Chapter 2

The production of speech: the physiological aspect

2.1 The speech chain

Speech is the result of a highly complicated series of events. The communication in sound of such a simple concept as *It's raining* involves a number of activities on the part of the speaker. In the first place, the formulation of the concept will take place in the brain; this first stage can be said to be psychological or psycholinguistic. The nervous system transmits this message to the organs of speech which will produce a particular pattern of sound; thus the second stage is physiological or, more precisely, articulatory. The movement of the organs of speech will create disturbances in the air; these varying air pressures can be investigated and constitute the third stage in the chain, the physical or, more precisely, acoustic. Since communication generally requires a listener as well as a speaker, these stages will be reversed at the listening end: the sound waves will be received by the hearing apparatus and information transmitted along the nervous system to the brain, where the linguistic interpretation of the message takes place. Phonetic analysis has often ignored the role of the listener. But any investigation of speech as communication must ultimately be concerned with both the production and the reception ends.

However, our first concern is with the second stage, speech production, or articulation. For this reason, we now examine how the various organs behave to produce the sounds of speech. References are made to the videos on the website where the reader can view the organs in action.

2.2 The speech mechanism

Humans possess, in common with many other animals, the ability to produce sounds. Humans differ from other animals in being able to organise the range of sounds which they can emit into a highly efficient system of communication. Non-human animals only rarely progress beyond the stage of using the sounds they produce as a reflex of certain basic stimuli to signal fear, hunger, sexual excitement and the like.[1] Nevertheless, like other animals, man uses organs for speaking whose original physiological function developed before vocal communication was acquired; in particular, organs situated in the respiratory tract.

2.2.1 Sources of energy—the lungs

The most usual source of energy for our vocal activity is provided by an airstream expelled from the lungs. There are languages which possess sounds not requiring lung (pulmonic) air for their articulation and, indeed, in English we have one or two extralinguistic sounds, such as that we write as *tut-tut* and the noise of encouragement made to horses, which are produced without the aid of the lungs; but all the essential sounds of English use lung air for their production. Our utterances are, therefore, partly shaped by the physiological limitations imposed by the capacity of our lungs and by the muscles which control their action. We are obliged to pause in articulation in order to refill our lungs with air and this will to some extent condition the division of speech into intonational phrases (see §11.6.1.1). In those cases where the airstream is not available for the upper organs of speech (as when, after the surgical removal of the larynx, lung air does not reach the mouth but escapes from an artificial opening in the neck) a new source of energy, stomach air, may be employed. A new source of this kind imposes more restrictions than those exerted by the lungs and variation of energy is less efficiently controlled.

A number of techniques are available for the investigation of the activity during speech of the lungs and their controlling muscles. At one time air pressure within the lungs was observed by the reaction of an air-filled balloon in the stomach. On the basis of such evidence from a gastric balloon, it was claimed that syllables were formed by chest pulses.[2] Such a primitive procedure was replaced by the technique of electromyography, which demonstrated the electrical activity of those respiratory muscles most concerned in speech, notably the internal intercostals; this technique disproved the relationship between chest pulses and syllables.[3] X-ray photography and CT scans can reveal the gross movements of the ribs and hence by inference of the surrounding muscles, although the technique of Magnetic Resonance Imaging (MRI) is now preferred on medical and safety grounds.[4]

2.2.2 The larynx and the vocal cords

The airstream provided by the lungs undergoes important modifications in the upper parts of the respiratory tract before it acquires the quality of a speech sound. First of all, at the top end of the TRACHEA or windpipe, it passes through the LARYNX, containing the vocal folds or, as they are more commonly called, the VOCAL CORDS (see Fig. 1).

The larynx is a casing, formed of cartilage and muscle, situated in the upper part of the trachea. Its forward portion is prominent in the neck below the chin and is commonly called the 'Adam's apple'. Housed within this structure from back to front are the vocal cords, two folds of ligament and elastic tissue which may be brought together or parted by the rotation of the arytenoid cartilages (attached at the posterior end of the cords) through muscular action. The inner

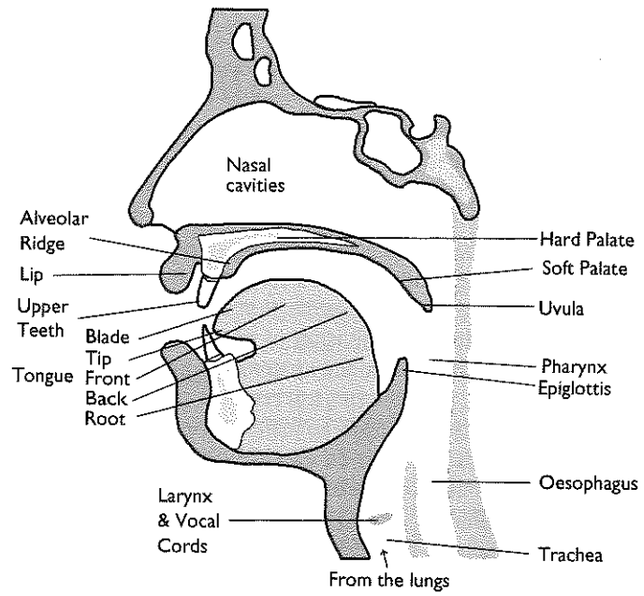

Figure 1 Organs of speech.

edge of these cords is typically about 17 to 22 mm long in males and about 11 to 16 mm in females.[5] The opening between the cords is known as the GLOTTIS. Biologically, the vocal cords act as a valve which is able to prevent the entry into the trachea and lungs of any foreign body, or which may have the effect of enclosing the air within the lungs to assist in muscular effort on the part of the arms or the abdomen. In using the vocal cords for speech, the human being has adapted and elaborated upon this original open-or-shut function in the following ways (see Fig. 2):

