PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 947–963
doi:10.1098/rstb.2007.2152
Published online 7 September 2007

Basic auditory processes involved in the analysis of speech sounds

Brian C. J. Moore*

Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK

This paper reviews the basic aspects of auditory processing that play a role in the perception of speech. The frequency selectivity of the auditory system, as measured using masking experiments, is described and used to derive the internal representation of the spectrum (the excitation pattern) of speech sounds. The perception of timbre and distinctions in quality between vowels are related to both static and dynamic aspects of the spectra of sounds. The perception of pitch and its role in speech perception are described. Measures of the temporal resolution of the auditory system are described and a model of temporal resolution based on a sliding temporal integrator is outlined. The combined effects of frequency and temporal resolution can be modelled by calculation of the spectro-temporal excitation pattern, which gives good insight into the internal representation of speech sounds. For speech presented in quiet, the resolution of the auditory system in frequency and time usually markedly exceeds the resolution necessary for the identification or discrimination of speech sounds, which partly accounts for the robust nature of speech perception. However, for people with impaired hearing, speech perception is often much less robust.

Keywords: hearing; frequency selectivity; timbre; pitch; temporal resolution

1. INTRODUCTION

This paper reviews selected aspects of auditory processing, chosen because they play a role in the perception of speech. The review is concerned with relatively basic processes, many of which are strongly influenced by the operation of the peripheral auditory system and which can be characterized using simple stimuli such as pure tones and bands of noise. More central processes of auditory pattern analysis are described elsewhere in this volume. The resolution of the auditory system in frequency and time is characterized and its role in determining the internal representation of speech sounds is described. It turns out that the resolution of the auditory system in frequency and time, as measured in psychoacoustic experiments, usually markedly exceeds the resolution necessary for the identification or discrimination of speech sounds. This partly accounts for the fact that the speech perception is robust and resistant to distortion of the speech and background noise. However, hearing impairment usually leads to a reduced ability to analyse and discriminate sounds, and background noise then has much more disruptive effects.

2. FREQUENCY SELECTIVITY

(a) *The concept of the auditory filter*

Frequency selectivity refers to the ability to resolve the sinusoidal components in a complex sound, and it plays a role in many aspects of auditory perception, including the perception of loudness, pitch and timbre. Fletcher (1940), following Helmholtz (1863), suggested that frequency selectivity can be modelled by considering the peripheral auditory system as a bank of bandpass filters with overlapping passbands. These filters are called the 'auditory filters'. Fletcher thought that the basilar membrane within the cochlea (see Young 2008) provided the basis for the auditory filters. Each location on the basilar membrane responds to a limited range of frequencies, so each different point corresponds to a filter with a different centre frequency.

Frequency selectivity can be most readily quantified by studying masking, which is the process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound. The following assumptions are made about a listener trying to detect a sinusoidal signal in a broadband noise background.

(i) The listener makes use of an auditory filter with a centre frequency close to that of the signal. This filter passes the signal but removes a great deal of the noise.
(ii) Only the components in the noise which pass through the filter have any effect in masking the signal.
(iii) The threshold for detecting the signal is determined by the amount of noise passing through the auditory filter; specifically, threshold is assumed to correspond to a certain signal-to-noise ratio at the output of the filter.

This set of assumptions is known as the 'power spectrum model' of masking (Patterson & Moore 1986), since the stimuli are represented by their long-term power spectra, i.e. the short-term fluctuations in

*bcjm@cam.ac.uk

One contribution of 13 to a Theme Issue 'The perception of speech: from sound to meaning'.

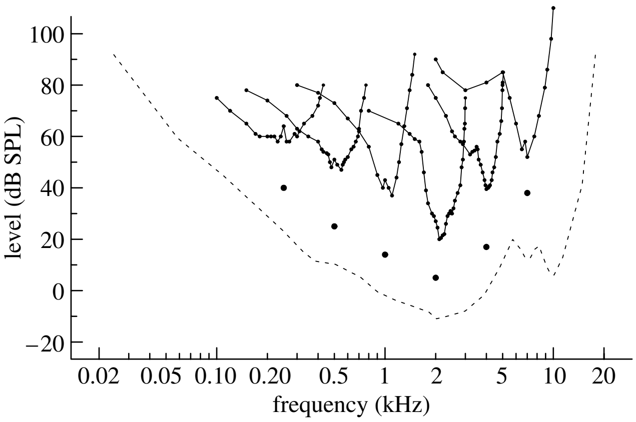

Figure 1. Psychophysical tuning curves (PTCs) determined in simultaneous masking, using sinusoidal signals at 10 dB SL. For each curve, the solid circle below it indicates the frequency and level of the signal. The masker was a sinusoid which had a fixed starting phase relationship with the 50 ms signal. The masker level required for threshold is plotted as a function of masker frequency on a logarithmic scale. The dashed line shows the absolute threshold for the signal. Data from Vogten (1978).

the masker are ignored. Although the assumptions of the model are not always valid (Moore 2003*a*), stimuli can be found for which the assumptions are not strongly violated.

The question considered next is 'what is the shape of the auditory filter?' In other words, how does its relative response change as a function of the input frequency? Most of the methods for estimating the shape of the auditory filter at a given centre frequency are based on the assumptions of the power spectrum model of masking. The threshold of a signal whose frequency is fixed is measured in the presence of a masker whose spectral content is varied. It is assumed, as a first approximation, that the signal is detected using the single auditory filter which is centred on the frequency of the signal, and that threshold corresponds to a constant signal-to-masker ratio at the output of that filter. Both the methods described below measure the shape of the filter using this technique.

(b) *Psychophysical tuning curves*

One method of measuring the shape of the auditory filter involves a procedure which is analogous in many ways to the determination of a neural tuning curve (see Young 2008), and the resulting function is called a psychophysical tuning curve (PTC). To determine a PTC, the signal is fixed in level, usually at a very low level, say, 10 dB above absolute threshold (called 10 dB sensation level, SL). The masker can be either a sinusoid or a narrowband noise.

For each of several masker frequencies, the level of the masker needed just to mask the signal is determined. Because the signal is at a low level, it is assumed that it produces activity primarily at the output of a single auditory filter. It is assumed further that at threshold the masker produces a constant output power from that filter, in order to mask the fixed signal. Thus, the PTC indicates the masker level required to produce a fixed output power from the auditory filter as a function of frequency. Normally, a filter characteristic is determined by plotting the output from the filter for an input varying in frequency and fixed in level. However, if the filter is linear, then the two methods give the same result. Thus, assuming linearity, the shape of the auditory filter can be obtained simply by inverting the PTC. Examples of some PTCs are given in figure 1.

One problem in interpreting PTCs is that, in practice, the listener may use the information from more than one auditory filter. When the masker frequency is above the signal frequency, the listener might do better to use the information from a filter centred just below the signal frequency. If the filter has a relatively flat top, and sloping edges, this will considerably attenuate the masker at the filter output, while only slightly attenuating the signal. Using this filter the listener can improve performance. This is known as 'off-frequency listening' or 'off-place listening', and there is good evidence that humans do indeed listen 'off-frequency' when it is advantageous to do so (Johnson-Davies & Patterson 1979; O'Loughlin & Moore 1981*b*). The result of off-frequency listening is that the PTC has a sharper tip than would be obtained if only one auditory filter was involved (O'Loughlin & Moore 1981*a*).

Another problem with PTCs is that they can be influenced by the detection of beats, which are amplitude fluctuations caused by the interaction of the signal and the masker. The rate of the beats is equal to the difference in frequency of the signal and masker. Beats of a low rate are more easily detected than beats with a rate above approximately 120 Hz (Kohlrausch *et al.* 2000), and slow beats provide a detection cue which results in an increase in the masker level required for threshold for masker frequencies adjacent to the signal frequency. This results in a PTC which has a sharper tip than the underlying auditory filter (Kluk & Moore 2004). This sharpening effect is greatest when a sinusoidal masker is used, but it occurs even when the masker is a narrowband noise (Kluk & Moore 2004).

