5
Phonetics
The sounds humans make when speaking
Andrew Butcher

5.1 Introduction
Phonetics is traditionally defined as the scientific study of speech sounds, their articulation, transmission and reception. It deals with substance – with physical events which take place in time. Phonology, on the other hand, is the study of how speech sounds are used in language. It deals with form – with mental targets and symbolic representations. I use the term phonology in the same way as Hulst (see Chapter 6) uses the term ‘phonemics’. Unlike Hulst, I regard phonetics and phonology as separate disciplines and thus I have no overarching term to encompass the two. But there can be no doubt that they are inextricably linked, and a central question for phonetic research (sometimes known as ‘the invariance question’) concerns the nature of the phonetics/phonology interface: how are the discrete, static, context-free mental targets (phonemes) translated by the speaker into a continuous, dynamic, context-sensitive stream of sound and how does the listener retrieve those same mental targets from the continuous stream? Traditional descriptive phonetics relies on the fact that human beings are capable of doing this, and that literate speakers of languages with alphabetic writing systems in particular become aware of phonemes at an early age. The objects of our description are thus chunks of the speech stream which we perceive as corresponding to phonemes. That this is not as straightforward as it seems becomes apparent if one attempts to discover such chunks in a language one does not speak.

Phonetics is probably the oldest of the linguistic sciences, with a tradition that goes back at least to the India of the fifth century BCE and the work of Pāṇini on the sounds of Sanskrit. Interest in systems of notation continued in Europe in the eighteenth and nineteenth centuries, often with a prescriptive goal. Modern experimental phonetics began towards the end of the nineteenth century with the invention of the phonograph. The discipline is often divided into the subfields of articulatory phonetics – concerning itself principally with the physiology of speech production; acoustic phonetics – concentrating on the transmission of the speech signal through the air; and auditory phonetics – investigating the perception of speech by human listeners. But this is somewhat misleading: no serious researcher in either speech production or speech perception would work without knowledge of or reference to
acoustic phonetics. Similarly the study of acoustic phonetics for its own sake would seem to be a somewhat esoteric exercise.

5.2 The basics

Traditional descriptive phonetics is based on how speech sounds are produced in physiological terms, rather than on their measurable acoustic qualities. An assumption is made that we can identify stable and repeatable patterns of simultaneous gestures within an utterance which correspond to single phonemes in the speaker’s brain. We distinguish between these sound segments and the prosodies – mainly variations in pitch, loudness and duration – which can extend over longer stretches of speech. In acoustic phonetics the same assumptions are made regarding patterns in a wave form or spectrogram and we now have algorithms which will automatically segment large corpora of speech signals on this basis.

Much speech perception research is similarly focused on investigating how we perceive individual sounds. Ironically a great deal of research is also devoted to ‘connected speech processes’ – i.e. what happens when we put the putative sound units back together in a natural utterance. As a general rule, the description of any speech sound comprises three components (although not all will be explicitly stated): the airstream mechanism, the state of the vocal folds and the position of the articulators within the vocal tract – the pharynx and oral and nasal cavities.

5.2.1 Airstream mechanisms – the powerhouse of speech

In order to produce audible speech a power source is needed and this comes in the form of a moving stream of air. This is sometimes known as initiation. All the languages of the world use a system of initiation which relies on the movement of air outwards from the lungs – the egressive pulmonic airstream mechanism. A vanishingly small number of languages (perhaps two) have been claimed to use an ingressive pulmonic mechanism to realise linguistically contrastive sounds, although individual speakers in many cultures may produce isolated syllables on an intake of breath. About a quarter of the world’s languages, however, require the use of additional, non-pulmonic mechanisms to intersperse linguistically contrastive consonant sounds within the basic pulmonic egressive stream. One such mechanism involves a vertical movement of the closed larynx, either down, to draw air in, or up, to push air out. This is the glottalic airstream mechanism. Ingressive sounds produced in this way are known as implosives and occur in about 10 per cent of the world’s languages, mainly in equatorial Africa and in south-east Asia. Glottalic egressive sounds are called ejectives; these are found in another 18 per cent or so of the world’s languages. They are particularly common in the Americas and also in the languages of the Caucasus. A third mechanism achieves inward air flow by using the tongue as a suction pad. This is traditionally known as the velaric ingressive airstream mechanism, but the term velaric is increasingly being replaced by the more appropriate term lingual. It is a mechanism commonly used paralinguistically in many cultures (English tut-tut – expression of disapproval), but used to signal linguistic contrasts in only about 2 per cent of the world’s languages – all of them in southern and eastern Africa. Sounds made in this way are commonly called clicks. There are thus four airstream mechanisms in common use linguistically.

But the universal basic powerhouse for human speech is the respiratory system. This system, consisting of the lungs, ribcage, diaphragm and abdominal muscles, has the primary function of keeping the body alive by replenishing the oxygen in the bloodstream and
removing carbon dioxide. The lungs themselves are two spongy masses without any muscles. Expanding and contracting them in order to move air in and out is thus dependent on there being a vacuum between the outside of the lung wall and the inside of the ribcage and diaphragm; when the latter move, the lungs must move with them. Breathing in is an active process, requiring the contraction of various muscles to increase the volume of the chest cavity. There are two main mechanisms for achieving this: we can contract and thereby flatten the diaphragm, which is dome shaped when at rest, or we can raise the ribcage relative to the backbone by using the external intercostal muscles (it is easy to learn to control these two mechanisms separately). When the volume of the lungs increases, the pressure within becomes lower than the prevailing atmospheric pressure and air flows into the lungs until the pressures are equalised. The average total lung capacity of human adults is about 6 litres in men and about 4.5 litres in women; the vital capacity, which is the amount of air which can be moved in and out of the lungs, is about 1.5 litres less than this in each case. During tidal (relaxed) breathing, we utilise only a fraction of this capacity – about half a litre – breathing in until our lungs are about 45 per cent full and breathing out again until they are about one-third full – our resting expiratory level. When speaking, we commonly breathe in about twice as much air as this and we tend to do so much more quickly. The times spent breathing in and out – the inspiratory and expiratory fractions – in tidal breathing are approximately equal (in 40–50 per cent, out 50–60 per cent); the corresponding fractions in conversational speech are extremely variable, but average out to about 15 per cent of the time spent breathing in and about 85 per cent of the time speaking. These two facts about the timing of breathing (greatly increased variability and greatly increased expiratory fraction in speech) indicate that the respiratory mechanism becomes very much subject to the requirements of the task of speaking. In tidal breathing, expiration is achieved through relaxation pressure – largely driven by the elastic recoil of the soft tissues with some help from gravity. In speech the descent of the ribcage may be braked by the action of the external intercostal muscles and the diaphragm held in the relaxed convex position by the action of the major abdominal muscles (external obliques, transverse abdominis, rectus abdominis) compressing the internal organs below it. The outward flow of air from the lungs is thus very precisely controlled by a delicate balance of (active) muscular pressure and (passive) relaxation pressure. The imperative for this precise control is to maintain a constant overpressure beneath the vocal folds. The vibration of the vocal folds is known as phonation and this is the main generator of the sound waves which carry speech. In order for phonation to begin the pressure below the folds (subglottal pressure) must be between 200 and 1,000 Pascals (Pa) greater than the pressure above (depending on the desired pitch). A somewhat lower pressure differential is required to maintain phonation once it has been initiated. Increased loudness is achieved through increasing the volume velocity of air from the lungs. Normally this will also result in an increase in pitch (frequency of vibration). However, fine control of this parameter is performed by the intrinsic muscles of the larynx. We return to this in the following section.