(1) The glottis may be held tightly closed, with the lung air pent up below it. A 'glottal stop' [ʔ] is produced when this closure is suddenly released and occurs in English, e.g. as an energetic onset to a vowel as in *apple* [ʔapl̩] or when it reinforces /p,t,k/ as in *clock* [klɒʔk] or even replaces them, as in *cotton* [kɒʔn̩]. It may also be heard in defective speech, such as that arising from cleft palate, when [ʔ] may be substituted for the stop consonants [p,t,k], which, because of the nasal air escape associated with cleft palate, cannot be articulated with proper compression in the mouth cavity.
(2) The glottis may be held open as for normal breathing and for voiceless sounds like [s] in *sip* and [p] in *peak*.
(3) The most common action of the vocal cords in speech is as a vibrator set in motion by lung air, which produces voice, or, more technically, PHONATION;

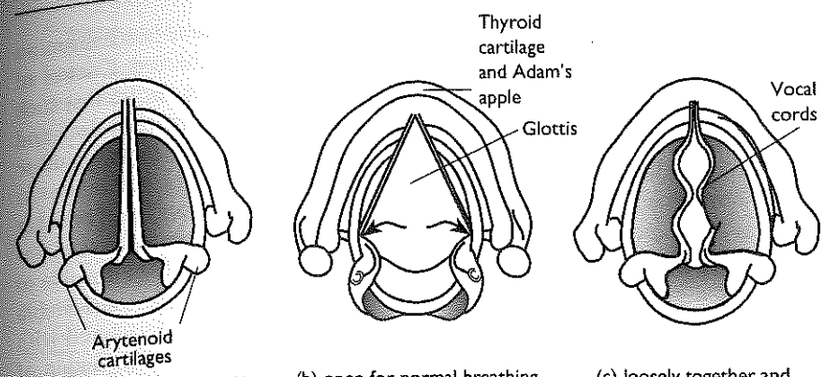

Figure 2 The states of the vocal cords as seen from above.

this vocal cord vibration is a normal feature of all vowels or of such a consonant as [z] in *zip* which makes them VOICED as compared with VOICELESS as [s] in *sip*. In order to achieve the effect of voice the vocal cords are brought sufficiently close together that they vibrate when subjected to the appropriate air pressure from the lungs. This vibration is caused by compressed air forcing the opening of the glottis and the resultant reduced air pressure permitting the elastic cords to come together again.[6] The vibration may be felt by touching the neck in the region of the larynx or by putting a finger over each ear flap when pronouncing a vowel or [z] for instance. In the typical speaking voice of a man, this opening and closing action is likely to be repeated between 100 and 150 times in a second, i.e. there are that number of cycles of vibration (called hertz, which is abbreviated to Hz); in the case of a woman's voice, this frequency of vibration might well be between 200 and 325 Hz. We are able, within limits, to consciously vary the speed of vibration of our vocal cords in order to change the pitch of the voice; the more rapid the rate of vibration, the higher the pitch (an extremely low rate of vibration being partly responsible for what is usually called creaky voice). Normally the vocal cords come together rapidly and part more slowly, the opening phase of each cycle thus being longer than the closing phase. This gives rise to 'modal' (or 'normal') voice which is used for most English speech. Other modes of vibration result in other voice qualities, most notably breathy and creaky voice which are used contrastively in a number of languages and may be used in English by some individuals and in some styles (see §5.8). Moreover, we are able, by means of variations in pressure from the lungs, to modify the size of the puff of air which escapes at each vibration of the vocal cords; in other words, we can alter the amplitude of the vibration, with a corresponding change of loudness of the sound heard by a listener. The

normal human being soon learns to manipulate his glottal mechanism so that most delicate changes of pitch and loudness are achieved. Control of this mechanism is, however, very largely exercised by the ear, so that such variations are exceedingly difficult to teach to those who are born deaf, and a derangement of pitch and loudness control is liable to occur among those who become totally deaf later in life.

(4) One other action of the larynx should be mentioned. A very quiet whisper may result merely from holding the glottis in the voiceless position throughout speech. But the more normal whisper, by means of which we are able to communicate with some ease, involves energetic articulation and considerable stricture in the glottal region. Such a whisper may in fact be uttered with an almost total closure of the glottis and an escape of air in the region of the arytenoid[7] cartilages.

The simplest way of observing the behaviour of the vocal cords is by the use of a laryngoscope, which gives a stationary mirrored image of the glottis. Using stroboscopic techniques, it is possible to obtain a moving record and high-speed films have been made of the vocal cords, showing their action in ordinary breathing, producing voice and whisper, and closed as for a glottal stop. The modern technique of observation is to use fibre-optic endoscopy coupled if required with a tiny videocamera.

2.2.3 The resonating cavities

The airstream, having passed through the larynx, is now subject to further modification according to the shape assumed by the upper cavities of the pharynx and mouth, and according to whether the nasal cavities are brought into use or not. These three cavities function as the principal resonators of the voice produced in the larynx.