(c) *The notched-noise method*

Patterson (1976) described a method of determining auditory filter shape which limits off-frequency listening and appears not to be influenced by beat detection. The method is illustrated in figure 2. The signal (indicated by the bold vertical line) is fixed in frequency, and the masker is a noise with a spectral notch centred at the signal frequency. The deviation of each edge of the notch from the centre frequency is denoted by Δf. The width of the notch is varied and the threshold of the signal is determined as a function of notch width. Since the notch is symmetrically placed around the signal frequency, the method cannot reveal asymmetries in the auditory filter, and the analysis assumes that the filter is symmetric on a linear frequency scale. This assumption appears not unreasonable, at least for the top part of the filter and at moderate sound levels, since PTCs are quite symmetric around the tips. For a signal symmetrically placed in a notched noise, the optimum signal-to-masker ratio at the output of the auditory filter is achieved with a filter centred at the signal frequency, as illustrated in figure 2.

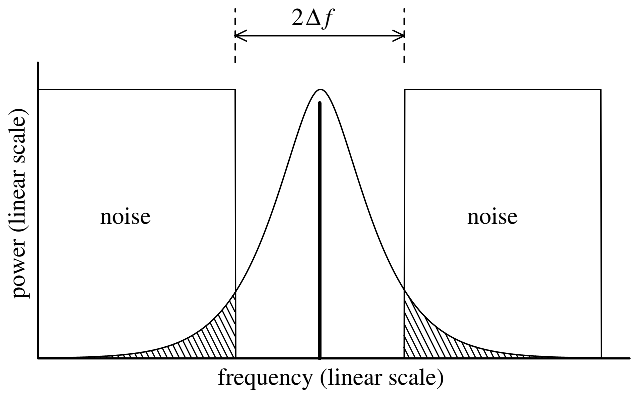

Figure 2. Schematic illustration of the technique used by Patterson (1976) to determine the shape of the auditory filter. The threshold of the sinusoidal signal (indicated by the bold vertical line) is measured as a function of the width of a spectral notch in the noise masker. The amount of noise passing through the auditory filter centred at the signal frequency is proportional to the shaded areas.

As the width of the spectral notch is increased, less and less noise passes through the auditory filter. Thus the threshold of the signal drops. The amount of noise passing through the auditory filter is proportional to the area under the filter in the frequency range covered by the noise. This is shown as the shaded areas in figure 2. Assuming that threshold corresponds to a constant signal-to-masker ratio at the output of the filter, the change in signal threshold with notch width indicates how the area under the filter varies with Δ*f*. The area under a function between certain limits is obtained by integrating the value of the function over those limits. Hence by differentiating the function relating threshold to Δ*f*, the relative response of the filter at that value of Δ*f* is obtained. In other words, the relative response of the filter for a given deviation, Δ*f*, from the centre frequency is equal to the slope of the function relating signal threshold to notch width, at that value of Δ*f*.

A typical auditory filter derived using this method is shown in figure 3. It has a rounded top and quite steep skirts. The sharpness of the filter is often specified as the bandwidth of the filter at which the response has fallen by a factor of two in power, i.e. by 3 dB. The 3 dB bandwidths of the auditory filters derived using the notched-noise method are typically between 10 and 15% of the centre frequency. An alternative measure is the equivalent rectangular bandwidth (ERB), which is the bandwidth of a rectangular filter that has the same peak transmission as the filter of interest and passes the same total power for a white noise input. The ERB of the auditory filter is a little larger than the 3 dB bandwidth. In what follows, the mean ERB of the auditory filter determined using young listeners with normal hearing and using a moderate noise level is denoted $\mathrm{ERB_N}$ (where the subscript N denotes normal hearing). An equation describing the value of $\mathrm{ERB_N}$ as a function of centre frequency, F (in hertz), is (Glasberg & Moore 1990)

$$\mathrm{ERB_N} = 24.7(0.00437F + 1). \tag{2.1}$$

Sometimes it is useful to plot psychoacoustical data on a frequency scale related to $\mathrm{ERB_N}$, called the $\mathrm{ERB_N}$-number scale. For example, the value of $\mathrm{ERB_N}$ for a

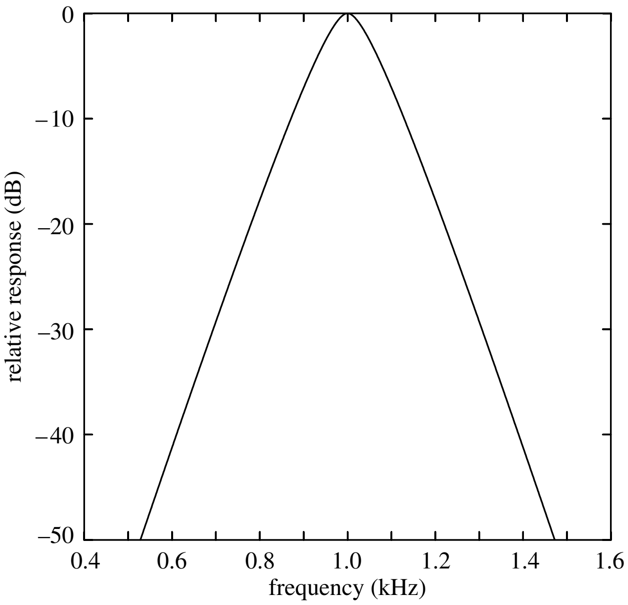

Figure 3. A typical auditory filter shape determined using the notched-noise method. The filter is centred at 1 kHz. The relative response of the filter (in decibels) is plotted as a function of frequency.

centre frequency of 1 kHz is approximately 132 Hz, so an increase in frequency from 934 to 1066 Hz represents a step of one $\mathrm{ERB_N}$-number. The formula relating $\mathrm{ERB_N}$-number to frequency is (Glasberg & Moore 1990)

$$\mathrm{ERB_N}\text{-number} = 21.4 \log(0.00437F + 1), \tag{2.2}$$

where F is the frequency in hertz. Each one-$\mathrm{ERB_N}$ step on the $\mathrm{ERB_N}$-number scale corresponds approximately to a constant distance (0.9 mm) along the basilar membrane (Moore 1986). The $\mathrm{ERB_N}$-number scale is conceptually similar to the Bark scale (Zwicker & Terhardt 1980), which has been widely used by speech researchers, although it differs somewhat in numerical values.

The notched-noise method has been extended to include conditions where the spectral notch in the noise is placed asymmetrically about the signal frequency. This allows the measurement of any asymmetry in the auditory filter, but the analysis of the results is more difficult, and has to take off-frequency listening into account (Patterson & Nimmo-Smith 1980). It is beyond the scope of this paper to give details of the method of analysis; the interested reader is referred to Patterson & Moore (1986), Moore & Glasberg (1987), Glasberg & Moore (1990, 2000) and Rosen *et al.* (1998). The results show that the auditory filter is reasonably symmetric at moderate sound levels, but becomes increasingly asymmetric at high levels, the low-frequency side becoming shallower than the high-frequency side. The filter shapes derived using the notched-noise method are quite similar to inverted PTCs (Glasberg *et al.* 1984), except that PTCs are slightly sharper around their tips, probably as a result of off-frequency listening and beat detection.

(d) *Masking patterns and excitation patterns*

In the masking experiments described so far, the frequency of the signal was held constant, while the

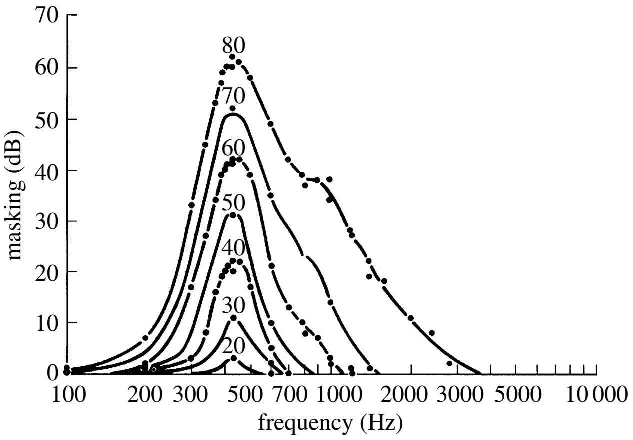

Figure 4. Masking patterns for a narrowband noise masker centred at 410 Hz. Each curve shows the elevation in threshold of a pure-tone signal as a function of signal frequency. The overall noise level in dB SPL for each curve is indicated in the figure. Data from Egan & Hake (1950).

masker was varied. These experiments are most appropriate for estimating the shape of the auditory filter at a given centre frequency. However, in many experiments the masker was held constant in both level and frequency, and the signal threshold was measured as a function of the signal frequency. The resulting functions are called masking patterns or masked audiograms.