The production of a consonant using the glottalic airstream mechanism involves three stages:

1. the enclosure of a part of the vocal tract by tightly closing three valves – the vocal folds, the soft palate against the pharyngeal wall and an articulatory closure forward of this, involving the tongue or lips;
2. raising or lowering the larynx in order to decrease or increase the volume of this enclosed cavity; and
3. the release of the articulatory closure, allowing air to flow from the area of high pressure to the area of low pressure.

Ejective sounds involve raising the tightly closed larynx by some 5–10 mm. The resulting pressure within the oral cavity varies from language to language and from speaker to speaker, with the more posterior articulations resulting in smaller cavities and higher pressures. Bilabial articulations can have pressures as low as 500 Pa, while pressures as high as 900 Pa have been measured in uvular articulations. It will be obvious that this mechanism is incompatible with vocal fold vibration (voicing). Implosive sounds, however, are commonly voiced, as lowering the larynx, as well as decreasing the pressure in the mouth, also increases the pressure in the trachea, causing air to leak through the glottis and set the vocal folds in motion, if the latter are not tightly closed. This means that the intraoral pressure decrease is usually minimal and often non-existent, with no air flow in or out, leading to the conclusion that voiced implosives are perhaps distinguished from their pulmonic counterparts by ‘lack of explosion’ rather than by ingressive air flow.

The velaric airstream mechanism also involves the enclosure of a small space within the vocal tract, with subsequent volume change and articulatory release. In this case the posterior boundary of the space is a closure between the tongue dorsum and the soft palate or velum – hence the traditional term for the mechanism. In a very few languages there is a contrast between clicks made with a closure at the front part of the velum and those made with a closure further back, at the uvula. The anterior boundary is formed by the front part of the tongue, which is positioned anywhere between the front of the hard palate and the back of the teeth, with a complete seal between both lateral edges of the tongue and the upper teeth and gums. In fact at the start of the process there is typically no space at all between the tongue and the roof of the mouth. A cavity is then formed by lowering the centre of the tongue whilst leaving the seals intact. Within this tiny cavity there is very high degree of negative pressure (−1,500 to −2,000 Pa), so that the subsequent release of the articulators results in an inrush of air and a relatively loud click sound. Since the main initiator of the air flow for these sounds is the tongue, it seems more logical to refer to this as the lingual airstream mechanism, but if we do so, we must acknowledge that bilabial clicks are made in a different way, using the cheeks rather than the tongue as the initiator, making this a buccal airstream mechanism. To change the volume of the cavity the cheeks may be sucked in (an ‘air kiss’) or blown out (French pouah – the so-called ‘audible shrug’) before the release of the lip closure. The small size and anterior location of the cavity means that the remainder of the vocal tract is free to produce a variety of accompaniments, including various types of phonation and nasality.

5.2.2 Phonation – the sound source of speech

The larynx has a far more important and widely occurring function in speech than acting as an airstream initiator. It is the means by which the audible tone is generated that is the main carrier of acoustic information for the listener. Phonation occurs when the air flow generated by the pulmonic egressive airstream mechanism passes through the glottis – the space between the vocal folds, which are situated within the larynx. The larynx is a cartilaginous structure, suspended from the hyoid bone, which is anchored by a number of muscles within the neck. Its primary functions are to prevent foreign objects from entering the lungs and to provide abdominal fixation in muscular exertion, such as required for lifting heavy objects and for childbirth. The suprahyoid muscles contract to move the larynx upwards for glottalic
egressive sounds and the infrahyoid muscles (the so-called ‘strap’ muscles) pull the larynx down for glottalic ingressive sounds. It consists of nine cartilages in all, but in describing phonation we need only refer to four: the ring-shaped cricoid cartilage, positioned at the top of the trachea, the shield-shaped thyroid cartilage which sits ‘astride’ the cricoid, and the paired arytenoid cartilages which are positioned within the larynx on the posterior rim of the cricoid. The vocal folds are symmetrical folds of mucous membrane, about 1.25–2.50 cm in length, running from front to back across the larynx, attaching posteriorly to the vocal processes of the arytenoid cartilages and anteriorly to the inner surface of the thyroid. If they are held lightly together, the egressive pulmonic airstream will raise the subglottal pressure until they are forced apart. As the air passes between the vocal folds they are drawn together again by a combination of tissue elasticity and reduced air pressure until they are once again touching. At this point the pressure begins to increase again and the whole cycle repeats itself.

The rate and the mode of vibration is controlled by the intrinsic laryngeal muscles. In order to increase the rate of vibration, the folds must be stiffened. This is achieved in two ways: the muscles running within the folds – known as the vocalis muscles – can be tensed, thereby stiffening the ‘body’ of the fold itself, or the muscles running between the cricoid and thyroid cartilages at the front of the larynx (the cricothyroid muscles) can be contracted. This has the effect of pulling the two major cartilages together anteriorly, pivoting at the cricothyroid joint, thereby stretching the mucous membrane ‘cover’ of the folds. Normal or modal voice requires relatively high tension in the body and less tension in the cover of the vocal folds. We perceive the rate of vibration of the vocal folds as the pitch of the voice. Languages use pitch variation in two main ways: in all languages pitch variation is used to signal differences in meaning at the level of the phrase or sentence, such as differences between statements and questions, differences in focus on words within the sentence and differences in syntactic structure. In addition to this function, about half the world’s languages – mainly concentrated in sub-Saharan Africa, south-east Asia and parts of the Americas – use pitch variation to signal lexical or grammatical differences between words (so-called ‘tone languages’). If the vocalis muscles are relaxed and a high degree of longitudinal tension is applied to the vocal fold covers by the cricothyroid muscles alone, the folds will no longer vibrate in the usual modal register, but will move into falsetto register, with the stretched edges vibrating like strings, rather than closing fully at each cycle. This is used in singing, but rarely in speech and never contrastively in language.