2.2.3.1 The pharynx

The pharyngeal[8] cavity (see Fig. 1) extends from the top of the trachea and oesophagus, past the epiglottis and the root of the tongue, to the region at the rear of the SOFT PALATE. It is convenient to identify these sections of the PHARYNX by naming them: laryngopharynx, oropharynx, nasopharynx. The shape and volume of this long chamber may be considerably modified by the constrictive action of the muscles enclosing the pharynx, by the movement of the back of the tongue, by the position of the soft palate which may, when raised, exclude the nasopharynx, and by the raising of the larynx itself. The position of the tongue in the mouth, whether it is advanced or retracted, will affect the size of the oropharyngeal cavity; the modifications in shape of this cavity should, therefore, be included in the description of any vowel. It is a characteristic of some kinds of English pronunciation that certain vowels, e.g. the [a] vowel in *sad*, are

articulated with a strong pharyngeal contraction. Additionally, a constriction may be made between the lower rear part of the tongue and the wall of the pharynx so that friction, with or without voice, is produced, such fricative sounds being a feature of a number of languages, e.g. Arabic.[9]

The pharynx may be observed by means of a laryngoscope or fibre-optic nasendoscopy and its constrictive actions are revealed by MRI. **(See video 4.14–21)**

The escape of air from the pharynx may be effected in one of three ways:

(1) The soft palate may be lowered, as in normal breathing, in which case the air may escape through the nose and the mouth. This is the position taken up by the soft palate in articulation of the French NASALISED vowels in such a phrase as *un bon vin blanc* [œ̃ bõ vɛ̃ blɑ̃], the particular quality of such vowels being achieved through the resonance of the nasopharyngeal cavities. There is no absolute necessity for nasal airflow out of the nose, the most important factor in the production of nasality being the sizes of the posterior oral and nasal openings (some speakers may even make the nasal cavities vibrate through nasopharyngeal mucus or through the soft palate itself).[10]

(2) The soft palate may be lowered so that a NASAL outlet is afforded to the airstream, but a complete obstruction is made at some point in the mouth, with the result that, although air enters all or part of the mouth cavity, no oral escape is possible. A purely nasal escape of this sort occurs in such nasal consonants as [m,n,ŋ] in the English words *ram*, *ran*, *rang*. In a snore and some kinds of defective speech, this nasal escape may be accompanied by friction or a trill between the rear side of the soft palate and the pharyngeal wall.

(3) The soft palate may be held in its raised position, eliminating the action of the nasopharynx, so that the air escape is solely through the mouth. All English sounds, with the exception of the nasal consonants mentioned in (2), usually have this oral escape. Moreover, if for any reason the lowering of the soft palate cannot be effected, or if there is an enlargement of the organs enclosing the nasopharynx or a blockage brought about by mucus, it is often difficult to articulate either nasalised vowels or nasal consonants. In such speech, typical of adenoidal enlargement or the obstruction caused by a cold, the French phrase mentioned above would have its nasalised vowels turned into their oral equivalents and the English word *morning* would have its nasal consonants replaced by [b,d,g] becoming [bɔɪdɪg]. On the other hand, an inability to make an effective closure by means of the raising of the soft palate—either because the soft palate itself is defective or because an abnormal opening in the roof of the mouth gives access to the nasal cavities—will result in the general nasalisation of vowels and the failure to articulate such oral stop consonants as [b,d,g]. This excessive nasalisation (or hypernasality) is typical of such a condition as cleft palate.

The action of the soft palate can be observed by MRI. **(See video 6.4ff.)** The pressure of the air passing through the nasal cavities may be measured at the nostrils or within the cavities themselves.

2.2.3.2 *The mouth*

Although all the cavities so far mentioned play an essential part in the production of speech sounds, most attention has traditionally been paid to the behaviour of the cavity formed by the mouth. Indeed, in many languages the word for 'tongue' is used to refer to our speech and language activity. Such a preoccupation with the oral cavity is due to the fact that it is the most readily accessible and easily observed section of the vocal tract. The shape of the mouth determines finally the quality of the majority of our speech sounds. Far more finely controlled variations of shape are possible in the mouth than in any other part of the speech mechanism.

The only boundaries of the mouth which are relatively fixed are, in the front, the teeth; in the upper part, the hard palate; and, in the rear, the pharyngeal wall. The remaining organs are movable: the lips, the various parts of the tongue and the soft palate with its pendant uvula (see Fig. 1). The lower jaw is capable of very considerable movement; its movement will control the gap between the upper and lower teeth and also to a large extent the disposition of the lips. Movement of the lower jaw is also one way of altering the distance between the tongue and the roof of the mouth.

It is convenient for descriptive purposes to divide the roof of the mouth into three parts: moving backwards from the upper teeth, first, the teeth ridge (adjective: ALVEOLAR)[11] which can be clearly felt behind the teeth; second, the bony arch which forms the hard palate (adjective: PALATAL) and which varies in size and arching from one individual to another; and finally, the soft palate (adjective: VELAR) which, as we have seen, is capable of being raised or lowered, and at the extremity of which is the uvula (adjective: UVULAR).[12] All these parts can be readily observed by means of a mirror.

(1) Of the movable parts, the lips (adjective: LABIAL), constitute the final obstruction to the airstream when the nasal passage is shut off. The shape which they assume affects very considerably the shape of the total cavity. They may be shut or held apart in various ways. When they are held tightly shut, they form a complete obstruction or occlusion to the airstream, which may either be momentarily prevented from escaping at all, as in the initial sounds of *pat* and *bat*, or may be diverted through the nose by the lowering of the soft palate, as in the initial sound of *mat*. If the lips are held apart, the positions they assume may be summarised under five headings:

(a) held sufficiently close together over all their length that friction occurs between them. Fricative sounds of this sort, with or without voice, occur in many languages and the voiced variety [β] is sometimes wrongly used by foreign speakers of English for the first sound in the words *vet* or *wet*;

(b) held sufficiently far apart for no friction to be heard, yet remaining fairly close together and energetically spread. This shape is taken up for vowels like that in *see* when energetically enunciated and is known as the SPREAD lip position;
(c) held in a relaxed position with a lowering of the lower jaw. This is the position taken up for the vowel of *sat* and is known as the NEUTRAL position;
(d) tightly pursed, so that the aperture is small and rounded, as in a precise or energetic enunciation of the vowel of *do*, or more markedly in the French vowel of *doux*. This is the CLOSE ROUNDED position;
(e) held wide apart, but with slight projection and rounding, as in the vowel of *got*. This is the OPEN ROUNDED position.