Masking patterns show steep slopes on the low-frequency side (when the signal frequency is below that of the masker), of between 55 and 240 dB per octave. The slopes on the high-frequency side are less steep and depend on the level of the masker. Figure 4 shows a typical set of results, obtained using a narrowband noise masker centred at 410 Hz, with the overall masker level varying from 20 to 80 dB SPL in 10 dB steps (data from Egan & Hake 1950). Note that on the high-frequency side the curve is shallower at the highest level. Around the tip of the masking pattern, the growth of masking is approximately linear; a 10 dB increase in masker level leads to roughly a 10 dB increase in the signal threshold. However, for signal frequencies well above the masker frequency, in the range from approximately 1300 to 2000 Hz, when the level of the masker is increased by 10 dB (e.g. from 70 to 80 dB SPL), the masked threshold increases by more than 10 dB; the amount of masking grows nonlinearly on the high-frequency side. This has been called the 'upward spread of masking'.

The masking patterns do not reflect the use of a single auditory filter. Rather, for each signal frequency the listener uses a filter centred close to the signal frequency. Thus, the auditory filter is shifted as the signal frequency is altered. One way of interpreting the masking pattern is as a crude indicator of the excitation pattern of the masker (Zwicker & Fastl 1999). The excitation pattern is a representation of the effective amount of excitation produced by a stimulus as a function of characteristic frequency (CF) on the basilar membrane (see Young 2008) and is plotted as effective level (in decibels) against CF. In the case of a masking sound, the excitation pattern can be thought of as representing the relative amount of vibration produced by the masker at different places along the basilar membrane. The signal is detected when the excitation it produces is some constant proportion of the excitation produced by the masker at places with CFs close to the signal frequency. Thus, the threshold of the signal as a function of frequency is proportional to the masker excitation level. The masking pattern should be parallel to the excitation pattern of the masker, but shifted vertically by a small amount. In practice, the situation is not so straightforward since the shape of the masking pattern is influenced by factors such as off-frequency listening, the detection of beats and combination tones (Moore *et al.* 1998; Alcántara *et al.* 2000) and by the physiological process of suppression (Delgutte 1990; see also Young 2008).

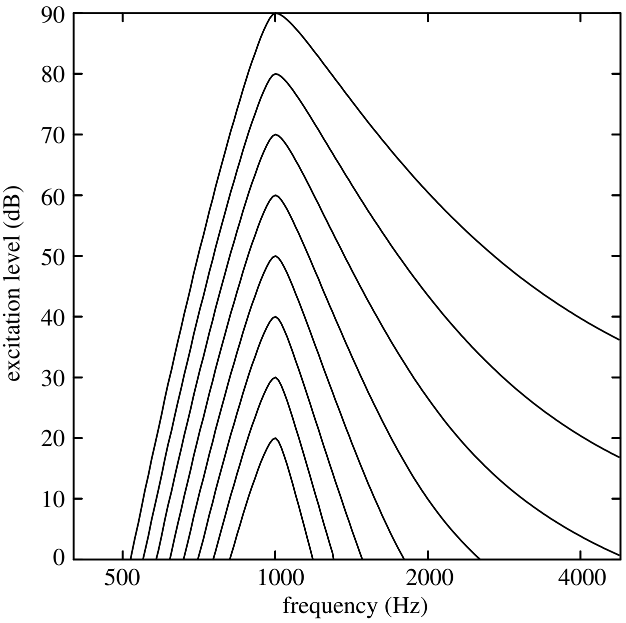

Figure 5. Excitation patterns for a 1000 Hz sinusoid at levels ranging from 20 to 90 dB SPL in 10 dB steps.

Moore & Glasberg (1983*b*) have described a way of deriving the shapes of excitation patterns using the concept of the auditory filter. They suggested that the excitation pattern of a given sound can be thought of as the output of the auditory filters plotted as a function of their centre frequency. To calculate the excitation pattern of a sound, it is necessary to calculate the output of each auditory filter in response to that sound and to plot the output as a function of the filter centre frequency. The characteristics of the auditory filters are determined using the notched-noise method described earlier. Figure 5 shows excitation patterns calculated in this way for 1000 Hz sinusoids with various levels. The patterns are similar in form to the masking patterns shown in figure 4. Software for calculating excitation patterns can be downloaded from http://hearing.psychol.cam.ac.uk/Demos/demos.html.

It should be noted that excitation patterns calculated as described above do not take into account the physiological process of suppression, whereby the response to a given frequency component can be suppressed or reduced by strong neighbouring frequency component (Sachs & Kiang 1968; see also Young 2008). For speech sounds having spectra with strong peaks and valleys, such as vowels, suppression

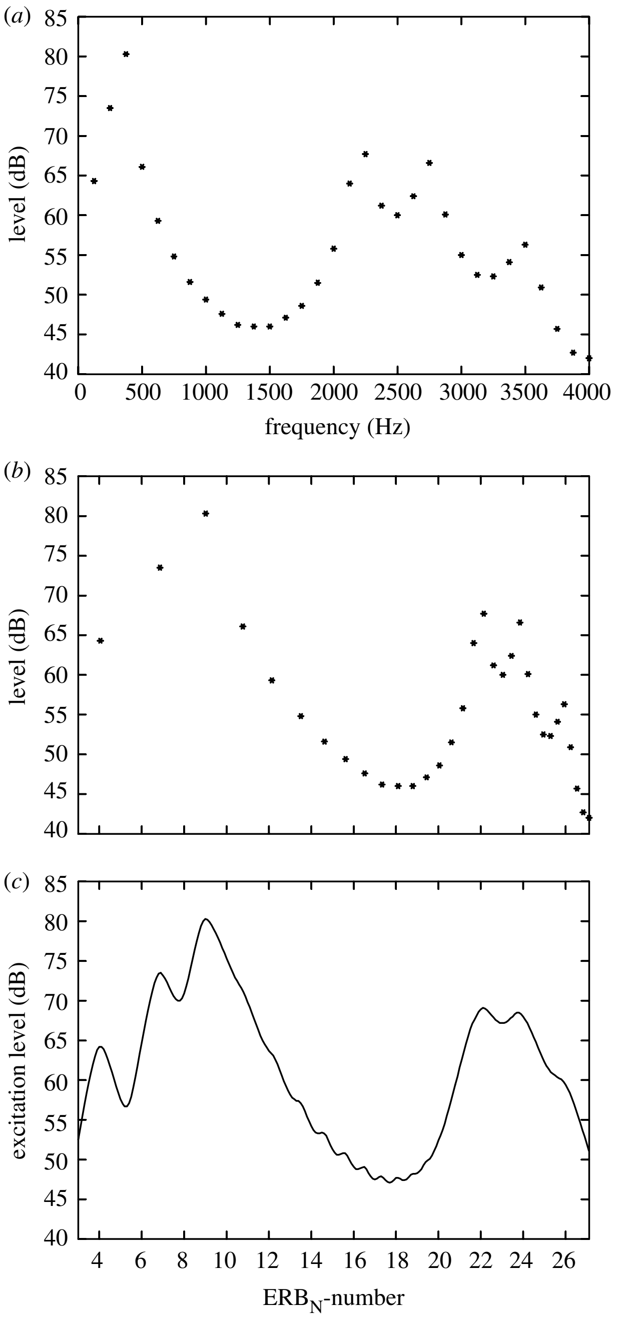

Figure 6. (*a*) The spectrum of a synthetic vowel /I/ plotted on a linear frequency scale. (*b*) The same spectrum plotted on an ERB$_N$-number scale. (*c*) The excitation pattern for the vowel plotted on an ERB$_N$-number scale.

may have the effect of increasing the peak-to-valley ratio of the excitation pattern (Moore & Glasberg 1983*a*). Also, the calculated excitation patterns are based on the power spectrum model of masking and do not take into account the effects of the relative phases of the components in complex sounds. However, it seems probable that excitation patterns provide a reasonable estimate of the extent to which the spectral features of complex sounds are represented in the auditory system.

