Exercises

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

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

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

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

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

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

8 Speech Perception

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

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

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

8.1 Introduction

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

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

because of the capacity of listeners to perceive an invariant phonological structure from extremely variable speech signals which are rich with multilayered information. Lindblom (1986) and Stevens (1989) provide stimulating discussion and overview of this issue, particularly in relation to the perception and production of vowel sounds.

A simple example of this richness and variability which can nevertheless produce an invariant phonological percept is a phrase such as 'is that your ticket?' uttered by four speakers, say a young adult female, a young adult male, a very young child and a very old male. As listeners we are not only able to perceive the phonological structure of this phrase as produced by four quite different voices; and even without seeing the speakers we can usually identify their age and sex as well, at least to the point of distinguishing female speakers from male, very young from elderly, and so on. But, more than that, if our four speakers were to repeat this phrase several times, we can probably judge, from the speech signal alone, whether they are now getting angry or remaining patient or becoming overpolite, and we achieve this without undermining our perception of the phonological structure. Yet these 'repetitions' of the same phonological structure by different speakers under different conditions will actually vary substantially in their acoustic signal and its spectral time-course.

The second problem has already been alluded to above, namely the rather fluid relationship between our reliance on high-level linguistic and contextual knowledge and our response to the acoustic cues in the acoustic signal itself. Despite some uncertainty here, we do know that listeners can determine phonological structure when relying almost entirely on the acoustic speech signal alone: all of us are, after all, able to write down recognizable representations for the pronunciation of nonsense words or proper names which we have not heard before; and with training, professionals can make reasonably accurate phonetic transcriptions of unfamiliar speech patterns in linguistic field work or clinical sessions.

The preceding chapter (section 7.9) has already introduced some of the basic perceptual properties of sound waves to explain the psychoacoustic basis for the units of measurement used to quantify amplitude and frequency. In this chapter we examine the perception of speech more generally, concentrating on acoustic-phonetic aspects of the processes which underlie our capacity to identify the phonological structure of speech.

8.2 The auditory system

The human auditory system is generally considered to consist of two broad components, the peripheral and central systems. Our concern is mainly with the peripheral system and its properties in processing the acoustic signals of speech. Figure 8.2.1 shows the structure of the peripheral system.

The peripheral system has three parts, the outer, middle and inner ears. The outer ear comprises the PINNA or AURICLE and the auditory MEATUS or outer ear canal. The pinna makes little or no contribution to our basic hearing

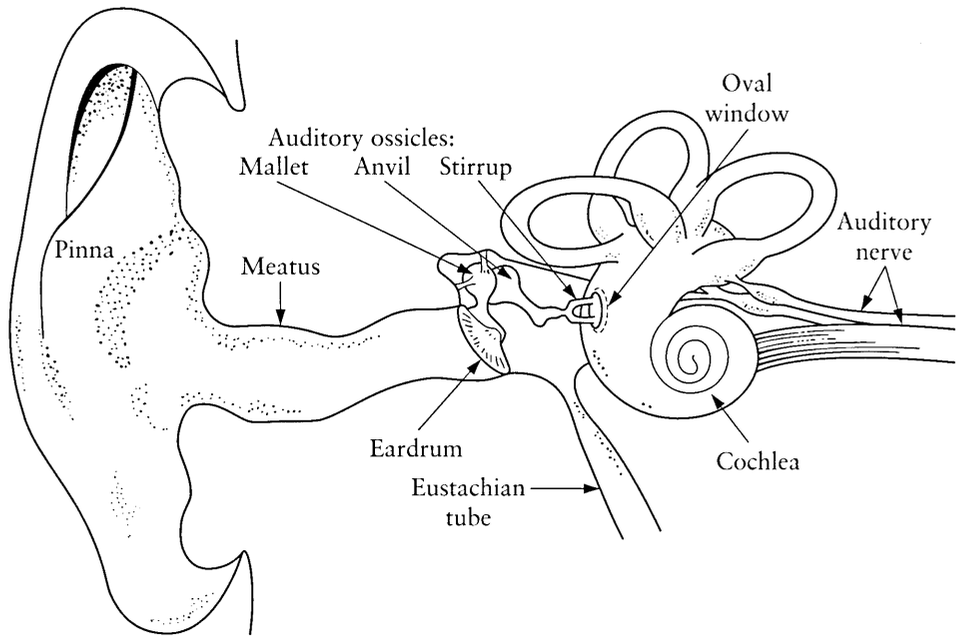


Figure 8.2.1 The structure of the peripheral auditory system

acuity, but serves to protect the entrance to the ear canal and does seem also to contribute to our ability to localize sounds, especially at higher frequencies. (The topic of auditory localization lies outside our linguistic concerns here, but it is worth noting that our ability to localize a source of sound is important in enabling us to be selective, for example in a crowded room where many people are talking and we are trying to listen to one speaker only. This ability is of course greatly enhanced by our having two ears.)

The pinna connects to the outer ear canal, a short tube of variable shape between 25 and 53 mm long which provides the pathway for acoustic signals to the middle ear. The canal has two major functions. The first is the obvious one of providing physical protection to the complex and not very robust mechanical structures of the middle ear. The second is to act as a tube resonator (section 7.13 above) which favours the transmission of high-frequency sounds between 2,000 and 4,000 Hz. This function is important to speech perception and particularly supports the perception of fricative sounds, as their identity is often encoded in aperiodic energy in this region of the acoustic spectrum. The resonance in the auditory meatus also contributes to our general hearing acuity between 500 and 4,000 Hz, which is the range of frequencies containing the major cues to phonological structure.

The middle ear consists of a cavity within the skull structure containing the EARDRUM (a membrane at the inner end of the outer ear canal), a set of three interconnected bones, known as the mallet, the anvil and the stirrup (together termed the AUDITORY OSSICLES), and associated muscle structure. The function of the middle ear is to transform the sound pressure variations in air that arrive

at the outer ear into equivalent mechanical movements. This process of transformation begins at the eardrum membrane, which is deflected by air pressure variations reaching it via the canal. The resulting movement is transmitted to the auditory ossicles, which act as an ingenious mechanical lever system to convey these movements to the oval window at the interface to the inner ear and the cochlear fluids beyond.

The lever action of the ossicles, and the fact that the eardrum has a much larger surface area than the oval window, ensure efficient transmission of acoustic energy between 500 and 4,000 Hz, effectively maximizing the sensitivity of the ear in this frequency range. The musculature associated with the auditory ossicles also works to protect the ear against damage from excessively loud sounds by an action known as the acoustic reflex mechanism. This mechanism comes into action when sounds of around 90 dB and greater reach the ear: the musculature contracts and repositions the ossicles to reduce the efficiency of sound transmission to the oval window (Borden and Harris 1980, Moore 2003).

The middle ear is connected to the pharynx by a narrow tube known as the EUSTACHIAN TUBE. This provides an air pathway which opens when necessary to equalize background air pressure changes between the outer and middle ear structures.

The inner ear is a complex structure encased within the skull, and our discussion here will focus on the COCHLEA, which is responsible for converting mechanical movement into neural signals: the mechanical movement conveyed to the oval window by the auditory ossicles is transformed into neural signals that are transmitted to the central nervous system. Essentially, the cochlea is a coil-like structure terminating in a window with a flexible membrane at each end. Figure 8.2.1 shows the general form of the cochlea, and figure 8.2.2 shows a cross-section through it.

Internally, the cochlea is divided by two membranes, one of which, the BASILAR MEMBRANE, is central to hearing. When movements (caused by sound vibrations) occur at the oval window, they are transmitted through the cochlear fluid and cause displacement of the basilar membrane. The basilar membrane is stiffer at one end than the other, and this means that the way in which it is displaced depends on the frequency of the incoming sound. High-frequency sounds will cause greater displacement at the stiff end; with decreasing frequency, maximum displacement moves progressively towards the less stiff end.

Attached along the basilar membrane is the ORGAN OF CORTI, a complex structure containing many hair cells. It is the movement and excitation of these hair cells which transform basilar membrane displacement into neural signals. Because the membrane is displaced at different places depending on frequency, the cochlea and its inner structures are able to transform sound intensity and frequency into neural signals. But it must be emphasized that the ultimate neural representation of frequency information is not dependent on the location of maximum basilar membrane displacement alone, and our understanding of the way in which frequency is encoded through the auditory system is incomplete.