Other modes of vibration – producing different voice qualities or phonation types – are controlled by muscles at the back and sides of the larynx. In particular, the interarytenoid muscles pull the arytenoid cartilages together along the rim of the cricoid, applying adductive tension to the folds; the lateral cricoarytenoid muscles swivel the arytenoids on the cricoid rim, bringing the vocal processes together, applying medial compression to the folds. The only muscles which abduct (open) the vocal folds are the posterior cricoarytenoids, which pull the arytenoids apart along the cricoid rim. Full abduction is required for voiceless consonants, especially if aspirated; about two-thirds of the world’s languages make a contrast between consonants made with modal voice and consonants made with abduction of the vocal folds (voiced and voiceless respectively). Partial abduction with little longitudinal tension may also be used during phonation to create a breathy voice quality (also known as murmur). This is used to contrast with modal voice in some languages in vowels and nasals and, more rarely, in the release of stops. A high degree of adductive tension coupled with high medial compression produces a phonation type known as creak (also laryngealisation or vocal fry). With the arytenoids tightly together, only the front, ligamental portion of the
vocal folds can vibrate, and only then when the longitudinal tension is very low – leading to a lower frequency of vibration. A number of languages contrast creaky voice with modal voice in vowels and sonorants. A few languages make a three-way contrast between, breathy, creaky and modal voiced vowels. Very high adductive tension and medial compression will result in complete glottal closure – a glottal stop. As with other glottal ‘articulations’ (see §5.2.3) this may be analysed as a segment in some languages and as a prosody in others.

5.2.3 Articulation – the shaping of sounds

Once a flow of air has been generated – and in the majority of cases modulated by the vibration of the vocal folds – it will be modified by its passage through the vocal tract. The nature of this modification allows us to define two major categories of sound in languages: consonants and vowels. Consonants are defined as sounds made by forming an identifiable constriction somewhere in the vocal tract. We can describe or label such sounds in terms of where the constriction is made – the place of articulation – and how it is made – the manner of articulation – as well as specifying the accompanying state of the glottis – most often simply voiced or voiceless. A consonant can be articulated anywhere between the glottis and the lips and its place of articulation is generally identified in relation to a nearby relatively stable landmark in the vocal tract, as shown in Figure 5.1. In some cases it is necessary to refer to the part of the tongue involved in the articulation or even its overall shape. Sounds made with a constriction between both lips (1 in Figure 5.1) – as both consonants in the English word map – are known as bilabial; sounds made with the tongue tip against the upper incisors (2) – as in thigh – are known as dental and sounds made by

![Figure 5.1 Schematic sagittal cross-section of the human vocal tract, showing places of articulation](image-url)
pressing the lower lip against the upper teeth (as both consonants in five) are known as
labiodental. The bony ridge in which the teeth sit is known as the alveolar ridge and sounds
which are made with the tongue tip articulating against the part of this ridge just behind the
upper incisors (3) are called alveolar sounds; there are a quite a few of these in English,
including all of the consonants in the words nose, tide and loose. Sounds made articulating the
body of the tongue against the bony part of the roof of the mouth – the hard palate (4) – are
called palatal sounds. The only true palatal sound in English is the initial consonant of the
word yes, but other palatal sounds occur in European languages – such as in French (agneau),
Italian (gli) and German (ich). However, English does have consonants which are made at the
very front of the palate, just behind the alveolar ridge; these sounds are known as palato-
alveolar and are found in words such as ash and edge. Another type of sound is also made in
this area, but involves the tip of the tongue (or even the underside) bent back to the front of the
palate; these sounds are known as retroflex and are not found in British or Australian English.
They occur in many Indian languages (Hindi, Tamil), including Indian English, which
pronounces t, d, n and l in this way. The soft posterior portion of the palate is known as the
velum (5); sounds which are articulated here are called velars (both consonants in English
gang). Further back still, the back of the tongue can be articulated against the end of the soft
palate (6) to form uvular consonants. These sounds are not found in English, but the r sounds
in standard French and German are both pronounced at this place of articulation (French rue,
German roh). If further differentiation is needed, the tip and blade of the tongue can be
distinguished from the tongue body (the blade is the part of the tongue extending about 15–
20mm back from the tip, and normally lying beneath the alveolar ridge when the tongue is at
rest). Sounds made with the tongue tip (7) are apical; sounds made with the blade (8) are
laminal; and sounds made with the tongue body (9) are dorsal. A limited range of sounds can
be articulated by narrowing the upper pharynx (10) or articulating the epiglottis (11) against
the rear wall of the pharynx. These pharyngeal and epiglottal sounds are not found in English;
they sound very similar and are not in contrast with one another in any language. They occur
in Hebrew and Arabic (the word for the Arabic lute, ‘ud, begins with a pharyngeal sound). As
mentioned above, the vocal folds (12) are sometimes regarded as an articulator, as in the
production of h (air blown through an open vocal tract), the voiced version of this sound
(breathy voice in an open vocal tract), and the glottal stop.