Variations of these five positions may be encountered, e.g. in the vowel of *saw*, for which a type of lip-rounding between open rounded and close rounded is commonly used. It will be seen from the examples given that lip position is particularly significant in the formation of vowel quality. English consonants, on the other hand, even including [p,b,m,w] whose primary articulation involves lip action, will tend to share the lip position of the adjacent vowel. In addition, the lower lip is an active articulator in the pronunciation of [f,v], a light contact being made between the lower lip and the upper teeth.

(2) Of all the movable organs within the mouth, the tongue is by far the most flexible and is capable of assuming a great variety of positions in the articulation of both vowels and consonants. The tongue is a complex muscular structure which does not show obvious sections; yet, since its position must often be described in considerable detail, certain arbitrary divisions are made. When the tongue is at rest, with its tip lying behind the lower teeth, that part which lies opposite the hard palate is called the FRONT and that which faces the soft palate is called the BACK, with the region where the front and back meet known as the CENTRE (adjective: CENTRAL). These areas together with the ROOT (which forms the front side of the pharynx) are sometimes collectively referred to as the body of the tongue. The tapering section in front of the body and facing the teeth ridge is called the blade (adjective: LAMINAL)[13] and its extremity the tip (adjective: APICAL). The edges of the tongue are known as the rims.

Generally, in the articulation of vowels, the tongue tip remains low behind the lower teeth. The body of the tongue may, however, be 'bunched up' in different ways, e.g. the front may be the highest part as when we say the vowel of *keen* **(see video 11.5)**; or the back may be most prominent as in the case of the vowel in *zoo* **(see video 2.28)**; or the whole surface may be relatively low and flat as in the case of the vowel in *father* **(see video 10.18)**. Such changes of shape can be felt if the above words are said in succession. These changes, together with the variations in lip position, have the effect of modifying very considerably the size of the mouth cavity and of dividing this chamber into two parts: that part of the cavity which is in the forward part of the mouth behind the lips and that which is in the rear in the region of the pharynx.

The various parts of the tongue may also come into contact with the roof of the mouth. Thus, the tip, blade and rims may articulate with the teeth as for the *th* sounds in English **(see video 12.16)**, or with the upper alveolar ridge as in the case of /t,d,s,z,n/ **(see videos 3.15,23, 2.17, 5.16, 6.7)** or the apical contact may be only partial as in the case of /l/ (where the tip makes firm contact whilst the rims make none—see the PALATOGRAM of /l/ in Figures 48 and 49) or intermittent in a trilled /r/ as in some forms of Scottish English. In some languages, notably those of India, Pakistan and Sri Lanka, the tip contact may be retracted to the very back of the teeth ridge or even slightly behind it; the same kind of RETROFLEXION, without the tip contact, is typical of some kinds of English /r/, e.g. those used in south-west England and by some speakers in the U.S.

The front of the tongue may articulate against or near to the hard palate. Such a raising of the front of the tongue towards the palate is an essential part of the [ʃ,ʒ] **(see videos 1.1, 11.23)** sounds in English words such as *show* and *measure*, being additional to an articulation made between the blade and the alveolar ridge; or again, it is the main feature of the [j] sound initially in *yield*.

The back of the tongue can form a total obstruction by its contact with the front side of the soft palate, the nasopharynx being closed in the case of [k,g], e.g. in *guard* **(see video 12.1)** and open for [ŋ] as in *hang* **(see video 10.4)**; or again, there may merely be a narrowing between the soft palate and the back of the tongue, so that friction of the type occurring finally in the Scottish pronunciation of *loch* is heard. And finally, the uvula may vibrate against the back of the tongue, or there may be a narrowing in this region which causes uvular friction, as at the beginning of the French word *rouge*.

It will be seen from these few examples that, whereas for vowels the tongue is generally held in a position which is convex in relation to the roof of the mouth, some consonant articulations, such as the southern British English /r/ in *ride* **(see video 4.2)** and the /l/ in *crawl* **(see video 2.9)**, will involve the 'hollowing' of the body of the tongue so that it has, at least partially, a concave relationship with the roof of the mouth.

Moreover, the surface of the tongue, viewed from the front, may take on various forms: there may be a narrow groove running from back to front down the mid line as for the /s/ in *see*, or the grooving may be very much more diffuse as in the case of the /ʃ/ in *ship*.

(3) The oral speech mechanism is readily accessible to direct observation as far as the lip movements are concerned as are many of the tongue movements which take place in the forward part of the mouth. Although a lateral view of the shape of the tongue over all its length and its relationship with the palate and the velum can be obtained by MRI, it should not be assumed that pictures of the articulation of, say, the vowel in *cat* will show an identical tongue position when pronounced by different people. Not only is the sound itself likely to be slightly different from one individual to another, but, even if the sound is for all practical purposes the 'same', the tongue positions may be different, since the shape of the mouth cavity is not identical for any two speakers; and, in any case,

two sounds judged to be the same may be produced by the same individual with somewhat different articulations. When, therefore, we describe an articulation in detail, it should be understood that such an articulation is typical for the sound in question, but that variations are to be expected.

Palatography, showing the extent of the area of contact between the tongue and the roof of the mouth, has long been a practical and informative way of recording tongue movements. At one time the palate was coated with a powdery substance, the articulation was made, and the 'wipe-off' subsequently photographed. But the modern method uses electropalatography, whereby electrodes on a false palate respond to any tongue contact, the contact points being simultaneously registered on a visual display. This has the advantage of showing a series of representations of the changing contacts between the tongue and the top of the mouth during speech. Electropalatograms of this sort are used to illustrate the articulations of consonants in Chapter 9.

