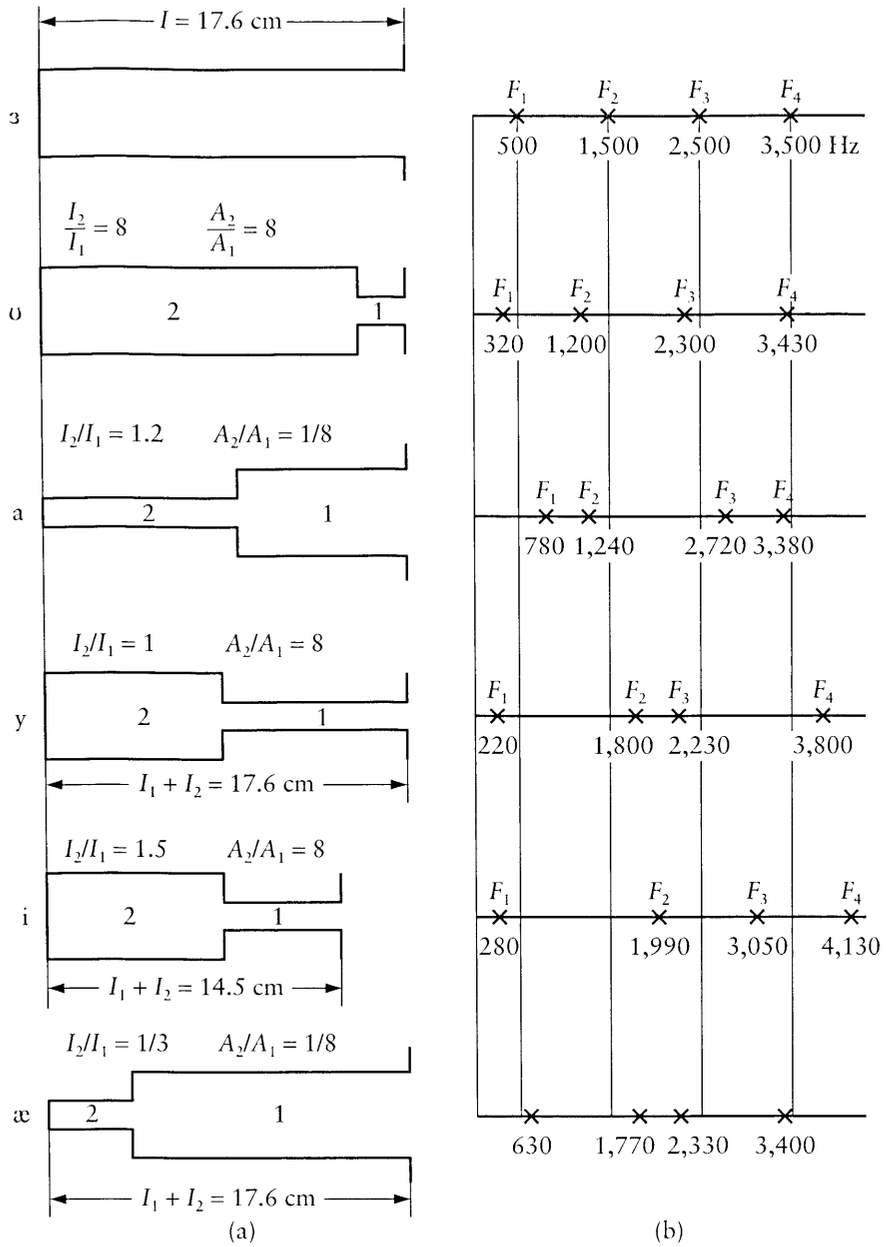


Figure 7.13.5 Two-tube resonators approximating the vocal tract for various vowels: (a) resonator dimensions; (b) formant pattern
Adapted from: Fant 1960, p. 66.

in these spectra which distinguishes these vowels from each other. Note that the absolute values of the formant frequencies are not crucial, but their relative relationships are, reflecting the inherent systemic character of phonological contrast and its encoding in the acoustic speech signal.

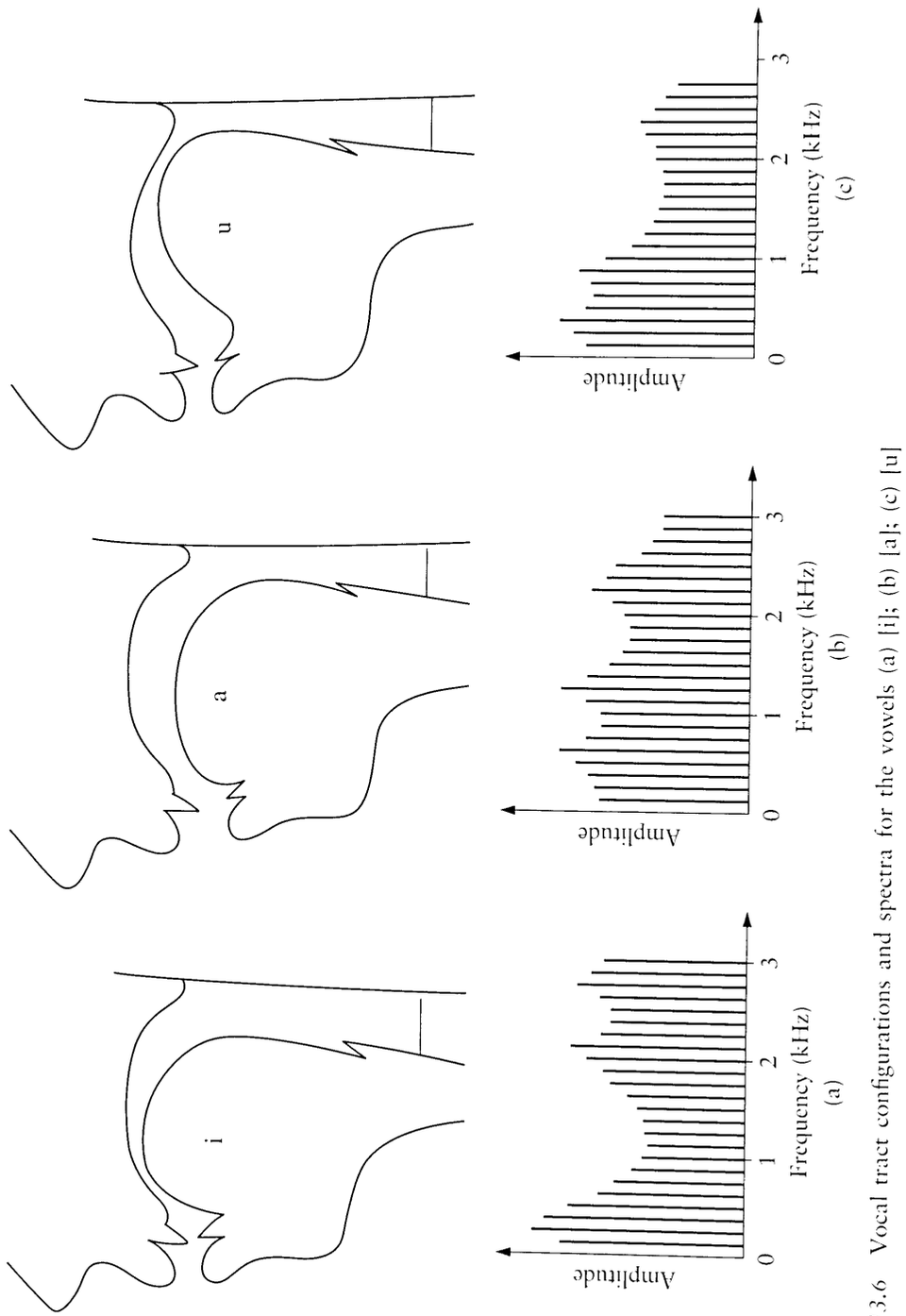


Figure 7.13.6 Vocal tract configurations and spectra for the vowels (a) [i]; (b) [a]; (c) [u]

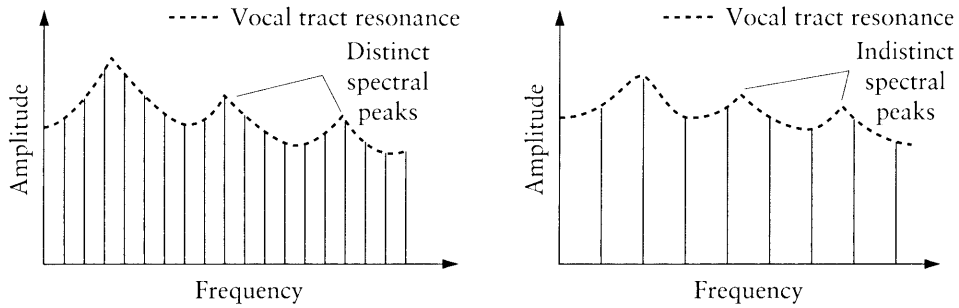


Figure 7.13.7 Effect of harmonic spacing on spectral peak structure

It is important to understand that the frequencies of the tract resonance peaks do not necessarily coincide exactly with harmonic energy lines from the voice source. They will coincide only when the resonance frequency is some multiple of the frequency of phonation (F_0); since both phonation frequency and resonance patterns change continuously in speech, the frequencies will not be consistently or systematically related. There are two important consequences of this. The first is that the frequency of a given formant (i.e. its frequency of maximum energy amplitude) may not coincide exactly with the frequency of tract resonance. The second is that for voices with high F_0 ranges (notably those of children and some females), there may be very few harmonic lines within the amplitude-enhancing range of a given tract resonance peak, and hence the formant which results may not be distinctly defined. Figure 7.13.7 illustrates two extremes.

So far we have ignored nasal vowels – vowels in which the oral–pharyngeal resonator system is coupled with the resonator system of the nasal cavities by the lowering of the soft palate. When nasal coupling occurs, the additional resonator system modifies the relatively simple resonance patterns found for oral vowels: some resonances, notably the lowest, are enhanced and others weakened by so-called ANTIRESONANCES in the compound resonator system. Although the nasal cavities are anatomically stable, their overall geometry is affected by physiological factors (section 6.8 above) and, as Fant (1960) points out, the contribution of nasal resonance is therefore rather unpredictable. From the listener's viewpoint, the most important aspect of nasal coupling is that it distorts that basic vowel spectrum. It is the relative difference between the distorted and undistorted spectra which is relevant in the perception of a contrast. Figure 7.13.8 shows (a) the complex resonant cavity system of a typical nasalized vowel, and (b) the kind of spectrum which may result.

Finally, it is worth reminding readers that for vowels at least (whether nasalized or not), it is the relative distribution of the resonant peaks in the vocal tract that matters. Thus if whisper rather than normal voice phonation is used as a sound source, vowel identity can still be preserved. In whispered vowels, the formants are peaks of aperiodic energy, demonstrating that it is the energy peaks themselves which reflect tract filter characteristics and which allow the hearer to discern the identity of the vowels.

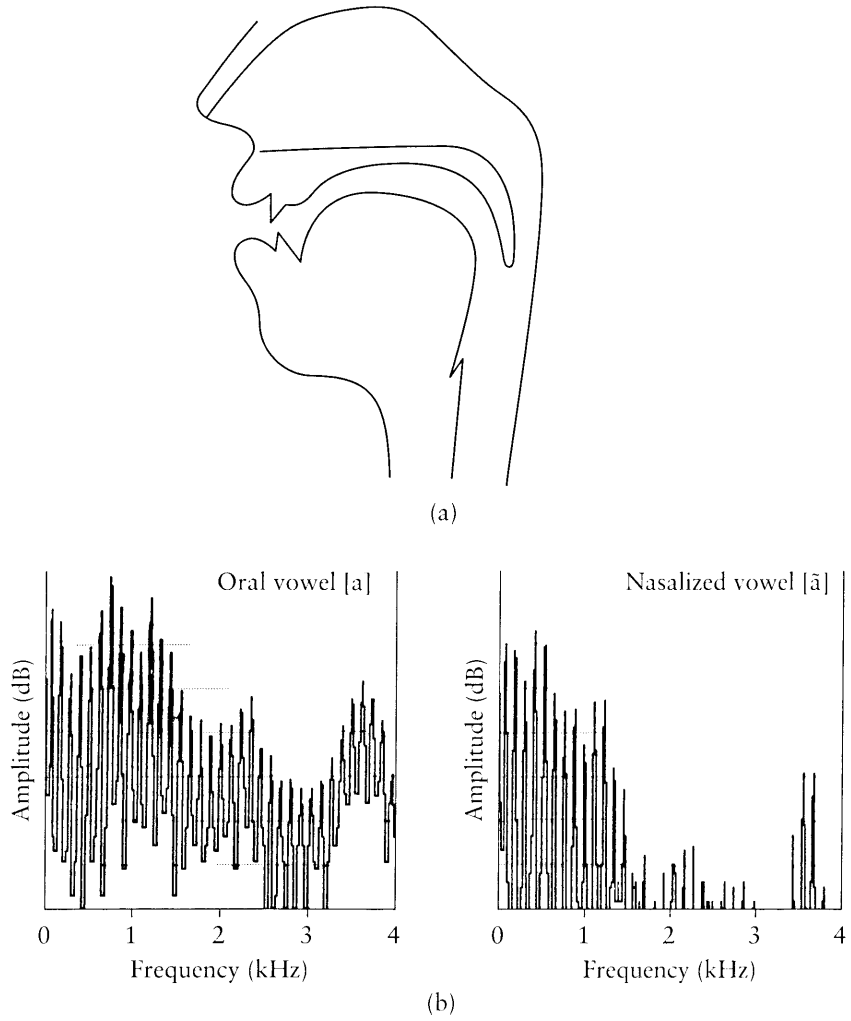


Figure 7.13.8 Nasalized vowel: (a) vocal tract resonance system; (b) comparative oral and nasal vowel spectra for [a]

7.14 Spectrographic analysis of speech

Our account so far has been concerned with static vowel sounds: all the spectral information we have presented has been in the form of SPECTRAL SECTIONS. In effect we have taken slices of speech, to show the distribution of acoustic energy across the frequency spectrum in a specific portion of time. Such an analysis gives no information about any changes during the time course. Single spectral sections are certainly very useful, but given that speech is a dynamic process, it is essential to have a spectral analysis which also displays the changes

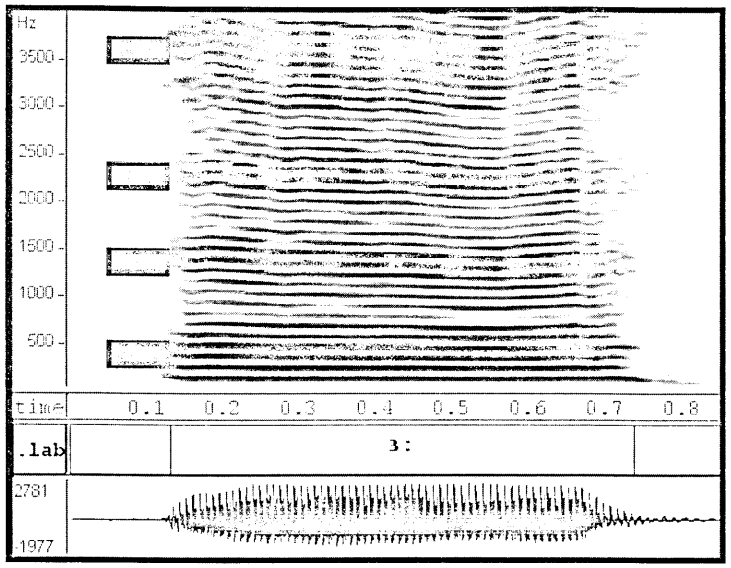


Figure 7.14.1 Spectrogram of the vowel [ɜ:]

in the speech spectrum over time. Indeed, the development of the first instruments which could do this, at the end of World War II, proved to be one of the great landmarks in experimental phonetics.

There are several ways in which time-varying spectral energy can be displayed, the challenge being that of portraying three continuously variable dimensions on a flat (two-dimensional) display. The classic format is known as the SPEECH SPECTROGRAM: frequency is represented on the vertical axis of the display, and time on the horizontal axis, while the magnitude of acoustic energy is shown by the intensity (darkness or brightness) of the display. Figure 7.14.1 shows such a display for the central vowel [ɜ:], produced with a constant fundamental frequency of around 125 Hz. The frequency axis usually has variable scaling but is commonly set to 0–8 kHz or less. In the spectrogram of figure 7.14.1, the harmonic structure of the speech signal can be seen very clearly, the darkest harmonic lines indicating the peaks of energy of the formants. These energy peaks are marked by the black bars on the left-hand side of the spectrogram.

The instrument traditionally used to produce an analysis of this sort is the SPEECH SPECTROGRAPH. First described by Koenig et al. (1946), it has been refined technically over the years but with little change in principle. These analog hardware-based machines were available commercially in several forms for many years, but today, purpose-built hardware of this sort has been largely replaced by digital signal processing (DSP) software packages which will run on a variety of computers and workstations including most current PCs. These packages generally include facilities for editing primary recordings of speech, spectrographic analysis and pitch analysis, and often allow the integrated

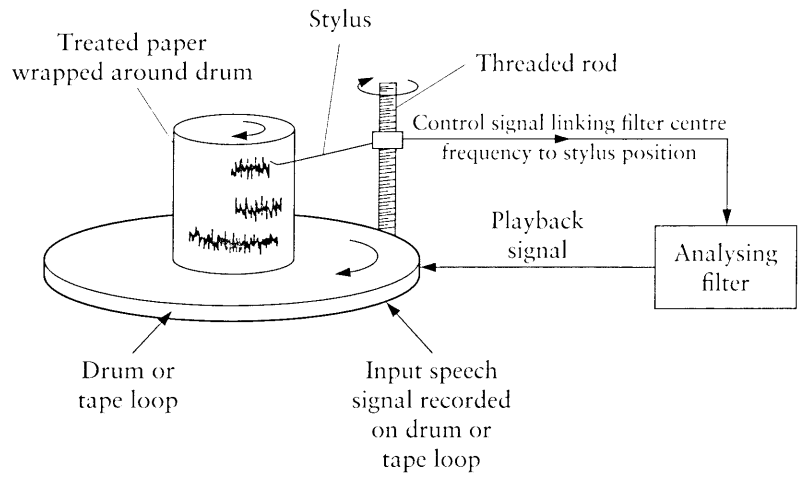


Figure 7.14.2 The speech spectrograph

recording and management of segmentation and labelling data. All of the spectrograms used as illustrations in this text were produced with a PC-based software package called Wavesurfer, developed at the Royal Institute of Technology in Stockholm. Other widely used acoustic phonetics software options that run on different computer platforms include Praat, developed by Paul Boersma at the University of Amsterdam, and the EMU suite of tools developed by Cassidy and Harrington (2001).