Early research on speech perception took little account of the basic perceptual properties of the ear. Rather, it tried to correlate the perceptual properties of the speech signal with the kind of representation of a linear time-varying spectrum of the kind we have already examined in chapter 7, especially section 7.14. By

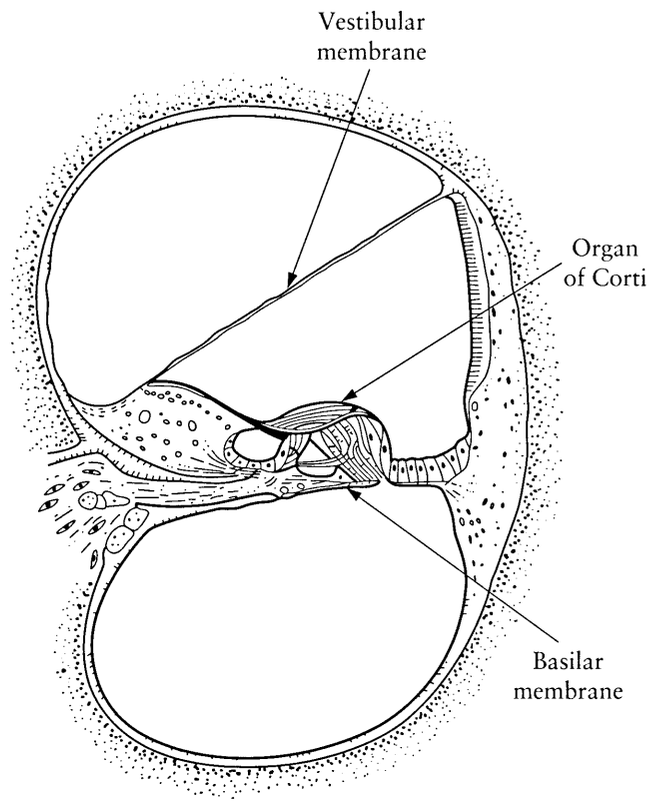


Figure 8.2.2 Cross-section through the cochlea
Adapted from: Denes and Pinson 1963, p. 71.

about 1980 researchers had realized that it was important to understand the analytical effects of the human auditory system on speech signals and that it was unwise to treat listeners as though they were simply processing information in the same way as a conventional spectrograph.

For this reason, the following section offers a brief review of the basic psychophysical properties of the auditory system in respect of frequency, time and amplitude, as they affect speech signals. For each of these three aspects of the signal, the most striking property of the human auditory system is that it is nonlinear.

8.3 Psychophysical properties of the auditory system

In section 7.9 above we showed that the auditory system is capable of making discriminations between successive changes in the frequency of an acoustic signal of about 0.5 per cent below about 1,000 Hz (figure 7.9.1). This ability is very

important for our detection of cues to intonation and word tone encoded in speech signal fundamental frequency patterns. The magnitude of the just noticeable difference (JND) also depends upon the way in which the test stimuli are presented. See Zwicker and Fastl (1990) for a review of work in this area.

Our ability to discriminate differences in the centre frequencies of formants in speech signals is about an order of magnitude poorer, with JNDs at around 5 per cent. This reflects the more complex nature of the signal. Nevertheless, this level of discrimination is substantially better than that required to encode and distinguish phonological contrast between acoustically similar vowels and sonorant consonants. O'Shaughnessy (1987) provides a useful overview of work on formant discrimination.

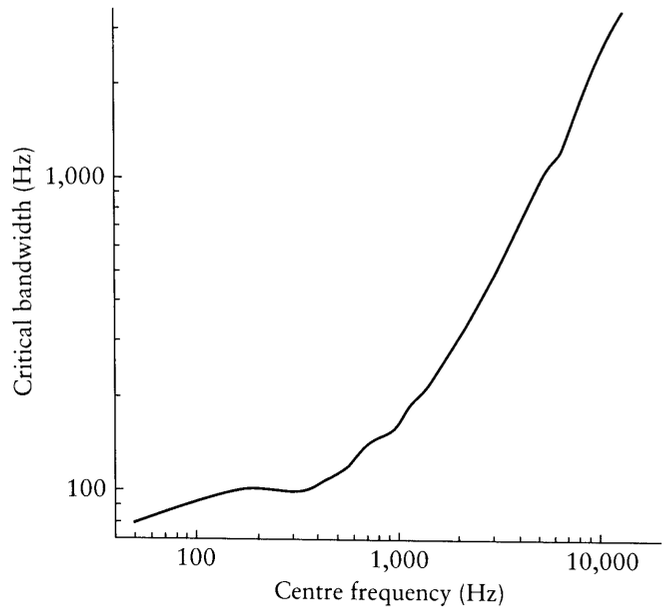
A further important property of the auditory system is its frequency selectivity – its capacity to resolve the contiguous frequency components of a complex acoustic signal such as speech. This aspect of the auditory system was first investigated in the 1920s and has been a continuing object of inquiry since. The most common method of measuring this property is to use a constant amplitude stimulus consisting of a narrow band of noise which is progressively increased in bandwidth until the listener can detect a change in loudness. As long as the listener hears no loudness change with bandwidth change, it is assumed that the auditory system is unable to resolve the increase in noise bandwidth; but when the bandwidth exceeds the limits of the auditory system resolution, this is detected as a loudness change. This psychophysical measure of frequency resolution is known to correspond with the neurophysiological frequency resolving capability of the cochlea.

As with other psychophysical measures, frequency resolution data vary somewhat with stimulus structure and presentation methods. Moore (2003) describes these and the results obtained. Most commonly, frequency resolution is expressed in terms of critical bands (or Bark), specifying the limiting bandwidth of acoustical energy which can be resolved at any frequency. Figure 8.3.1 shows the most commonly cited results of Zwicker (1962).

Figure 8.3.1 shows that the auditory system has quite fine frequency resolution to about 500 Hz; above this, the resolution broadens approximately logarithmically. In terms of speech signals, this means that we are able to resolve harmonic information in sounds such as vowels and sonorant consonants up to about 500 Hz, and phonologically relevant spectral peaks up to about 3,000 Hz. Broad-band fricative noise information in the range 3,000 to 5,000 Hz (which encompasses all the essential information in the speech signal spectrum) is more crudely resolved. As might be expected, these resolution characteristics correlate well with the progressively broader frequency domain encoding of phonologically contrastive information for nonresonant sounds.

Frequency is interwoven with time in speech signals: we respond to phonological encoding in the spectral time-course of the speech signal which reflects its characteristically dynamic nature. Time is important both in the encoding and perception of short-term acoustic events in stops and affricates and in the much longer-term encoding of prosodic information.

Temporal processing may be considered from two perspectives. The first concentrates on the interval over which the auditory system integrates information, and the second is concerned with the ability of the auditory system to



Figue 8.3.1 Critical bandwidth plotted against its centre frequency
Adapted from: Zwicker 1962.

detect gaps in otherwise apparently continuous acoustic signals. We have noted that the frequency resolution of the auditory system increases nonlinearly with increasing frequency; but there is no simple relationship between filter bandwidth and temporal resolution as is found in the electronic or software filters used in speech signal analysis (section 7.13 above).

The temporal integration of short-term signals by the auditory system is of direct relevance to the detection of very weak acoustic information. It appears that the threshold of audibility for sounds decreases progressively up to 200 ms and is unchanged thereafter, which suggests that stop bursts and other rapid onsets make substantial demands on the auditory system. But this generalization needs to be treated with caution because different test stimuli and protocols used by a number of investigators have yielded varying data in investigations of temporal integration.

Temporal acuity – demonstrated by the ability of listeners to distinguish between two successive acoustic events – also varies depending upon the stimuli and test protocols used. Pisoni (1977) found listeners able to distinguish temporal differences between 500 Hz and 1,500 Hz signal at minimum relative differences of 20 ms. Moore et al. (1993) investigated the ability of listeners to detect gaps in a signal consisting of a sinusoidal wave. The just distinguishable gap (or 'gap detection threshold') was roughly constant at around 6 to 8 ms for test signals in the range 400 to 2,000 Hz: outside this frequency range the gap detection threshold rises to around 18 ms. Other techniques for measuring gap detection threshold have yielded figures as low as 2 ms at test frequencies around 8,000 Hz. Overall, it appears that whatever the measurement

methods, the auditory system is capable of resolving the rapid onsets and acoustic energy gaps associated with obstruent consonants in running speech.

Another important form of temporal performance is the detection of spectral change in complex signals. The most common and significant form of change in speech occurs in the transitional movements of formants at the onset and coda of syllables (section 7.17 above). Perceptual experiments using complex synthesized speech-like signals with varying rates of frequency change suggest that rapid changes below about 30 ms are temporally integrated in the auditory system and heard as a single broader bandwidth signal. Extensive investigation of this area indicates that the ability to discriminate short duration frequency transitions is greater where a contiguous steady state signal follows. The relatively rapid formant transitions of around 50 ms for voiced stops in speech are, in perceptual and phonological terms, close to the relevant limits of the auditory system's processing ability. There are also suggestions by Jamieson (1987) that 50 ms may be close to an optimal level of salience for formant transition rates.