Articulators used in speech

Articulator	*Relevance*
Airstream	Usually pulmonic
Vocal folds	Closed, wide apart, or vibrating
Soft palate	Lowered (giving nasality) or raised (excluding nasality)
Tongue	Back, centre, front, blade, tip and/or rims raised
Lips	Neutral, spread, open-rounded, close-rounded

Notes

1 But see, for example, Fouts & Mills (1999), for the training of chimpanzees to use a rudimentary language (but signed rather than spoken).
2 Stetson (1951).
3 Ladefoged (1967).
4 In this chapter and in Chapters 4, 8 and 9, videos on the companion website are referenced as, for example, video 12.4, where 12 is the number of the video and 4 is the relevant point on the video. See Cruttenden (2013) for some discussion of the use of these videos in teaching.
5 Clark *et al.* (2007: 178).
6 According to the 'Bernoulli Effect'.
7 Pronounced /aˈrɪtənɔɪd/ or /arɪˈtiːnɔɪd/.
8 Pronounced /ˈfarɪŋks/ and /fəˈrɪnʤɪəl/ or /fərɪnˈʤiːəl/.
9 Recent research has shown that the pharyngeal consonants in Arabic are more correctly described as epiglottal (Ladefoged & Maddieson, 1996: 167).
10 Laver (1980: 77ff).
11 Pronounced /alviˈəʊlə/ or /alˈviːələ/.
12 Pronounced /ˈjuːvjuːlə/.
13 Pronounced /ˈlaminəl/ and /ˈeɪpɪkəl/.

Chapter 3

The sounds of speech: the acoustic and auditory aspects

3.1 Sound quality

To complete an act of communication, it is not sufficient that our speech mechanism should produce sounds; these have to be heard and interpreted, after transmission through a medium, normally the air, which is capable of conveying sounds. We now examine the nature of the sounds which we hear, the characteristics of the transmission phase of these sounds and the way in which these sounds are perceived by a listener.

When we listen to a continuous utterance, we perceive an ever-changing pattern of sound. When it is a question of our own language, we are not conscious of all the complexities of pattern which reach our ears: we tend consciously to perceive and interpret only those sound features which are relevant to the intelligibility of our language. Nevertheless, despite this linguistic selection which we ultimately make, we are aware that this changing pattern consists of variations of different kinds: of SOUND QUALITY—we hear a variety of vowels and consonants; of PITCH—we appreciate the melody, or intonation, of the utterance; of AMPLITUDE—some sounds or syllables sound 'louder' than others; and of LENGTH—some sounds will be longer to our ears than others. These are judgements made by a listener about a sound continuum from a speaker and, if the sound from the speaker and the response from the listener are made in the same linguistic system, then the utterance will be meaningful for speaker and listener alike. The transmission phase links the listener's impressions of changes of quality, pitch, amplitude and length with the articulatory activity on the part of the speaker. But an exact correlation between the production, transmission and reception phases of speech is not always easy to establish.

The production of sound requires some kind of energy and frequently this energy makes something vibrate. In the case of human speech the thing vibrating is usually the vocal cords which are energised by air pressure from the lungs. Any such sound produced in the larynx is then modified by the resonating chambers of the pharynx, the mouth and, in some cases, the nasal cavities. The listener's impression of sound quality will be determined by the way in which the speaker's vibrator and resonators function together.

Speech sounds, like other sounds, are conveyed to our ears by means of waves of compression and rarefaction of the air particles (the commonest medium of communication). These variations in pressure, initiated by the action of the vibrator, are propagated in all directions from the source, the air particles themselves vibrating at the same rate (or frequency) as the original vibrator. In speech, these vibrations may be of a complex but regular pattern, producing 'tone' such as may be heard in a vowel sound; or they may be of an irregular kind, producing 'noise', such as in the consonant /s/ or there may be both regular and irregular vibrations present, i.e. a combination of tone and noise, as in /z/. In the production of normal vowels, the vibrator is provided by the vocal cords; in the case of many consonant articulations, however, a source of air disturbance is provided by constriction at a point above the larynx, with or without accompanying vocal cord vibrations.

Despite the fact that the basis of all normal vowels is the glottal tone, we are all capable of distinguishing a large number of vowel qualities. Yet the glottal vibrations in the case of [ɑː] are not very different from those for [iː], when both vowels are said with the same pitch. The modifications in quality which we perceive are due to the action of the supraglottal resonators which we have previously described. To understand this action, it is necessary to consider a little more closely the nature of the glottal vibrations.

It has already been mentioned that the glottal tone is the result of a complex, but mainly regular, vibratory motion. In fact, the vocal cords vibrate in such a way as to produce, in addition to a basic vibration over their whole length (the FUNDAMENTAL FREQUENCY), a number of overtones or HARMONICS having frequencies which are simple multiples of the fundamental or first harmonic. Thus, if there is a fundamental frequency of vibration of 100 Hz, the upper harmonics will be of the order of 200, 300, 400 Hz. Indeed, there may be no energy at the fundamental frequency, but merely the harmonics of higher frequency such as 200, 300, 400 Hz. Nevertheless, we still perceive a pitch which is appropriate to a fundamental frequency of 100 Hz, i.e. the fundamental frequency is the highest common factor of all the frequencies present, whether or not it is present itself.

The number and strength of the component frequencies of this complex glottal tone will differ from one individual to another and this accounts at least in part for the differences of voice quality by which we are able to recognise speakers. But we can all modify the glottal tone so as to produce vowels as different as [iː] and [ɑː], so that despite our divergences of voice quality we can convey the distinction between two words such as *key* and *car*. This variation of quality, or timbre, of the glottal tone is achieved by the shapes which we give the resonators above the larynx—the pharynx, the mouth and the nasal cavities. These chambers are capable of assuming a very large number of shapes, each of which will have a characteristic vibrating resonance of its own. Those harmonics of the glottal tone which coincide with the chamber's own resonance are very considerably amplified. Thus, certain bands of strongly reinforced harmonics are characteristic

of a particular arrangement of the resonating chambers which produces, for instance, a certain vowel sound. Moreover, these bands of frequencies will be reinforced whatever the fundamental frequency. In other words, whatever the pitch on which we say, for instance, the vowel [ɑː], the shaping of the resonators and their resonances will be very much the same, so that it is still possible, except on extremely high or low pitches, to recognise the quality intended. It is found that, for male speakers, the vowel [iː] has one such characteristic band of strong components in the region of 280 Hz and another at about 2,200 Hz, while for female speakers these bands of energy are at about 300 Hz and 2,700 Hz (see §8.6).