At each place of articulation, a number of different types of constriction can be made. To
begin with, there are three degrees of narrowness which can be employed. These are
definable in aerodynamic terms. A stop is made with complete closure of the articulators,
completely interrupting the flow of air from the mouth. Air pressure builds up behind the
closure and is audibly released when the articulators are parted. A plosive is a pulmonic
egressive oral stop, but the two terms are often used interchangeably (all the consonants in
bide, pike and gate). A fricative requires ‘close approximation’ of the articulators, which
means that they must be close enough together for the prevailing airstream to become
turbulent – i.e. produce audible friction – at the point of constriction (all the consonants in
five, size, thigh and shy). If the airstream is slightly reduced or the constriction is slightly
wider, such that no friction is produced, this is said to be ‘open approximation’ and the result
is an approximant (all the consonants in yell, row and war). However, manner of articulation
is not just a question of degree of constriction; there are a number of other variables which
cut across these categories to produce further sound distinctions. Chief among these is the
oral/nasal distinction – which in theory can be applied to all articulations from lips to uvula,
but in languages is found chiefly amongst stops and, in some cases, vowels. The velum can
be raised to close off the nasal cavity from the mouth or it can be lowered to allow air from
the lungs to pass through the nose. Oral sounds are made with a raised velum; nasal sounds are made with a lowered velum. Nasal stops are commonly referred to simply as nasals; there are three such sounds in English (sum, sun and sung). A second major distinction, applying only to lingual articulations, is between sounds where the airstream is channelled centrally along the tongue and those where the air is channelled over one or both sides of the tongue. The latter are known as laterals and can be produced as fricatives or approximants (there is only one in English, an alveolar approximant, as in hull, but Welsh has a voiceless lateral fricative, as in llan). At a few places in the vocal tract the articulators are sufficiently flexible to be set in vibration by the airstream in exactly the same way as the vocal folds are set in motion for phonation. These sounds are known as trills, and can be produced at the lips (paralinguistic brr – expression of cold), at the tongue tip (stereotypical Scottish English r) and at the uvula (standard French and German r). The tongue tip can also be struck against the alveolar ridge to produce a tap (casual pronunciation of ought to or order, which may sound the same in Australian English). An affricate consists of a short stop articulation followed by a short fricative (both consonants in charge). Whether or not such sequences are regarded as single segments is largely a phonological question. For example the final sounds in English plaits are regarded as constituting two segments, whereas the final consonants of German Platz are regarded as one. This is not because of any phonetic differences, but because German ts behaves like a single consonant (appearing initially in words such as Zeit, for example), whereas English ts behaves as a sequence of stop+fricative, which cannot occur word-initially.

The above descriptive framework is not commonly used for vowels. By definition, these sounds involve a more open configuration of the vocal tract (what would be the manner of articulation?), they involve the whole of the body of the tongue (where would be the place of articulation?) and they are almost always voiced (this part of the description would be redundant). Fortunately the contour of the tongue when articulating vowels is almost always convex; traditionally, therefore, we describe vowels in terms of the notional position of the highest point of the tongue horizontally – whether at the back, front or centre of the mouth – and vertically – whether it is close, half-close, half-open or open in relation to the roof of the mouth. In addition to this we need to state whether the lips are rounded or unrounded. In British and Australian English the vowels of beat, be and bat are front unrounded vowels; put and port have back rounded vowels and bird has an (unrounded) central vowel. Beat and boot are close, pet and port are half-close; pot is half-open, and putt and part have open vowels. Although this framework appears to be based on articulation, it is probably more accurate to say that it is auditorily based; there are also, as we shall see, some fairly close acoustic correlates.

The cardinal vowel system is a system of reference vowels which define the periphery of the available ‘space’ for the articulation of vowels. With considerable practice, the vowels of any language can be described by a skilled listener according to their position within this (two-dimensional) space, with reference to the cardinal vowels. The space is traditionally represented as a quadrilateral as in Figure 5.2, with the edges of the quadrilateral representing the limits of tongue movement. Two anchor points are established aerodynamically – the closest and most front vowel that can be articulated without producing (palatal) friction [i] and the most open and furthest back vowel that can be produced without (pharyngeal) friction [u]. The remaining vowels along the front and back edges of the space are positioned at intervals which are, supposedly, auditorily equidistant from one another. Of the eight primary cardinal vowels, the first five (on the front and bottom edges) are unrounded; the four on the back edge of the space are said to have increased lip rounding from open to
5.3 Speech production

5.3.1 Goals of speech production research

In real speech the sets of gestures required for neighbouring consonants and vowels are rarely produced strictly sequentially. For example, the English word fence consists of four phonemes – three consonants and a vowel. But in many people’s pronunciation the velum will open during the articulation of the vowel, rather than at the same moment that the tongue tip reaches the alveolar ridge for the articulation of [n], so the vowel will be nasalised. This is easier to explain if we regard the tongue tip raising gesture and the velum lowering gesture as produced in parallel by two of the subsystems we have outlined above and therefore not necessarily precisely synchronised. We have identified three major vocal tract subsystems: the laryngeal, the velic and the oral. At the boundary between the final two
consonants in *fence* a change is required in all three: in passing from [n] to [s] we need to change from voiced to voiceless in the laryngeal subsystem, from nasal to oral in the velic subsystem and from stop to fricative in the oral subsystem. Most people pronounce this sequence as [nts] (to rhyme with *vents*), because they make the first and second changes together, but the final change is delayed. They change from voiced to voiceless and from nasal to oral, but because they delay the change from stop to fricative, the result is a voiceless oral stop [t] separating the other two consonants. We can further divide the oral subsystem into three ‘sub-subsystems’: the lips, the tongue tip and blade, and the tongue body. The concept of coarticulation – of overlapping gestures being produced in parallel by separate subsystems – is a more transparent way of describing what happens in connected speech than the linear sequential approach. The latter was perpetuated by the dominance of generative phonology and its offshoots (including ‘natural’ phonology) which required phonetics to explain how individual speech sounds are ‘glued’ back together again by means of such ‘processes’ as assimilation, elision, epenthesis and vowel reduction. In the last thirty years nonlinear approaches, especially articulatory phonology and feature geometry, have found a rather more comfortable alignment with a parametric approach to phonetics. A major goal of speech production research remains, however, the explication of the invariance question from the encoding end: how discrete, static, context-free mental targets are transformed by the speaker into continuous, dynamic, context-sensitive gestures. This can be broken down into a number of sub-questions:

1. What are the units of speech production: are they equivalent in size to phonological features, phonemes, syllables, words – or something else entirely? Is there a single unit or do we invoke different-sized units at different stages of the process?
2. How is serial ordering achieved? The ordering of units is crucial to how the utterance will be perceived: how does the system ensure that the units emerge in the right order?
3. How does the system cope with the enormous number of degrees of freedom? Each occurrence of the ‘same’ sound requires a different set of muscle movements: how are the various vocal tract subsystems coordinated in order to produce the appropriate gesture (motor equivalence)?
4. (related to the above) How does the system take into account context sensitivity? Starting points, endpoints and trajectories of gestures will differ according to segmental and prosodic context.