Turning to the amplitude of the speech signal, we note that the auditory system accommodates an extremely wide range of sound intensities. The system responds to differences in intensity logarithmically, a fact recognized by the development of the decibel scale as described earlier in section 7.9. The minimum JND depends as always on the measurement methods and stimuli, but data from Florentine et al. (1987) indicate figures of around 1 dB for high-intensity stimuli at frequencies below 10 kHz, and 3 to 4 dB at more moderate intensities and frequencies above 10 kHz. These levels of acuity are well beyond those required to decode speech signals.

The actual perceived loudness of sound for a constant intensity stimulus varies considerably with frequency. The threshold of intensity at which sound can be detected varies by about 70 dB between 20 Hz and 15,000 Hz. This is why some stereo systems have a so-called loudness control to boost up very low and very high frequencies when the system is being played at low sound levels. Some of this variability in auditory system acuity is a consequence of the frequency selective propagation of sound in the auditory canal. Fortunately, over the range 500 to 5,000 Hz, which contains most of the phonologically relevant information for speech, the auditory system has its lowest threshold of detectable intensity and thus is relatively uniform in sensitivity. Moore (2003) is an accessible textbook that summarizes this research, as well as key aspects of hearing and perception.

8.4 Speech intelligibility

The three basic dimensions of acoustic signals which we have been considering – frequency, time and intensity – and the related performance of the human auditory system have often been investigated by task-specific test signals designed to probe performance limits in the one dimension under investigation.

In our everyday perception of normal speech signals, however, we attend to the totality of a complex signal encoding actual language and we can use some top-down processing as well as bottom-up. This is, of course, highly relevant to our capacity and performance as listeners, and a brief review of this area follows. Most of the literature examining general speech intelligibility has focused either on whole words and syllables or on consonants, because of the interest in communication which has motivated the research. Vowels, the most intelligible component of syllables, have received more attention in later and more phonetically oriented studies.

In the first half of the twentieth century, telecommunications engineers embarked on extensive testing of the intelligibility of speech. One question of primary interest was to find out what band of frequencies had to be transmitted to ensure that speech was intelligible. An extensive set of investigations using filters to attenuate frequencies above and below a defined cut-off showed that most of the phonologically important information that ensures intelligible speech is contained in the band of frequencies between 300 Hz and 3,500 Hz. This is the typical passband used for telecommunications systems. The telephone system is a good example of an effective trade-off: the provision of a wider passband would have little cost benefit other than improving general fidelity and making speaker identification easier.

Differences in acoustic encoding among segments are such that not all sounds require even this passband, while some sounds will benefit from transmission of an even wider band of high frequencies. For example, back vowels such as /u/ gain little from frequencies above 2,500 Hz, whereas fricatives such as /f/ and /s/ would be more intelligible if telephones passed frequencies up to 5,000 Hz. Fletcher (1953) and O'Neill (1975) are useful summaries of the classical work in this area.

In addition to frequency passband, the effects of the intensity of presentation on intelligibility were also extensively studied in the same period. Typically these studies have shown that the intelligibility of monosyllabic words moves from about 10 per cent intelligibility to about 90 per cent intelligibility with an increase of 40 dB in stimulus presentation level (figure 8.4.1). These figures should be taken as a general guide only, because, as always, the actual figures obtained depend upon the particular stimuli chosen and the experimental protocol used.

The choice of stimuli is indeed crucial to the nature and results of speech intelligibility tests. If the speech materials used to test intelligibility are, for example, meaningful sentences, we do not rely on acoustic information alone to identify words. For instance, in a sentence such as 'the baker burned the bread', the word 'bread' is fairly predictable from the context and it is unlikely we would confuse it with similar sounding words such as 'bed' or 'pet' or 'brad'. On the other hand, if an intelligibility test asks us to identify nonsense syllables such as 'gup', 'dar' and 'oosh', there is little opportunity to use top-down linguistic knowledge to complement the available acoustic information. It is therefore not surprising that tests which include a linguistic context and offer substantial predictability produce higher scores for a given set of conditions (such as filtering or masking) than those involving meaningless syllables. Much of the earlier work in studying intelligibility failed to take real account

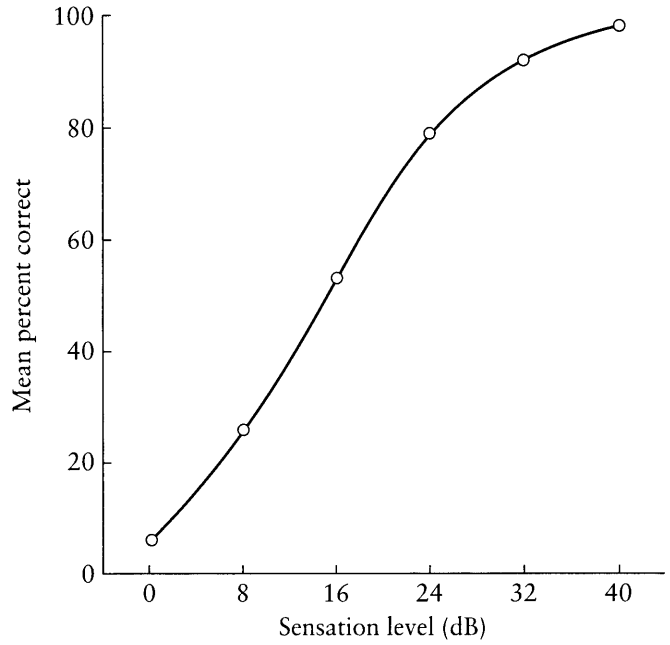


Figure 8.4.1 Performance Intensity Function for a set of monosyllables
Source: Robert Mannell, Macquarie University. Based on data in Kopra et al. 1968.

of these effects. Similarly, tests which use forced choice answers also result in higher scores than those which leave the listener without any options for a potentially correct response.

Figure 8.4.1 shows a typical graph (known as a Performance Intensity Function) of the progressive increase in the intelligibility of monosyllabic words with an increasing level of intensity. As with studies of the effects of a reduced frequency passband, intensity studies reveal different outcomes for different classes of speech sounds. The absolute intensity level at which the speech is presented can markedly affect intelligibility. Kent et al. (1979) examined the phonetically selective effects of intensity of presentation in some detail, and showed that sonorant and strong fricative consonants such as /w/ and /s/ require a markedly lower intensity level to be reliably recognized than do weak fricatives such as /v/ and voiceless stops such as /k/ and /t/.

Another method of investigating the effects of intensity on speech intelligibility is by the use of a competing signal to mask the speech. This has a practical merit, because it removes the artificiality of simply manipulating intensity. Instead it introduces the sort of competing signal which listeners encounter in the real world. Such competing signals may be as varied as the background noise in a jet aircraft cabin, the propagation noise of a radio communications link, or the babble of voices at a party.

In investigations of this kind, the masker is most commonly a broad-band noise signal with either a uniform frequency–intensity distribution, or a profile approximating the long-term averaged frequency–intensity spectrum of a number of speakers (of the kind shown in figure 7.19.1 above). The noise and the speech signals are mixed in precisely computed signal-to-noise ratios and presented to

listeners. The classic investigation in this area is by Miller and Nicely (1955), whose very comprehensive data have been extensively quoted and reanalysed in the literature of experimental phonetics. They showed, as might be expected, that voiceless sounds generally, and fricatives in particular, show greater losses in intelligibility than voiced sounds, especially sonorant sounds such as nasals. This demonstrated that nasality and voicing were the most robust phonetic features under masked listening conditions and that features such as place of articulation, duration and affrication are much less robust.

A series of later investigations, of which Pickett (1957), Pickett and Rubenstein (1960), Busch and Eldridge (1967), Williams and Hecker (1968) and Clark (1983) are examples, demonstrate that the effects of masking are generally explained by the relationship of the frequency–intensity profile of the masker to that of the speech sound under examination. Figure 8.4.2 illustrates the phonetically selective nature of band-limited uniform noise on various consonant classes in English.