3.2 The acoustic spectrum

This complex range of frequencies of varying intensity which go to make up the quality of a sound is known as the ACOUSTIC SPECTRUM, those bands of energy which are characteristic of a particular sound are known as the sound's FORMANTS. Thus, formants of [ɑː] are said to occur, for female speakers, in the regions around 700 and 1,300 Hz.

Such complex waveforms can be analysed and displayed as a SPECTROGRAM (see Fig. 3). Originally this display required a special instrument, a spectrograph, but nowadays it is generally done by computer. The spectrogram consists of a three-dimensional display: frequency is shown on the vertical axis, time on the horizontal axis and the energy at any frequency level either by the density of blackness in a black and white display, or by colours in a colour display. Thus the concentrations of energy at particular frequency bands (the formants) stand out very clearly. Fig. 3 shows, in the spectrogram of *Manchester music shops*, the extent to which utterances are not neatly segmented into a succession of sounds but, on the contrary, there is considerable overlap. Such spectrographic analysis provides a great deal of acoustic information in a convenient form.

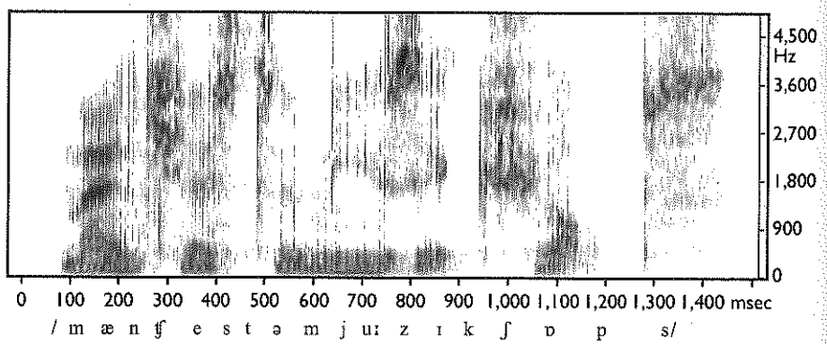

Figure 3 Spectrogram of the phrase *Manchester music shops* as said by a male speaker of GB.

Much of the information given on a spectrogram is, in fact, irrelevant to our understanding of speech and we need to establish which elements of the spectrum are essential to speech communication. For instance, two, or at most three, formants appear to be sufficient for the correct identification of vowels. As far as the English vowels are concerned, the first three formants are all included in the frequency range 0–4,000 Hz, so that the spectrum above 4,000 Hz would appear to be largely irrelevant to the recognition of our vowels. In both older (analogue) telephone systems with a frequency range of approximately 300–3,000 Hz and in newer (digital) systems with a range up to around 4,000 Hz, we have little difficulty in identifying the sound patterns of speakers and even in recognising individual voice qualities. Indeed, when we are dealing with a complete utterance in a given context, where there is a multiplicity of cues to help our understanding, a high degree of intelligibility may be retained even when there are no frequencies above 1,500 Hz.

As one would suspect, there appear to be certain relationships between the formants of vowels and the cavities of the vocal tract (i.e. the shapes taken on by the resonators, notably the relation of the oral and pharyngeal cavities). Thus, the first formant appears to be low when the tongue is high in the mouth: e.g. [iː] and [uː] have high tongue positions and have first formants for both men and women around 280–330 Hz, whereas [ɑː] and [ɒ] have their first formants in the region 600–800 Hz, their tongue positions being relatively low. On the other hand, the second formant seems to be inversely related to the length of the front cavity: thus [iː], where the tongue is raised high in the front of the mouth, has a second formant around 2,200–2,700 Hz, whereas [uː], where the tongue is raised at the back of the mouth (and lips are rounded), has a relatively low second formant around 1,200–1,400 Hz. (For averages for the first and second formants of all GB vowels for male and female speakers see Table 3 and Figs 10 and 11 in Chapter 8.)

It is also confirmed from spectrographic analysis that a diphthong, such as that in *my*, is indeed a glide between two vowel elements (reflecting a perceptible articulatory movement), since the formants bend from those positions typical of one vowel to those characteristic of another (see Table 4 and Fig. 9 in Chapter 8).

For many consonant articulations (e.g. the initial sounds in *pin*, *tin*, *kin*, *thin*, *fin*, *sin*, *shin*, in which the glottal vibrations play no part) there is an essential noise component, deriving from an obstruction or constriction within the mouth, approximately within the range 2,000–8,000 Hz (see Chapter 9, Fig. 32). This noise component is also present in the corresponding articulations in which vocal fold excitation is present, as in the final sounds of *ruse* and *rouge*, where we are dealing with sounds which consist of a combination of glottal tone and noise. Detailed acoustic data concerning vowel and consonant articulations in GB is given in Chapters 8 and 9.

Spectrographic analysis also reveals the way in which there tends, on the acoustic level, to be a merging of features of units which, linguistically, we treat separately. Thus, our discrimination of [f] and [θ] sounds depends only partly

on the frequency and duration of the noise component but also upon a characteristic bending of the formants of the adjacent vowel. Indeed, in the case of such consonants as [p,t,k], which involve a complete obstruction of the airstream and whose release is characterised acoustically by only a very brief burst of noise, the TRANSITION between the noise of the consonant and the formant structure of the vowel appears to be of prime importance for our recognition of the consonant (see §9.2.2(3) and Fig. 32).