Figure 7.14.2 shows the way the traditional analog instrument is organized. The stretch of speech which we wish to analyse is first recorded on a magnetic drum, or in some machines on a loop of magnetic recording tape. Usually no more than 2.5 seconds of speech can be recorded. To perform the analysis, the sample is repeatedly replayed by rotating the drum or running the tape loop. To save time and to minimize certain technical limitations in the analysis, this replaying is at a much higher speed than normal speech. During each rotation of the drum or tape loop, the replayed speech is passed through a bandpass filter to achieve a spectrum analysis. This filter allows only the energy from the range of frequencies within its bandwidth (or bandpass) to pass through. Figure 7.14.3 illustrates the principle of bandpass filtering.

The key property of the filter is that it can be electronically moved across the range of frequencies to be analysed. On the very first replay of the recorded speech, it is effectively located so that the centre of its band is at the lowest frequency to be analysed. The output from the filter is thus only the energy in the speech sample that falls within that small range of frequencies. Mechanically linked to the rotating drum or tape loop is a smaller drum around which is wrapped a sheet of specially treated paper. The paper is actually a sandwich with an inner layer of a carbon-based powder between two layers of paper. Beside the magnetic drum and paper drum assembly is a tall threaded metal

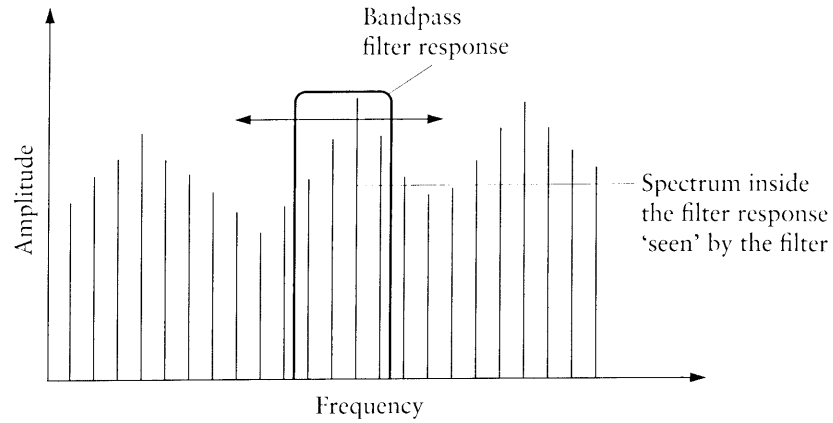


Figure 7.14.3 Bandpass filtering

rod, also mechanically linked to the rotating magnetic drum, usually via a rubber drive belt. Mounted on the threaded rod is an assembly, also threaded, containing a wire stylus. When the threaded rod rotates, this stylus assembly moves up the rod, just as a nut moves along a bolt if the bolt is rotated.

The output of the bandpass analysing filter is processed to yield a signal representing the overall intensity of the energy from the filter. The signal is transformed into a high voltage and fed to the wire stylus, which rests on the paper wrapped around the drum. The voltage at the stylus, representing the intensity of the energy in the speech sequence, burns into the carbon-loaded paper: the greater the stylus voltage, the greater the degree of burning and the blacker the paper at that point on the rotating drum. With each rotation, the stylus assembly moves up the paper and marks a new section. As the bandpass filter is linked electronically to the stylus position, it too moves to a higher frequency range with each rotation.

Thus the stylus and the analysing bandpass filter spiral continuously up the frequency scale, while the paper is synchronized so that the speech recorded on the drum is repeatedly passed through the filter, but always at a new band of frequencies. The stylus marks on the paper the relative intensities of all the frequency components in the speech sequence, and does so in correct timing, matching the pattern of the speech signal itself.

The frequency scale may be adjusted to display the range of speech frequencies required. For vowels the range is commonly set at 0–5 kHz, as there is little of phonological interest above this range. For fricatives, which often exhibit substantial high frequency aperiodic energy, 0–8 kHz may be more appropriate.

The other important variable in this form of analysis is the frequency resolution of the spectral analysis, which is set by the bandwidth of the analysing filter. If the filter has a very narrow bandwidth it will be able to pinpoint the energy from the individual harmonics, but if the bandwidth is wide in relation to harmonic spacing, the filter will not reveal comparable detail. In traditional

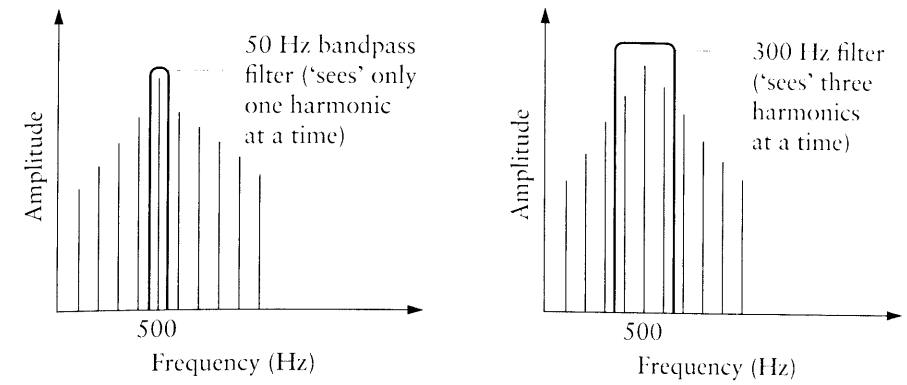


Figure 7.14.4 Effects of narrow (50 Hz) and broad (300 Hz) bandpass filters

analog instruments of the kind just described, the width of the narrow and broadband filters was usually set to 50Hz and 300Hz respectively. These bandwidths were chosen as a practical compromise for the reliable analysis of typical male speakers with relatively low-pitched voices. Fortunately, modern software-based spectral analysis of speech is not constrained by these limited analysis options, because most DSP packages allow almost infinite variability in setting analysis bandwidths. Figure 7.14.4 shows how an F_0 of 100 Hz would be 'seen' by bandpass filters of 50 Hz and 300 Hz respectively.

From figure 7.14.4 it is evident that the 50 Hz filter easily resolves the individual harmonics of the signal, and would do so down to an F_0 of nearly 50 Hz. Below this frequency, the harmonics would start to become too closely spaced for them to be identified individually. (In fact very few speakers have F_0 values below this value in their speech.) The 300 Hz filter will obviously be unable to resolve the harmonics unless F_0 approaches or exceeds 300 Hz.

In most analysis applications attention will be focused either on the formant structure of the speech or on the F_0 patterns. It was for this reason that hardware-based analog spectrographs were usually provided with at least two analysis filters, known simply as narrow and wide, in which the narrow-band filter was used for analysis of F_0, and the wide-band filter for formant analysis. The same principle applies in using DSP software packages, but with the great advantage that the filter bandwidths may be set at optimal widths for the voice being analysed. As noted earlier, this development has allowed much more effective research to be undertaken on a wider range of speakers, especially those with higher-pitched voices.

Wide-band filters effectively smear the harmonic structure so that only the overall energy seen by the filter across the span of its bandwidth (e.g. 200 or 300 Hz) is detected, thus enhancing the visual display of the formant structure. See figure 7.14.5 for narrow- and broad-band analyses of the vowel of figure 7.14.3; but note that in this case F_0 falls from the region of 140 Hz to around 90 Hz.

There is another effect of varying the filter bandwidth, which influences the time domain aspect of the analysis. All filters take a finite time to respond at

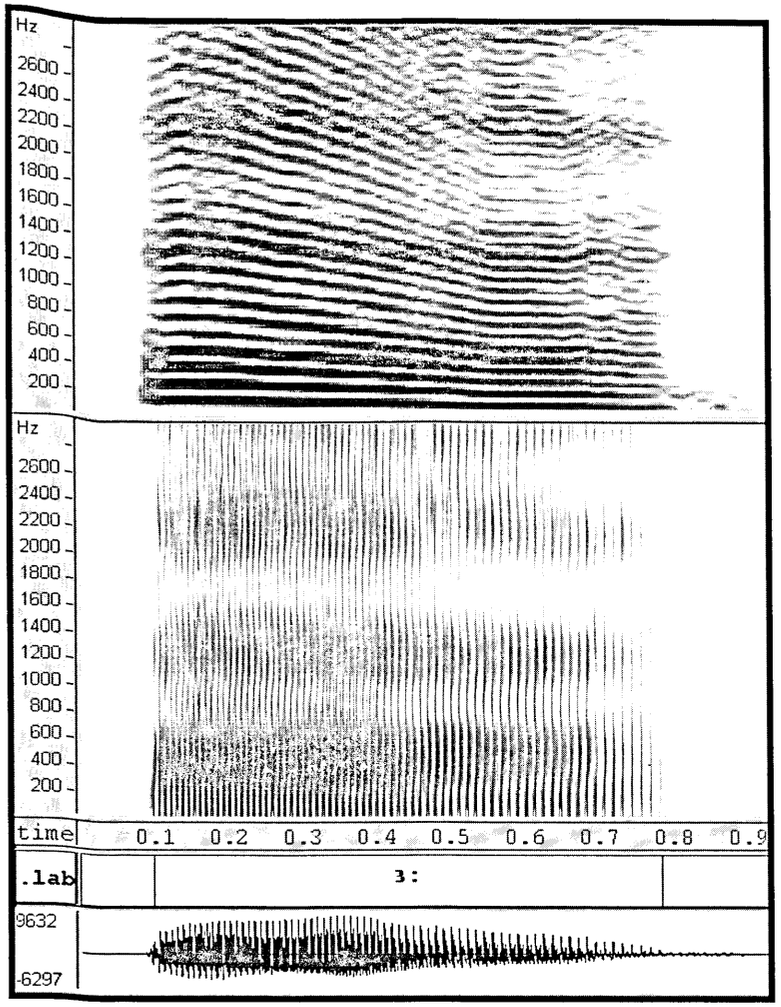


Figure 7.14.5 (a) Narrow- and (b) wide-band analyses of the vowel [ɜ:]

the output to energy at the input; this response time T_r, in seconds, is approximately the reciprocal of the filter bandwidth:

T_r = 1/filter bandwidth.

Thus the response times of 50 Hz and 300 Hz filters will be about 20 ms (one fiftieth of a second) and 3.3 ms (one three-hundredth of a second) respectively.

This means that there is a direct trade-off between frequency resolution and time resolution. Fortunately, this works largely to our advantage in speech analysis, because F_0-related analysis does not usually involve the examination of rapid changes of spectral energy in the time course of the speech; in particular, changes occurring over less than 20 ms are unlikely to be of interest. On the other hand, rapid changes are often relevant in formant-related analysis, where the wide-band filter is able to respond far more quickly (to energy changes occurring over more than 3.3 ms). The time- and frequency-resolution properties of the filters used for analysis are easily seen in figure 7.14.5. In the narrow-band analysis, the harmonics are clearly visible and come closer together on the frequency scale as F_0 falls; in the wide-band analysis, no harmonics can be seen but the location of formant energy is distinct. In addition, the wide-band display shows vertical lines which correspond to the individual pulses of phonation, and these can be seen to be more widely separated on the time scale as F_0 falls.

A comparison of the representation of F_0 in these two displays is another reminder of the importance of balancing different perspectives: no one view can be singled out as the sole objective representation. Wide-band analysis is used more commonly, because it provides detailed information about the dynamic properties of the speech spectrum that relate to its segmental structure. Narrow-band analysis, on the other hand, gives little segmental information, but provides unambiguous information about pitch patterns in speech.

Spectral displays of the sort shown above thus provide easily quantified and quite accurate information about the general pattern of energy distribution in the spectrum over time. In particular they highlight the location of formant peaks and other high-amplitude energy, and for many purposes this is all that is required. But they do not provide readily quantified data on the amplitude and shapes of energy distribution. This applies especially to wide-band spectrograms because of the filter's smearing effect on harmonic structure. Most spectrographic analysis equipment and software packages include a facility for making SPECTRAL SECTIONS which do give detailed amplitude information – based on the narrow-band analysis filter – about any given point on the waveform.

Although the use of computers and special-purpose digital hardware has overtaken much of the traditional analog instrumentation in speech research, it remains true of spectrum analysis, as of other areas of speech technology, that the underlying principles and aims do not really change. Most forms of spectral analysis relevant to our concerns perform a Fourier analysis (section 7.8 above): what is distinctive about digital analysis is only that it uses discrete samples. That is, analog instruments process a continuous speech waveform, but digital systems perform an analysis on discrete samples of that waveform. The general procedure is known as the DISCRETE FOURIER TRANSFORM or DFT.

Figure 7.14.6 shows the process of DIGITIZING a waveform: the sampling is discrete on both the amplitude axis and the time axis. The accuracy of the digital representation depends on the number of discrete steps or samples taken on each axis. The digital encoding of amplitude is known as QUANTIZATION. Since each amplitude value is represented as a number made up of bits (binary digits, i.e. zeros or ones), the total number of steps on the amplitude axis will always be a power of two (2, 4, 8, 16, 32, 64 and so on). Sufficient steps must be used to ensure that we do encode the range of amplitudes in the waveform being analysed. In speech, we are normally interested in frequency components of the spectrum of a given speech waveform over an amplitude range between 40 dB and 60 dB, and the amplitude quantization must contain 1,024 steps to

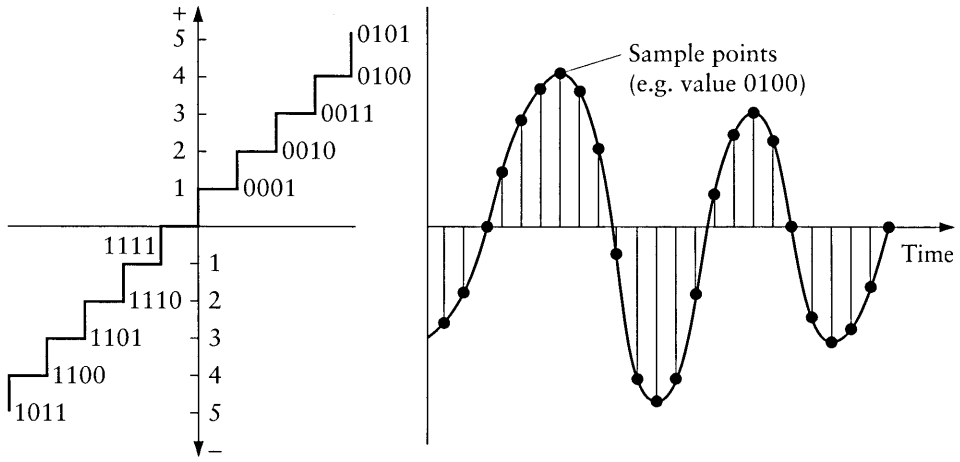


Figure 7.14.6 Waveform sampling: digital values for the magnitude of the waveform at each sample point may be read from the quantization scale

encode a 60 dB range. The figure 1,024 corresponds to a 10-bit binary number, and we thus need a 10-bit word per sample. (If this seems large, it is worth noting that hi-fi music is usually encoded on compact discs using 16 bits per sample.) The digital encoding of the instantaneous values of the amplitude at regular discrete intervals of time along the speech time domain waveform is known as the process of SAMPLING. The number of such samples taken per unit time is known as the SAMPLING RATE.

The maximum frequency (F_{max}) encoded is directly determined by the sampling rate. If T is the time in seconds between successive samples, then F_{max} is the reciprocal of $2T$:

$$F_{max} = 1/2T.$$

In other words, the frequency of the sampling rate must be double that of the highest frequency component which we wish to encode digitally. For example, to include all the frequency components up to 5,000 Hz we must have a sampling rate of 10,000 samples per second, so that $T = 10$ microseconds. Again, this is quite a modest level of accuracy compared to the sampling rates in excess of 40,000 per second used for compact discs and hi-fi digital tape recording.

The properties of the DFT spectrum analysis are directly related to sampling rate. To obtain a spectral section of the kind shown in figure 7.14.7, we select the place of interest in the waveform and use a set of samples on the time axis giving amplitude values around that point to make the spectral analysis (DFT) calculations. A single sample obviously cannot be used, because the Fourier analysis must have access to properties of the waveform over time. For this reason, a WINDOW on the time axis is needed to capture a precisely known interval of the waveform information. This yields a set of samples. To avoid

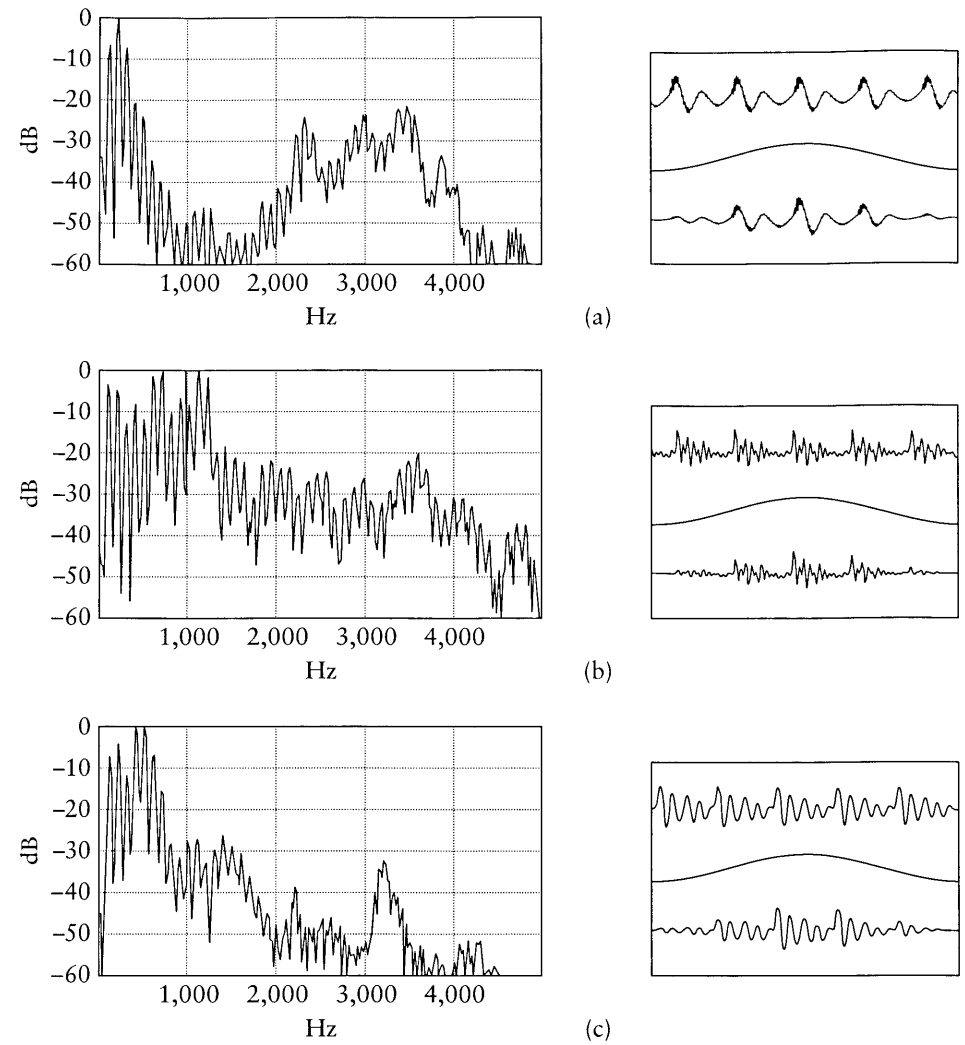


Figure 7.14.7 FFT analyses (512-point) of the vowels (a) [i:], (b) [a:]; (c) [ɔ:], together with their windowed waveforms used in the analysis

introducing artefacts into the analysis, the samples in the window either side of its centre can be amplitude weighted in one of several ways. The most common of these, known as the HAMMING WINDOW, progressively reduces the amplitude of the samples either side of the centre, using a cosine law. The width of the window is defined by the number of samples it contains. Where no amplitude weighting is used the window is said to be rectangular.