(e) *Excitation pattern of a vowel sound*

Figure 6*a* shows the spectrum of a synthetic vowel, /I/ as in ‘bit’, plotted on a linear frequency scale; this is the way that vowel spectra are often plotted. Each point represents the level of one harmonic in the complex sound (the fundamental frequency was 125 Hz). Figure 6*b* shows the same spectrum plotted on an ERB$_N$-number scale; this gets somewhat closer to an auditory representation. Figure 6*c* shows the excitation pattern for the vowel, plotted on an ERB$_N$-number scale; this is still closer to an auditory representation. Several aspects of the excitation pattern are noteworthy. Firstly, the lowest few peaks in the excitation pattern do not correspond to formant frequencies, but rather to individual lower harmonics; these harmonics are resolved in the peripheral auditory system and can be heard out as separate tones under certain conditions (Plomp 1964*a*; Moore & Ohgushi 1993). Hence, the centre frequency of the first formant is not directly represented in the excitation pattern; if the frequency of the first formant is relevant for vowel identification (see Diehl 2008), then it must be inferred from the relative levels of the peaks corresponding to the individual lower harmonics.

A second noteworthy aspect of the excitation pattern is that, for this specific vowel, the second, third and fourth formants, which are clearly separately visible in the original spectrum, are not well resolved. Rather, they form a single prominence in the excitation pattern, with only minor ripples corresponding to the individual formants. Assuming that the excitation pattern does give a reasonable indication of the internal representation of the vowel, the perception of this vowel probably depends more on the overall prominence than on the frequencies of the individual formants. For other vowels, the higher formants often lead to separate peaks in the excitation pattern (figure 8).

(f) *Frequency selectivity in cases of impaired hearing*

In the developed countries, the most common cause of hearing loss is damage to the cochlea. This is usually associated with reduced frequency selectivity; the auditory filters are broader than normal (Moore 1998). As a result, the excitation patterns of complex sounds, such as vowels, are ‘blurred’ relative to those for normally hearing listeners. This makes it more difficult to distinguish the timbres of different vowel sounds. It also leads to increased susceptibility to masking by background sounds. For example, when trying to listen to a target talker in the presence of an interfering talker, a hearing-impaired person will be less able than normal to take advantage of differences in the short-term spectra of the two talkers, as described by Darwin (2008).

3. ACROSS-CHANNEL PROCESSES IN MASKING

The discrimination and identification of complex sounds, including speech, require comparison of the outputs of different auditory filters. This section reviews data on across-channel processes in auditory masking and their relevance for speech perception.

(a) *Comodulation masking release*

Hall *et al.* (1984) were among the first to demonstrate that across-filter comparisons could enhance the detection of a sinusoidal signal in a fluctuating noise masker. The crucial feature for achieving this

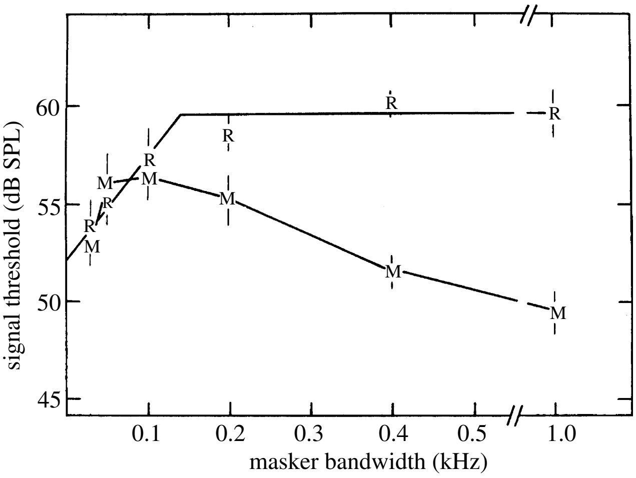

Figure 7. The points labelled 'R' are thresholds for detecting a 1 kHz signal centred in a band of random noise, plotted as a function of the bandwidth of the noise. The points labelled 'M' are the thresholds obtained when the noise was amplitude modulated at an irregular, low rate. Reproduced with permission from Hall *et al.* (1984) and *J. Acous. Soc. Am.*

enhancement was that the fluctuations should be correlated across different frequency bands. One of their experiments was similar to a classic experiment of Fletcher (1940). The threshold for detecting a 1000 Hz, 400 ms sinusoidal signal was measured as a function of the bandwidth of a noise masker, keeping the spectrum level constant. The masker was centred at 1000 Hz. They used two types of masker. One was a random noise; this has irregular fluctuations in amplitude and the fluctuations are independent in different frequency regions. The other was a random noise which was modulated in amplitude at an irregular, low rate; a noise lowpass filtered at 50 Hz was used as a modulator. The modulation resulted in fluctuations in the amplitude of the noise which were the same in different frequency regions. This across-frequency correlation was called 'comodulation' by Hall *et al.* (1984). Figure 7 shows the results of this experiment.

For the random noise (denoted by R), the signal threshold increases as the masker bandwidth increases up to approximately 100–200 Hz, and then remains constant, a result similar to that of Fletcher (1940). The value of ERB$_N$ at this centre frequency is approximately 130 Hz. Hence, for noise bandwidths up to 130 Hz, increasing the bandwidth results in more noise passing through the filter. However, increasing the bandwidth beyond 130 Hz does not substantially increase the noise power passing through the filter, so threshold does not increase. The pattern for the modulated noise (denoted by M) is quite different. For noise bandwidths greater than 100 Hz, the signal threshold decreases as the bandwidth increases. This suggests that subjects can compare the outputs of different auditory filters to enhance signal detection (see, however, Verhey *et al.* 1999). The fact that the decrease in threshold with increasing bandwidth occurs only with the modulated noise indicates that fluctuations in the masker are critical and that these need to be correlated across frequency bands. Hence, this phenomenon has been called 'comodulation masking release' (CMR).

It seems probable that across-filter comparisons of temporal envelopes are a general feature of auditory pattern analysis, which may play an important role in extracting signals from noisy backgrounds or separating competing sources of sound (see Darwin 2008). As pointed out by Hall *et al.* (1984): 'Many real-life auditory stimuli have intensity peaks and valleys as a function of time in which intensity trajectories are highly correlated across frequency. This is true of speech, of interfering noise such as 'cafeteria' noise, and of many other kinds of environmental stimuli'. However, the importance of CMR for speech perception remains controversial. Some studies have suggested that it plays only a very minor role in the detection and identification of speech sounds in modulated background noise (Grose & Hall 1992; Festen 1993), although common modulation of target speech and background speech can lead to reduced intelligibility (Stone & Moore 2004). For synthetic speech in which the cues are impoverished compared to normal speech (sine-wave speech; see Remez *et al.* 1981), comodulation of the speech (amplitude modulation (AM) by a sinusoid) can markedly improve the intelligibility of the speech, both in quiet (Carrell & Opie 1992) and in background noise (Carrell 1993). The AM may help because it leads to perceptual fusion of the components of the sine-wave speech, so as to form an auditory object (see Darwin 2008).

(b) *Profile analysis*

Green and colleagues (Green 1988) have carried out a series of experiments demonstrating that, even for stimuli without distinct envelope fluctuations, subjects are able to compare the outputs of different auditory filters to enhance the detection of a signal. They investigated the ability to detect an increment in the level of one component in a complex sound relative to the level of the other components; the other components are called the 'background'. Usually the complex sound has been composed of a series of equal-amplitude sinusoidal components, uniformly spaced on a logarithmic frequency scale. To prevent subjects from performing the task by monitoring the magnitude of the output of the single auditory filter centred at the frequency of the incremented component, the overall level of the whole stimulus was varied randomly from one stimulus to the next, over a relatively large range (typically 40 dB). This makes the magnitude of the output of any single filter an unreliable cue to the presence of the signal.