3.2.1 *Fundamental frequency: pitch*

Our perception of the pitch of a speech sound depends directly upon the frequency of vibration of the vocal cords. Thus, we are normally conscious of the pitch caused by the 'voiced' sounds, especially vowels; pitch judgements made on voiceless or whispered sounds, without the glottal tone, are limited in comparison with those made on voiced sounds, but may be induced by producing friction in the larynx or pharynx and varying the shape of the resonating cavities in the usual way.

The higher the glottal fundamental frequency, the higher our impression of pitch. A male voice may have an average pitch level of about 120 Hz and a female voice a level in the region of 220 Hz.[1] The pitch level of voices, however, will vary a great deal between individuals and also within the speech of one speaker, the total range of one male speaking voice being liable to have a range of anything between 80 and 350 Hz. Following adolescence men's voices become lower until the forties after which they rise continuously into old age; women's voices generally become lower up to the time of the menopause, remain much the same for twenty years or so and then possibly rise slightly.[2] Yet our perception of frequency extends further than the limits of glottal fundamental frequency, since our recognition of quality depends upon frequencies of a much higher order. In fact, the human ear perceives frequencies from as low as 16 Hz to about 20,000 Hz and in some cases even higher. As one becomes older, this upper limit may fall considerably, so that at the age of 50 it may extend no higher than about 10,000 Hz. As we have seen, such a reduced range is no impediment to perfect understanding of speech, since a high percentage of acoustic cues for speech recognition fall within the range 0–4,000 Hz.

Our perception of pitch is not, however, solely dependent upon fundamental frequency. Variations of intensity on the same frequency may induce impressions of a change of pitch; and conversely, tones of very high or very low frequency, if they are to be audible at all, require greater intensity than those in a middle range of frequencies.[3]

Instrumental measurement of fundamental frequency based on signals received through a microphone employs two general methods. The first is to count the number of times that a particular pattern is repeated within a selected segment of a waveform such as that provided on an oscillogram. The second is to track the progress of the fundamental frequency on a spectral display like that provided

on a spectrogram, or, alternatively, to track the progress of a particular harmonic and divide by the relevant number. Nowadays various computer programs are available[4] which average the results from a range of measurements based on the two general methods noted above. But even with such sophisticated programs there are still likely to be occasional mistakes like octave jumps (each doubling in Hz representing an octave).

A third method of fundamental frequency extraction involves direct measurement of the vibration of the vocal cords either by glottal illumination or by electroglottography. One well-known technique in the latter class involves using a LARYNGOGRAPH.[5] With this technique electrodes are attached to the outside of the throat and the varying electrical impedance is monitored and projected onto a visual display. The signal generated by the variation in impedance can also be stored, enabling this technique to be used outside the laboratory.

Measures of fundamental frequency do not always correspond to our auditory perception of pitch. Besides the dependence on intensity mentioned above, different segments affect the fundamental frequency in different ways: for example, other things being equal, an [i] will have a higher fundamental frequency than an [a], and a [p] will produce a higher frequency on a following vowel than a [b]. Such (slight) changes in frequency will generally be undetectable by the ear. As in many other cases of instrumental measurement, we still have to use our auditory perception to interpret what instruments tell us.

3.2.2 *Intensity: loudness*

Our sensation of the relative loudness of sounds may depend on several factors, e.g. a sound or syllable may appear to stand out from its neighbours—be 'louder' —because a marked pitch change is associated with it or because it is longer than its neighbours. It is better to use a term such as PROMINENCE to cover these general listener impressions of variations in the perceptibility of sounds. More strictly, what is 'loudness' at the receiving end should be related to INTENSITY at the production stage, which in turn is related to the size or AMPLITUDE of the vibration. An increase in amplitude of vibration, with its resultant impression of greater loudness, is brought about by an increase in air pressure from the lungs. As we shall see (§10.2), this greater intensity is not in itself usually the most important factor in rendering a sound prominent in English. Moreover, all other things being equal, some sounds appear by their nature to be more prominent or sonorous than others, e.g. the vowel in *barn* has more carrying power than that in *bean* and vowels generally are more powerful than consonants.

The judgements we make concerning loudness are not as fine as those made for either quality or pitch. We may judge which of two sounds is the louder, but we find it difficult to express the extent of the difference. Indeed we generally perceive and interpret only gross differences of overall loudness, despite the fact that we recognise different vowels by reacting to characteristic regions of intensity in the spectrum.

3.2.3 Duration: length

In addition to affording different auditory impressions of quality, pitch and loudness, sounds may appear to a listener to be of different length. Clearly, whenever it is possible to establish the boundaries of sounds or syllables, it will be possible to measure their duration on traces provided by oscillograms or spectrograms. Such delimitation of units, in both the articulatory and acoustic sense, may be difficult, as we shall see when we deal with the segmentation of the utterance. But, even when it can be done, variations of duration in acoustic terms may not correspond to our linguistic judgements of length. We shall, for instance, refer later to the 'long' vowels of English such as those of *bean* and *barn*, as compared with the 'short' vowel in *bin*. But, in making such statements, we shall not be referring to absolute duration values, since the duration of all vowels will vary considerably from utterance to utterance, according to factors like whether the utterance is spoken rapidly or slowly, whether the syllable containing the vowel is accented or not and whether the vowel is followed by a voiced or voiceless consonant. In the English system, however, we know that no more than two degrees of length are ever linguistically significant and all absolute durations will be interpreted in terms of this relationship. This distinction between measurable duration and linguistic length provides another example of the way in which our linguistic sense interprets from the acoustic material only that which is significant.