Duration, reflected in the timing of the components of a syllable, is phonologically important. Early investigations of duration showed that rapid periodic interruptions to a continuous speech signal – by turning it on and off in rapid succession for equal intervals of time – affect intelligibility. When interruptions to the speech signal approach intervals of 500 ms, intelligibility falls to near zero; but when the duration is reduced to 200 ms or less, intelligibility approaches 100 per cent. Predictably, when the duration of the interruptions exceeds 500 ms, the effects on intelligibility are confounded by the nature of

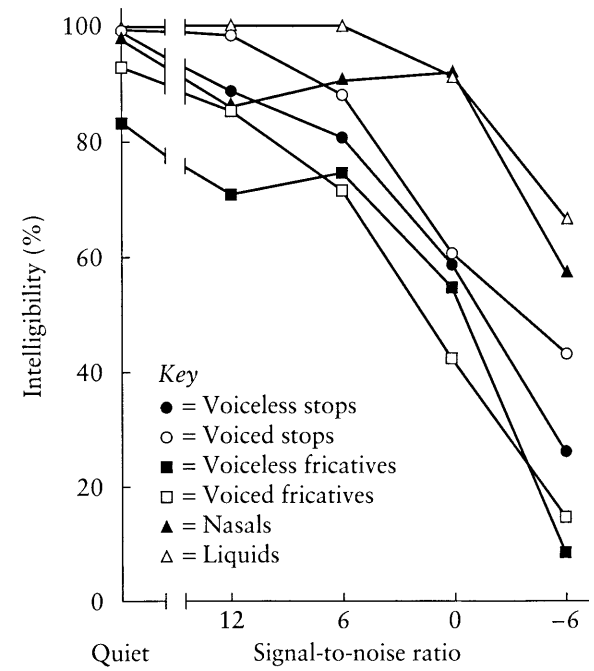


Figure 8.4.2 Effects of masking on consonant identification
Source: Clark 1983.

the test materials and by listeners' ability to use top-down sensitivity to the context being established by continuous speech.

Simple signal interruption is, of course, a relatively crude measure of the contribution of duration to intelligibility. Studies of the effects of time compression on speech indicate that the formant transitions in the onset and codas of syllables tolerate very little compression, but that the effects are quite variable on other parts of syllabic structure. Duration as part of the phonological structure of speech naturally plays a role in intelligibility: words of longer duration are typically more intelligible than shorter ones. In monosyllables, this is a function of the overall perceptual salience of the phonological structure of the syllable itself. Thus the word *hoof* is likely to be less intelligible than the word *rage*. The first word not only has a much shorter syllabic nucleus, it also has consonants at the onset and coda which are acoustically weak and relatively easily masked. Polysyllabic words are more complex, because the prediction of their intelligibility will depend on a mixture of duration, phonological structure and lexical familiarity. Consider for example the word *secretary*: there are not many English polysyllabic words beginning with similar sounds and having a similar stress pattern (such as *secondary*, *secular* and *sacrament*). Put any of these similar words into a reasonably genuine context (such as 'who's the departmental secretary?' or 'what kind of secondary school did you attend?') and the chances of mishearing them are quite low. Our familiarity with particular words in particular contexts thus introduces a significant top-down component into the recognition process. Data illustrating some of these effects can be found in Rubenstein et al. (1959) and Schultz (1964).

8.5 Acoustic-phonetic perception

Many general speech intelligibility studies have been motivated by what might be described as global interests in the properties of the speech signal in the context of the adequacy of communications systems or the impairment of hearing. As facilities for acoustic analysis, synthesis and signal processing have improved, researchers have investigated the detailed phonetic aspects of speech perception with the object of discovering how the cues to perceived phonological structure are encoded in the acoustic signal itself.

The pioneering studies of acoustic cues to the perception of phonological structure were undertaken at the Haskins Laboratories, using the painted spectrogram technique described earlier in section 7.17. These studies showed some of the ways in which formants and other spectral patterns encode the phonetic identity of segments in the time and frequency structure of the syllable. For details see Cooper et al. (1952) and Delattre et al. (1955). These early experiments demonstrated, among other things, the value of speech synthesis as a tool in the investigation of speech perception. With synthetic speech, the spectrum can be manipulated in a controlled fashion to check the perceptual significance of its dynamic spectral parameters.

Using synthesized speech, researchers from the Haskins group and elsewhere have shown that if a parameter is changed in equal increments from a value encoding a reliable percept of one segment, to a value encoding a reliable percept of another, listeners reach a point of sudden change in their perception from one segment to the other. There is no significant region of indecisiveness in the perception of sounds synthesized in the region of intermediate values. In other words, listeners do not gradually change their opinions on the identity of the stimulus in line with the progressive changes in the signal, but make a quite sudden changeover. The most striking form of this effect occurs when voice onset time (VOT) is delayed in stop consonants. If the delay is increased in small steps (say 10 ms) from around zero to about 100 ms after the release of the occlusion, English-speaking listeners continue to hear the stop as voiced up to about 20 or 30 ms (and perhaps up to 40 ms for velar stops), always depending on the particular stimulus properties. The next 10 ms increment then brings a switch in judgement and the stop is heard as voiceless. Figure 8.5.1 shows the effect, using idealized data.

This effect is known as CATEGORICAL PERCEPTION. Its presence in speech perception is not surprising, given that phonological organization is a matter of discrete options; in the context of acoustic and auditory analysis, it is appropriate to describe such perception as categorical. A further illustration emerges when listeners are asked to identify pairs of stimuli from a continuum as 'same' or 'different'. In general, we are not sensitive to differences within a series of values which we commonly count as occurrences of the same sound. As shown in figure 8.5.1(b), it is only around the VOT value at which listeners identify a change from voiced to voiceless that they can reliably hear a difference between pairs of stimuli. In other words, discrimination is weaker within the boundaries of a perceptual category, and sharper at or near the boundary. This again demonstrates the fundamental principle of functional contrastiveness. The effect has also been illustrated for formant transition frequencies and durations. (See Studdert-Kennedy 1976, Pickett 1980, Lieberman and Blumstein 1988 and Raphael 2005 for further discussion of this field of research.)

Category boundaries are of course language-dependent, at least to some extent. Thus English commonly has marked aspiration (delayed VOT) on stops, serving as a cue to their voicelessness, and shows larger VOT categories than languages like French (in which voiceless stops are generally not aspirated) or Thai (in which there is a three-way phonological distinction of voiceless aspirated, voiceless and voiced stops). There is also evidence that where more than one cue determines a category choice, trading relations may exist among the cues. For example, Repp (1979) has shown that aspiration, duration and intensity may be traded against each other in establishing the boundary of the voicing category in English.

Studies of animal perception suggest that categorical perception is not specific to human speech and hearing, but perhaps partly a consequence of general psychophysical boundary effects. If so, categorical perception need not be taken to be uniquely phonologically motivated: it may be that language capitalizes, as it were, on a basic psychoacoustic capability to optimize the phonetic processing of stimuli. See Diehl et al. (2004) for a good discussion of this.

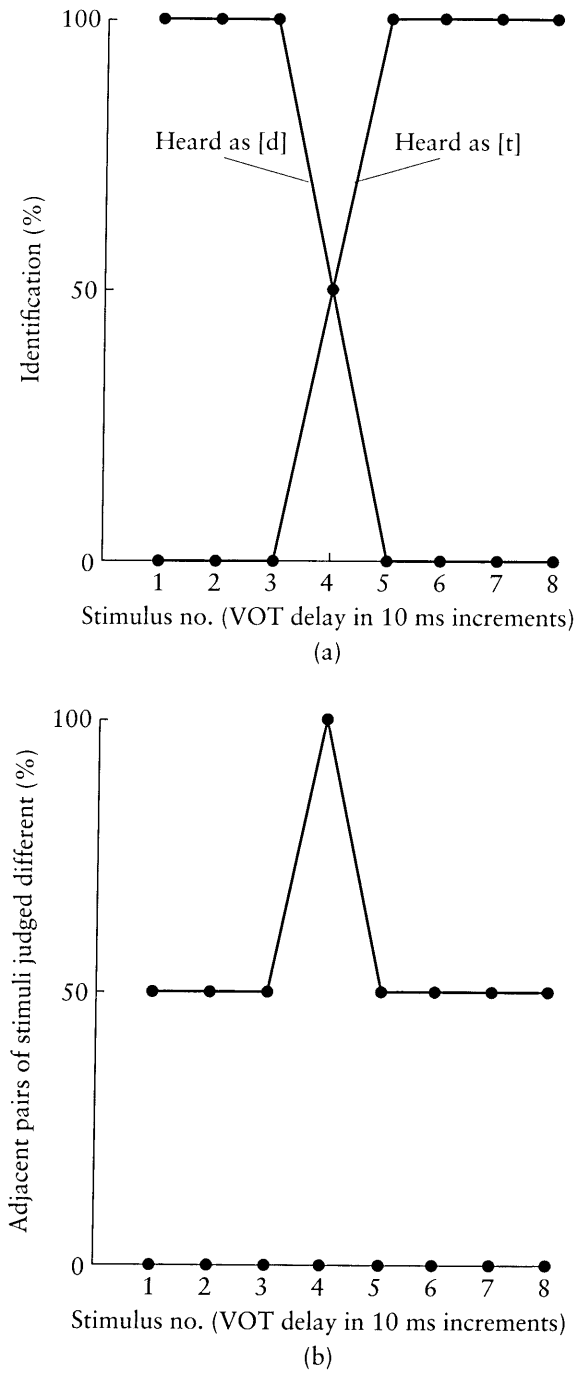


Figure 8.5.1 Perceptual responses to VOT delay: (a) identification; (b) discrimination