Subjects were able to detect changes in the relative level of the signal of only 1–2 dB. Such small thresholds could not be obtained by monitoring the magnitude of the output of a single auditory filter. Green and colleagues have argued that subjects performed the task by detecting a change in the shape or profile of the spectrum of the sound; hence the name 'profile analysis'. In other words, subjects can compare the outputs of different auditory filters, and can detect when the output of one changes relative to that of others, even when the overall level is varied. This is equivalent to detecting changes in the shape of the excitation pattern.

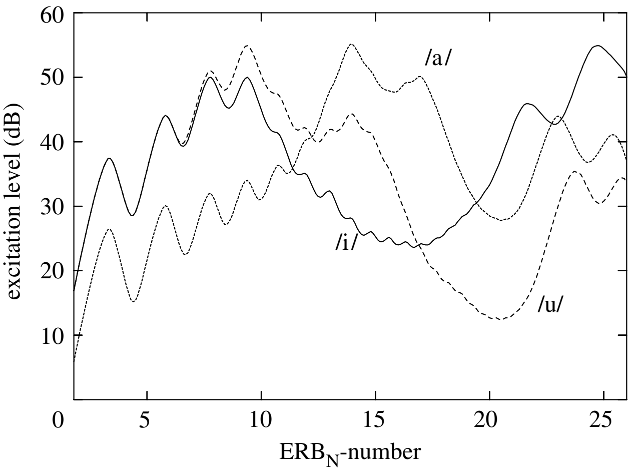

Figure 8. Excitation patterns for three vowels, /i/, /a/ and /u/, plotted on an ERB$_N$-number scale.

Speech researchers will not find the phenomenon of profile analysis surprising. It has been known for many years that one of the main factors determining the timbre or quality of a sound is its spectral shape (see §4). Our everyday experience tells us that we can recognize and distinguish familiar sounds, such as the different vowels, regardless of the levels of those sounds. When we do this, we are distinguishing different spectral shapes in the face of variations in overall level. This is functionally the same as profile analysis. The experiments on profile analysis can be regarded as a way of quantifying the limits of our ability to distinguish changes in spectral shape. In this context, it is noteworthy that the differences in spectral shape between different vowels result in differences in the excitation patterns evoked by those sounds which are generally far larger than the smallest detectable changes as measured in profile analysis experiments. This is illustrated in figure 8, which shows excitation patterns for three vowels, /i/, /a/ and /u/, plotted on an ERB$_N$-number scale. Each vowel had an overall level of approximately 58 dB SPL. It can be seen that the differences in the shapes of the excitation patterns are considerable.

(c) *Modulation discrimination interference*

In some situations, the detection or discrimination of a signal is impaired by the presence of frequency components remote from the signal frequency. Usually, this happens when the task is either to detect modulation of the signal or to detect a change in depth of modulation of the signal. Yost & Sheft (1989) showed that the threshold for detecting sinusoidal AM of a sinusoidal carrier was increased in the presence of another carrier, amplitude modulated at the same rate, even when the second carrier was remote in frequency from the first. They called this modulation detection interference (MDI). They showed that MDI did not occur if the second carrier was unmodulated.

Moore *et al.* (1991) determined how thresholds for detecting an increase in modulation depth (sinusoidal AM or frequency modulation) of a 1000 Hz carrier frequency (the target) were affected by modulation of carriers (interference) with frequencies of 230 and 3300 Hz. They found that modulation increment thresholds were increased (worsened) when the remote carriers were modulated. This MDI effect was the greatest when the target and interference were modulated at similar rates, but the effect was broadly tuned for modulation rate. When both the target and interfering sounds were modulated at 10 Hz, there was no significant effect of the relative phase of modulation of the target and interfering sounds. A lack of effect of relative phase has also been found by other researchers (Moore 1992; Hall *et al.* 1995).

The explanation for MDI remains unclear. Yost & Sheft (1989) suggested that MDI might be a consequence of perceptual grouping; the common AM of the target and interfering sounds might make them fuse perceptually, making it difficult to 'hear out' the modulation of the target sound (see Darwin 2008). However, certain aspects of the results on MDI are difficult to reconcile with an explanation in terms of perceptual grouping (Moore & Shailer 1992). One would expect that widely spaced frequency components would only be grouped perceptually if their modulation pattern was very similar. Grouping would not be expected, for example, if the components were modulated out of phase or at different rates, but, in fact, it is possible to obtain large amounts of MDI under these conditions.

An alternative explanation for MDI is that it reflects the operation of 'channels' specialized for detecting and analysing modulation (Kay & Mathews 1972; Dau *et al.* 1997*a*,*b*). Yost *et al.* (1989) suggested that MDI might arise in the following way. The stimulus is first processed by an array of auditory filters. The envelope at the output of each filter is extracted. When modulation is present, channels that are tuned for modulation rate are excited. All filters responding with the same modulation rate excite the same channel, regardless of the filter centre frequency. Thus, modulation at one centre frequency can adversely affect the detection and discrimination of modulation at other centre frequencies.

The purpose of the hypothetical modulation channels remains unclear. Since physiological evidence suggests that such channels exist in animals (Schreiner & Urbas 1986; Langner & Schreiner 1988), we can assume that they did not evolve for the purpose of speech perception. Nevertheless, it is possible, even probable, that speech analysis makes use of the modulation channels. There is evidence that AM patterns in speech are important for speech recognition (Steeneken & Houtgast 1980; Drullman *et al.* 1994*a*; Shannon *et al.* 1995). Thus, anything that adversely affects the detection and discrimination of the modulation patterns would be expected to impair intelligibility. One way of describing MDI is: modulation in one frequency region may make it more difficult to detect and discriminate modulation in another frequency region. Thus, it may be the case that MDI makes speech recognition more difficult in situations where there is a background sound that is modulated, such as one or more people talking (Brungart *et al.* 2005).

4. TIMBRE PERCEPTION

Timbre is usually defined as 'that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar' (ANSI 1994). The distribution of energy over frequency is one of the major determinants of timbre. However, timbre depends upon more than just the frequency spectrum of the sound; fluctuations over time can play an important role, as discussed below.

Timbre is multidimensional; there is no single scale along which the timbres of different sounds can be compared or ordered. Thus, a way is needed for describing the spectrum of a sound which takes into account this multidimensional aspect and which can be related to the subjective timbre. For steady sounds, a crude first approach is to look at the overall distribution of spectral energy. The 'brightness' or 'sharpness' (von Bismarck 1974) of sounds seems to be related to the spectral centroid. However, a much more quantitative approach has been described by Plomp and colleagues (Plomp 1970, 1976). They showed that the perceptual differences between different steady sounds, such as vowels, were closely related to the differences in the spectra of the sounds, when the spectra were specified as the levels in 18 1/3-octave frequency bands. A bandwidth of 1/3 octave is slightly greater than the $\mathrm{ERB_N}$ of the auditory filter over most of the audible frequency range. Thus, timbre is related to the relative level produced at the output of each auditory filter. In other words, the timbre of a steady sound is related to the excitation pattern of that sound.

It is probable that the number of dimensions required to characterize the timbre of steady sounds is limited by the number of $\mathrm{ERB_N}$s required to cover the audible frequency range. This would give a maximum of approximately 37 dimensions. For a restricted class of sounds, such as vowels, a much smaller number of dimensions may be involved. It appears to be generally true, both for speech and non-speech sounds, that the timbres of steady tones are determined primarily by their magnitude spectra, although the relative phases of the components may also play a small role (Plomp & Steeneken 1969; Patterson 1987).