A commonly used version of the DFT algorithm is the FAST FOURIER TRANSFORM (FFT), so called because it provides a rapid method of calculating DFTs. These normally work with windows with numbers of samples which are a power of 2, and hence are known as 64-, 128-, 256-, 512- or 1,024-point FFTs.

The effective time resolution (T_r) of the DFT is the number of points in the DFT multiplied by the time in seconds between successive samples; and the frequency resolution (F_r) is the reciprocal of T_r. If N is the number of points, and T the time in seconds between successive samples, then

$$T_r = N \times T$$

and $$F_r = 1/T_r.$$

Thus for a 512-point FFT (DFT), performed on a sample digitized at 10,000 samples per second, the width of the window will be 51.2 ms, which will also be the worst-case time resolution. The frequency resolution will be 19.5 Hz consisting of 256 equally spaced points. (The reason why the analysis contains only 256 points and not 512 is that only the points in the lower half of the transform contain information related to the frequency spectrum below half the sampling frequency, i.e. 0–5,000 Hz, the actual encoded frequency range of signal.) Figure 7.14.7 shows 512-point FFTs for the vowels [i:], [a:] and [ɔ:], with the Hamming windowed waveform samples on which the analysis was performed.

There are two major advantages of digital analysis of speech signals. The first is that once the signal has been digitally encoded and stored, it can be edited, processed, measured, manipulated and filed with far greater efficiency than is possible with analog instruments and an ordinary tape recorder. The second is that the analysis itself can be more easily varied to give optimum time- and frequency-resolution properties.

The FFT example in figure 7.14.7 shows a very narrow-band spectral analysis. If less frequency resolution is required in a spectral section, the simplest procedure is to reduce the number of points in the FFT. As the trading relationship between time and frequency resolution in the analysis applies to the DFT just as it does to the analog spectrograph, reducing the number of points will reduce the width of the analysis window. Time resolution will then become finer, which may or may not be an advantage. If the window becomes shorter than the pitch period of the waveform, the location of its centre will become more critical, and may markedly affect the spectral shape yielded by the analysis. It is also important to understand that as the number of points in the DFT is reduced, the interpolation on the frequency scale becomes coarser.

If all that is needed is a smoothed outline of the spectral energy, and time resolution is not critical, a CEPSTRALLY SMOOTHED narrow-band DFT may be more desirable. In this process, further spectrum analysis and processing are performed on the spectrum itself by treating it as though it were yet another signal waveform. This yields a smoothed spectral envelope undisturbed by voicing ripple, and preserves more frequency domain detail and interpolation than is possible by merely reducing the number of points of analysis. Figure 7.14.8 shows spectral sections for the vowel [i:], with (a) a 512-point FFT, (b) a 128-point FFT, and (c) a cepstrally smoothed 512-point FFT. All these FFTs were produced by a signal editing and processing package operating at 10,000 samples per second, and greater than 10-bit quantization.

DFT techniques are used to generate spectrograms by taking a large number of spectral sections side by side and overlapping them. This procedure of

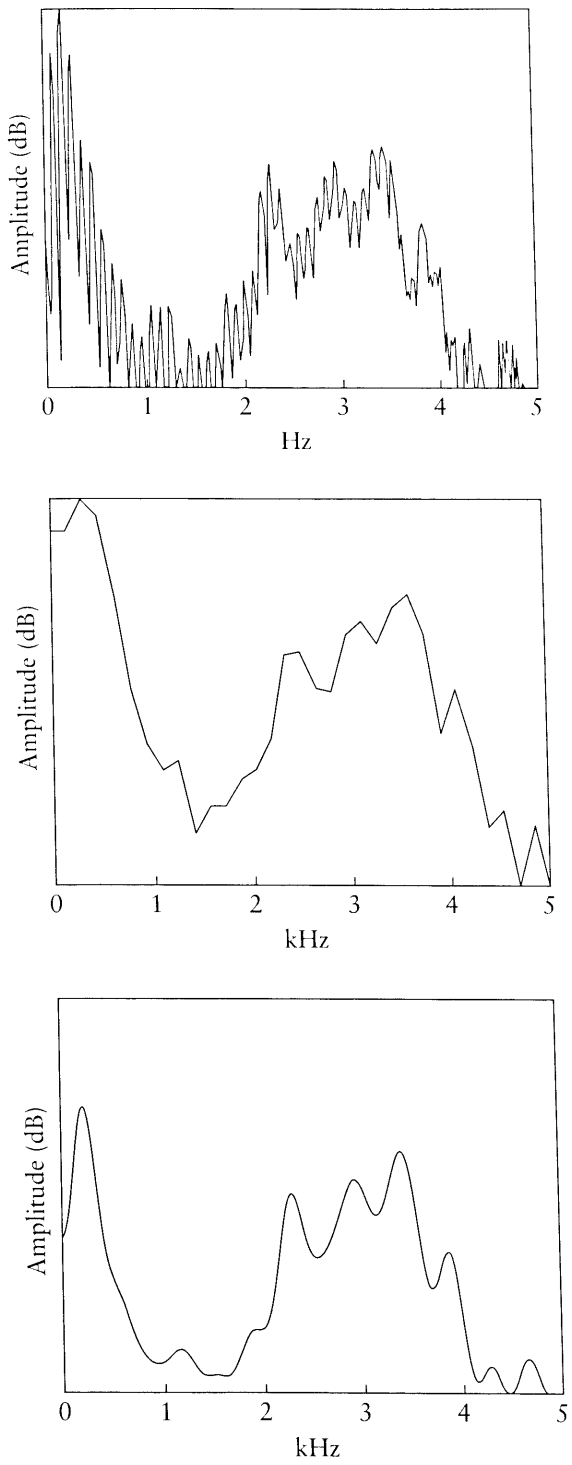


Figure 7.14.8 Spectral sections of the vowel [i:]: (a) 512-point FFT; (b) 128-point FFT; (c) cepstrally smoothed 512-point FFT

course differs from that of the traditional analog spectrograph, in which each sweep (or drum rotation) covers a small frequency range of the full duration of the speech being analysed; here, each analysis covers the whole frequency range. The case of the narrow-band spectrogram is simple enough: an FFT is taken every 5 or 10 ms along the waveform (giving reasonable overlap) and then displayed in the conventional way with frequency on the vertical axis and time on the horizontal axis. The amplitude of the spectral energy is given by intensity of blackness, or in some cases by colour (either in hard copy or on a visual display screen). The broad-band spectrogram presents more of a problem, because a simple reduction of the number of FFT points results in a frequency axis with rather poor interpolation. This is overcome by performing equivalent narrow-band FFTs of, say, 512 points, but actually using only the centre of the sample window for data and filling the outer parts of the window with zeros. The result provides the excellent frequency interpolation of the 512-point analysis, yet has the fine time resolution and broad frequency resolution of an analysis with many fewer points. Figure 7.14.9 shows examples of narrow- (25Hz) and broad-band (200Hz) DFT spectrograms of the vowels [iː], [aː] and [ɔː].

Although Fourier-based spectrographic analysis of speech is the most common technique for examining the properties of the speech spectrum, linear prediction coefficient analysis (LPC) has proved increasingly popular in recent

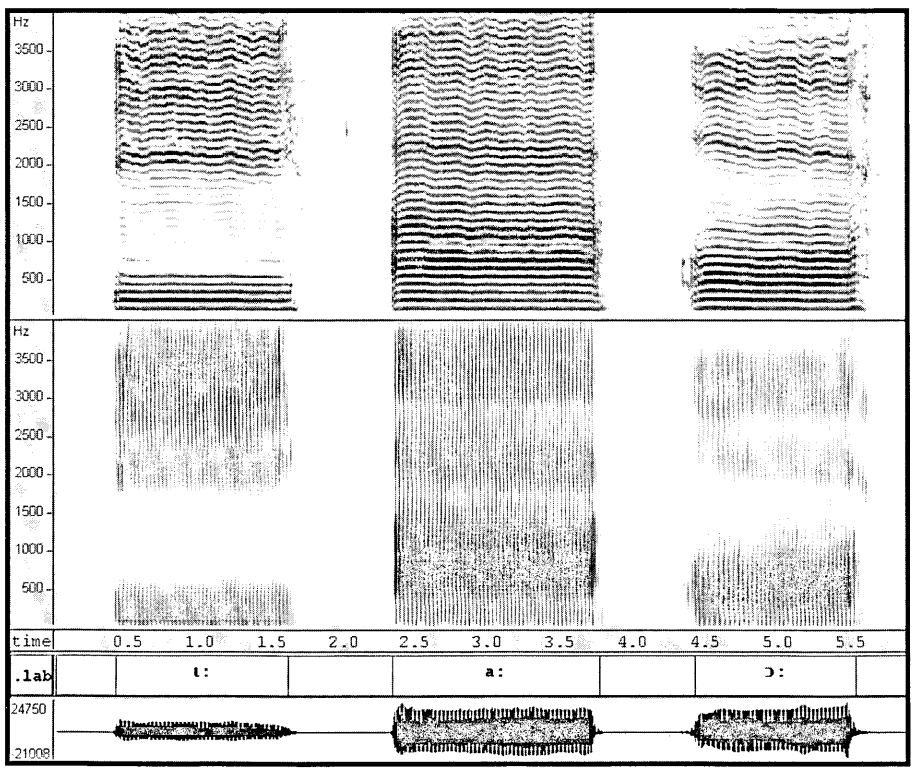


Figure 7.14.9 Narrow- (top) and broad- (bottom) band DFT spectrograms of the vowels [iː] [aː] [ɔː]

years. LPC analysis represents the speech signal in terms of a set of coefficients which aim to predict the signal from its past time domain values with minimum error. These coefficients may be used to produce a spectral representation of that signal. In essence this takes the form of a vocal tract filter frequency response, including the effects of the slope of the voice source spectrum and radiation, which would produce a time domain speech waveform the same as that of the waveform being analysed. The result is a pseudo-spectral section rather similar in appearance to a cepstrally smoothed DFT section, which can show the formant structure of the speech very clearly (figure 7.14.10).

Although very useful for the analysis of vowels (and some approximants) because of the clarity with which it can identify formant locations, LPC analysis must be treated with some caution. In its conventional form, it is based on resonances only in the vocal tract, and some sounds such as nasals and fricatives have more complex tract frequency response properties which are not properly accounted for by some forms of LPC analysis. There may be some occasions when the LPC analysis will generate a spectrum which does not correspond to a DFT-calculated spectrum, even though it is a computationally valid alternative tract frequency response for the input waveform.

Much more could be said about the details of spectral analysis, and readers wishing for more information should consult Fry (1979) or Pickett (1980) on the analog spectrograph, and Witten (1982) on digital signal analysis. Those with a mathematical background will find further material on speech signal analysis in Markel and Gray (1976), Wakita (1976), Rabiner and Schafer (1978) and O'Shaughnessy (1987).

To avoid complexity, we have concentrated on spectrographic analysis of simple static vowels. But one of the important properties of the spectrogram is its dynamic portrayal of the time-varying spectrum, and we include here (figure 7.14.11) a spectrogram of the phrase *human speech* segmented to show the parts of the spectrum which characterize its phonological structure (in so far as a simple serial segmentation is capable of doing this). Further information about the acoustic properties of speech sounds in the context of the spectral dynamics of the speech signal can be found in the following sections of this chapter.

The conventional spectrographic display is not the only means of providing amplitude information on a time-varying spectrum. An alternative sometimes used in speech research is a geometric projection of spectral slices. This can provide very useful detail over short periods of time, but is somewhat difficult to read for long stretches of speech. Rabiner and Schafer (1978, p. 314) and Lieberman and Blumstein (1988, p. 194) provide typical examples.

7.15 Acoustic properties of vowel quality

It has long been recognized that the auditory distinctions in individual vowel quality which enable us to give them phonologically distinct labels are predominantly determined by the frequency distributions of the first three formants (section 7.13 above). In the nineteenth century, researchers such as Willis and

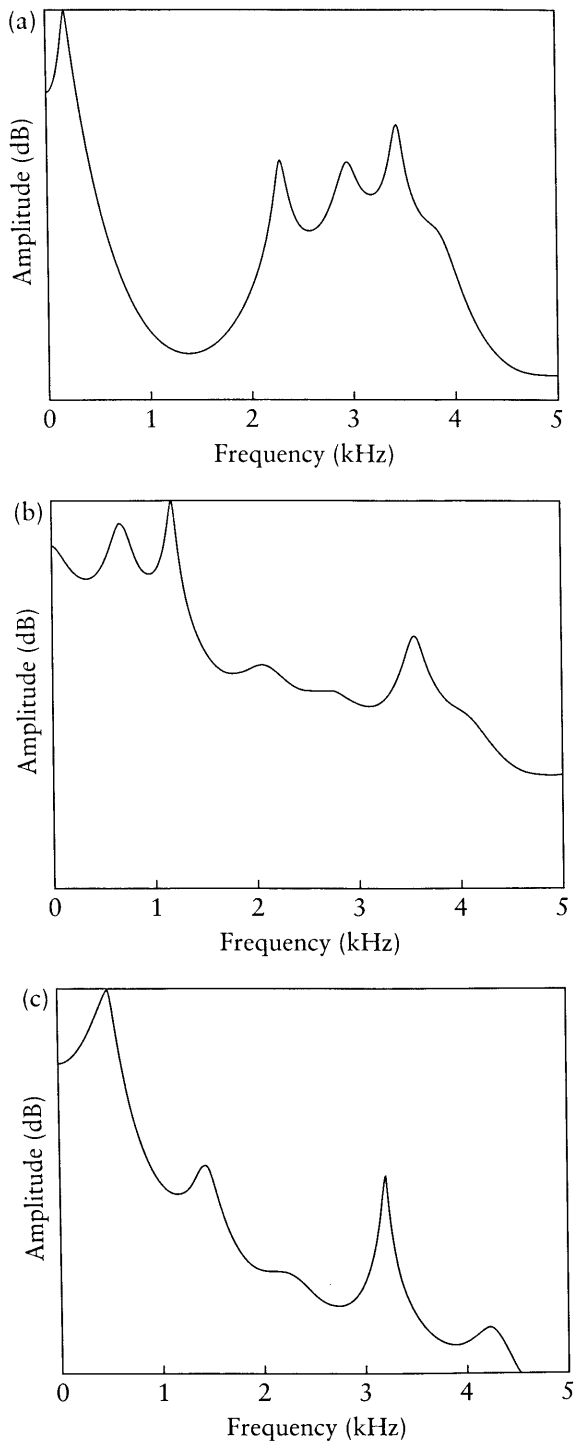


Figure 7.14.10 Linear prediction coefficient sections (P = 15) for the vowels (a) [i:]; (b) [a:]; (c) [ɔ:]

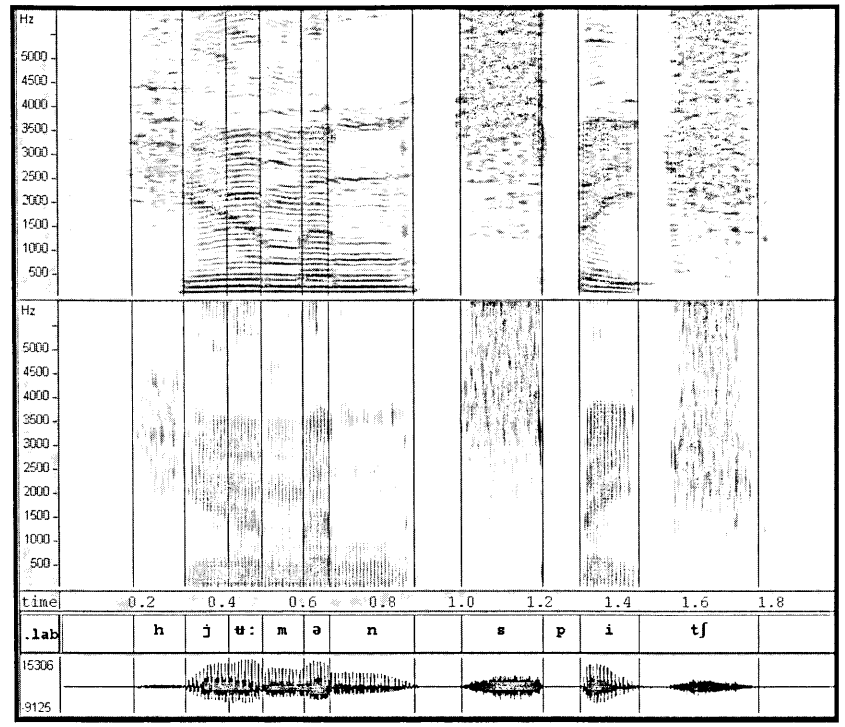


Figure 7.14.11 Segmented and labelled narrow- and wide-band spectrograms of the phrase *human speech*

Helmholtz recognized the role of tract resonance and formant structure – although the term 'formant' was not then current – as did Miller, Stumpf, Paget and others in the first half of the twentieth century. Stewart (1922) succeeded in demonstrating the validity of the source and filter model for speech acoustics with a primitive electrical analog of the vocal tract. This used a buzzer as a periodic electrical impulse generator for a voice (larynx) source, and simple resonant electrical circuits (each with a single resonance peak) for the tract filter, to produce identifiable synthetic vowel sounds. Stewart's success pointed to the role of the formant in defining vowel quality. Delattre et al. (1952) and Miller (1953) confirmed this with extensive perceptual studies using much more sophisticated vowel synthesis techniques, and their investigations have been followed by many others, using methods involving both production and perception.

Vowel quality is, however, not just a matter of static formant values. Miller (1953), Lindblom and Studdert-Kennedy (1967), Millar and Ainsworth (1972) and others have shown that we also depend on the overall dynamic pattern of syllable structure to supplement formant information in establishing the phonological identity of vowel sounds. Relevant factors include F_0 and formant movements next to consonants.