8.6 Vowel perception

The prime importance of the values of the first three formants in the encoding of vowel quality was confirmed in the early Haskins experiments (section 7.15 above). It has also been shown by Carlson et al. (1975) that accurate percepts can be obtained from synthetic vowels using only two formants, where F_2 is adjusted upwards to compensate for the absence of the high-frequency energy of the upper formants. Peterson and Barney (1952) recorded natural vowels in words beginning with /h/ and ending with /d/ from a range of speakers (men, women and children), analysed the formant structure of these, and conducted perceptual studies using the same recordings. Their analysis, and later work by Shepard (1972), showed that where perceptual confusions occurred, they were generally well correlated with acoustic proximity as defined by the three lowest formants. Their data also show a remarkable degree of variability among supposedly identical vowels and overlap between apparently different vowels. Work on Australian English by Bernard and Mannell (1986) and Cox (1999) demonstrates comparable variability and overlap. Figure 8.6.1 shows the variability of a number of Australian English vowels and the overlap among them when their formants are plotted against each other.

These data reveal an important aspect of vowel perception, namely the crucial importance of the systemic nature of the formant-specified acoustic relationships: we distinguish vowels from each other, and are less concerned with their absolute values. We have already had cause to note that there is significant diversity in the acoustic properties of the vowels of children, women and men, arising from differences in vocal tract length, as well as further diversity due to differences among individuals in their vocal tract and in the habitual settings of their speech organs (section 7.15 above). As a consequence, researchers have formulated mathematical algorithms for normalizing data variance, particularly that which results from variations in vocal tract length.

Ladefoged and Broadbent's experiment demonstrated that formant frequencies only determine phonological identity within a vowel system (Ladefoged and Broadbent 1957, and section 7.15 above). Using synthetic speech they showed that if the complete vowel system in a sentence was shifted except for the test vowel, listeners would reliably normalize if the systemic shift effectively placed the vowel within the bounds of the acoustic specification of a phonologically distinct vowel. Thus the vowel of *head* could be made to be heard as the vowel of *hid* if the formant frequencies of all the other vowels were lowered. It seems that listeners can normalize to a new speaker within the first few words that they hear. (See further discussion in Holmes 1986 and Pisoni 1997.)

Nevertheless, vowels also seem to differ from many consonants in being identified along a continuum of values rather than categorically. Fry et al. (1962) used synthetic speech and the labelling and discrimination techniques previously applied to consonants to generate a continuum of vowels with precise increments in formant values. They were unable to find the same categorical shift in labelling or the same peaks in discrimination. This suggests some

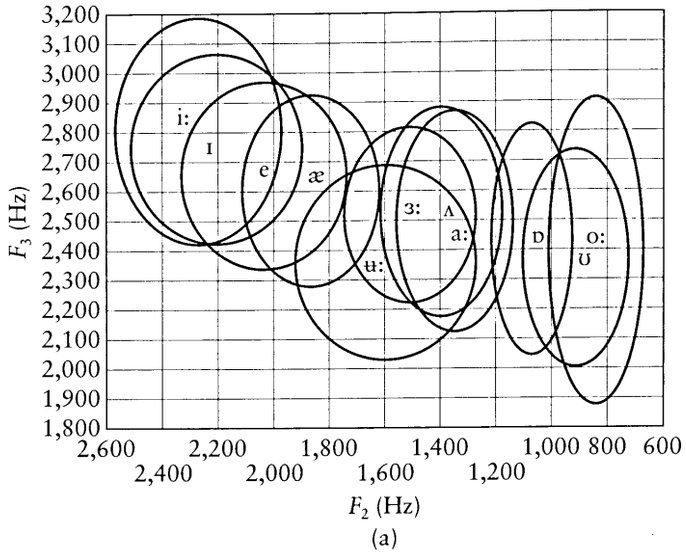


Figure 8.6.1 (a) F_2/F_3 plot (2 standard deviation ellipses) of the monophthongs produced by 172 adult male speakers of Australian English; (b) F_1/F_2 plot (2 standard deviation ellipses) of the monophthongs produced by 172 adult male speakers of Australian English

justification for the tradition of describing most consonant sounds in terms of a discrete set of production categories, but characterizing the acoustic and articulatory possibilities in vowel production in terms of continua.

Our discussion here has followed most researchers in concentrating on the first few formants as the acoustic determinants of vowel identity. But it has been persuasively argued by Strange et al. (1983) that when listeners identify the vowels of natural speech, as opposed to experimentally constrained synthetic stimuli, they also depend upon the dynamic coarticulatory transitional information in the formant structure of the syllable. While this has been challenged by some (e.g. Harrington and Cassidy 1994), it seems highly likely that listeners do, in normal situations, gain extra information in this way.

8.7 Consonant perception

As discussed earlier in section 7.17, the work by the Haskins group using painted spectrograms to synthesize stimuli provided basic evidence of the principal acoustic cues to place of articulation in stops, and to the voiced–voiceless distinction. Extensive work at Haskins and elsewhere using synthetic speech has provided detailed knowledge of most of the acoustic features of consonants. This includes the notion of the formant locus, and the role of noise bursts as cues to voicing and place of articulation in stops. Figure 8.7.1 illustrates the

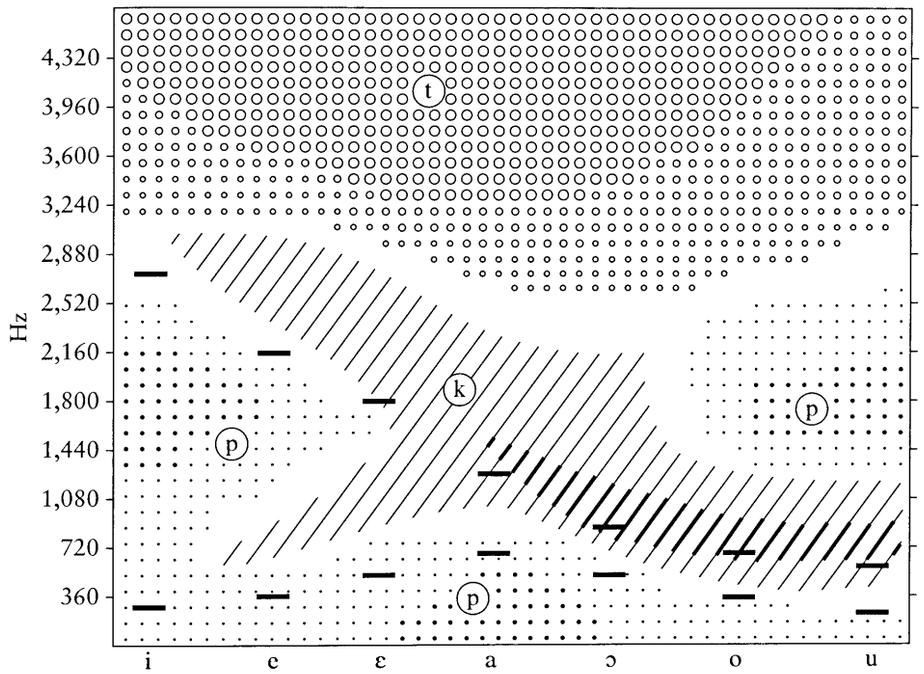


Figure 8.7.1 The role of the noise burst spectrum. Twelve different noise bursts (their centre frequencies shown along the vertical axis) were combined with seven different vowels and presented to 30 listeners. The zones show the dominant perceptions of the burst as [p], [t] or [k] according to its frequency and the following vowel *Adapted from*: Cooper et al. 1952.

role of the noise burst spectrum and its coarticulatory relationship with the following vowel in CV syllables, in well-known data from Cooper et al. (1952). The Haskins work also showed that transitions and the rate of change of formant transitions at the onset and coda of a syllable had a significant role in distinguishing stops from sonorant consonants (Liberman et al. 1956), in identifying approximants (O'Connor et al. 1957), and in identifying nasals and stops (Liberman et al. 1954). Harris (1958) also showed that the spectral structure of the noise in fricatives provides a major cue to their perception.