A ‘coproduction’ approach to speech production assumes that gestures have their own temporal structure and can overlap with each other in time, the degree of overlap being controlled at the plan stage, rather than being an incidental outcome of the execution of the plan. Thus gestures are not ‘modified’ or ‘influenced’ by other gestures, they simply overlap with one another – they are ‘coproduced’ – and the degree of overlap depends on the extent to which different or identical subsystems are involved.

### 5.3.2 Techniques in speech production research

Since speech is a dynamic and multisystem process, phonetic research must employ a variety of dynamic instrumental techniques to capture it. Most of these techniques are based around a multichannel speech data acquisition system, and large amounts of data can be processed using specially developed software for semi-automatic segmentation, labelling and statistical processing of the parallel signals.
Volume velocity of the air flow during speech is normally registered by means of a device known as a pneumotachograph. Oral and nasal air flow are best recorded via separate airtight masks in order to eliminate the risk of leakage between the two channels. As the air exits the mask, the pressure drop across a nylon gauze screen is registered by a variable reluctance differential pressure transducer. Intraoral pressure can also be measured by inserting a catheter through the nose and into the pharynx. Aerodynamic data are particularly useful for investigating obstruent voicing contrasts, for verifying the direction of air flow (egressive or ingressive), and for determining the degree and timing of nasalisation.

The most commonly used and least invasive technique for observing the laryngeal subsystem is electroglottography (EGG), also known as (electro)laryngography, which records the degree of vocal fold contact over time. A pair of electrodes placed externally either side of the larynx measures the variations in impedance to a very small electrical current passed between them. As well as producing an excellent ‘clean’ signal for deriving a pitch contour, the technique is useful for investigating subtle differences in the timing of vocal fold vibration and distinguishing phonation types such as breathiness and glottalisation, although waveforms must always be interpreted with care. EGG is also a useful way of looking at glottal stops and glottalised consonants, which are commonly accompanied by a single sharp rise in the signal.

Techniques for recording gestures of the oral subsystems are usually more invasive, or at least more uncomfortable. Electropalatography (EPG) is a method of registering the contact between the tongue and the roof of the mouth, using an electropalate – a denture-like acrylic plate (minus teeth), which is individually moulded to fit the palate of each speaker and is held in place by stainless steel clasps around the teeth. Mounted on the underside of this plate are a number of electrodes (>60), each of which is connected to a thin copper wire. The wires are collected into two bundles which emerge at the corners of the mouth and are connected to a control unit, and thence to the computer. During recording the speaker holds a hand electrode connected to the control unit, which supplies a small AC signal to the body; this signal is conducted from the tongue to the palate electrodes whenever contact is made. The changing patterns of linguo-palatal contact can be displayed in real time on the computer screen, usually together with a simultaneously recorded acoustic signal, and stored for subsequent analysis. A range of data reduction procedures can then be performed to extract a number of quantitative measures of contact patterns in the form of single numerical values. EPG is a useful tool for the study of the tongue tip/blade and tongue body subsystems, but can give no information on articulations further back than the front edge of the velum, no information on labial articulations and no information where there is no actual contact between the articulators (e.g. vowels).

Electromagnetic mid-sagittal articulography (EMA) is in many ways complementary to EPG. This is what is known as a flesh-point tracking technique, which tracks the movement of the tongue, lips and jaw and velum in the two dimensions of the mid-sagittal plane. Small sensors are stuck on to the surface of the articulators of interest and speakers wear a helmet that contains a number of transmitter coils which produce alternating low strength magnetic fields. These in turn generate currents in the sensors as they move through the fields enabling them to be tracked by computer. More recently 3D systems have been developed, allowing for calculation of the XYZ coordinates of the sensors, giving more complete and accurate tongue visualisation than 2D models. These systems use external fixed transmitters, thus doing away with the need for a helmet. They provide high quality information on the spatio-temporal coordination of the supralaryngeal subsystems, including labial and velar/uvular articulations, as well as movements of the velum and the jaw. Since they do not rely on
actual contact between articulators for the registration of a signal, they can also provide accurate information on vowel gestures.

A technique which is less invasive, but still provides 3D information on the oral subsystems even without articulatory contact, is ultrasound. It relies on hundreds of piezoelectric crystals, which each in turn emit and receive ultra-high-frequency sound waves. Given that different media have different sound transmission properties, these waves are reflected back at different delays and with different amplitudes. Specifically, the ultrasound wave is reflected back very strongly at the intersection between the soft tissue of the tongue and the air layer above. Since muscle tissue and air have very different sound transmitting properties, the edge of the tongue can be seen as a bright white line on a good ultrasound image. The currently used real-time B-scan arrays have sample rates up to 124 Hertz (Hz, i.e. each crystal is sampled 124 times a second). As well as tracking the time course of overall tongue shape, ultrasound enables the observation of parts of the vocal tract other techniques cannot reach, especially tongue root and pharynx wall, as well as non-sagittal lingual measures such as lateral release and tongue grooving. If data are obtained in both planes, 3D images of the tongue can be constructed.

Other techniques used in the study of speech production include electromyography (EMG), which measures electrical activity in the muscles of the articulators. There are also various endoscopy-based techniques. A fibre-optic nasendoscope, for example, introduced through the nostril, with a light source and camera can be used to record the movements of the velic subsystem and, if inserted further through the velo-pharyngeal port, can be positioned over the vocal folds to study the laryngeal subsystem. In videofluoroscopy a speaker is placed between an X-ray source and an image intensifier coupled to a video camera, allowing the images to be recorded and played on a monitor. This technique is used more commonly in research into swallowing and velopharyngeal function. Magnetic resonance imaging (MRI) relies on detecting a radio-frequency signal emitted by excited hydrogen atoms in the body, using energy from an oscillating magnetic field applied at the appropriate resonant frequency. Until recently this has been anything but a ‘dynamic’ technique, as the speaker must maintain the vocal tract shape of the target sound for several seconds while the image is produced. The last few years has seen the advent of real-time MRI, which allows the continuous monitoring of moving tissues in real time, offering interesting possibilities for future use in speech research.

5.4 Speech acoustics

5.4.1 Goals of acoustic phonetic research

The overall goal of acoustic phonetics is to understand the relation of vocal tract shape and vocal tract subsystem gestures to the sounds that result. We seek to describe speech acoustically in sufficient detail to be able to characterise the nature of the crucial distinctions in speech: first and foremost, what distinguishes one sound from another in normal speakers of any language; but also what distinguishes a pathological speaker from the normal population (and possibly one pathology from another), what distinguishes one individual speaker from others, and what is the range of ‘accidental’ variation. Unlike speech production techniques, acoustic recording is totally non-invasive and can indeed take place without the knowledge of the speaker. As well as its use in linguistic description and the development of linguistic theory, acoustic phonetics has a number of practical applications, including clinical speech pathology (assessment, diagnosis and treatment), speech technology
(telephony, speech synthesis and automatic speech recognition) and in forensic science (especially speaker comparison).