The sounds comprising any utterance will have varying durations and we will have the impression that some syllables are longer than others. Such variations of length within the utterance constitute one manifestation of the rhythmic delivery which is characteristic of English and so is fundamentally different from the flow of other languages, such as French, where syllables tend to be of much more even length.

As already mentioned, the absolute duration of sounds or syllables will depend, among other things, upon the speed of utterance. An average rate of delivery might contain anything from about 6 to 20 sounds per second, but lower and much higher speeds are frequently used without loss of intelligibility. The time required for the recognition of a sound will depend upon the nature of the sound and its pitch, vowels and consonants differing considerably in this respect, but it seems that a vowel lasting only about 4 msecs may have a good chance of being recognised.

3.2.4 'Stress'

We have purposely avoided the use of the word 'stress' in this chapter because this word has been used in different and ambiguous ways in phonetics and linguistics. It has sometimes been used as simply equivalent to loudness, sometimes as meaning 'made prominent by means other than pitch' (i.e. by loudness or length) and sometimes as referring just to syllables in words in the lexicon

and meaning something like 'having the potential for accent on utterances'. Throughout this book we will avoid use of the term 'stress' altogether, using prominence as the general term referring to segments or syllables, SONORITY as the particular term referring to the carrying power of individual sounds and ACCENT as referring to those syllables which stand out above others, either in individual words or in longer utterances.

3.3 Hearing

Our hearing mechanism must be thought of in two ways: the physiological mechanism which reacts to the acoustic stimuli—the varying pressures in the air which constitute sound; and the psychological activity which, at the level of the brain, selects from the gross acoustic information that which is relevant in terms of the linguistic system involved. In this way, measurably different acoustic stimuli may be interpreted as being the 'same' sound unit. As we have seen, only part of the total acoustic information seems to be necessary for the perception of particular sound values. One of the tasks which confront the phonetician is the disentanglement of these relevant features from the mass of acoustic material that modern methods of sound analysis make available. The most fruitful technique for discovering the significant acoustic cues is that of SPEECH SYNTHESIS, controlled by listeners' judgements. After all, the sounds [ɑː] and [s] are [ɑː] and [s] only if listeners recognise them as such. Thus, it has been established that only two formants are necessary for the recognition of vowels, because machines which generate sound of the appropriate frequency bands and intensity produce vowels which are correctly identified by listeners.

Listeners without any phonetic training can, therefore, frequently give valuable guidance by their judgements of synthetic qualities. But it is important to be aware of the limitations of such listeners, so as to be able to make a proper evaluation of their judgements. A listener's reactions are normally conditioned by his experience of handling his own language. Thus, if there are only five significant vowel units in his language, he is liable to allow a great deal more latitude in his assessment of what is the 'same' vowel sound than if he has twenty. An Englishman, for instance, having a complex vowel system and being accustomed to distinguishing such subtle distinctions as those in *sit*, *set*, *sat*, will be fairly precise in his judgement of vowel qualities. A Spaniard, however, whose vowel system is made up of fewer significant units, is likely for this reason to be more tolerant of variation of quality. Or again, if a listener is presented with a system of synthetic vowels which is numerically the same as his own, he is able to make allowance for considerable variations of quality between his and the synthetic system and still identify the vowels correctly—by their 'place' in a system rather than by their precise quality; this is partly what he does when he listens to and understands his language as used by a speaker of a different dialect.

Our hearing mechanism also plays an important part in monitoring our own speech; it places a control upon our speech production which is complementary

to our motor, articulatory, habits. If this feedback control is disturbed, e.g. by the imposition of an artificial delay upon our reception of our own speech, disturbance in the production of our utterance is likely to result. Those who are born deaf or who become deaf before the acquisition of speech habits are rarely able to learn normal speech completely; similarly, a severe hearing loss later in life is likely to lead eventually to a deterioration of speech, although not down to the same level as those born deaf.

Notes

1 Fant (1956).
2 Hollien & Shipp (1972); Russell *et al.* (1995).
3 Denes & Pinson (1993: 101ff).
4 For example, *Praat* (Boersma & Weenink, 2011).
5 Abberton & Fourcin (1984).

Chapter 4

The description and classification of speech sounds

4.1 Phonetic description

We have considered briefly both the mechanism which produces speech sounds and also some of the acoustic and auditory characteristics of the sounds themselves. We have seen that a speech sound has at least three stages available for investigations—the production, transmission and reception stages. The most convenient and brief descriptive classification of speech sounds relies either on articulatory criteria or on auditory judgements, or on a combination of both. Those sounds which are commonly known as consonants are most easily described mainly in terms of their articulation, whereas the description of vowels depends more on auditory impressions.

4.2 Vowel and consonant

Two types of meaning are associated with the terms VOWEL and CONSONANT. In one type of definition consonants are those segments which, in a particular language, occur at the edges of syllables, while vowels are those which occur at the centre of syllables. So, in *red*, *wed*, *dead*, *lead*, *said*, the sounds represented by <r,w,d,l,s> are consonants, while in *beat*, *bit*, *bet*, *but*, *bought*, the sounds represented by <ea,i,e,u,ou> are vowels. This reference to the functioning of sounds in syllables in a particular language is a phonological definition. But once any attempt is made to define what sorts of sounds generally occur in these different syllable-positions, then we are moving to a phonetic definition. This type of definition might define vowels as median (air must escape over the middle of the tongue, thus excluding the lateral [l]), oral (air must escape through the mouth, thus excluding nasals like [n]), frictionless (thus excluding fricatives like [s]), and continuant (thus excluding plosives like [p]); all sounds excluded from this definition would be consonants. But difficulties arise in English with this definition (and with others of this sort) because English /j,w,r/, which are consonants phonologically (functioning at the edges of syllables), are vowels phonetically. Because of this these sounds are often called semi-vowels. The reverse type of difficulty is encountered in words like *sudden* and *little* where