Since this early work, the technology of speech analysis and synthesis has become far more sophisticated, accurate and flexible, and there is now a large body of literature on the acoustic cues to a variety of consonants. In general, these studies accord with the consonantal acoustic properties described from the point of view of production in chapter 7 above. Fant (1973), Shoup and Pfeiffer (1976), Pickett (1980), O'Shaughnessy (1987) and Raphael (2005) provide extensive reviews of the work on perceptual features of consonants.

8.8 Units of perception

The phoneme as a unit of linguistic processing generally, and of perceptual processing in particular, continues to be defended by many researchers. Work by Warren (1970, 1984), for example, has demonstrated that when segments are excised from the stream of speech and replaced by noise, listeners will report hearing the correct missing segment. They presumably restore the segment by top-down contextual prediction.

Much of the classic work in speech perception has chosen to focus on investigating cues in the acoustic signal which encode the identity of phonological segments. Yet, as even the very early Haskins work showed, these cues are not generally discrete or invariant. Nor does the syllable simply consist of a concatenation of discrete, isolated phonological features and segments. Rather, the features and segments may overlap with each other, and are materially influenced by the phonological structure and context of the syllables in which they are produced. (See section 4.1 above, Fant 1973, and Fowler and Smith 1986.)

This has led some researchers to consider the syllable as the primary unit of production and perception. (Compare the argument in section 7.17 above that the acoustic structure of segments can be properly understood and described only within the context of the acoustic syllable.) Studdert-Kennedy (1976) describes the syllable as a 'symbiosis of consonant and vowel' which acts as the effective vehicle for the transmission of linguistic information. The greater salience of the syllable than the segment is also suggested by a speech error experiment by Tent and Clark (1980), in which listeners detected syllable-level errors far more readily than segment errors. Crompton (1982) also argues from speech error data that the syllable is the primary unit in which articulatory patterns are stored; if, as many researchers believe, there are direct links between production and perception, this also has implications for perception.

By contrast, Blumstein and Stevens (1979, 1980) sparked a major debate by arguing, on evidence from both production and perception, that in certain segments, notably stops, there were invariant spectral cues in the acoustic signal. Other researchers have argued for the existence of subsegmental units of perception in the form of phonological features. Such work has often used multivariate statistical data reduction to obtain the necessary supporting evidence. The work of Miller and Nicely (1955) is an early example of this approach. Their primary data were presented segmentally in confusion matrices, of the kind shown in figure 8.8.1: here the test segments presented to listeners are shown in the rows of the matrix and the sounds heard by the listeners in the

		SPOKEN CONSONANT																						
		p	t	k	b	d	g	ʧ	dʒ	f	θ	s	ʃ	h	v	ð	z	ʒ	m	n	ŋ	l	r	w	j
PERCEIVED CONSONANT	p	3		3						1															
	t	1	3	6				4		1		2													
	k	10	7	4						2															
	b		1		1												1								
	d					9						2	1				1								
	g			1		2	1		2																
	ʧ							3					11												
	dʒ				3	3	9		8																
	f	1	2					1		9		7													
	θ									1							1								
	s		1					2				2	1												
	ʃ							2					1												
	h			1																					
	v				10					1					15		1						1	1	
	ð												1				1					1			
	z					1						2					4						1		
	ʒ						1																		
	m																		13						
	n																		2	15					
	ŋ																								
	l																					12	1		
	r							1									5					2	10	7	5
	w																						1	7	
	j						4		5														1		10
	null		1		1			2									1								
	%	20	20	27	7	60	7	20	53	60		13	7		100		27		87	100		80	67	47	67

Figure 8.8.1 Confusion matrix for consonants in a CV frame heard in noise of equal intensity. The data are from 15 subjects listening to masked natural speech
Source: Robert Mannell, Macquarie University.

columns. Such forms of presentation are useful in allowing an immediate view of the pattern of perceptual errors. Simple visual inspection may, however, fail to reveal important underlying patterns.

Further analysis of these data suggests that there are regular underlying relationships between the listening conditions and the intelligibility of certain phonological features, as shown in figure 8.8.2.

Researchers using statistical analysis techniques, such as multidimensional scaling and hierarchical clustering, have conducted further analyses of intelligibility data to provide visualizations of the perceptual properties of features and segments in relation to listening conditions. Shepard (1972), Wang and Bilger (1973) and Singh (1975) provide detailed accounts of such studies, including the statistical methods used. Other approaches include that of Wickelgren (1966), who undertook feature-based analyses of short-term memory

	SPOKEN MANNER					
PERCEIVED MANNER		ST	AF	FR	NS	AP
	ST	58	20	15		
	AF	13	37	15		
	FR	22	17	63		7
	NS				100	
	AP	4	20	7		93
	null	2	7	1		

	SPOKEN VOICING		
PERCEIVED VOICING		+V	–V
	+V	93	10
	–V	4	88
	null	2	3

	SPOKEN PLACE			
PERCEIVED PLACE		LD	AL	VL
	LD	75	15	10
	AL	3	66	53
	VL	20	10	20
	null	2	9	17

Key
Manner (all consonants)
ST = Stops
AF = Affricates
FR = Fricatives
NS = Nasals
AP = Approximants and semi-vowels

Voicing (stops, affricates and fricatives only)
+V = Voiced
–V = Unvoiced

Place (stops, affricates and fricatives only)
LD = Labial and dental
AL = Alveolar and postalveolar
VL = Velar

Figure 8.8.2 Effective intelligibility of selected features. The figures are percentages of inter-class confusions (intra-class confusions are not included)
Source: Robert Mannell, Macquarie University.

error in consonant recall. He concluded that feature-based analyses had greater explanatory power for the data than segments alone, and that the explanatory power of some feature sets was greater than others.

Other researchers have sought evidence for the existence of specific perceptual feature detection mechanisms, prompted by the more general evidence of functionally specific neural auditory detectors in cats and other animals. Eimas and Corbit (1973) conducted a series of experiments which demonstrated that it was possible to shift phonetic category boundaries by repeatedly presenting a stimulus at one end of a feature continuum (such as VOT). Their hypothesis was that this effect might be explained by the fatiguing of the relevant feature detector and the increased sensitivity of the contrasting feature detector, causing a shift in category boundary. Research since then has not revealed substantial evidence to advance this claim, which is now regarded with some caution.

Much of the investigation of the perceptual units of speech described so far has relied on manipulation of the spectral time-course of the speech signal, either by reprocessing natural speech or by parametric manipulation of formant coded synthetic speech. A different approach has been to present listeners with a natural speech recording in which the time-domain waveform has been 'gated' so that only a precise fraction of the signal is heard by the listener. The duration of this gated fraction is usually progressively increased to a point at which the signal is likely to be reliably identified by most listeners. In its simplest form such an experiment might align the start of the gate with the start of the consonant stimulus in, say, a CV syllable, and then lengthen the duration of the gate incrementally until the consonant is reliably identified. For initial stops it has been shown that the first 10 to 15 ms of the release burst is often all that is needed for accurate identification. Studies of various classes of consonants reveal that sounds such as fricatives, which have less rapid changes of spectrum after release, require longer gate times for reliable identification; although for most sounds the required duration remains well under 80 ms. Interestingly, these results indicate that identification of consonants does not always rely on formant transition from the acoustic nucleus of the syllable, although in some instances this does improve the reliability of identification.

What then is the basic perceptual unit of speech? No simple answer can be given, because there is no clear evidence pointing to just one unit. It is clear that we can perceive some features, such as voicing, without correctly identifying the segment in which that feature is present. On the other hand, in some instances, cues to segment identity are distributed across the entire syllable, or at least across more than one segment; for example, the voicing of postvocalic fricatives in English is often detected from the length of the preceding vowel, not from any strong presence of periodicity caused by phonatory modulation of the fricative noise. In general, we may say that the syllable provides the normal acoustic structure of the continuous speech. Cues to phonological structure may be distributed across the syllable in various ways that allow us to perceive both phonological features and segments. But the syllable is not always an absolutely essential structure for the communication of all information about phonological features or segments.

8.9 Prosodic perception

The term 'prosodic' is used here to refer to linguistic information of the kind often described as rhythm and intonation (which will be dealt with in detail in chapter 9 below). The chief acoustic parameters of relevance here are duration, fundamental frequency and intensity. As we have already seen, these features may also encode phonological information within segments and syllables, but we are now concerned more with their functions across longer stretches of speech. (This is another reminder that acoustic cues often serve more than one function: the encoding of speech is complex and multilayered.)