5.4.2 Basic principles: pressure waves

Sound consists of a series of pressure waves that produce a sensation in the human ear; these pressure waves are produced by the oscillation of air molecules. In the case of speech the source of the sound is the vibration of the vocal folds, which modulates (‘chops up’) the flow of air from the lungs. This produces tiny variations of between 20 $\mu$Pa (= 20 millionths of a Pascal) and 20 Pa around an average atmospheric pressure of just over 100,000 Pa. These variations repeat themselves on average 110 times a second in adult men and at about twice that rate in women. The rate of repetition is known as the fundamental frequency, and is measured in Hertz (cycles per second). This corresponds to what we hear as the pitch of a sound. The volume of sound (what we hear as loudness) is measured in two ways: one is the sound pressure level (SPL), which is the force necessary to displace the air molecules per unit area (measured in Pascals) and intensity, which is the power transmitted per unit area (and is measured in Watts per square centimetre). Both of these ways of measuring produce inconveniently large ranges of numbers when applied to human hearing – in the case of SPL from 0.00002 Pa to 200 Pa and in the case of intensity from $10^{-16}$ W/cm$^2$ to $10^{-2}$ W/cm$^2$ (which means that the intensity of sound that causes pain is one hundred trillion times the intensity of the quietest sound we can hear). It is much more convenient (and much closer to the way our hearing works) to use a ratio scale, with a unit called the decibel (dB). On this scale a sound that is ten times the intensity of the threshold of hearing is 10 dB, but a sound that is one million times threshold level is only 60 dB – the level of typical conversation.

Waves that repeat the same pattern of pressure variation for a number of cycles are known as periodic waves. Since this is the type of wave produced by the vibrating vocal folds, such wave patterns (waveforms) are characteristic of voiced sounds. Waves that show random pressure variation are known as aperiodic waves. These are characteristic of fricatives and some stop releases. Silent intervals, mainly the result of complete closure of the articulators, are also an important part of the speech signal. Periodic waves, aperiodic waves, and silent intervals are each visible in the upper part of Figure 5.3.

5.4.3 Basic principles: harmonics and formants

The simplest kind of periodic waveform is known as a sine wave. It describes the kind of motion exhibited by a perfect pendulum or spring – known as simple harmonic motion – where the object in question is moved away from its point of rest (thereby storing up potential energy) and then released. It moves with constant acceleration until it reaches its maximum velocity as it passes its point of rest, thence moving with constant deceleration until it reaches a point opposite its point of release, where it moves once more with constant acceleration back to its point of rest, whereupon the next cycle begins. We can describe any sine wave in terms of just two numbers: the amplitude of vibration and the frequency of vibration. This is how air molecules behave when a sound is transmitted, but only sounds of a certain type – so-called pure tones – produce simple harmonic motion. The periodic vibrations produced by the laryngeal subsystem are much more complex. Nevertheless, like any complex periodic wave, they can be analysed in terms of a finite number of sine waves and thus described as a finite series of pairs of numbers, corresponding to the frequency and amplitude of each sine wave component or harmonic. In speech, the first harmonic is the
Phonetics

fundamental frequency and the others (around thirty in the adult male voice) are all at whole multiples of the fundamental frequency. This means that we can represent a complex periodic waveform at any moment in time by a graph of intensity versus frequency: this is called a sound spectrum. The spectrum shows all of the harmonics and their intensity as generated at the larynx (see the bottom of Figure 5.4). When the fundamental frequency decreases, the first harmonic will move down the frequency axis and, as they are whole multiples of it, all the higher harmonics will move closer together. When the fundamental increases, the opposite occurs; thus the harmonics of a woman’s voice are roughly twice as far apart as those of a man’s. Different voice qualities will result in different relative amplitudes among the harmonics. Modal voice usually produces a spectrum which falls in amplitude from the first harmonic at a rate of roughly \(-12 \text{ dB per octave}\). The first and second harmonics (H1 and H2) will be roughly equal in amplitude, although H1 is likely to be somewhat lower. In creaky voice the spectral tilt is much shallower and the amplitude of H1 will be considerably lower than that of H2. In breathy voice the spectral tilt is much steeper than in modal voice and the amplitude of H2 is lower than that of H1.

If we want to show changes in the spectrum over time, we have to introduce a third dimension. The way this is usually done is to plot time along the horizontal axis and frequency on the vertical axis, with intensity shown by the darkness of the shading. Areas of
intense energy are shown as black; less intense energy appears in a corresponding shade of grey. This is called a sound spectrogram – an example of one is shown in the lower half of Figure 5.3. According to the acoustic theory of speech production, the picture we see here is the product of a source and a filter: the source, as we have already seen, is the complex periodic sound made by the vocal folds, consisting of a series of harmonics. The filter is formed by the oral and nasal subsystems. Figure 5.4, consisting of component sine waves or

![Diagram showing the acoustic theory of speech production](image)

**Figure 5.4** The pressure wave produced by the vibrating larynx
harmonics, shows the pressure wave produced by the vibrating larynx which is the source of the speech signal. From bottom to top, it passes through the vocal tract which acts as a filter, with a set of resonant frequencies. Harmonics in the source wave which correspond to these frequencies pass through unmodified; those in between these frequencies have their amplitude reduced. The result is the output wave emerging from the lips. A filter is a mechanism which, because of its shape, will attenuate (reduce the energy of) sounds at certain frequencies and let through sounds at other frequencies – the latter are the resonances of the filter (which can also be called a resonator). Using the same dimensions as for a spectrum (frequency on the horizontal axis, intensity on the vertical axis), we can draw a resonance curve to describe the effect that any filter will have on a source sound (see the middle of Figure 5.4). The peaks show the frequencies of the resonances and the troughs show the frequencies at which sound will be attenuated. The combination of source and filter produces a spectrum in which the harmonics vary in amplitude according to the vocal tract resonances (see top of Figure 5.4). The peaks are called formants and their frequency and amplitude will vary as the shape of the vocal tract changes over time during speech. They can be seen as three or four broad bands of energy on the spectrogram in Figure 5.3 (5) and (7). It is important to realise that the source component (from the laryngeal subsystem) and the filter component (from the vocal tract subsystems) vary independently of one another – i.e. the fundamental frequency goes up and down independently of whether the formants are going up or down. One of the crucial measurement tasks in acoustic phonetics is to separate out the contribution of these two components in the output signal.