Differences in spectral shape are not always sufficient to allow the absolute identification of an 'auditory object', such as a musical instrument or a speech sound. One reason for this is that the magnitude and phase spectrum of the sound may be markedly altered by the transmission path and room reflections (Watkins 1991). In practice, the recognition of a particular timbre, and hence of an auditory object, may depend upon several other factors. Schouten (1968) has suggested that these include (i) whether the sound is periodic, having a tonal quality for repetition rates between approximately 20 and 20 000 periods/s, or irregular, and having a noise-like character; (ii) whether the waveform envelope is constant, or fluctuates as a function of time, and in the latter case what the fluctuations are like; (iii) whether any other aspect of the sound (e.g. spectrum or periodicity) is changing as a function of time; and (iv) what the preceding and following sounds are like.

A powerful demonstration of the last factor may be obtained by listening to a stimulus with a particular spectral structure and then switching rapidly to a stimulus with a flat spectrum, such as white noise. A white noise heard in isolation may be described as 'colourless'; it has no pitch and has a neutral timbre. However, when a white noise follows immediately after a stimulus with spectral structure, the noise sounds 'coloured'. The coloration corresponds to the inverse of the spectrum of the preceding sound. For example, if the preceding sound is a noise with a spectral notch, the white noise has a pitch-like quality, with a pitch value corresponding to the centre frequency of the notch (Zwicker 1964). It sounds like a noise with a small spectral peak. A harmonic complex tone with a flat spectrum may be heard as having a vowel-like quality if it is preceded by a harmonic complex having a spectrum which is the inverse of that of a vowel (Summerfield *et al.* 1987).

The cause of this effect is not clear. Three types of explanation have been advanced, based on adaptation in the auditory periphery (see Young 2008), perceptual grouping (see Darwin 2008) and comparison of spectral shapes of the preceding and test sounds (Summerfield & Assmann 1987). All may play a role to some extent, depending on the exact properties of the stimuli. Whatever the underlying mechanism, it appears that the auditory system is especially sensitive to *changes* in spectral patterns over time (Kluender *et al.* 2003). This may be of value for communication in situations where the spectral shapes of sounds are (statically) altered by room reverberation or by a transmission channel with a non-flat frequency response.

Perceptual compensation for the effects of a non-flat frequency response has been studied extensively by Watkins and co-workers (Watkins 1991; Watkins & Makin 1996*a*,*b*). In one series of experiments (Watkins & Makin 1996*b*), they investigated how the identification of vowel test sounds was affected by filtering of preceding and following sounds. All sounds were edited and processed from natural speech spoken with a British accent. Listeners identified words from continua between /Itʃ/ and /ɛtʃ/ (itch and etch), /æpt/ and /ɒpt/ (apt and opt), or /sləʊ/ and /fləʊ/ (slow and flow). The parts of the stimuli other than the vowels (e.g. the /tʃ/ or the /pt/) were filtered with complex frequency responses corresponding to the difference of spectral envelopes from the endpoint test sounds (the vowels). An example of a 'difference filter' is shown in figure 9*c*. The shift in the phoneme boundary of the vowels was used to measure perceptual compensation for the effects of the spectral distortion of the consonants. When the words were presented without a precursor phrase, the results indicated perceptual compensation. Thus, information from the consonants modified the perception of the preceding or following vowels. When a precursor phrase 'the next word is' was used, and was filtered in the same way as the consonant, the effects were larger. However, the large effects were somewhat reduced when the precursor phrase was filtered but the following consonant was not. This clearly indicates a role for sounds following a vowel. The effects of following sounds found in these experiments clearly indicate that factors other than adaptation play a role. Presumably, these effects reflect relatively central perceptual compensation mechanisms.

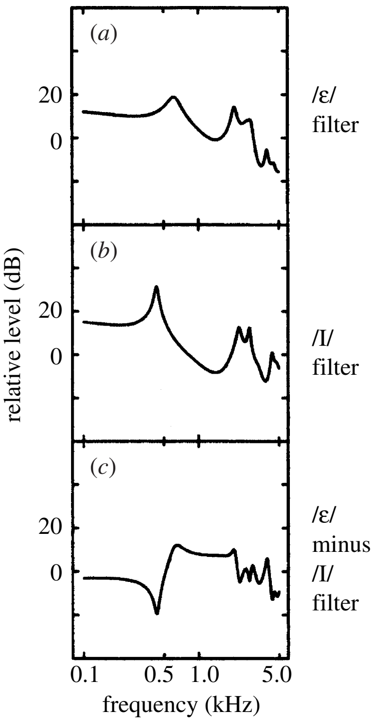

Figure 9. Illustration of the filters used by Watkins & Makin (1996*a*). (*a*,*b*) 'Filters' corresponding to the spectral envelopes of the vowels '/ɛ/' and '/I/', respectively. (*c*) Filter corresponding to the difference between the spectral envelopes of the vowels '/ɛ/' and '/I/'.

Overall, the results described in this section indicate that the perceived timbre of brief segments of sounds can be strongly influenced by sounds that precede and follow those segments. In some cases, the observed effects appear to reflect relatively central perceptual compensation processes.

5. THE PERCEPTION OF PITCH

Pitch is usually defined as 'that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from low to high' (ANSI 1994). In other words, variations in pitch give rise to a sense of melody. For speech sounds, variations in voice pitch over time convey intonation information, indicating whether an utterance is a question or a statement, and helping to identify stressed words. Voice pitch can also convey information about the sex, age and emotional state of the speaker (Rosen & Fourcin 1986). In some languages ('tone' languages), pitch and variations in pitch distinguish different lexical items.

Pitch is related to the repetition rate of the waveform of a sound; for a pure tone this corresponds to the frequency and for a periodic complex tone to the fundamental frequency, F0. There are, however, exceptions to this simple rule. Since voiced speech sounds are complex tones, this section will focus on the perception of pitch for complex tones.

(a) *The phenomenon of the missing fundamental*

Although the pitch of a complex tone usually corresponds to its F0, the component with frequency equal to F0 does not have to be present for the pitch to be heard. Consider, as an example, a sound consisting of short impulses (clicks) occurring 200 times/s. This sound has a low pitch, which is very close to the pitch of a 200 Hz sinusoid, and a sharp timbre. It contains harmonics with frequencies 200, 400, 600, 800, etc. Hz. However, if the sound is filtered so as to remove the 200 Hz component, the pitch does not alter; the only result is a slight change in the timbre of the note. Indeed, all except a small group of mid-frequency harmonics can be eliminated, and the low pitch still remains, although the timbre becomes markedly different.

Schouten (1970) called the low pitch associated with a group of high harmonics the 'residue'. He pointed out that the residue is distinguishable, subjectively, from a fundamental component which is physically presented or from a fundamental which may be generated (at high sound pressure levels) by nonlinear distortion in the ear. Thus, the perception of a residue pitch does not require activity at the point on the basilar membrane which would respond maximally to a pure tone of similar pitch. Several other names have been used to describe residue pitch, including 'periodicity pitch', 'virtual pitch' and 'low pitch'. This paper will use the term low pitch. Even when the fundamental component of a complex tone is present, the low pitch of the tone is usually determined by harmonics other than the fundamental. Thus, the perception of a low pitch should not be regarded as unusual. Rather, low pitches are normally heard when listening to complex tones, including speech. For example, when listening over the telephone, the fundamental component for male speakers is usually inaudible, but the pitch of the voice can still be easily heard.

(b) *The principle of dominance*

Ritsma (1967) carried out an experiment to determine which components in a complex sound are most important in determining its pitch. He presented complex tones in which the frequencies of a small group of harmonics were multiples of an F0 which was slightly higher or lower than the F0 of the remainder. The subject's pitch judgements were used to determine whether the pitch of the complex as a whole was affected by the shift in the group of harmonics. Ritsma found that: 'For fundamental frequencies in the range 100–400 Hz, and for SLs up to at least 50 dB above threshold of the entire signal, the frequency band consisting of the third, fourth and fifth harmonics tends to dominate the pitch sensation as long as its amplitude exceeds a minimum absolute level of about 10 dB above threshold'.