Duration illustrates the point, for it signals various things. As indicated in sections 8.6 and 8.7 above, it contributes to segmental contrasts in English, in the distinction between long and short vowels, in the VOT distinction between voiceless and voiced stops, and in the encoding of postvocalic consonant voicing in the length of the preceding segment. Across longer stretches of speech, duration is a measure of speaking rate. There is, however, no direct relationship between the overall rate of utterance and the durations of syllabic and segmental structures within the stream of speech. A number of studies of speaking rate have shown that as the rate increases, the speaker preserves those aspects of the acoustic structure which are valuable for encoding segmental and prosodic structure; at the same time, listeners are able to compensate for increased coarticulatory effects and for the spectral and temporal contraction of less important information. See O'Shaughnessy (1987) for a useful general overview, and Allen and Miller (2001), who show that speaking rate has a different effect on the location of phonetic boundaries from other effects like lexical status.

The speech rhythm of a language such as English is perceived in the durational interplay of prominent (or 'stressed') syllables and weaker or less prominent ones. English has traditionally been considered to have an isochronous pattern of rhythm, that is a pattern in which prominent syllables seem to occur at roughly equal intervals, regardless of the number of weak syllables occurring between the prominent ones (see section 9.3 below). Roach (1982), Buxton (1983) and Dauer (1983) all suggest that despite this perceptual effect, the speech production evidence for isochrony in English is rather weak. Buxton concludes that it is likely that other factors, such as distributed coarticulatory acoustic cues, may contribute to the strength of the isochrony percept.

In investigating temporal patterning, phoneticians have tried to identify the point or points (so-called P-CENTRES) in a stream of speech which are perceived to be the location of prominence or stress. Experimental evidence suggests that these perceived locations depend on syllable duration and total syllabic structure, rather than on the particular segmental constituents of the syllable. Morton et al. (1976) showed that if a series of syllables is spaced so that there is equal time between successive syllable onsets, listeners do perceive a pattern of isochrony. But if the spacing is based on p-centres, a much stronger effect of rhythmicality is perceived.

While the interest in p-centres in rhythm research has waned in recent years, there is still a reluctance to abandon totally the role of linguistic rhythm in speech perception. Roach (1982) and Dauer (1983) also suggest that in languages like English which show vowel reduction in unstressed syllables and which have heavy syllables (i.e. those that contain a long vowel and/or a consonant coda), inter-stress intervals tend to be perceived as more or less equal. See also the work by Cutler and colleagues (reviewed in Cutler et al. 1997), who provide some good evidence that fundamental rhythmic characteristics of a language may influence word-processing strategies in these languages. However, Klatt (1976) and Crystal and House (1988) also claim that syllable timing patterns in English words can reliably be modelled on the basis of segment duration characteristics and stress patterns, without any specific reference to 'isochrony'.

Cutler (2005) reviews the research on stress perception in general. Much of this work has focused on the relevant contributions of pitch, loudness and length to the perception of linguistic prominence, with often conflicting results (see also sections 9.2 and 9.6 below). Some studies claim that pitch is the major cue to stress, some claim loudness, duration or even spectral tilt. Cutler makes the very important point that current research really elaborates and extends earlier findings of the 1950s and 1960s. Crucial to this debate is whether researchers are conflating word stress and postlexical intonational prominence or accent, because cues may be different depending on the level of stress concerned (see also section 9.2 below). Cutler also points out that there are often language-specific differences in the perception of prominence. For example, in Welsh (Williams 1985), prominence is cued by duration, but the role of F_0 is equivocal. Once again, it is important to bear in mind the nature of prominence across languages, before any hard-and-fast conclusions can be made about stress and prominence perception. Beckman (1986) also provides a good background to this often perplexing area of research.

Our sensitivity to small changes in pitch, its consequent strong perceptual salience, and our capacity to control pitch in speech production are discussed in some detail in sections 7.9 and 7.19 above and 9.2 below. As with duration, pitch provides several layers of information to the listener. It is a major contributor to voice quality, it helps us to identify the sex and age of a speaker, and it can in some cases be a means of distinguishing among individual speakers. It even seems to be the case that listeners make judgements about the personality, attitude and even truthfulness of speakers on the basis of pitch information; Cooper and Sorenson (1981) give a useful overview of studies that have investigated these global (and largely nonlinguistic) aspects of information which listeners derive from fundamental frequency. Mullenix (1997) also discusses in some detail the problem of perceptual adjustment to voice.

Despite the significance of fundamental frequency as a cue, Brown et al. (1980) report that even trained listeners have difficulty in making accurate estimates of the magnitude of pitch movement in prominent syllables. In fact, the experience of many introductory classes in phonetics and phonology shows that some students are initially unable to identify the direction of perceived pitch movement in a prominent syllable consistently, much less its magnitude

relative to other points of pitch-based prominence in the same speech sequence. This limited ability to make accurate judgements about local detail in pitch patterns is, of course, unsurprising given the enormous variability among speakers in their production of fundamental frequency patterns. What has long been established is that listeners do make very effective use of the dynamics of fundamental frequency patterns as the basis for judgements about contrasts that are relevant within the particular language (such as stressed versus unstressed or querying tone versus determinate). Further details can be found in chapter 9 below, in overviews such as Lehiste (1970), Gandour (1978) and Vaissière (2005), and in specific treatments of perception of intonational events by Ladd and Morton (1997) and Kohler (2004). Cutler et al. (1997) is a good summary of a wide range of research that has explored the link between prosody and language processing.

8.10 Word recognition

This chapter has so far concentrated on phonetically motivated approaches to understanding the perception of speech, based on our bottom-up processing of the acoustically encoded cues in the spectral time-course of speech. Less central to phonetics and phonology and of more significance in cognitive psychology is work on the cognitive processes involved in the recognition of words – how listeners process phonological structure sequentially and how they access lexical information from memory.

We have already noted earlier in this chapter the effects of top-down influences such as context and word familiarity in mediating reliable perception. Warren's phonemic restoration effect, described in section 8.8, is an example of top-down processing making use of context and the listener's linguistic knowledge base. Work by Samuel (1981) also shows that words in common use show a stronger restoration effect and that the effect is stronger for word-final segments than for word-initial segments. However, Samuel (1996) also suggests these effects may be somewhat fragile and dependent on the nature of the replacement sound and replaced phoneme.

One well-known way of investigating this question as a matter of cognitive processing is a speech shadowing task in which subjects repeat what they hear as quickly as possible after it is spoken. Marslen-Wilson (1985) and Marslen-Wilson and Welsh (1978) have shown that skilled listeners are able to shadow a speaker so closely that words can be recognized as little as 200 ms after their onset. In such rapid shadowing, the listener has generally had too little time to respond to the acoustic cues alone and must therefore be making top-down predictions as well.

Some further insight into this process comes from Aitchison and Straf (1982), who compared adults' and children's errors in retrieving words. Although this experimental work investigates retrieval rather than direct perception, it suggests that children rely far more on macrophonetic aspects of the word being recalled (such as rhythm and the location of the stressed syllable) but that adults rely more on initial consonants (perhaps implying that adults have more recourse to their extensive mental lexicon). There remains considerable debate about the processes involved in lexical access, and about the roles of top-down and bottom-up processing, and the way in which these are integrated in the overall perceptual task. In a discussion of evidence from the literature, Marslen-Wilson (1989b) concludes that linguistic contextual information does not necessarily override or unreasonably constrain the use of bottom-up information from the speech signal itself. See Luce and McLennan (2005) for a good summary of this debate.

8.11 Models of speech perception

Research into speech perception still awaits the development and verification of a comprehensive explanatory model, although several models have been proposed. We concentrate here on those constructed from an essentially phonetic and phonological perspective.

An early version of the ANALYSIS BY SYNTHESIS model is described by Stevens and Halle (1967). The model is computational in approach and assumes, in essence, that listeners perform a spectral analysis of the incoming speech signal, resolving it into features and parameters which are then stored. The acoustically analysed information is then further analysed to provide an estimate (which may also be mediated by higher-order information) of the phonological structure of the input. This estimate or trial form of the phonological structure is operated on by a phonological rule system to generate a hypothesized utterance which is compared with an appropriate neural auditory representation of the analysed input. If the match is good, the hypothesis is taken to be correct and accepted. If the match is poor, the process is iterated until an acceptable match is obtained.

The MOTOR THEORY (MT) is one of the oldest, best-known and most widely criticized of the phonetically based models of perception. Its basic hypothesis is that we decode the perceived acoustic signal in terms of stored articulatory patterns which can generate an acoustic signal with the same linguistic percept. The theory gained currency through the proposals of Liberman et al. (1967). A more recent version of MT (Liberman and Mattingly 1985) maintains that stored articulatory patterns have a more abstract status as underlying forms representing articulatory intentions which are directly perceived by the listener. Defenders of this theory have yet to provide an explanation of how the model works in detail, and of how the storage and accessing of the underlying articulatory information are accomplished.