There is an approximate correlation between the frequency of the lowest two formants (F1 and F2) and the position of the tongue in the mouth. Broadly speaking tongue height is inversely correlated with first formant frequency: high vowels have low F1 and low vowels have a high F1. Tongue position on the front–back dimension is correlated with second formant frequency: front vowels have high F2 and back vowels have low F2. Lip rounding lowers second formant frequency. In high front vowels (such as in pea) the tongue forms a narrow front cavity and a wide back cavity, resulting in a low F1 and a high F2 (see Figure 5.3 (7)). Open back vowels (such as in pod) have a wide front cavity and a narrow back cavity, which gives a high first formant and a low(ish) second formant. Close back rounded vowels (such as in pool) have two cavities of similar shape and volume, producing a low F1 and a low F2. Plotting vowels on a chart with F1 as the horizontal axis and F2 as the vertical axis shows an acoustic relationship between vowels which is remarkably similar to the traditional cardinal vowel chart (especially if the axes are flipped, so that the origins are at the top right), as can be seen in Figure 5.5.

5.4.4 Basic principles: formant transitions, noise and anti-formants

Formants also play a role in distinguishing between consonants. The rapid changes in vocal tract shape which occur when the articulators move from a consonant to a vowel or vice versa give rise to correspondingly rapid movements of the formants. These movements are known as formant transitions; they can easily be identified on a spectrogram, and they help differentiate between various manners of articulation. Broadly speaking, approximants have rather slow transitions, whereas stops have much faster transitions. Going from a consonant into a vowel, voiceless stops have a much abbreviated F1 transition (see Figure 5.3 (4)), laterals show an abrupt increase in F1, and nasals show jumps up in both F1 and F2. These characteristics are shown in mirror image as the articulators move from a vowel to a consonant. But it is in distinguishing places of articulation that second formant transitions
Figure 5.5 A formant chart: plotting the first formant frequency against the second formant frequency of a set of vowels produces a configuration very similar to that of the traditional vowel quadrilateral (cf. Figure 5.2)

play their most important role. Every combination of consonant place of articulation and adjacent vowel will result in a different F2 slope; the initial consonant–vowel transition in English *deep*, for example, will be slightly rising, while that of *dart* will be falling (see Figure 5.6b). This is an example of the context sensitivity issue discussed above. It can be shown, however, that if all the F2 transitions for a given place of articulation are extrapolated in time towards the consonant, they all converge at about the same frequency. This is known as the *locus* frequency and it appears to be one of the major perceptual cues to place of articulation. In general, bilabial sounds have a low-frequency locus (approximately 750 Hz in English), alveolar sounds have a mid-frequency locus (around 1,800 Hz in English) and allophones of velars before front vowels (*key*, *care*) have a high-frequency locus (around 3,000 Hz in English). Back allophones of velars (*cot*, *cook*) have a low- to mid-frequency locus (around 1,200 Hz in English). The third formant frequency is not immensely important for the differentiation of most speech sounds – with one exception: retroflex sounds have an extremely low F3 and this helps to distinguish them from their neighbours, especially the alveolars. The F3 transition at the beginning of the vowel in English *raft*, for example is extremely steep – rising from below 1,500 Hz to the 2,500 Hz target for the vowel within a few milliseconds.

There is one other major source of sound in speech other than the vibration of the vocal folds. When flowing in a wide, unconstricted space, the molecules of a liquid or gas flow smoothly in straight lines, like the water in a broad flat river bed. There are two ways to speed up the motion of the fluid and cause its flow to become turbulent. We can create channel turbulence by forming a constriction in part of the vocal tract (like a narrow river gorge) or we can create wake turbulence by introducing an obstacle into the stream, such as the upper incisors (like a rock in the river). This will cause aperiodic vibration or *noise* (like white water). All fricatives have channel turbulence, as do the voiceless stops of English at their release (see Figure 5.3 (3)); the so-called strident fricatives (such as in *see* and *she*) also
Figure 5.6 Perceptual equivalence: schematic spectrograms of synthesised speech demonstrating context sensitivity in speech perception. (a) The same burst of noise, centred at 1,500 Hz, is perceived as the release of a ‘p’ when followed by a high front vowel but as a ‘t’ when followed by an open central vowel. (b) In order for listeners to hear a ‘d’ a completely different F2 transition is required before a high front vowel than before a low central vowel.

have wake turbulence also (see Figure 5.3 (1)). The acoustic theory of speech production holds for these sounds too: the source sound generated by the turbulence is filtered by the resonances of the cavities in front of it. Meanwhile the cavities behind it, with no opening to the outside and no sound source of their own, behave as ‘side branch resonators’, which actually absorb energy from the source, forming valleys in the spectrum called anti-formants or zeroes.

All nasal stops have the pharynx plus the nasal cavity as the main resonator. They thus have a so-called nasal formant (N1) at a very low frequency (below 400 Hz) and higher nasal formants at odd multiples of this. In these sounds it is the oral cavity which is the side branch resonator. In laterals the supra-lingual space is the side branch resonator. In nasalised vowels, the pharynx and oral cavity form the main resonator just as in oral vowels, but the nasal cavities form a side branch resonator, giving these sounds great spectral complexity, which makes them more difficult to distinguish from one another than oral vowels.
5.5 Speech perception

5.5.1 Goals of speech perception research

Speech perception research is ultimately seeking answers to the invariance question from the decoding end: how continuous, dynamic, context-sensitive gestures are transformed by the listener into discrete, static, context-free mental targets. Again this can be broken down into sub-questions which mirror those of speech production research.

1. What are the units of speech perception: are they equivalent in size to phonological features, phonemes, syllables, words – or something else entirely? Is there a single unit or do we invoke different-sized units at different stages of the process?

2. How do we remember the order of phonemes that make up a word: how is the serial structure of speech represented neurally? Perceptual processing of one unit may depend on information from a preceding unit. The ordering of units is crucial to how the utterance will be perceived.

3. How does the system take into account context sensitivity? Starting points, endpoints and trajectories of gestures will differ according to segmental and prosodic context. There will be multiple cues to the same percept and the same cue may signal more than one percept (perceptual equivalence) (see Figure 5.6).