Before the spectrograph became available in the 1940s, the acoustic analysis of speech was so laborious and so restricted by equipment limitations that formant

structure and its relationship to auditory qualities of speech sounds had been very little explored in natural speech. From 1946 the spectrograph brought a dramatic change. It was soon recognized that if the first two formant frequencies of vowels were plotted against each other on axes with appropriate scaling and direction, the result was a vowel map which bore a remarkable resemblance to that of a traditional auditory map of vowel quality (section 2.7 above). The earliest published account of this mapping relationship seems to be that of Essner (1947), although work by Joos (1948) is probably more widely known. Ladefoged (1967, ch. 2) and Catford (1981) provide some details of this history.

Given the problems of providing a reliable auditory description of vowel quality (section 2.7 above), the availability of an ostensibly objective technique of acoustic analysis, free from the bias of the human observer, was an important step in phonetic and phonological description. The basic technique for obtaining such an acoustic map of vowel quality is to plot F_2 on the horizontal axis, with values increasing from right to left, and F_1 on the vertical axis, with values increasing from top to bottom. In addition, the frequency scale of F_1 must be at least double that of F_2 to ensure that the resulting map has an appropriate aspect ratio. Figure 7.15.1 shows plots of the so-called pure (monophthongal) vowels of Australian and New Zealand English, using formant data from Bernard (1985).

Figure 7.15.1 reveals the vowels in a standard auditory arrangement, distributed from front to back horizontally and from high to low vertically. In this diagram, vowel fronting is represented as proportional to the value of F_2, while vowel height is inversely proportional to the value of F_1. Thus back vowels are seen to have lower F_2 values, and high vowels are seen to have lower F_1 values. The diagram incidentally shows some characteristics of Australian English, in particular the acoustic similarity of [a:] (as in *calm* or *heart*) and /ʌ/ (as in *come* or *hut*): this is part of the evidence for the assertion that these two vowels are distinguished purely by duration in Australian English (Bernard 1967), although in certain transcription systems for Australian English, the vowel of *come* or *hut* is represented as [ɐ]. It also shows the very substantial front vowel shift that has occurred in New Zealand English.

Although plots of this kind are extremely useful, they do not match a cardinal vowel diagram quite as closely as might be hoped. This was recognized even in the 1940s and Joos (1948) used logarithmic scales for the formant axes, while Delattre (1951) commented at length on the problem of relating articulatory and auditory vowel descriptions to acoustic descriptions, in the context of logarithmic plots of the kind used by Joos. In his extensive study of vowel quality Ladefoged (1967) converts formant data to the mel or pitch scale in an attempt to move closer to a perceptually oriented and hence auditorily realistic acoustic map. Ladefoged notes that, even with this transformation, a two-formant plot does not adequately display the auditory differences between vowels at the extreme high and back areas of the vowel space; elsewhere it is reasonably satisfactory.

Lindau (1978) and Ladefoged (2006, esp. p. 212), on the strength of quite sophisticated statistical analyses, replace the F_2 dimension by the difference between F_2 and F_1 (i.e. $F_2 - F_1$). The claim is that this difference is more directly

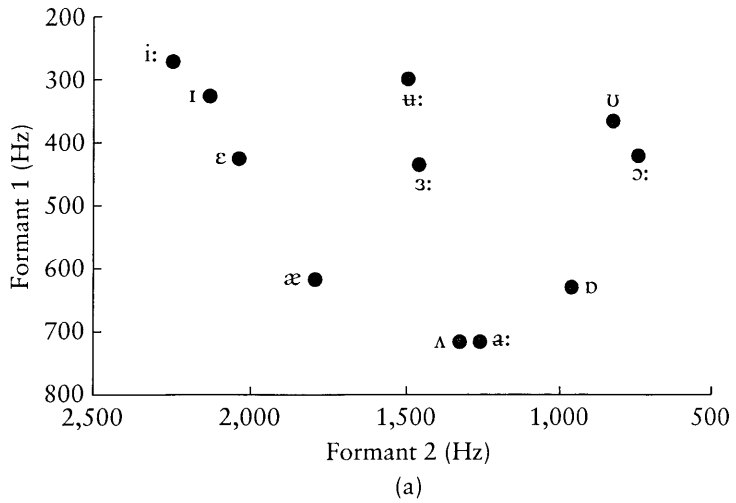


Figure 7.15.1 Acoustic mapping of vowels to correspond to auditory map: (a) Australian English; (b) New Zealand English

related to the auditory concept of 'frontness' or 'backness' than F_2 alone. They also retain a pitch (mel) scaling of the frequency values on both axes. Figure 7.15.2 shows the vowels of figure 7.15.1(a) replotted in this way, and it can be seen that the front-to-back scaling is now a little more like an auditory plot, particularly for the back vowels.

Catford (1981) has proposed another solution to the problem. He warps the frequency scaling and angular axis relationships of the formant plot to fit a traditional cardinal vowel chart (in which all the vowels are either unrounded or rounded, thus removing one variable from the mapping problem). Although this is a very interesting idea, and provides quite a good fit between Catford's acoustic and auditory data, it rests on the assumption that a cardinal vowel

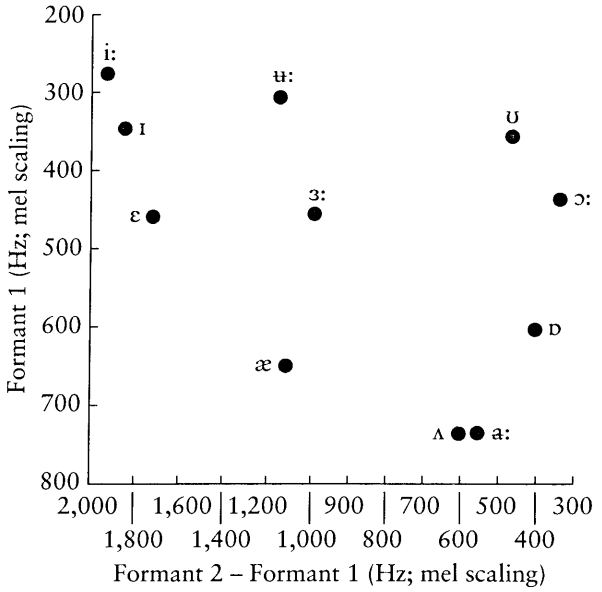


Figure 7.15.2 Alternative form of vowel mapping

diagram is a consistent and (presumably) reasonably linear map of auditory quality which can be taken as a standard. Given an inevitable component of arbitrariness in the choice of the original cardinal vowels, and the somewhat apostolic nature of their subsequent preservation and use, a scientific defence of this assumption is likely to be difficult. The separation of vowels into rounded and unrounded categories has some merit acoustically, but is not without drawbacks, as it is often of phonological interest to map all the vowels of a language in a single plot which portrays their systemic relationships. The reality is that most languages exhibit a mix of unrounded and rounded vowel sounds, and that lip position is often intermediate between fully rounded and fully unrounded.

Two other strategies that have been adopted over the years involve converting the first two formant frequencies to perceptually sensitive scales, either weighted Bark (critical bands; see section 8.3 below and Zwicker 1962) or ERB (equal rectangular bandwidth) scales (Moore 2003). By contrast, a recent neuroimaging study of 12 speakers showed reasonably good evidence for cortical vowel spatial representations corresponding to traditional F_1–F_2 frequencies for German (Obleser et al. 2003). However, irrrespective of whether vowels are plotted using conventional acoustic frequency or alternative auditory measures, Ladefoged's (1967) observation still holds, in that any kind of two-dimensional vowel map will encounter difficulties, particularly where vowels having similar height and fronting values but different lip positions are involved. The underlying acoustic problem for the two-formant plot is that it does not account for F_3, which also contributes to vowel quality. For example, in moving from cardinal vowel 5 to cardinal vowel 6, there may be a fall in F_3 of several hundred Hz, but this is not shown on a normal two-formant plot. Vowel qualities between 5 and 6 may well require more than the first two formants

to provide an adequate mapping of their real auditory relationships. Fant (1968, 1973) has addressed this problem, partly from the perspective of speech synthesis, and proposed the use of a weighted F_2, which would take account of F_3 as F_2 increases in frequency; but this approach too has its limitations and is not commonly used. It is also important to note that none of these schemes takes account of the normalization of vocal tract length. Auditory maps of vowel quality appear to generalize across a number of speakers and thus imply some such normalization; but it is not clear how data from diverse voices (males, females, children) are actually to be reconciled.

For practical purposes, it is often most useful to accept the basic formant plot as it stands, and to add to it an F_3 axis as an extension of the F_1 axis, thus making a dual plot which will show the relative importance of the contributions of all three formants in any set of distinctions. As will be seen later, this strategy is also of value in consonant mapping. Figure 7.15.3 shows the vowel data of figure 7.15.1 replotted in this fashion, and it can be seen that F_3 does indeed provide useful additional information (note in particular the central and back vowels). A true three-dimensional representation of the vowel space based on this principle is described by Broad and Wakita (1977).

We return to the dynamic aspects of the continuous speech spectrum. The formant data used in figure 7.15.1 are taken from the targets of nominally pure vowels, that is, vowels having a single stable articulatory and auditory target

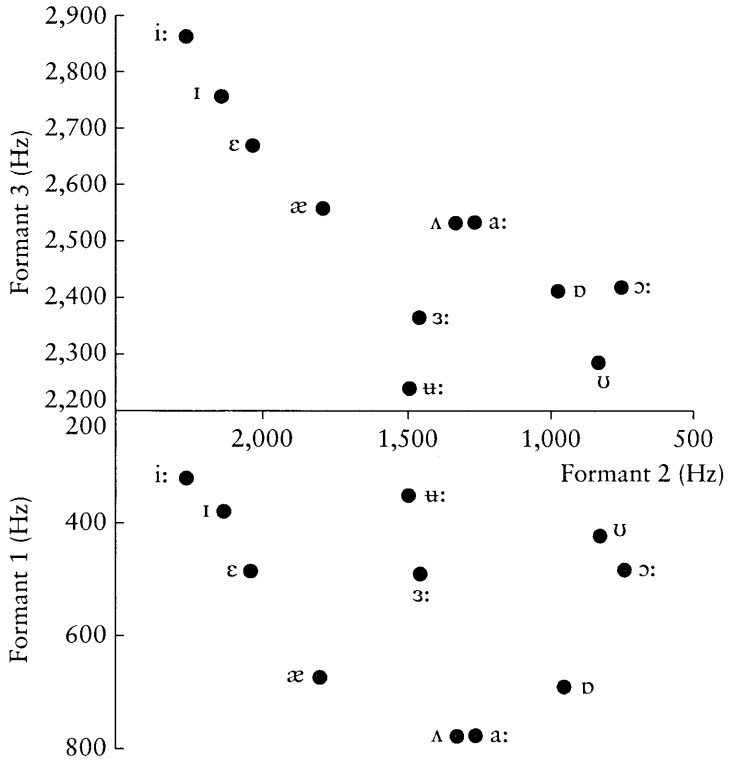


Figure 7.15.3 Vowel mapping using three formants

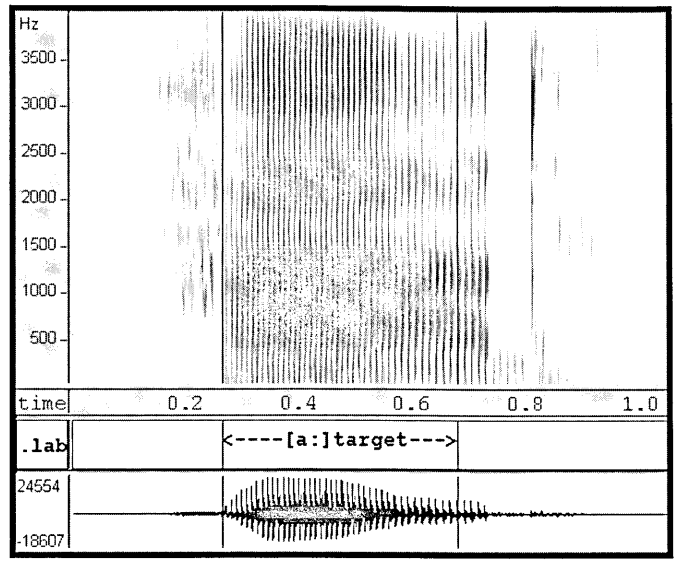


Figure 7.15.4 Acoustic vowel target in the word *hard*

value (section 2.7 above). The acoustic target of a vowel can be recognized on a spectrogram by the stable spectral structure in a syllable nucleus – although this is not always an entirely simple matter. Figure 7.15.4 provides an example that is relatively straightforward, using the word *hard* [ha:d] to illustrate the principle. The word [ha:d] happens to have a long, low vowel with a stable target of substantial duration. The target is identified by the spectral sequence in which the formants are parallel to the time axis and thus not changing. It is usual to take the target formant values from about the centre of this sequence. (In this spectrogram there are other spectral changes relating to the consonants in the peak and coda, but these will be ignored for the moment.) Making accurate target estimations is less straightforward when the syllable peak is of short duration and there is insufficient time for a stable target to be established by the articulators, which usually results in target undershoot; or it may be that the effects of preceding and following consonants hinder a stable target. Figure 7.15.5 shows broad-band spectrograms of two words, *kit* and *bag*, which illustrate some of the problems in identifying vowel targets.

In the spectrogram of *kit* it can be seen that the formants move towards a peripheral vowel position, but never stabilize there. The best estimate of the vowel target is taken as shown, at the (acoustically) most peripheral position. In *bag*, the consonant influence again prevents a stable target, and the area shown is in the middle region of the syllable nucleus, where the effects of the two consonants merge into a vowel target region. Comparable problems of analysing vowel targets within the dynamic spectra of syllables have been extensively discussed by Stevens et al. (1966).

Acoustic mapping of vowels need not be confined to static information in the spectrum. Glides and diphthongal vocalic movements can be shown very clearly using a formant plot. Figure 7.15.6 shows (a) spectrograms of the

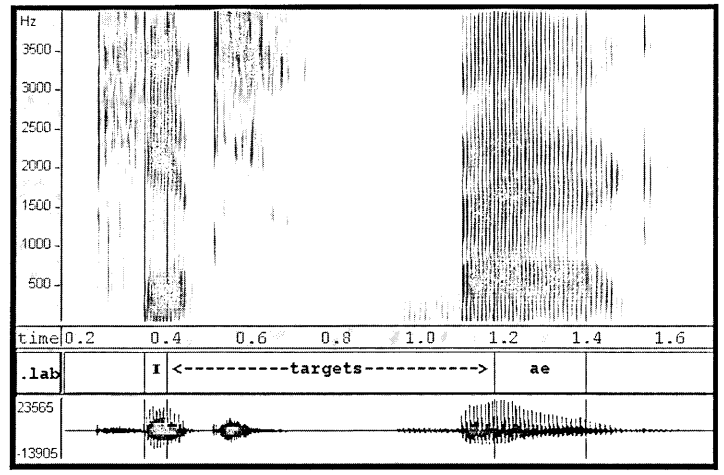


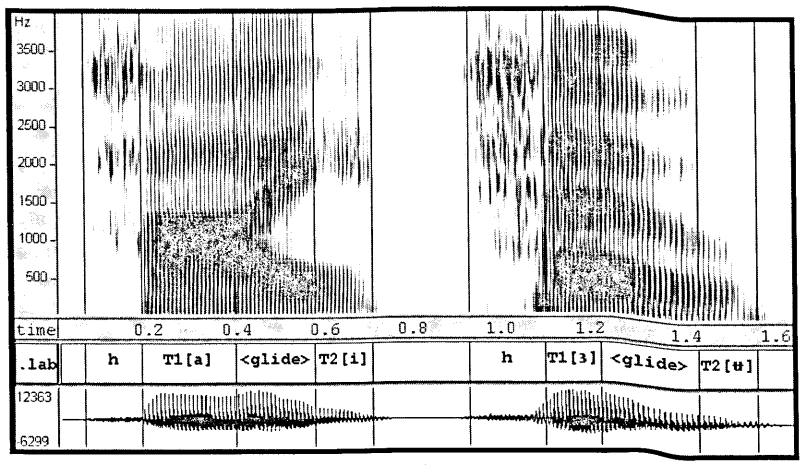
Figure 7.15.5 Acoustic vowel targets in the words *kit* and *bag*

diphthongs in the words *high* and *hoe* articulated slowly in Australian English, and (b) two-formant plots of their targets. In (a), the two spectral targets of the diphthongs can be seen, as can the smoothly changing spectrum between targets which characterizes these diphthongs both articulatorily and acoustically. In (b), these targets have been plotted and lines drawn between them to show the direction of their spectral movements. This suggests a useful general correspondence with auditory impressions of these sounds.

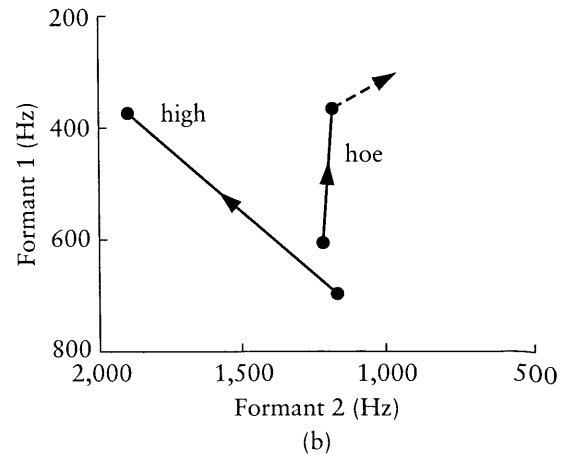
The classic study of vowel mapping was done quite soon after the appearance of the speech spectrograph, by Peterson and Barney (1952), who analysed a set of ten monophthongal vowels from each of 76 men, women and children. The two-formant plots of these vowels showed an appreciable amount of overlap between adjacent vowels, suggesting that the absolute acoustic discrimination between some vowels was not particularly good.

The explanation for this is twofold. Firstly, speakers do vary in the phonetic (acoustic) realizations of their vowels, but they normally maintain their systemic contrasts. Thus two speakers may vary substantially in the shape of their vowel space, and the formant values for, say, [ɛ] in one speaker may be close to those for [æ] in another; but both speakers will make adequate acoustic distinction between [ɛ] and [æ] within their own vowel system. Secondly, speakers differ substantially in the length of their vocal tract (section 7.11 above), and these variations have an inherent influence on the patterns of formant values in ways which speakers cannot really control. There is therefore no absolute acoustic discrimination among adjacent or nearby vowels that applies to a number of different speakers. As language users we have little difficulty in coping with this, as we are able, with only a very small speech sample from an individual, to normalize to the vowel system of that speaker. In other words we are accustomed to adjusting the map from speaker to speaker.