One of the claims of MT is that the perceptual mechanism invokes specialized phonetic processes that are specific to speech. An alternative set of theories has emerged that does not propose specialized 'speech' modes of perception. One of these theories, proposed in the 1980s, is DIRECT REALISM (see Fowler 1986,

1996 for a detailed description). Unlike MT, the direct realists suggest that rather similar perceptual strategies to those used in visual perception, for example, are deployed in the perception of speech. There is a direct relationship between signal and percept, with the object of perception being the event that produces the signal. These events are the articulatory gestures. There has been a long, passionate debate between the direct realists and another group of speech perception researchers who support a more general auditory and perceptual learning approach (e.g. Diehl and Kluender 1989, Kuhl 1992, Jusczyk 1993). Like direct realists, proponents of the general auditory approach (championed most vigorously by Diehl and colleagues) do not invoke a special speech perception mode, but presume that speech sounds are perceived using the same mechanisms of audition and perceptual learning that have evolved in humans or human ancestors to handle other classes of environmental sounds (Diehl et al. 2004, p. 154). Unlike proponents of direct realism and MT, supporters of the general auditory approach believe that the recovery of meaningful phonetic and linguistic units occurs via the processing of (psycho)acoustic cues without direct reference to any underlying articulatory gestures.

Taking a more computational or engineering approach to speech perception that does not assume any role for articulation, Klatt (1979, 1981) has proposed a highly signal-driven model called LEXICAL ACCESS FROM SPECTRA (LAFS). This model assumes a very large store of spectral patterns or templates as the basis for identifying all familiar words held in the listener's memory. It avoids any postulation of stored segmental representation or of segmentally organized analysis of incoming speech, and thus bypasses many of the problems of context-sensitive variability in the spectral representation of segmental sequences, both within words and across word boundaries. Decisions about phonetic identity are made using spectral distance metrics which allow the match between the input spectrum and competing spectral templates to be scored, the candidates compared and a choice made. This model presupposes very powerful analysis, storage, access and decision processes in any computational realization of it. It does, however, address in a direct way some of the realities of dealing with natural speech which are swept away by various forms of cognitive or linguistic abstraction in other models.

The TRACE model is probably the best known of the models of speech perception and recognition inspired by work on connectionist models of cognition. While it is probably more a model of spoken word recognition rather than speech perception as such, the model, as described by Elman and McClelland (1986), depicts a procedure that begins by generating spectral slices from the input signal every 5 ms. These form the input to a set of interconnected processing elements, known within connectionist models as nodes, which act as feature detectors. Connections to the nodes are either excitatory or inhibitory, and the features themselves are defined in terms of spectral properties. Progressive slices of analysed speech will either inhibit or cumulatively excite a given node and so identify a particular feature. The feature nodes are in turn connected to a set of segmental detection nodes, and the same basic process is repeated to accumulate a decision which will identify a particular segment. In turn, the outputs of the segmental nodes are connected to a set of word

detection nodes. Interconnections between nodes are weighted to adjust their level of contextual influence on node output.

Luce and McLennan (2005) provide an insightful discussion of Trace and other models of spoken word recognition, including SHORTLIST (e.g. Norris 1994), a descendant of Trace, PARSYN (Luce et al. 2000) and the DISTRIBUTED COHORT model or DCM (e.g. Gaskell and Marslen-Wilson 2002). While the details of these vary somewhat in their fundamental architecture, and in whether they have mediated access or direct access to phonetic features of a token, what they have in common is that an initial stage of normalization or recoding takes place at a prior level of speech processing that removes the elements of talker variability or 'noise' associated with tokens.

In a more or less parallel development, the years since 1995 have witnessed an increased interest in exemplar-based models of speech perception (see Johnson 1997, Pisoni 1997), which exploit indexical information (e.g. talker-specific detail) as well as information about the token from overlapping acoustic signals. A distinction is often drawn between this approach and abstractionist speech perception models, that is, most models based on analysis by synthesis perception, which essentially normalize out differences between speakers 'prior to identification of linguistic categories' (Johnson 1997, p. 145). In essence, an exemplar view suggests that a perceptual category is effectively all remembered instances or memory traces of that category, which may be marshalled to 'perceive and categorize novel stimuli' (Pisoni 1997, p. 28). Indexical traits such as dialect or variety, talker identity and so on are not treated as superfluous in the perceptual process, but instead form part of the mental or neural representation of the particular category (see Johnson 1997, pp. 152–3, where he outlines a connectionist model that circumvents the obvious auditory storage problem that would arise if all auditory instances of a token were stored, for example). Exemplars are not the same as 'speech prototypes', as proposed by Kuhl (1992) in another attempt to deal with variability in the speech signal. Kuhl suggested that these speech prototypes effectively act as perceptual magnets in that 'they assimilate near neighbours by effectively reducing the perceived distance between the centre of the category and the outlying members of the category' (Kuhl 1992, p. 251). Prototypes represent an *abstraction* of an actual instance or averaged instances of a category.

In many respects the LAFS model proposed by Klatt (1979, 1981) is not too dissimilar from the exemplar approach, because there is a shared focus on the particularities of the speech signal, rather than on normalized or abstract elements of the signal. Criticized by some as being overly 'nonanalytic' or too 'analogic' (see Johnson 1997, p. 162, for counter-arguments), exemplar models are nevertheless influential not only in speech perception research, but in newer connectionist models of spoken word recognition (see Luce and McLennan 2005) and models of phonological representation (e.g. Pierrehumbert 2003).

Other perceptual models can be found in the literature, and the above review cannot do justice to this diverse field. For a comprehensive critique of traditional speech perception models, see Klatt (1989), and for a summary of current models of speech perception, see Diehl et al. (2004). A set of classic papers covering the main areas of speech perception research is included in Miller

et al. (1991). Strange (1995) contains a useful series of papers addressing the specific issue of cross-language perception of consonant and vowel contrasts, and Johnson and Hume (2001) is a diverse collection of papers focusing on the interplay of speech perception and phonology. Johnson and Mullennix (1997) present a selection of papers dealing with talker variability, and Pisoni and Remez (2005) offer a collection of papers covering all of the major areas of speech perception research.

8.12 Conclusion

A very large body of information about speech perception has been collected since the 1950s, and our knowledge of the basic acoustic correlates contributing to many phonological features and segments is now quite extensive. The failure to establish an unassailable case for a particular basic unit of perception probably reflects the fact that linguistic information is encoded in the speech signal at various levels and in ways that exploit interdependence and redundancy. It is evident, for example, that some features and segments can be reliably identified within tens of milliseconds from the onset of the syllable, while others rely on information distributed across the entire syllable, and even beyond.

In recent years, some researchers have turned their attention to audiovisual speech processing, with particular interest in the 'McGurk effect' (McGurk and Macdonald 1976). This effect refers to what happens when an auditory stimulus like /ba/ is dubbed on to a visual stimulus /ga/, but the listener perceives /da/ or /va/. This has expanded into interest in the contribution of facial movement to speech perception. See Bernstein (2005) for a useful general discussion of audiovisual speech perception in the context of phonetic processing in the brain, and Munhall and Tohkura (1998), Massaro and Cohen (1999) and Davis and Kim (2004) for specific examples of experimental research in this area.

The array of models of speech perception reflects the lack of a unified understanding of perceptual processes and of the complex interaction of its top-down and bottom-up aspects. This debate is still very much ongoing. None of the models of perception which have been computationally implemented has been demonstrated on more than a very limited set of test materials, although researchers are increasingly aware of the need to test their models in the context of the enormous variability among speakers and the complexities of rapid continuous speech which are the everyday reality of actual discourse.

Exercises

1 Describe the structure of the human ear, including the outer, middle and inner ear. What role does each part of the ear play in the process of hearing?
2 What does it mean to say that the auditory system is 'frequency selective'?
3 Explain the difference between top-down and bottom-up speech processing.

4 What is meant by 'categorical perception'?
5 What evidence could you use when arguing that we can understand what someone is saying without hearing every detail of the speech signal?
6 Why is it difficult to identify any particular unit of language (such as phoneme or syllable) as the basic unit of perception?
7 We hear with the brain, not with the ear. What evidence can be used to support this claim?
8 From the following list of names, choose pairs which you judge most likely to be confused with each other in a telephone conversation. Explain why.

Anderson, Cannon, Flanders, Flinders, Ford, Freeman, Hanson, Horne, Sanders, Sanderson, Shaw, Sleeman, Thorn