This finding has been broadly confirmed in other ways (Plomp 1967), although the data of Moore *et al.* (1984, 1985) show that there are large individual differences in which harmonics are dominant, and for some subjects the first two harmonics play an important role. Other data also show that the dominant region is not fixed in terms of harmonic number, but depends somewhat on absolute frequency (Plomp 1967; Patterson & Wightman 1976). For high F0s (above approx. 1000 Hz), the fundamental is usually the dominant component, while for very low F0s, approximately 50 Hz, harmonics above the fifth may be dominant (Moore & Glasberg 1988; Moore & Peters 1992). Finally, the dominant region shifts somewhat towards higher harmonics with decreasing duration (Gockel *et al.* 2005). For speech sounds, the dominant harmonics usually lie around the frequency of the first formant.

(c) *Discrimination of the pitch of complex tones*

When the F0 of a periodic complex tone changes, all of the components change in frequency by the same ratio, and a change in low pitch is heard. The ability to detect such changes is better than the ability to detect changes in a sinusoid at F0 (Flanagan & Saslow 1958) and can be better than the ability to detect changes in the frequency of any of the sinusoidal components in the complex tone (Moore *et al.* 1984). This indicates that information from the different harmonics is combined or integrated in the determination of low pitch. This can lead to very fine discrimination; changes in F0 of approximately 0.2% can often be detected for F0s in the range 100–400 Hz.

The discrimination of F0 is usually best when low harmonics are present (Hoekstra & Ritsma 1977; Moore & Glasberg 1988; Shackleton & Carlyon 1994). Somewhat less good discrimination (typically 1–4%) is possible when only high harmonics are present (Houtsma & Smurzynski 1990). F0 discrimination can be impaired (typically by about a factor of two) when the two sounds to be discriminated also differ in timbre (Moore & Glasberg 1990); this can be the situation with speech sounds, where changes in F0 are usually accompanied by changes in timbre.

In speech, intonation is typically conveyed by differences in the pattern of F0 change over time. When the stimuli are dynamically varying, the ability to detect F0 changes is markedly poorer than when the stimuli are steady. Klatt (1973) measured thresholds for detecting differences in F0 for an unchanging vowel (i.e. one with static formant frequencies) with a flat F0 contour, and also for a series of linear glides in F0 around an F0 of 120 Hz. For the flat contour, the threshold was approximately 0.3 Hz. When both the contours were falling at the same rate (30 Hz over the 250 ms duration of the stimulus), the threshold increased markedly to 2 Hz. When the steady vowel was replaced by the sound /ya/, whose formants change over time, thresholds increased further by 25–65%.

Generally, the F0 changes that are linguistically relevant for conveying stress and intonation are much larger than the limits of F0 discrimination measured psychophysically using steady stimuli. This is another reflection of the fact that information in speech is conveyed using robust cues that do not severely tax the discrimination abilities of the auditory system. Again, however, this may not be true for people with impaired hearing, for whom F0 discrimination is often much worse than normal (Moore & Carlyon 2005).

It should be noted that in natural speech the period (corresponding to the time between successive closures of the vocal folds) varies randomly from one period to the next (Fourcin & Abberton 1977). This jitter conveys information about the emotional state of the talker and is required for a natural voice quality to be perceived. Human listeners can detect jitter of 1–2% (Pollack 1968; Kortekaas & Kohlrausch 1999). Large amounts of jitter are associated with voice pathologies, such as hoarseness (Yumoto *et al.* 1982).

(d) *Perception of pitch in speech*

Data on the perception of F0 contours in a relatively natural speech context were presented by Pierrehumbert (1979). She started with a natural nonsense utterance 'ma-MA-ma-ma-MA-ma', in which the prosodic pattern was based on the sentence 'The baker made bagels'. The stressed syllables (MA) were associated with peaks in the F0 contour. She then modified the F0 of the second peak, over a range varying from below to above the F0 of the first peak. Subjects listened to the modified utterances and were required to indicate whether the first or second peak was higher in pitch. The results reflected what she called 'normalization for expected declination'; when the two stressed syllables sounded equal in pitch, the second was actually lower in F0. For first peak values of 121 and 151 Hz, the second peak had to be shifted over a range of approximately 20 Hz to change judgements from 75% 'second peak lower' to 75% 'second peak higher'. This indicates markedly poorer discriminability than found for steady stimuli. Similarly, 't Hart (1981) found that about a 19% difference was necessary for successive pitch movements in the same direction to be reliably heard as different in extent.

Hermes & van Gestel (1991) studied the perception of the excursion size of prominence-lending F0 movements in utterances resynthesized in different F0 registers. The task of the subjects was to adjust the excursion size in a comparison stimulus in such a way that it lent equal prominence to the corresponding syllable in a fixed test stimulus. The comparison stimulus and the test stimulus had F0s running parallel on either a logarithmic frequency scale, an $\mathrm{ERB_N}$-number scale, or a linear frequency scale. They found that stimuli were matched in such a way that the average excursion sizes in different registers were equal when the $\mathrm{ERB_N}$-number scale was used. In other words, the perceived prominence of F0 movements is related to the size of those movements expressed on an $\mathrm{ERB_N}$-number scale.

6. TEMPORAL ANALYSIS

Time is a very important dimension in hearing, since almost all sounds change over time. For speech, much of the information appears to be carried in the changes themselves, rather than in the parts of the sounds which are relatively stable (Kluender *et al.* 2003). In characterizing temporal analysis, it is essential to take account of the filtering that takes place in the peripheral auditory system. Temporal analysis can be considered as resulting from two main processes: analysis of the time pattern occurring within each frequency channel and comparison of the time patterns across channels. This paper focuses on the first of these.

A major difficulty in measuring the temporal resolution of the auditory system is that changes in the time pattern of a sound are generally associated with changes in its magnitude spectrum—the distribution of energy over frequency. Thus, the detection of a change in time pattern can sometimes depend not on

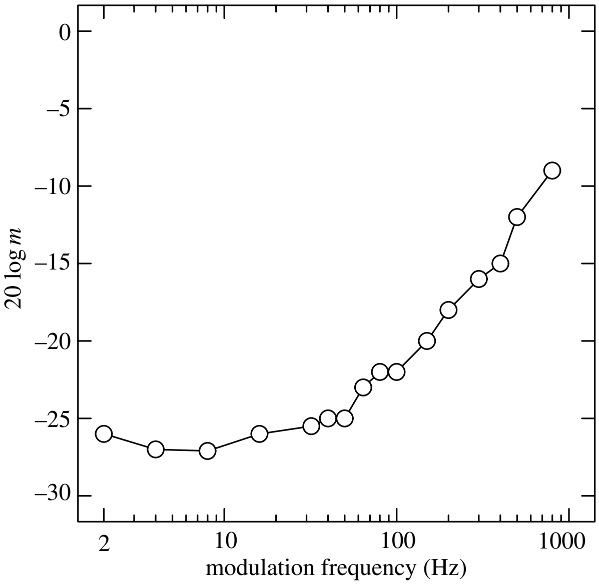

Figure 10. A temporal modulation transfer function (TMTF). A broadband white noise was sinusoidally amplitude modulated, and the threshold amount of modulation required for detection is plotted as a function of modulation rate. The amount of modulation is specified as 20 log *m*, where *m* is the modulation index. The higher the sensitivity to modulation, the more negative is this quantity. Data from Bacon & Viemeister (1985).

temporal resolution *per se*, but on the detection of the spectral change. Sometimes, the detection of spectral changes can lead to what appears to be extraordinarily fine temporal resolution. For example, a single click can be distinguished from a pair of clicks when the gap between the two clicks in a pair is only a few tens of microseconds, an ability that depends upon spectral changes at very high frequencies (Leshowitz 1971). Although spectrally based detection of temporal changes can occur for speech sounds, this paper focuses on experimental situations which avoid the confounding effects of spectral cues.

There have been two general approaches to avoiding the use of cues based on spectral changes. One is to use signals whose magnitude spectrum is not changed when the time pattern is altered. For example, the magnitude spectrum of white noise remains flat if a gap is introduced into the noise. The second approach uses stimuli whose spectra are altered by the change in time pattern, but extra background sounds are added to mask the spectral changes. Both of these approaches will be considered.

(a) *Within-channel temporal analysis using broadband sounds*

All the experiments described below use broadband sounds whose long-term magnitude spectrum is unaltered by the temporal manipulation being performed. For example, interruption or AM of a white noise does not change its long-term magnitude spectrum, and time reversal of any sound also does not change its long-term magnitude spectrum.