Indeed our ability to cope with different systems is not confined within a relatively homogeneous group, but may extend to quite radically different regional varieties or dialects. Despite many jokes about misunderstandings



(a)



(b)

Figure 7.15.6 Diphthongs in *high* and *hoe*: (a) spectrograms; (b) two-formant plots of targets

within the English-speaking world, the number of genuine confusions is relatively small. This is all the more impressive given that there are substantial discrepancies and overlaps among the regional varieties of English: the vowel in a New Zealander's *catch*, for example, may be close to the vowel of a Londoner's *ketch*; when an Australian says *clerk*, many North Americans may hear the vowel quality as equivalent to their own *clock*. But the point is of course that speakers maintain their own patterns of distinctiveness: a New Zealander distinguishes *catch* from *ketch*, and so does a Londoner, but they do so in different ways, with a different contrast of vowel quality. Thus normalization reflects the general principle that phonological distinctiveness is a matter of relative contrast within a system rather than a matter of absolute or universal phonetic values (cf. section 4.8 above). Ladefoged and Broadbent (1957) demonstrated the capacity for normalization with a rather ingenious experiment involving synthesized speech: in effect, they showed that the same sound

could be perceived as different vowels, depending on the listener's normalization triggered by what was uttered immediately before the sound in question. Taking an anecdotal instance of the principle, we may say that one and the same utterance may be heard as *clock* if listeners are expecting North American speech but as *clerk* if they are expecting Australian speech. It is thus inadvisable to combine the formant data from a variety of speakers, as Peterson and Barney did.

A number of algorithms have been proposed for performing mathematical normalizations of formant data to remove some of the sources of variance. One of the earliest is Fant's (1966) scaling of formant data in relation to vocal tract size. A number of other techniques have been developed since, of which Gerstman (1968) and Nearey (1977) are well-known examples. An extensive appraisal and review of several normalization procedures can be found in Disner (1980) and Johnson (2005). Disner observes that mere reduction of data variance does not of itself have any value if it does violence to the vowel quality relationships within or between languages.

Much of the research into the acoustic aspects of vowels has focused on the use of the linear time–frequency space properties of vowels, as provided fairly directly by conventional spectrographic analysis. In the past few years, there has been increasing interest in a more listener-based approach to analysis procedures and acoustic representations of vowel quality. Specifically, it is well known that the human auditory system has rapidly decreasing frequency resolution above 1 kHz, and that as a consequence our ability to discriminate individual peaks of formant energy becomes poorer at these higher frequencies. It is likely that current research will continue to generate standard techniques for both mapping and normalizing vowel data, based on transformation of the data into a spectral representation which models that in the human auditory system itself (see Chistovich et al. 1979, Bladon and Lindblom 1981, and Syrdal and Gopal 1986). Such procedures aim to focus our attention on the most perceptually relevant aspects of the data, but they are as yet controversial and incompletely understood.

To end this section, we offer an example of spectrographic analysis and mapping, dealing with the back vowels of Australian English and their context-sensitivity to a following velarized lateral. Figure 7.15.7 is a comparative two-

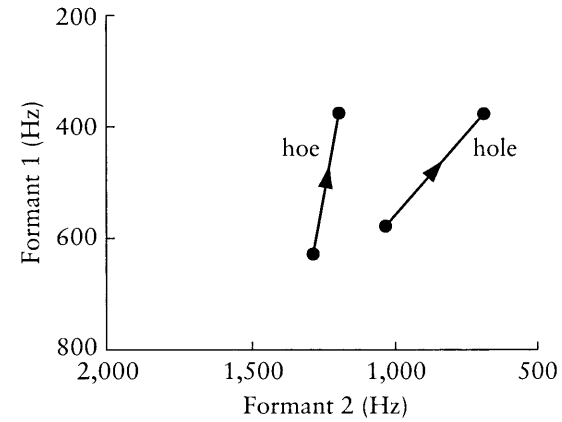


Figure 7.15.7 Two-formant plot showing effects of [l] on preceding [oʊ]
Adapted from: Bernard 1985, p. 328.

formant plot of data from Bernard (1985): it demonstrates the strong retraction effect of the lateral on the entire target and glide structure of the diphthong in the word *hole* compared with the word *hoe*.

7.16 The vocal tract filter in consonant production

The articulation of consonants is generally characterized by the constriction and partitioning of the oral–pharyngeal vocal tract, with the addition, in the case of nasal or nasalized consonants, of coupling of the nasal cavity system (sections 2.9 and 6.8 above). Approximant consonants are in many ways comparable to vowels, but other consonants have a more complex tract resonance system (although mention of nasalized vowels in section 7.13 above has anticipated some of the complexity). In this section we will consider static aspects of vocal tract filter properties in consonant production; in section 7.17 we will then relate these to the dynamic spectral patterns which encode the phonological features of consonants within the syllable.

The most vowel-like tract filter properties are found in approximants, which fall into two groups. The first, needing no further treatment here, consists of certain central approximants which have acoustic properties little different from very high vowels; they are classified as consonants more by their functional role in syllabic structure than by their acoustic properties. The most common examples are [w] and [j] (sections 2.12 and 3.11 above). The second group consists of those which partition or constrict the tract more radically than vocalic sounds, resulting in demonstrably different resonance properties. Common examples of these are laterals such as [l] and central approximants such as [ɹ].

Laterals divide the oral cavity into two around the location of the tongue occlusion. The oral cavity remains undivided both in front of and behind this point of occlusion. The analysis by Fant (1960) of [l] suggests that this complex divided resonator system yields a spectrum with low values for both F_1 and F_2, and a marked dip of energy in the spectrum in the region surrounding 2,000 Hz, caused by an antiresonance in the tract filter. Figure 7.16.1(a) shows a spectral section for [l] taken from the nonsense syllable *lah*.

An example of a central approximant is English [ɹ], as in *rag* or *ruck* (as pronounced by, say, Londoners or Australians rather than by Scottish speakers, who may use a flap or a trill). Central approximants have resonance patterns that deviate markedly from vocalic sounds. The principal feature of these sounds is a very low F_3, resulting from resonance associated with the anterior cavity formed by tongue tip and blade constriction. Figure 7.16.1(b) gives an example of a spectral section taken from [ɹ] in the nonsense syllable *rah*. The difference in F_3 value is often an important means of discriminating [ɹ] from other approximants such as [w] and [l].

The constrictions of fricative articulation produce tract resonance properties which differ even more from those of vowels. An important difference from

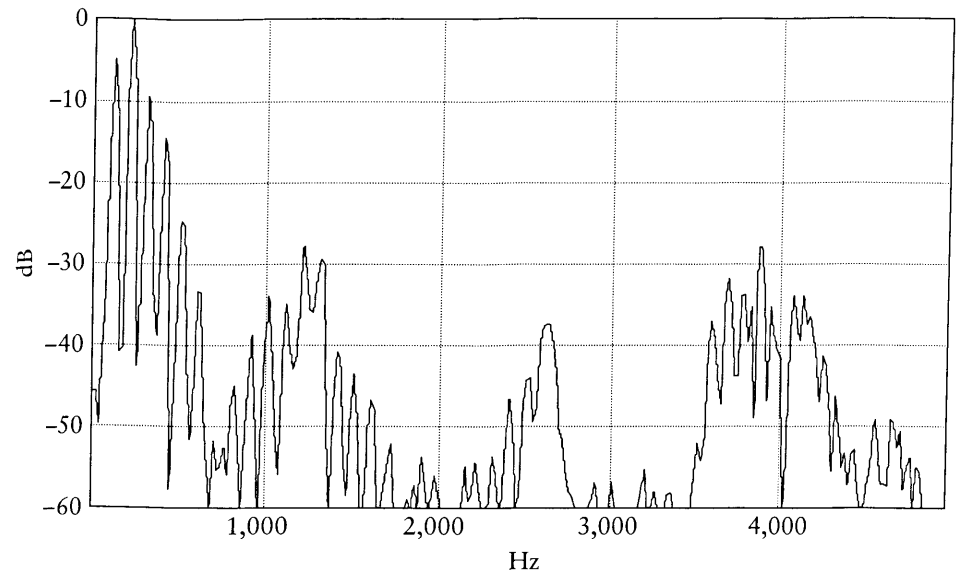


Figure 7.16.1 Spectral sections (512-point FFT) of the approximant consonants: (a) [l] in *lah*; (b) [r] in *rah*

other sounds is that the excitation source of a fricative is not necessarily at the glottis. The vocal tract cavity is effectively divided into two parts at the point of fricative constriction. Alveolar and postalveolar fricatives provide typical examples of this form of resonator system: according to the analysis of Heinz

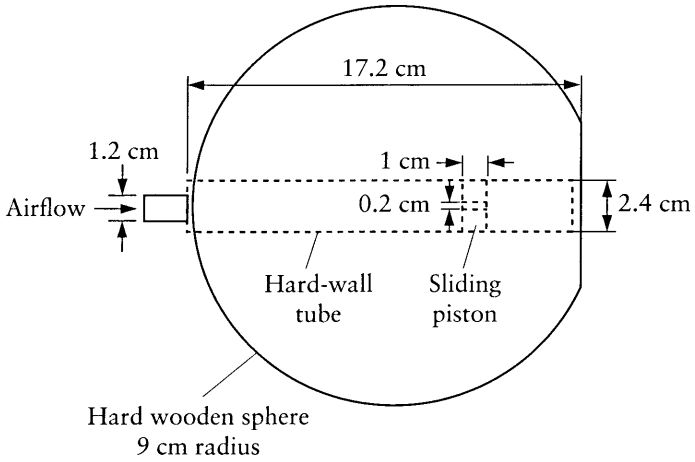
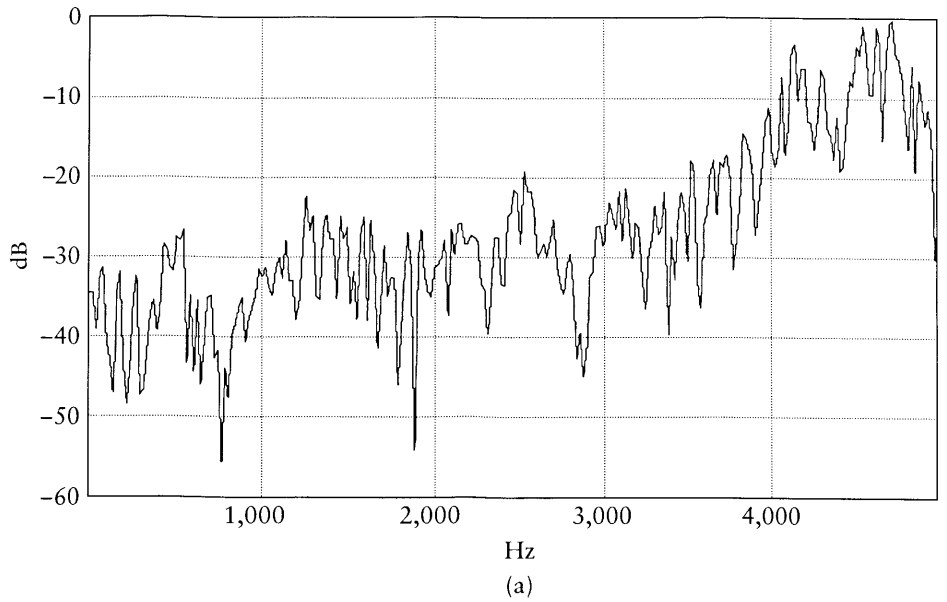


Figure 7.16.2 Model of fricative resonator system
Adapted from: Heinz 1958, p. 147.

and Stevens (1961), the cavity anterior to the constriction acts as a short closed pipe (quarter wave) resonator, coupled to the frication constriction which acts as an open pipe (half wave) resonator. Figure 7.16.2 shows a model of this system. The entire fricative source and filter system is quite short, typically less than 4 cm long for places of articulation anterior to the palate; and the large oral–pharyngeal cavity system behind the constriction is largely decoupled from the source and filter system. In voiceless sounds, the lowest resonance will also be heavily damped by the open glottis. Overall, when the effects of constriction antiresonance are taken into account, there is little resonance effect in voiceless fricatives below that resulting from the anterior system. The resultant fricative spectra generally exhibit a band of high-frequency, high-intensity energy, and very rapid energy attenuation at frequencies below those that are due to the anterior cavity resonance.

Although the cavity resonance effects in fricatives are more complex than those in vowels, the resultant high-frequency formant energy shows continuity with its associated syllable peak structure, as Heinz and Stevens (1961) point out. This reflects the principle of resonance continuity in the vocal tract, whatever the dynamic changes in its geometry during articulation. Where there is a marked step in fricative energy amplitude in the spectrum, the frequency region at which it occurs provides a strong static cue to the place of articulation of the fricative (see Strevens 1960, Heinz and Stevens 1961, Clark et al. 1982, and Karjalainen 1987). Figure 7.16.3(a) gives a spectrum for [s] taken from the nonsense syllable *sah*. The shape of the spectrum is also influenced by the effect of the upper front teeth, which deflect the fricative airstream, as Catford (1977) has shown.

What we have said about fricatives so far applies most clearly to fricatives with constrictions posterior to the teeth, because of the stronger spectrum-shaping influence of the anterior resonator on the frication noise source. It is



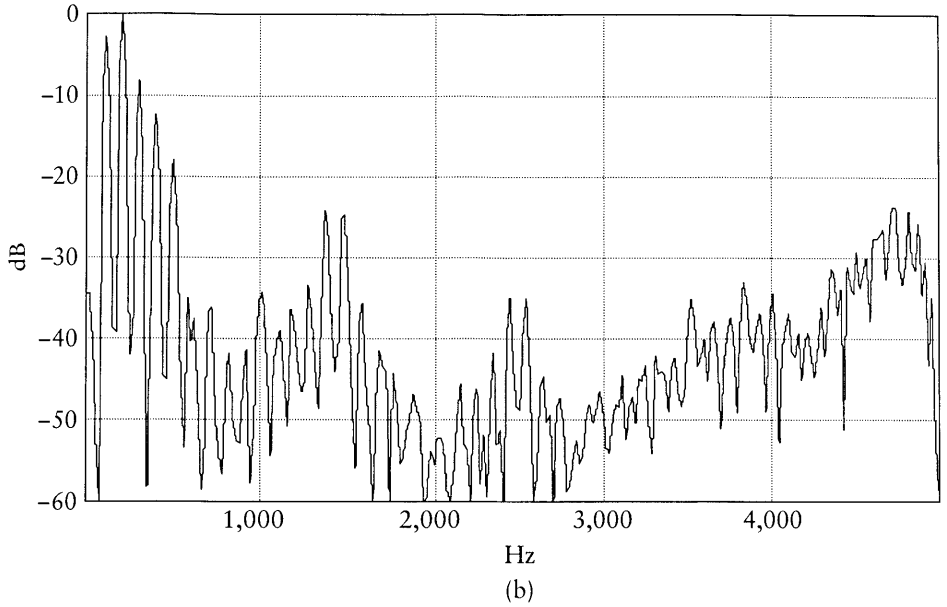


Figure 7.16.3 Spectral sections of fricative consonants: (a) [s] in *sah*; (b) [z] in *zah*

less easy to characterize the static spectra of dental and labial fricatives, whose resonant cavity effects are weaker in shaping the output of their frication source spectrum. The property of resonance continuity is of general importance to the perceptual integration of syllables, and will be mentioned again in the next section.

Voiced fricatives are more complex, because there are now two excitation sources, the phonating larynx and the frication constriction. As the glottis is not open, the lowest resonance will be less damped. Furthermore, phonation produces maximum energy at low frequencies, and predominantly periodic energy will be observed in this region, with characteristic fricative noise in the high frequencies now modulated by the phonation. Figure 7.16.3(b) gives a spectrum for [z] taken from the nonsense syllable *zah*.

Nasal consonants involve complete occlusion of the oral cavity, which is coupled to the nasal cavity as a side branch resonator. The nasal resonant cavity system itself cannot be systematically varied by the processes of articulation – although there are individual variations in the structure and geometry of the cavity itself, and various physiological effects (see sections 6.8 and 7.13 above). According to Fant's data (1960), the nasal passages in a male speaker form a resonator system about 12 cm long which couples into the oral–pharyngeal system some 7 cm from the glottis. The oral cavity thus forms a closed resonating chamber with a length determined by the place of nasal articulation. Figure 7.16.4 shows a simple model of the system.