4. How does the perceptual process cope with the enormous number of degrees of freedom? How do we recognise a sound from one speaker as being perceptually equivalent to a sound from another speaker, despite differences in gender, dialect and vocal tract size?

Up until fairly recently there have been basically two ways to approach the invariance question: either look harder for invariant units in the speech signal or ignore the signal and look for invariant units ‘higher up’ in the central nervous system. Theories of speech perception which take the first approach are sometimes referred to as passive theories as they invoke the sensory pathways only. The premise is that the production and perception of speech may well share a common set of units, but that production need not be referred to for perception. The invariant units are in our perception: we learn to produce them. One such model is the exemplar-based approach which claims that particular instances of speech sounds are stored in the memory of a listener. Incoming speech signals are compared with these exemplars in the simultaneous and interactive processes of speech perception and speaker recognition.

Theories that take the second approach are often called active theories of speech perception, as they involve the motor neural pathways in a conversion of acoustic–phonetic information to a speech representation via articulatory knowledge. This type of theory takes the view that, since production and perception share a common set of units, therefore they must be linked. The invariant units are in our production: we learn to hear them – an idea apparently supported by the recent discovery of mirror neurones. The best known of these theories is probably the motor theory of speech perception. Theories of this type usually invoke the idea of ‘duplex’ perception – that listening to speech switches on a specialised module for perceiving speech. There is no doubt that there are some apparently special aspects to speech perception. For example, the sequential order of natural-sounding speech sounds is accurately reported, whereas that of non-speech sounds is not. Furthermore, highly encoded stimuli, such as initial voiced stops, appear to be perceived categorically (i.e. discriminated better at phoneme boundaries than within them), whereas less encoded
stimuli, such as vowels, appear to be perceived continuously (i.e. discriminated equally well across the phonetic continuum). The preferred version of this theory nowadays is that, rather than a specific speech mode, there may be a specialised mode of listening for any acoustic signal that requires complex auditory processing.

There is, however, another approach to the invariance question, which does not look at variability in speech as a problem, but sees perception as ‘exploiting’ it. This is the approach taken by parallel distributed processing (PDP) theories. In contrast to the ‘bottom-up’ processing approach taken by passive theories and modularist active theories alike, PDP models of speech perception recognise the importance of ‘top-down’ processing, using multiple sources of information including acoustic cues, phonological context, syntactic/semantic context and the overlapping of information in time (right and left context effects). Listeners are said to make use of the so-called ‘lawful’ (i.e. rule-based) variability in speech: the invariant units are in our heads and we utilise all available information in perceiving the speech signal in terms of them. Furthermore, it is now well recognised that speech perception also integrates information from the visual modality when this is available.

Most researchers agree that there are two initial stages in the speech perception process prior to categorisation in terms of linguistic units. These are known by various names (1 sensory memory, echoic memory, sensory register, primary auditory analysis; 2 auditory short-term memory, working memory, pre-categorical acoustic storage, articulatory loop). Results from a number of different experimental paradigms point to a decay time for post-stimulatory sound traces in sensory memory of 200 (± 50) ms. This figure may have a physiological explanation in terms of the short-term adaptation time in single auditory nerve fibres. Auditory attributes of sounds (such as duration, pitch, loudness and spectral form) are stored in a short-term auditory memory. Results from serial recall experiments suggest a survival time in short-term memory of one to two seconds, during which active rehearsal of items takes place (the so-called ‘articulatory loop’). Research suggests that the survival time of sounds in this loop depends on the quality of information. As suggested above, there are no qualitative differences between vowels and consonants in this respect, but there are differences according to the ease of discrimination of items (and vowels typically tend to be more easily discriminable than consonants).

As to the question of units of speech perception, it would appear that the syllable may be the minimal unit of perception (needed for prosodic information and some consonantal cues). Signal chopping and alternating speech between ears is most disruptive to recognition at an interval equivalent to the duration of the syllable. This is not to say that the syllable is necessarily the basic unit of perception; in fact there may not be a single unit of perception, but rather a series or network of processing levels. Any one of these levels can be the focus of the listener’s attention, but listeners normally focus on the highest level – the meaning of the message.

5.5.2 Techniques in speech perception research

Until recently one of the main barriers to understanding speech perception has been our inability to observe the process directly. Researchers were limited to the behavioural paradigm commonly used in psychology, whereby subjects are presented with stimuli and asked to respond to them in one of a number of standard ways: identifying them (yes/no, forced choice labelling), discriminating between them, rating their similarity, etc. Stimuli may be natural or synthesised, presented under ideal conditions or in noise or subjected to attenuation, filtering, or gating (shortening), presented in one or both ears or alternating
between the two. We may present other sounds or images first in order to test the ‘priming’ effect on the subject’s perception. The data may be derived by measuring the accuracy of the responses or the reaction times of the subjects. Special paradigms have been developed to test speech perception in infants and prelingual children, who respond to a presentation by e.g. increasing their sucking rate or turning their head towards a visual stimulus. In recent years eye-tracking methodology has been used, as a more accurate way of measuring reaction times and also to follow the scan path of the subject in choosing between items on a screen.

The development of neurophysiological techniques has enabled researchers to observe the brain’s responses to speech stimuli unmediated by the overt behaviour of the listener (see Chapter 19). Using traditional presentation paradigms the responses of the brain can be measured employing a number of non-invasive methodologies. These show that responses at this level can be more sensitive and subtle than those measured via behavioural paradigms. For example, brain responses may reveal a sensitivity to the difference between two speech sounds which the subject may not show sensitivity to in a discrimination test. Methods used to measure neural responses to speech include electroencephalography (EEG), which measures electrical signals in biological tissue by means of a sensor net on the scalp, and magnetoencephalography (MEG), which records the magnetic signals corresponding to those electric currents by means of magnetometers. Both methods can be used to record event-related potentials (ERPs) in the brain. Such studies can indicate the timing, the degree of engagement (amplitude) and the functional equivalence of underlying processes (distribution across the scalp). Near-infrared spectroscopy (NIRS) is an optical imaging technique that estimates changes in neuronal activity through detecting regional changes in total haemoglobin concentration and oxygenation by measuring the transmission of near-infrared light through the tissue, thus yielding an index of local activation on the surface of the brain. These techniques are particularly useful in the study of infant (and even in utero) speech perception, and can be used to map the very early development of brain structures, and their interaction in speech perception.

Further reading