The static spectral properties of this complex resonator system are a set of relatively stable nasal tract formants with generally greater damping than those of the open oral tract. These formants are said to occur in the regions of 250 Hz, 1,000 Hz, 2,000 Hz and 3,000 Hz. (See Fant 1960, Minifie 1973, Pickett 1980, O'Shaughnessy 1987, and Lieberman and Blumstein 1988 for reviews of nasal formant characteristics.) The formants are not always clearly visible in standard displays and may be generally or selectively weakened by the coupling of the oral cavity resonator, which contributes both resonance and antiresonance effects to the spectrum. The overall result is a nasal consonant spectrum with a broad peak of low-frequency energy and rather weaker upper formant energy which provides quite strong spectral cues to the nasal manner of articulation, but rather weaker cues to the place of articulation. The point of articulation is also cued by the spectral dynamics of the syllable in which the

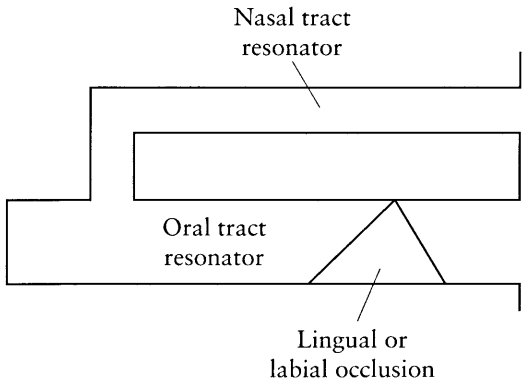
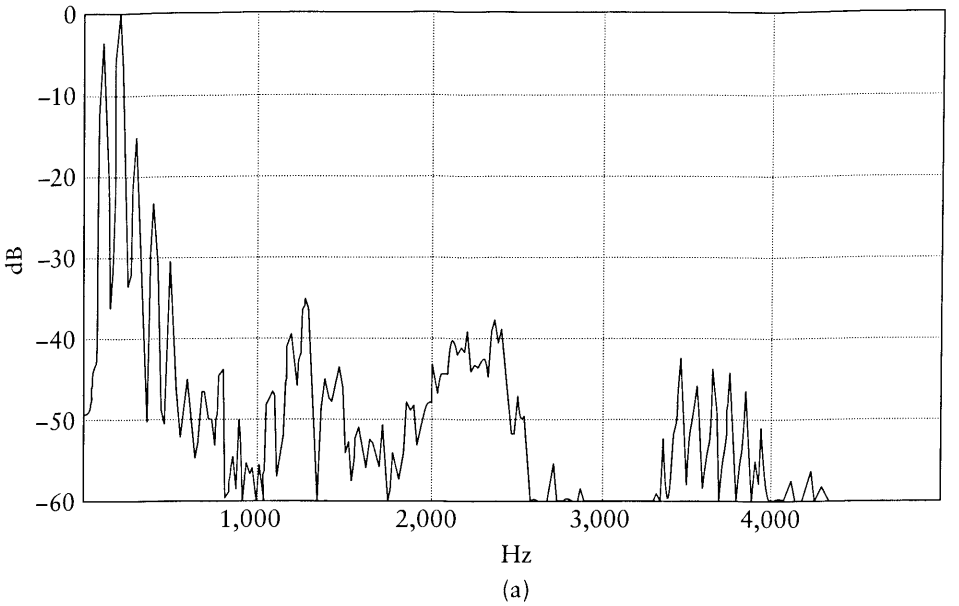


Figure 7.16.4 Model of nasal consonant resonator system



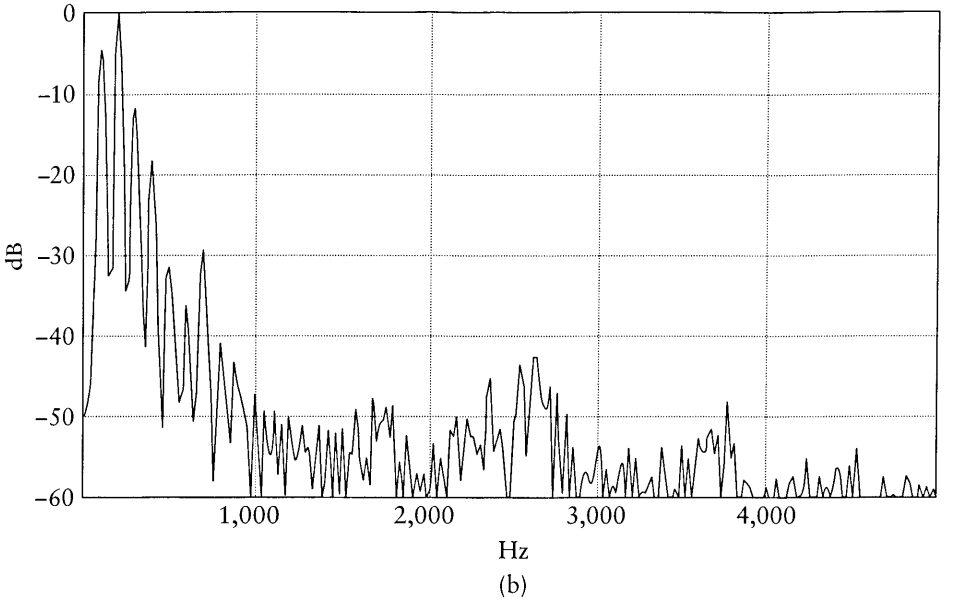


Figure 7.16.5 Spectral sections of nasal consonants: (a) [m] in *mah*; (b) [n] in *nah*

nasal consonant occurs. Figure 7.16.5 gives (a) the spectrum of [m] in the nonsense syllable *mah*, and (b) the spectrum of [n] in *nah*.

Stop consonants, by the nature of the articulation, do not have the same kind of stable constriction phase as approximants, fricatives and nasals. They are therefore characterized not so much by the typical spectrum of a 'steady

state' as by the dynamic spectral properties of the formation and release of the oral occlusion (discussed further in section 7.17 below). There is in a sense a stable state for stops, namely the occlusion phase itself, but the vocal tract is of course very strongly damped in this phase. In voiced stops, where there is some airflow to generate voicing during occlusion, a broad peak of low-frequency energy is seen as a 'voice bar' in spectrograms. Far more importance attaches to the release burst of a stop, which is the result of the momentary frication between the articulators as they part at the release of the occlusion. For a very brief period (typically less than 20 ms), the vocal tract is effectively producing a fricative. As with fricatives, occlusions anterior to the teeth have the least clearly defined spectra, with the weight of energy at the low end of the spectrum. It has been suggested that this is due, in part at least, to some contribution from the large anterior tract. Alveolar stops, and stops at locations posterior to this, produce noise spectra with energy distributions predominantly influenced by the length of the anterior cavity at the moment of release. Thus, not surprisingly, we find that the release burst of an alveolar stop has a spectrum comparable to that of the fricative [s], with the major energy occurring above 3 kHz. The velar stop burst has major energy distribution in the mid frequency range 1.5–2.5 kHz. Figure 7.16.6 shows the spectra of release bursts of the initial stops in *bah*, *dah* and *gah*.

The spectral properties of release bursts are significant cues to the place of articulation of stops, and have been described in detail by Halle et al. (1957) and Fischer-Jørgensen (1954), and by Blumstein and Stevens (1979) and Kewley-Port (1983), who suggest that their spectral properties are relatively uninfluenced by context. Perceptual studies by Blumstein and Stevens (1980) have tended to confirm these conclusions.

The differences between voiced and voiceless stops require reference to dynamic characteristics and the coordination of glottal and supraglottal articulatory activity (section 7.17 immediately below).

7.17 The acoustic properties of consonants in syllables

In concentrating so far on static aspects of acoustic representations, we have tried not to lose sight of the dynamic character of speech. While it is important to understand the properties of 'steady state' displays, it is characteristic of speech that the articulatory organs are constantly moving, often anticipating their next movement or adjusting to the simultaneous or partly overlapping activities of other articulators. As a consequence, the acoustic output is, by and large, not a series of steady states but a continuously varying signal. Oversimplifying somewhat, we can say that human hearing is attentive to changes in the signal, as much as to the nature of the signal at any point; and what we hear at any point may tell us as much about what has just happened or what is going to happen next as about what is actually happening at that point.

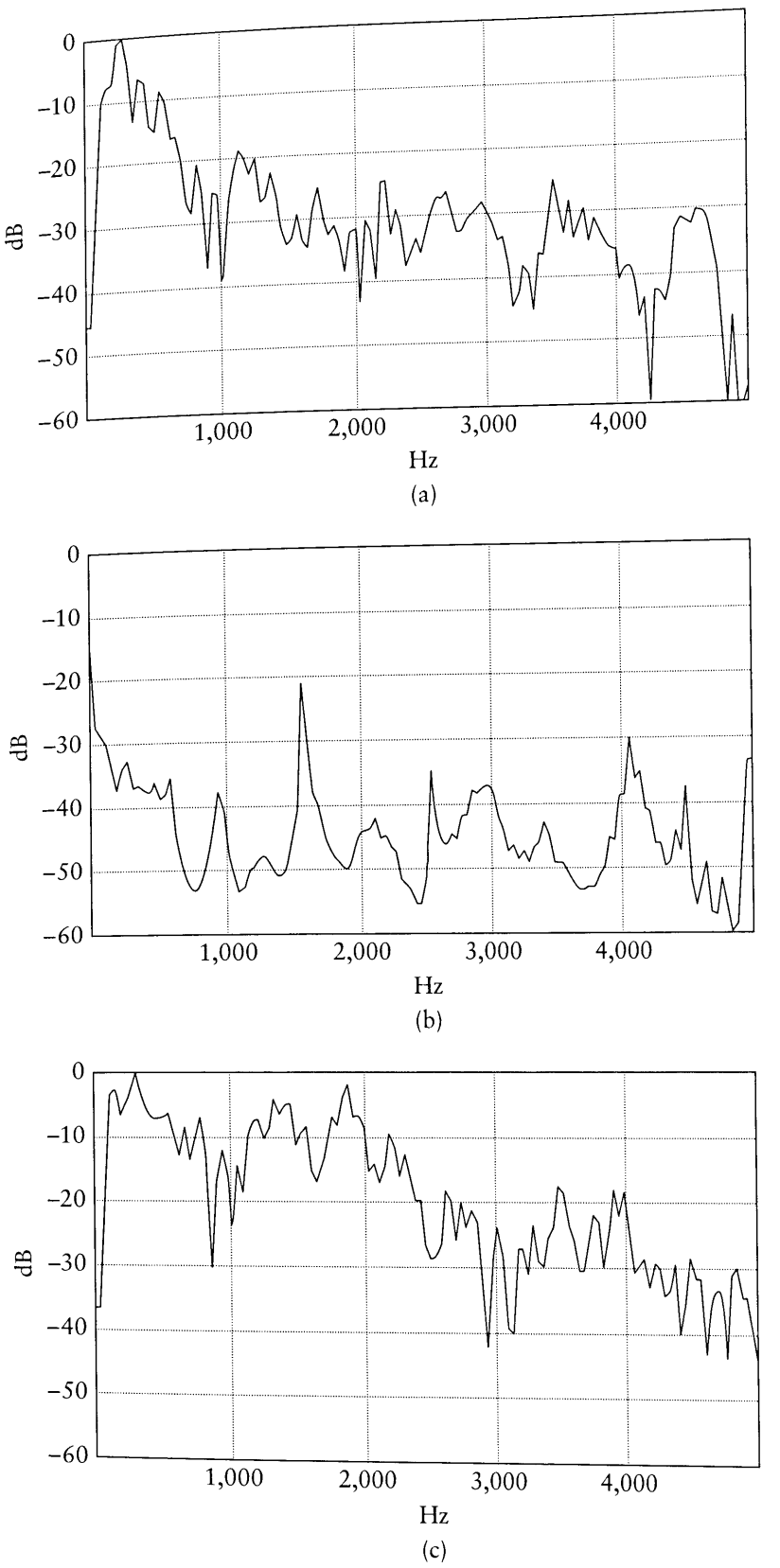


Figure 7.16.6 Spectral sections (256-point DFT) of release burst of plosives: (a) [b] in *bah*; (b) [d] in *dah*; (c) [g] in *gah*

This brings us back of course to the question of phonological encoding. Often our expectations are simply wrong: we look at the spelling, or even the usual phonetic transcription, of an English word such as *sent* [sɛnt] and suppose that it has four sounds, four successive pieces of information. In so doing we overlook the possibilities which are characteristically exploited in normal speech and hearing: that the nasality of the consonant [n] may be anticipated in, and even conveyed by, nasalization of the preceding vowel; that the timing of the nasal consonant itself may be a significant cue to the hearing of a following [t]; and so on. In short, it is useful to look at larger units of speech – in particular at syllables – to see how the properties of the vocal tract system, viewed as static properties, are actually integrated into a linear flow.

Syllabic organization, studied acoustically, naturally reflects the structural patterns discussed in section 3.1 and other sections of chapter 3 above. The peak or nucleus may display the strong and relatively simple formant patterns produced by vocalic resonance in an unobstructed tract, and the onset and coda usually have the more or less complex spectral patterning of the various vocal tract configurations associated with consonants. Syllables vary in their structure of course, but a CV syllable is the most useful starting point for description (figure 7.17.1).

Figure 7.17.1 has the format of a spectrogram in which only the first three formants (or comparable properties) are displayed. It is divided into two major components, the second of which is further divided into two, making three sections in all. The first of these, marked T_0 to T_1 in the figure, is a quasi-stationary complex spectrum consisting of formants or noise components or both (depending on the manner of articulation). Its precise nature is determined by the particular consonant constriction and excitation source(s) used. It is quasi-stationary in the sense that the formants are, in principle, constant over time, but actually vary with both manner and context of articulation. The sequence T_1 to T_2 is the first part of the syllable peak or nucleus, and is characterized by formant movement from the point of consonant constriction release at T_1

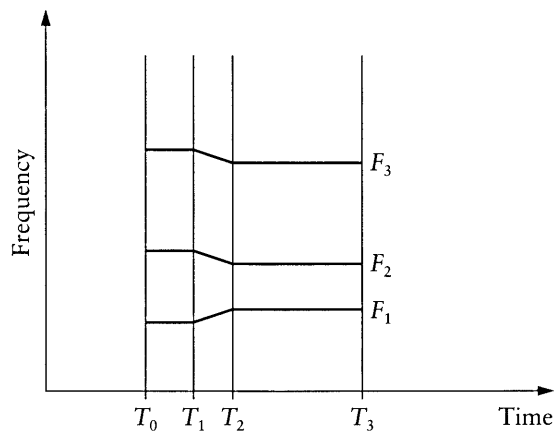


Figure 7.17.1 Idealized structure of a syllable

to the vowel target values at T_2. The formant movements in this sequence are known as the FORMANT TRANSITIONS, and they play an important role in encoding consonant-related information. Formant transitions are a classic illustration of the overlapping of acoustic encoding of phonological information: although located in the syllable peak, they provide important information about adjacent consonants. The formant movements reflect the rapid shift of articulatory position from consonant to vowel, in which the tract changes its consonant constriction shape to become an unobstructed resonant system, often creating rapid resonance changes during the process. Sequence T_2 to T_3 is the vowel target proper, with a nominally stable spectrum as described in section 7.15 above (unless of course the syllable nucleus is occupied by a diphthong).

Figure 7.17.1 shows that the formant structure is quite continuous through the coda and peak, which illustrates the principle of resonance continuity mentioned in section 7.16 above in connection with fricatives. Nevertheless, complex spectra such as those of nasals and fricatives are such that the continuity will not be observed for all resonances. Figure 7.17.1 also illustrates the detail with which durational analyses can be made on the speech signal using a spectrographic display (section 7.14 above). T_0–T_3 gives the total syllable duration, and T_1–T_3 gives the vowel duration, which normally includes the transition and target components of the nucleus.

The frequency changes occurring in the formant transition provide important information about the place of articulation of the preceding consonant and may contribute to information about its manner of articulation. The actual frequency values in the transition are usually determined both by the consonant place of articulation and by the acoustic target (i.e. formant patterns) of the following vowel. This is a classic example of context-sensitivity, in the form of coarticulation.

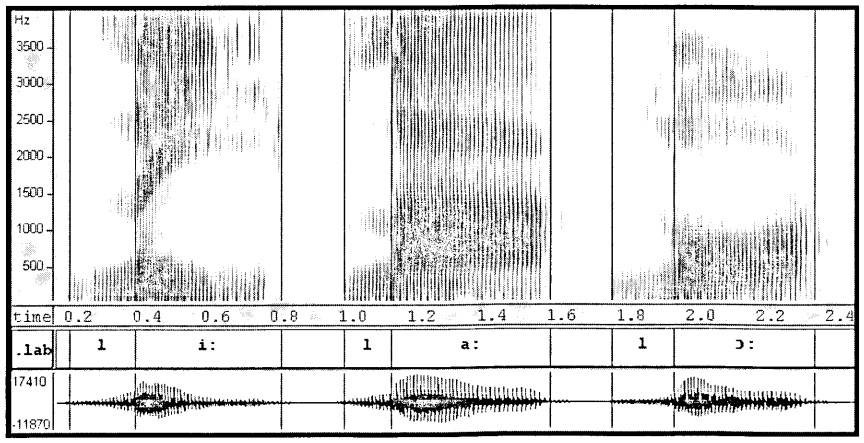
The pioneering work in exploring the dynamic spectral patterns of speech was undertaken during World War II by a research team at Bell Laboratories in the USA, and later published in a comprehensive form by Potter et al. (1947). The context of their work was an attempt to make it possible to read speech from spectrographic displays (hence 'visible speech' in their title). Potter and his colleagues tried to deal with context-dependent variability by the notion of the HUB, which they defined as the characteristic position of F_2. Recognizing that the position of the hub varied according to the strength of coarticulation effects, they included appropriate allowances in their recognition rules. Thus velars exhibit more hub variability than alveolars, and so on.

The next important step in understanding spectral structure was taken in perceptual studies initiated by Cooper et al. (1952) at the Haskins Laboratories in the USA. In this research, representations of speech spectrogram patterns were hand painted on clear acetate sheets, which were then used to generate synthetic speech. The researchers could thus manipulate the values and shapes of formant patterns and replay them on a special machine called a 'pattern playback'.

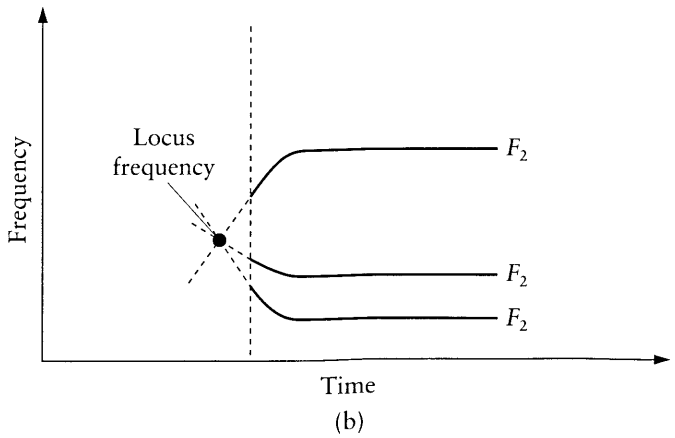
In a development of this work, Delattre et al. (1955) showed that F_2 in particular appears to 'point' towards a notional characteristic frequency for a given place of articulation, whatever the associated vowel. This observation was later

extended by Liberman et al. (1959) to include F_3. These notional frequencies were called the consonant LOCI, and the Haskins group demonstrated that once the dynamic spectral properties of syllable structure were understood, it was possible to formulate rules by which spectral patterns for any given utterance could be constructed. Using their pattern playback machine, they demonstrated that simple utterances could be produced using these rules without having to copy the corresponding spectral patterns of a human speaker. This was an important development, which has been followed by research designed to extend our understanding of rules relating phonological strings to the acoustic structures which realize them. The pioneering pattern playback experiments also illustrate the importance of relating the dimensions of production to those of perception in studying the phonological properties of speech.

Figure 7.17.2(a) contains spectrograms of the CV syllables *lee* [li:], *lah* [la:] and *law* [lɔ:]. In figure 7.17.2(b) the three F_2 patterns of part (a) have been combined in a single display to show how all three point towards a locus. The



(a)



(b)

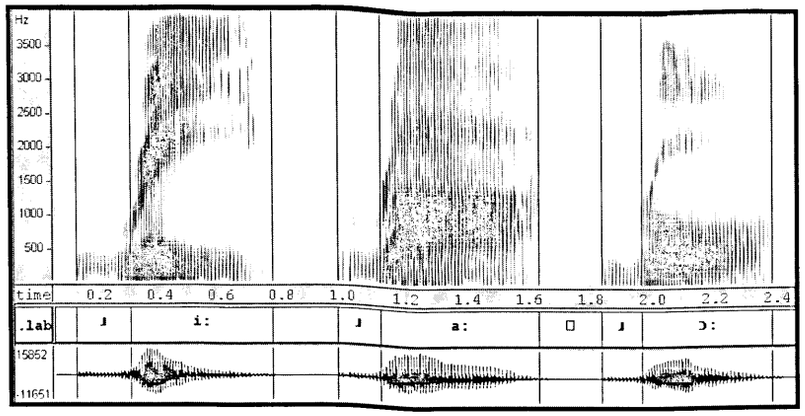
Figure 7.17.2 (a) Spectrograms of *lee*, *lah*, *law*; (b) locus effect of coarticulation on F_2

choice of [l] as the initial consonant in the three examples ensures a reasonably stable spectrum during the consonant constriction, and a strong formant structure illustrating the resonance continuity principle. Plotting formant movements in the time domain, as illustrated in figure 7.17.2(b), is a common and useful technique for the display and analysis of speech spectrum dynamics. It is, however, not always the best way of explaining how phonological distinctiveness is conveyed. The locus theory is a case in point. The theory assumes that each formant transition for a given sound points towards a specific frequency locus. In effect, the locus frequency is a notionally invariant point, mostly never reached because of coarticulation. Unfortunately, in the case of velar stops, investigation has shown that it is impossible to determine a single locus frequency for each formant and for all vowels. To overcome this problem, proponents of the theory have argued for two locus frequencies – one associated with front vowels, and one with back vowels – but this pragmatic solution has little in the way of true explanatory value.

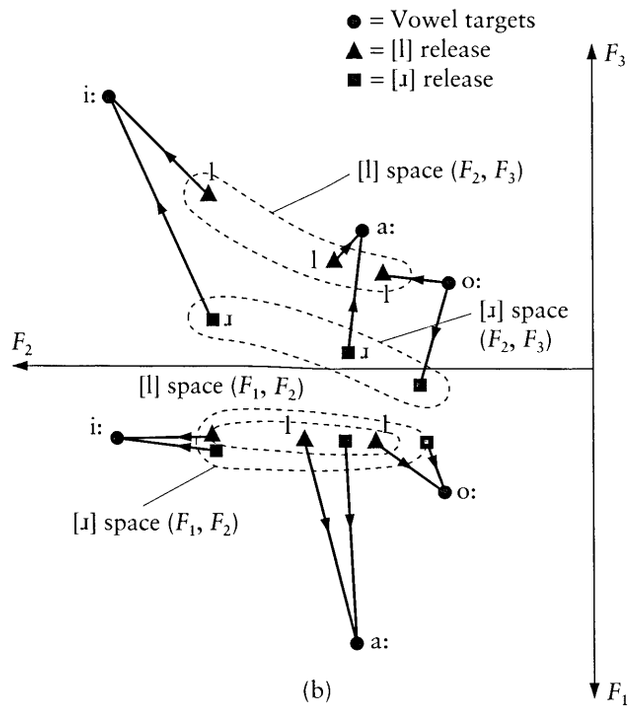
In his fundamental study of coarticulation, Öhman (1966) argued against the locus theory on the grounds that it was impossible to reconcile a single invariant acoustic 'target' with the observable phenomena of coarticulation. He proposed instead that the speech signal should be regarded as a sequence of vocalic structures upon which consonantal perturbations were imposed; and the strongest influence on a consonant was the following vowel. More recent work by Purcell (1979) on coarticulation using modern statistical analysis tends to confirm Öhman's work, which supports the general concept of syllabic structure centred on a vocalic peak.

It is from this vowel-based perspective that an alternative form of spectral data display has been developed. If the formant frequencies at the point of release or formation of the consonant constriction and the associated vowel target values are plotted on a combined F_1/F_2 and F_2/F_3 plot, then the true consonant formant patterns, in relation to their associated acoustic vowel space, are more easily portrayed. In this way we retain the general concept of consonant locus, but in the form of a 'locus space', plotted within a general frequency space whose axes are F_1, F_2 and F_3 independently of the time domain. The data for individual segments can then be seen in the context of associated sounds in the phonological system. The principle is illustrated in figure 7.17.3 with the consonant [ɹ] in the context of the vowels [i:], [a:] and [ɔ:] in CV syllables. Figure 7.17.3(a) shows spectrograms of the syllables *re*, *rah* and *raw* – which can be compared with the syllables containing [l] in figure 7.17.2(a) above – and figure 7.17.3(b) shows a formant plot (at consonant release and vowel target). The distinctive spectral spatial patterns which result give a clear indication of the ways in which the two consonants are acoustically distinct, despite their variability. None of this discounts the importance of the time domain, for it is also the rate of the formant transitions that distinguishes these consonants from stops, which have more rapid formation and release of occlusion (as also shown by Liberman et al. 1956, using the pattern playback technique).

Nasal consonants show strong low-frequency energy and weaker upper formant structure during their oral occlusion phase, as noted in section 7.16 above. There is a sudden increase in formant amplitude when the oral occlusion is



(a)



(b)

Figure 7.17.3 (a) Spectrograms of *re*, *rah*, *raw*; (b) locus space for [l] and [ɹ] with [i:] [a:] [ɔ:]

released, because the less damped oral tract suddenly becomes the main resonator system again. The formant transitions exhibit a pattern related to place of articulation in which the locus space for alveolars is much smaller than that for velars. Figure 7.17.4 shows spectrograms of the words *timer*, *finer* and *singer*, illustrating nasals at bilabial, alveolar and velar points of articulation.

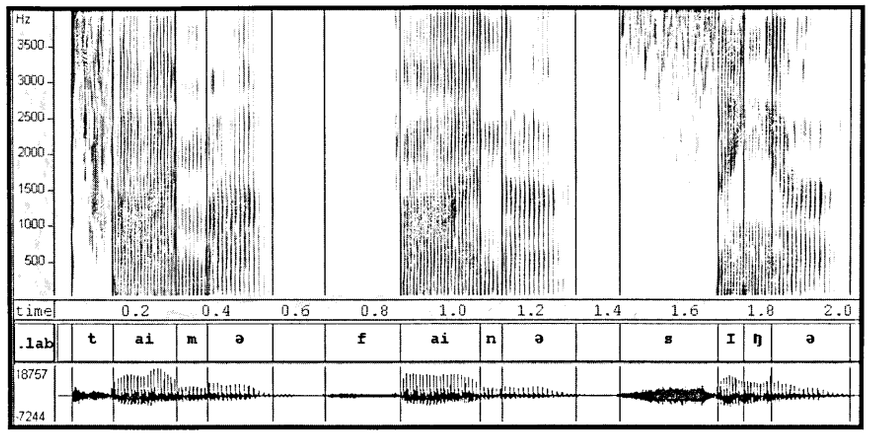
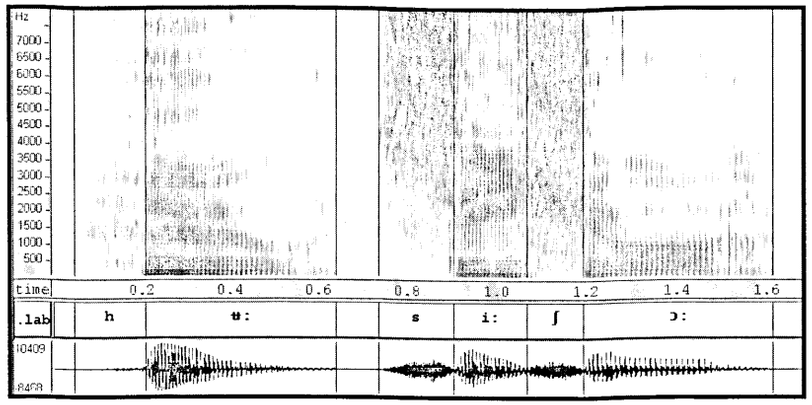


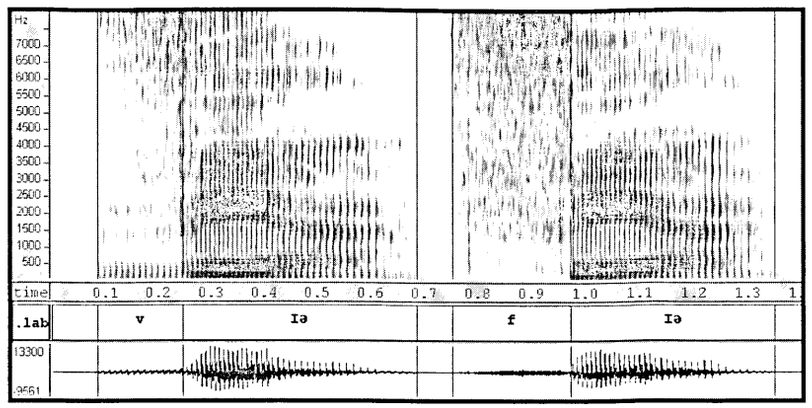
Figure 7.17.4 Spectrograms illustrating nasal consonants in *timer*, *finer* and *singer*

Fricatives show complex periodic spectral properties during their constriction (section 7.16 above). Spectrographic analysis reveals how the integration of glottal with supraglottal activity meets the aerodynamic demands of frication yet allows a rapid switchover to phonation. This is shown in its simplest form on the left-hand side of figure 7.17.5(a) for the fricative [h] in *who*: here the constriction is at the larynx, so that the fricative uses the full vocal tract resonator system, giving it a formant structure very similar to that of the following vowel. Only the source differs – aperiodic in the fricative, periodic in the vowel. Since no supraglottal articulators are involved in producing the [h] fricative, there are no appreciable formant transitions. Contrast this with the right-hand side of figure 7.17.5(a), which shows the two fricatives [s] and [ʃ] in the word *seashore*. The [s] shows little energy below 4,000 Hz because of the short anterior resonator system (section 7.16 above). The formant transitions from the [s] into the following [i:] vowel are clearly seen, and are similar in pattern to those of other alveolar sounds. Resonance continuity is preserved largely in F_4 and F_5. A similar structure can be seen in the intervocalic [ʃ] except that the fricative energy now extends down to around 2,500 Hz. As noted in section 7.16 above, the frequency at which the fricative noise is sharply attenuated is an important cue to place of articulation. In the example shown, the noise attenuation frequency is relatively stable for [s], but in [ʃ] it falls appreciably between the vowels because of their coarticulatory influence. The demands of the vowels are such that during the articulation of the [ʃ], the tongue and lips are already moving towards their positions for the following vowel.

Figure 7.17.5(b) shows the words *veer* and *fear*, illustrating the voicing contrast in fricative spectra. The low-frequency periodic voicing, and its modulating effects on the upper-frequency noise spectrum, can be seen clearly in [v]. Formant structure and spectrum shaping are not very apparent in fricatives such as these, since their constriction site is very close to the front end of the tract. It is also evident that there is less formant movement in the vocalic nucleus than in the examples of figure 7.17.5(a). This is due in part to the



(a)



(b)

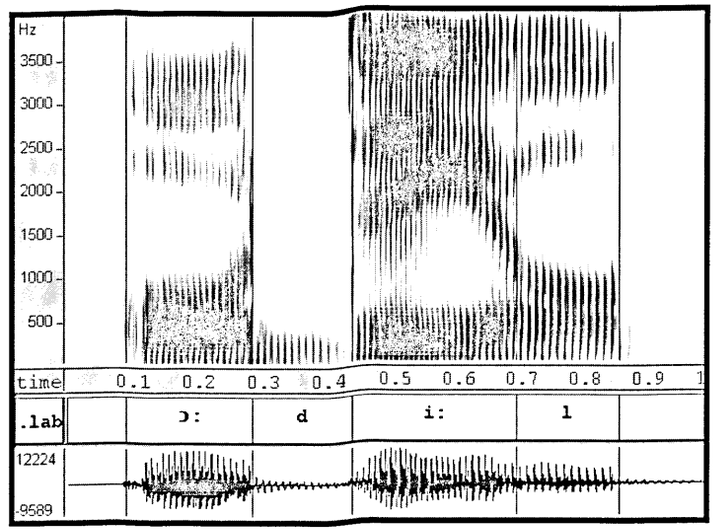
Figure 7.17.5 Spectrograms illustrating fricatives (a) *who* and *seashore*; (b) *veer* and *fear*

anterior location of the constriction, which causes less perturbation of vowel-related tract resonance properties than would occur with lingual fricatives.

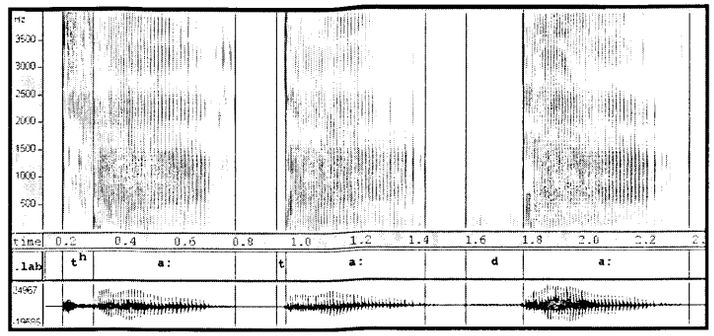
The last major class of sounds to be considered here is that of stops. These have three basic spectral components: an occlusion (which is silent in voiceless stops); a release burst composed of a short period of relatively stable frication energy (section 7.16 above); and, if a vowel follows, a transition into it characterized by rapid formant movement.

Stops generally exhibit strong coarticulation effects in their formant transitions, and it is the combination of the burst spectrum and the transitions which identifies their place of articulation. Their manner of articulation is identified by the relatively low frequency of F_1 at occlusion release, and the rapid rise of spectral energy thereafter. Fant (1973) includes an extensive analysis of stop acoustics, with a good example of the formant mapping described at the beginning of this section.

Figure 7.17.6(a) shows the stop [d] in the word *ordeal*, illustrating the three spectral components. The coarticulation influence on formant transitions at



(a)



(b)

Figure 7.17.6 Spectrograms of (a) *ordeal*; (b) [t^ha:] [ta:] [da:]

closure and release (where the consonant is flanked by different vowels) is analogous to that of figure 7.17.5(b). The low values of F_1 at the start and end of occlusion are also easily seen, as is the low-frequency 'voice bar' during the occlusion. Figure 7.17.6(b) shows the syllables [t^ha:], [ta:] and [da:]. These three kinds of stop (voiceless aspirated, voiceless unaspirated, and voiced) are phonologically distinctive in languages such as Thai and Burmese. The spectrograms illustrate the effects of voice onset time (section 2.16 above): in [t^ha:], phonation does not start until after the release of the occlusion, [ta:] shows phonation beginning at about the point of release, and [da:] has phonation starting during the occlusion. The timing of voice onset in stops is of great importance in the recognition of stops as voiced, voiceless or voiceless aspirated. Both our productive control of this timing and our perception of it have been studied extensively by Lisker and Abramson (1964, 1971), Slis and Cohen (1969) and Ladefoged (1971). There is ample evidence from the research that languages differ in the values of voice onset time used to signal voicing

contrast. In languages such as English and German, for example, aspiration is often a crucial feature of voiceless stops, distinguishing them from voiced stops (at least in some contexts), whereas in languages such as French and Dutch the contrast may be more truly one of presence or absence of voicing during the occlusion itself. In yet other languages, such as Hindi, stops may also have voiced aspiration (in contrast to voiceless aspiration), which demonstrates that speakers can exploit very complex coordination of laryngeal behaviour in relation to the supraglottal articulation to achieve phonological distinctions.

Phenomena such as we have been describing point to the danger of trying to locate acoustic features within discrete segments, and underline the importance of the syllable as a whole. It is worth repeating the point made in section 7.15 above: there is good perceptual evidence that just as consonantal information is partly specified by the coarticulatory dynamics of formant structure in the syllable peak, certain aspects of syllable structure as a whole contribute to the robustness of the perception of vowel identity (Strange et al. 1976). It has also been found by Lindblom (1963) and Stevens et al. (1966) that as syllable peaks are shortened, the formant transitions tend to be preserved at the expense of the more spectrally stable vowel target. This again argues for the importance of overall dynamic spectral patterns in phonological encoding.

This section has provided no more than a foundation for the study of complex spectral and temporal aspects of speech sounds, and the spectrographic examples have illustrated the segmentation and labelling process in a general and basic way. What was traditionally done by hand, by measurement and marking of hard copies of spectrograms, is increasingly being done by multipurpose speech editing and analysis software packages designed for the purpose, or on stand-alone and purpose-built speech analysis equipment using digital signal processing chip technology. Some systems also allow storage of the segmented and labelled spectrographic data for further analysis. But the technology does not of itself guarantee insight and analysis, and the understanding of basic principles remains essential.

More detailed accounts of the acoustic properties of speech sounds may be found in Fant (1960), Minifie (1973), Shoup and Pfeiffer (1976), Fry (1979), Pickett (1980), O'Shaughnessy (1987) and Stevens (2000). Johnson (2003) is an excellent, easy-to-read introduction to acoustic modelling of speech sounds and digital processing of speech. Likewise, Sawusch (2005) provides a concise summary of different types of speech signal processing with particular reference to speech synthesis. For more advanced approaches to speech signal processing, Harrington and Cassidy (1999) is recommended.

7.18 The relationship between articulatory and acoustic properties of speech production

Phonological description has always tended to be articulatory in orientation, for the obvious reason that the gestures and settings of articulatory organs are more easily observed than sound waves. Certainly it was possible to describe articulation – even if impressionistically – without much recourse to modern technology. Once spectrographic analysis became available, a natural and immediate step was to try to relate what were already known as articulatory properties to what were now being investigated as acoustic or spectral properties (see e.g. Delattre 1951).

We can think of the relationship between articulation and acoustics in terms of transformations which will derive acoustic properties from articulatory properties. The basic method of obtaining such transformations is by modelling the vocal tract, treating the supraglottal resonator system as a series of very short tube sections of fixed length and variable cross-sectional area. Before computers, this analog was realized as an electrical transmission line consisting of coils, capacitors and resistors which modelled each short section, including losses due to damping. These AREA FUNCTION ANALOGS, described in detail by Dunn (1950) and Fant (1960), typically had from 18 to 40 separate sections, each of adjustable area. In more recent times it has been possible to simulate a transmission line on computer, or to use models based on reflection coefficients, although there are limitations as some models do not deal well with tract losses (Kelly and Lochbaum 1962, Wakita 1976). The transformation from articulatory to acoustic is achieved by adjusting each section of the model to an appropriate area value, to approximate the vocal tract shape for a particular sound. To do this, of course, researchers really need accurate articulatory information (from X-rays or other such sources) to make the model approximate the cross-sectional area of the actual vocal tract shape as closely as possible. Figure 7.18.1 shows how the vocal tract can be analysed to this end.

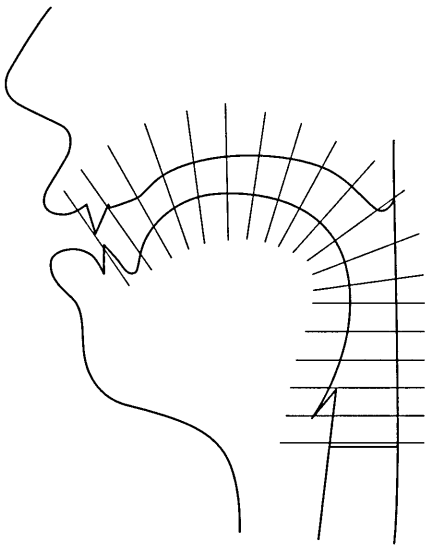


Figure 7.18.1 Derivation of vocal tract area function characteristics, showing cross-sections of the vocal tract at equal intervals between glottis and lips, used to specify the overall vocal tract area function which can then be used to determine the effective vocal tract filter frequency response

It is then possible to compute the effective frequency response of the vocal tract shape to which the model has been adjusted, and to obtain the corresponding formant frequency data. The process is reasonably accurate, but laborious. Researchers soon looked for more economical articulatory specifications, preferably in the form of a parametric articulatory model, a model with discrete articulatory values that could be, so to speak, overlaid on the vocal tract. The parameters or categories of this model should ideally be related to those of conventional phonetic description, and should at the same time be capable of specifying the cross-sectional areas of all sections of the area function model.

Stevens and House (1955) approached the task by treating the vocal tract as a tube with adjustable lip and tongue hump geometry. Fant (1960) used two- and three-tube compound resonator tubes as rudimentary approximations of vocal tract configurations. In both approaches, nomographs were supplied, from which one could predict formant values for a large combination of the input parameters to the models.

The limitation in these two approaches is that both used articulatory parametric overlays which were somewhat removed from phonetically valid measurements of the human vocal tract shape. Lindblom and Sundberg (1971) took the essential further step of devising an economical set of parametric measures specifying lip and tongue shape, and jaw and larynx height. This meant that a recognizable vocal tract could be defined. Like their predecessors, they provided nomographs which predicted formant frequency values for any combination of the parametric values defining vocal tract state. Again as in previous research, they used an area function analog to obtain the primary formant data with which they constructed their nomograms.

Lindblom and Sundberg's work suggests the following general relationships between articulatory and acoustic factors:

1 Jaw opening causes F_1 to rise quite markedly (all else being constant), usually in the context of controlling vowel height. It will cause F_2 to rise if the tongue is retracted up towards the soft palate: this effect is strongest when the lips are spread, but minimal in other articulatory positions. F_3 may rise sharply at moderate jaw apertures when the tongue is raised towards the palate region.
2 Tongue body movement in a general anterior-to-posterior direction causes a modest rise in F_1 (typically around 200 Hz) if the jaw is kept at a fixed opening (but the jaw is *not* normally kept in one position). Movement from anterior to neutral position results in a large drop in F_2 in all cases. From neutral to posterior position, F_2 will tend to rise with small jaw openings, but continue to fall with larger jaw openings.
3 Tongue body shape, which controls the degree of tract constriction (assuming a constant jaw position), has little effect on F_1 except that it results in a modest fall at maximum constriction if the tongue body is well forward. It has a strong effect on F_2, causing it to fall substantially as constriction increases if the tongue body is in neutral or posterior position. An anterior tongue body position combined with maximum constriction results in a sharp rise in F_2. F_3 is little affected by tongue body shape except for a modest fall at neutral and maximum constriction with an anterior tongue position.

4 Lip rounding has the general effect of lowering all formant frequencies, with the strongest effects observable on F_2 and F_3. The extent of the effect depends on what the tongue and jaw are doing at the same time.
5 Lowering of the larynx makes the vocal tract longer and tends to lower all formant frequencies; the degree of lowering of each formant partly depends on the overall state of the vocal tract. In general, larynx height influences F_2 and F_4 more than F_3.

Lindblom and Sundberg conclude that tongue height (maximum height of the tongue hump), despite its traditional importance in the description of vowels, does not relate directly or usefully to acoustic properties. In general agreement with Lindblom and Sundberg, Wood (1979) shows that formant frequency patterns in vowel production relate more directly to the location and degree of tongue constriction within the vocal tract.

Stevens (1972a) takes a more overtly phonological view of articulatory–acoustic relationships in his QUANTAL THEORY. He maintains that there are general states or regions of articulatory activity, within whose natural boundaries little change in the acoustic output of the tract can be achieved. On the other hand, a small shift beyond the boundary will produce a large (discontinuous) acoustic change. This argues that the relationship between vocal tract state and spectral properties is not linear. It is these 'step-wise' spectral changes which contribute to phonological distinctiveness, and Stevens suggests that articulation is organized to make optimum use of the vocal tract's ability to produce such changes. Discrete manners and places of articulation are located inside insensitive regions, which means that the acoustic properties of specific sounds are relatively tolerant to minor articulatory variability, but that large acoustic changes occur when we cross the boundaries of the regions.

An example of this principle is the sudden change from laminar to turbulent airflow when a constriction reaches a critical cross-sectional area and the sound thereby moves from vocalic or approximant mode to fricative mode. The way in which a shift of articulation from [s] to [ʃ] produces a sudden lowering of the fricative noise cut-off frequency is another illustration of the same principle. Similarly, it can be argued that the quasi-universal vowel triangle of [i], [a] and [u] is the preferred three-way vowel system because these vowels represent three stable and acoustically noncritical articulatory positions. Wood's X-ray studies (1979) appear to confirm the quantal hypothesis, as do Perkell's electromagnetic articulographic (EMA) studies (1996), and a study by Beckman et al. (1995) based on electro-microbeam data.

The difficulty of obtaining comprehensive data about dynamic articulation in speech – the research methods are often invasive and costly – makes it attractive to try to predict articulatory details from acoustic information. There is indeed continuing interest in acoustic-to-articulatory transformations and it is possible to predict the vocal tract shape for a given acoustic signal sample; but in many cases the prediction is not a unique solution. This is regrettable but hardly surprising, for it is an important attribute of the vocal tract that it can compensate for one or another articulatory constraint and still generate a required acoustic output. It is, for instance, possible to produce intelligible

speech with a fixed mandible, as when holding a pencil between the teeth. Although attempts have been made to produce acoustic-to-articulatory models (Ladefoged et al. 1978, Ladefoged and Harshman 1979, Carré 2004), the fact that transformations cannot be guaranteed to be unique has remained an underlying limitation.

7.19 Acoustic features of prosody

Prosodic or suprasegmental features – such as pitch and loudness – are reviewed in detail in chapter 9 below, but some basic acoustic aspects merit attention here. Of particular importance is the fundamental frequency (F_0), which carries a wealth of information, much of it describable as prosody or personal voice quality.

A gross but useful dimension of speech is its long-term spectral energy profile. This is derived simply by averaging a large number of spectral slices over a long sample of speech, usually at least several minutes. Obviously this measure gives no information about segmental details or even about intonation patterns, but it does provide some insight into voice quality and vocal effort. Standard data for English can be found in Dunn and White (1940). Figure 7.19.1 shows relevant long-term spectra.

In general, when speakers increase their vocal effort, there will be more high-frequency energy in their long-term spectrum, while reduced vocal effort means less. Often such a change in energy distribution will be revealed as a change in the high-frequency spectral slope (or rate of attenuation). Speakers do in

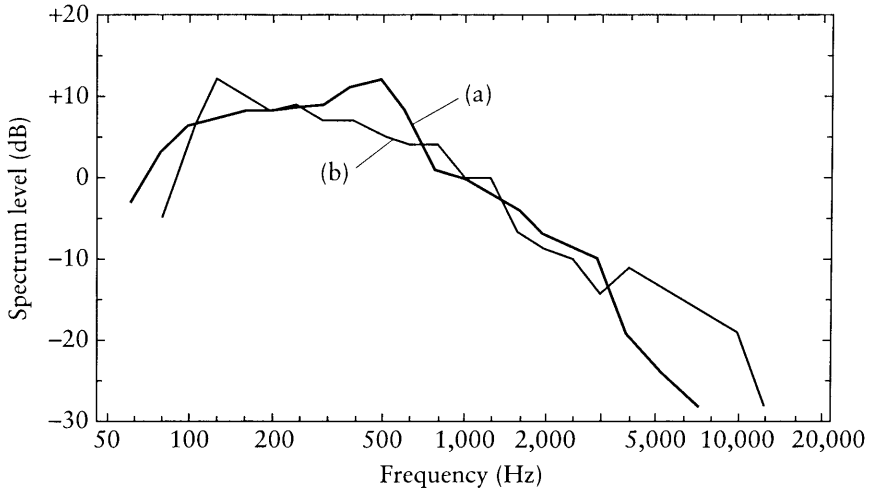


Figure 7.19.1 Long-term speech spectra: (a) data from six male American speakers; (b) data from five male Australian speakers
Source: (a) Dunn and White 1940.

any case differ from each other in the distribution of spectral energy within their speech, largely because they manage their phonatory and other vocal tract settings in different ways. The differences may be evident over relatively long stretches of speech, as in the case of voice quality and speaker identity characteristics described by Laver (1980) and Nolan (1983); or differences may be relatively short-term, reflecting the speaker's response to a specific communicative situation. Figure 7.19.2 is an example from Clark et al. (1987) showing the difference in voice quality between speech produced in a quiet environment and speech produced with considerable extra effort.

Long-term spectra are probably most useful in a comparative form, as shown in figure 7.19.2. They may also be of sociolinguistic interest where voice quality is characteristic of a regional or social group. In this connection, there may also be value in measures such as the mean value of F_0, which may have a typical range and distribution in particular communities.

Most useful of all is the F_0, or pitch contour. (The two terms are often used interchangeably, but 'pitch contour' is strictly speaking a perceptual measure only.) Measuring F_0 is not always easy, for several reasons. Firstly, the effects of vocal tract resonance in some speakers may make it hard to detect the F_0 pattern in the waveform by automatic methods. Secondly, speakers rarely produce 'ideal' phonatory patterns: in particular, the onset and offset of voicing is often weak and may have erratic periodicity, so that it is not always clear precisely where the F_0 pattern starts and finishes. Thirdly, certain segments, such as initial stops, may produce short-term perturbations in periodicity (section 9.2 below) which may not be accurately detected in F_0 measurement.

The techniques for measuring F_0 may be broadly divided into two types: time domain and frequency domain. In the first of these, the speech waveform is usually passed through a low bandpass filter to remove much of the high-frequency information which could obscure the periodic pattern. The resulting

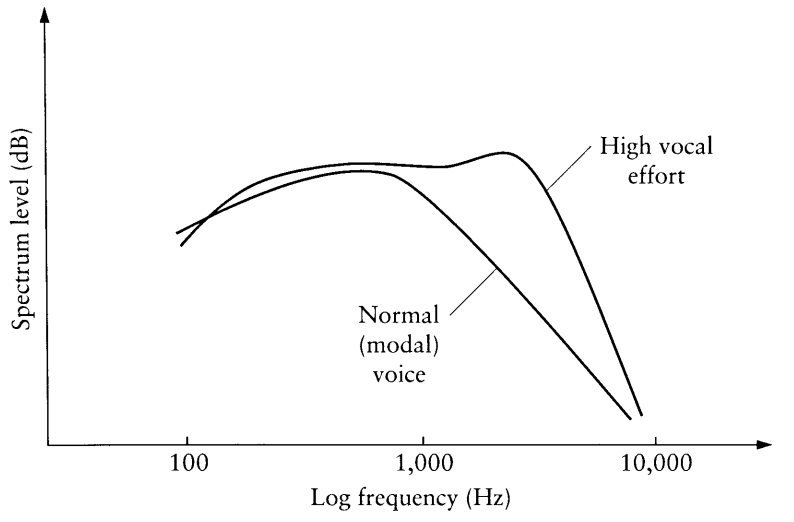


Figure 7.19.2 Long-term speech spectrum showing changes due to vocal effort

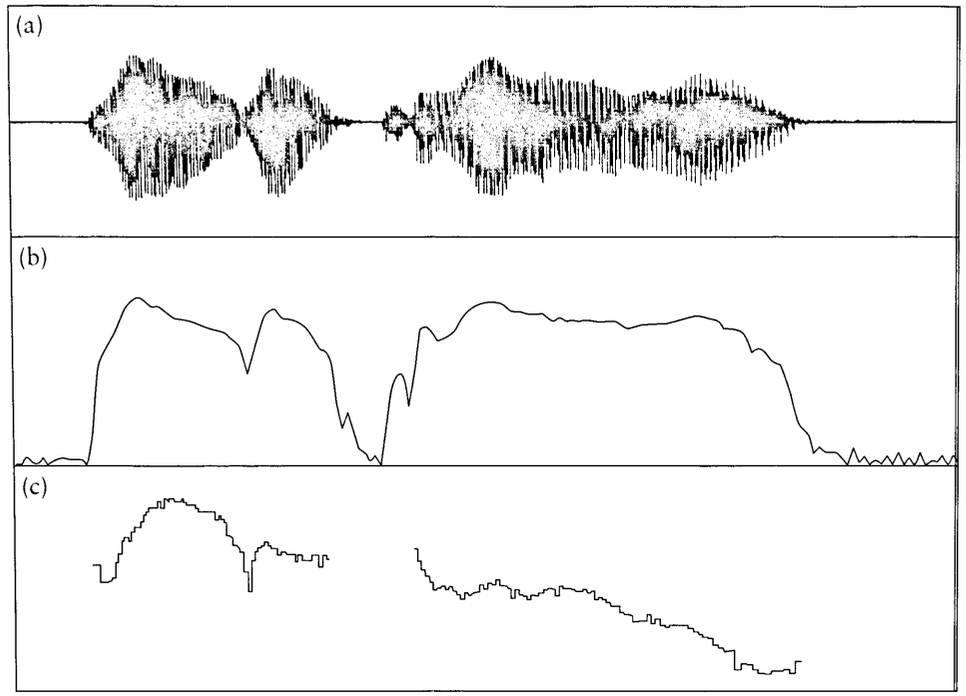


Figure 7.19.3 F_0 derived by analysis of time domain waveform *We went to Woolloomooloo*; (a) time domain waveform; (b) intensity contour; (c) F_0 contour

time domain waveform is then processed to identify the period of the speech wave, by detecting either the recurrent zero crossings or the peaks. F_0 is then easily determined as the reciprocal of the period. To display a continuous pitch trace, the cycle-by-cycle measures of F_0 are often smoothed, unless information on jitter or individual perturbations is required. An example of an unsmoothed trace is shown in figure 7.19.3.

The traditional equipment for this type of measurement used simple analog electronics. Computer-based methods are now being used, except where continuous real-time displays are needed. Some of the more sophisticated computer-based time domain analyses incorporate decision-making processes, in which alternative estimates of the pitch period can be checked against the estimates of adjacent pitches, to avoid anomalous decisions. A well-known and very successful example is Gold and Rabiner's time domain pitch algorithm (Gold and Rabiner 1969). With powerful algorithms such as this, there is far less need to pre-filter the signal before making pitch estimates, and the analysis can track rapid perturbations in F_0 much more accurately.

Frequency domain methods make the pitch estimates from the harmonic structure of the spectrum. They are inherently accurate, within the limits of the frequency resolution of the spectral analysis. If the resolution is too broad, the harmonic structure of the spectrum will not be adequately resolved, and if it is too narrow, its response time will be too slow to track rapid F_0 changes.

Figure 7.19.4 F_0 derived from a narrow-band spectrogram with equivalent segmented and labelled broad-band spectrogram of the utterance *He wanted to come*

The simplest form of frequency domain F_0 measurement uses a narrow-band spectrogram. The technique is to pick a single suitably clear harmonic (say between the third and seventh) and measure its frequency values. The F_0 value is then simply the harmonic frequency divided by the harmonic number. Figure 7.19.4 shows an example of this method using the third harmonic in which a series of F_0 values have been calculated by hand at points indicated by the black dots on the spectrogram.

Frequency domain methods are laborious, but as indicated earlier in the chapter, provide relatively unambiguous pitch estimates (using the method shown in figure 7.19.4). They are computationally demanding, which makes them less popular than time domain approaches. Digital signal analysis methods for F_0 are discussed in a general review of pitch measurement techniques in Hess (1983) and in Rabiner and Schafer (1978).

Intensity and time are also important in determining the suprasegmental characteristics of an utterance. For gross measures of intensity and time, see sections 7.6 and 7.7 above; and for detailed time measures, see the account of spectrographic segmentation in section 7.17. The broader phonological role of duration and intensity is taken up in chapter 9 below.

Exercises

1 Check that you understand what each of the following is.

   a. the amplitude and frequency of a vibration
   b. sinusoidal vibration
   c. damped vibration
   d. white noise
   e. fundamental frequency
   f. resonant frequency
   g. bandwidth
   h. spectral envelope
   i. formant
   j. the phase relationship of two waves

2 Identify and compare the principal acoustic properties of the following types of speech sounds.

   a. vowels
   b. approximant consonants
   c. nasal consonants
   d. plosives
   e. voiced fricatives
   f. voiceless fricatives

3 What is a decibel and how does it relate to sound pressure level and acoustic intensity?
4 What is a mel and how does it relate to frequency?
5 What is the phantom fundamental and what does it tell us about hearing?
6 Briefly explain each of the following.

   a. a resonance curve
   b. a discrete Fourier transform
   c. the concept of locus
   d. the concept of normalization

8 Speech Perception

Our ability to perceive – and understand – speech is quite remarkable. This chapter begins by drawing attention to the complexity of the perceptual task (8.1). It then describes the structure of the human ear (8.2) and the basic perceptual functioning of the ear (8.3).

The chapter then gives a brief account of research into speech intelligibility (8.4) and the perception of speech sounds (8.5) before dealing with particular phonological aspects in more detail: the perception of vowels is treated in 8.6 and the perception of consonants in 8.7, while 8.8 reviews discussion among researchers about the basic unit of perception, for example about whether the phoneme can be taken as a unit of speech perception. Section 8.9 turns to the perception of prosodic information, such as stress and pitch.

The chapter includes mention of work on word recognition – much of it usually considered to be research in psychology rather than phonetics (8.10). A brief overview of the principal models of speech perception that have been proposed by researchers (8.11) and concluding remarks (8.12) complete the chapter.

8.1 Introduction

Our recognition of linguistic units such as syllables and words and clauses depends on a number of factors. These include the acoustic structure of the speech signal itself, the context, our familiarity with the speaker, and our expectations as listeners. There is substantial evidence that much of our understanding of continuous speech involves a component of 'top-down' linguistic processing which draws on our personal knowledge base, and does not necessarily demand segment-by-segment processing of the acoustic signal to establish the phonological structure and arrive at its identity and meaning.

There are two central problems which are as yet not fully resolved in our total understanding of the processes leading to the perception of phonological structure in speech. The first is the highly variable and contextually sensitive relationship between the phonological structure and the acoustic cues embedded in the spectral time-course of the acoustic signal (sections 7.15 to 7.17 above). This is sometimes referred to in the literature as the invariance problem