The threshold for detecting a gap in a broadband noise provides a simple and convenient measure of temporal resolution. The gap threshold is typically 2–3 ms (Plomp 1964*b*). The threshold increases at very low sound levels, when the level of the noise approaches the absolute threshold, but is relatively invariant with level for moderate to high levels.

Ronken (1970) used pairs of clicks differing in amplitude as stimuli. One click, labelled A, had an amplitude greater than that of the other click, labelled B. Typically, the amplitude of A was twice that of B. Subjects were required to distinguish click pairs differing in the order of A and B: either AB or BA. The ability to do this was measured as a function of the time interval or gap between A and B. Ronken found that subjects could distinguish the click pairs for gaps down to 2–3 ms. Thus, the limit to temporal resolution found in this task is similar to that found for the detection of a gap in broadband noise. It should be noted that, in this task, subjects do not hear the individual clicks within a click pair. Rather, each click pair is heard as a single sound with its own characteristic quality. For example, the two click pairs AB and BA might sound like ‘tick’ and ‘tock’.

The experiments described above each give a single value to describe temporal resolution. A more general approach is to measure the threshold for detecting changes in the amplitude of a sound as a function of the rapidity of the changes. In the simplest case, white noise is sinusoidally amplitude modulated, and the threshold for detecting the modulation is determined as a function of modulation rate. The function relating threshold to modulation rate is known as a temporal modulation transfer function (TMTF; Viemeister 1979). An example of a TMTF is shown in figure 10 (Bacon & Viemeister 1985). The thresholds are expressed as 20 log *m*, where *m* is the modulation index ($m=0$ corresponds to no modulation and $m=1$ corresponds to 100% modulation). For low modulation rates, performance is limited by the amplitude resolution of the ear, rather than by temporal resolution. Thus, the threshold is independent of modulation rate for rates up to approximately 50 Hz. As the rate increases beyond 50 Hz, temporal resolution starts to have an effect; performance worsens, and for rates above approximately 1000 Hz the modulation is hard to detect at all. Thus, sensitivity to modulation becomes progressively less as the rate of modulation increases. The shapes of TMTFs do not vary much with overall sound level, but the ability to detect the modulation does worsen at low sound levels. Over the range of modulation rates important for speech perception, below approximately 50 Hz (Steeneken & Houtgast 1980; Drullman *et al.* 1994*a*,*b*), the sensitivity to modulation is rather good.

(b) *Within-channel temporal analysis using narrowband sounds*

Experiments using broadband sounds provide no information regarding the question of whether the temporal resolution of the auditory system varies with centre frequency. This issue can be examined by using narrowband stimuli that excite only one, or a small number, of auditory channels.

Green (1973) used stimuli where each stimulus consisted of a brief pulse of a sinusoid in which the level of the first half of the pulse was 10 dB different from that of the second half. Subjects were required to distinguish two signals, differing in whether the half with the high level was first or second. Green measured performance as a function of the total duration of the stimuli. The threshold was similar for centre frequencies of 2 and 4 kHz, and was between 1 and 2 ms. However, the threshold was slightly higher for a centre frequency of 1 kHz, being between 2 and 4 ms.

Performance in this task was actually a non-monotonic function of duration. Performance was good for durations in the range 2–6 ms, worsened for durations around 16 ms, and then improved again as the duration was increased beyond 16 ms. For the very short durations, subjects listened for a difference in quality between the two sounds—rather like the 'tick' and 'tock' described earlier for Ronken's stimuli. At durations around 16 ms, the tonal quality of the bursts became more prominent and the quality differences were harder to hear. At much longer durations, the soft and loud segments could be separately heard, in a distinct order. It appears, therefore, that performance in this task was determined by two separate mechanisms, one based on timbre differences associated with the difference in time pattern, and the other based on the perception of a distinct succession of auditory events.

Several researchers have measured thresholds for detecting gaps in narrowband sounds, either noises (Fitzgibbons 1983; Shailer & Moore 1983; Buus & Florentine 1985; Eddins *et al.* 1992) or sinusoids (Shailer & Moore 1987; Moore *et al.* 1993). When a temporal gap is introduced into a narrowband sound, the spectrum of the sound is altered. Energy 'splatter' occurs outside the nominal frequency range of the sound. To prevent the splatter being detected, the sounds are presented in a background sound, usually a noise, designed to mask the splatter.

Gap thresholds for noise bands decrease with increasing bandwidth but show little effect of centre frequency when the bandwidth is held constant. For noises of moderate bandwidth (a few hundred hertz), the gap threshold is typically approximately 10 ms. Gap thresholds for narrowband noises tend to decrease with increasing sound level for levels up to approximately 30 dB above absolute threshold, but remain roughly constant after that.

Shailer & Moore (1987) showed that the detectability of a gap in a sine wave was strongly affected by the phase at which the sinusoid was turned off and on to produce the gap (Shailer & Moore 1987). Only the simplest case is considered here, called 'preserved phase' by Shailer & Moore (1987). In this case, the sinusoid was turned off at a positive-going zero crossing (i.e. as the waveform was about to change from negative to positive values) and it started (at the end of the gap) at the phase it would have had if it had continued without interruption. Thus, for the preserved-phase condition it was as if the gap had been 'cut out' from a continuous sinusoid. For this condition, the detectability of the gap increased monotonically with increasing gap duration.

Shailer & Moore (1987) found that the threshold for detecting a gap in a sine wave was roughly constant at approximately 5 ms for centre frequencies of 400, 1000 and 2000 Hz. Moore *et al.* (1993) found that gap thresholds were almost constant at 6–8 ms over the frequency range 400–2000 Hz, but increased somewhat at 200 Hz, and increased markedly, to approximately 18 ms, at 100 Hz. Individual variability also increased markedly at 100 Hz.

Overall, the results of experiments using narrowband stimuli indicate that temporal resolution does not vary markedly with centre frequency, except perhaps for a worsening at very low frequencies (200 Hz and below). Gap thresholds for narrowband stimuli are typically higher than those for broadband noise. However, for moderate noise bandwidths, gap thresholds are typically approximately 10 ms or less. The smallest detectable gap is usually markedly larger than temporal gaps that are relevant for speech perception (for example, 'sa' and 'sta' may be distinguished by a temporal gap lasting several tens of milliseconds).

(c) *Modelling temporal resolution*

Most models of temporal resolution are based on the idea that there is a process at levels of the auditory system higher than the auditory nerve which is 'sluggish' in some way, thereby limiting temporal resolution. The models assume that the internal representation of stimuli is 'smoothed' over time, so that rapid temporal changes are reduced in magnitude but slower ones are preserved. Although this smoothing process almost certainly operates on neural activity, the most widely used models are based on smoothing a simple transformation of the stimulus, rather than its neural representation.

Most models include an initial stage of bandpass filtering, reflecting the action of the auditory filters. Each filter is followed by a nonlinear device. This nonlinear device is meant to reflect the operation of several processes that occur in the peripheral auditory system such as amplitude compression on the basilar membrane and neural transduction, whose effects resemble half-wave rectification (see Young 2008). The output of the nonlinear device is fed to a 'smoothing' device, which can be implemented either as a lowpass filter (Viemeister 1979) or (equivalently) as a sliding temporal integrator (Moore *et al.* 1988; Plack & Moore 1990). The device determines a kind of weighted average of the output of the compressive nonlinearity over a certain time interval or 'window'. This weighting function is sometimes called the 'shape' of the temporal window. The window is assumed to slide in time, so that the output of the temporal integrator is a weighted running average of the input. This has the effect of smoothing rapid fluctuations while preserving slower ones. When a sound is turned on abruptly, the output of the temporal integrator takes some time to build up. Similarly, when a sound is turned off, the output of the integrator takes some time to decay. The shape of the window is assumed to be asymmetric in time, such that the build up of its output in response to the onset of a sound is more rapid than the decay of its output in response to the cessation of a sound. The output of the sliding temporal integrator is fed to a decision device. The decision device may use different 'rules' depending on the task required. For

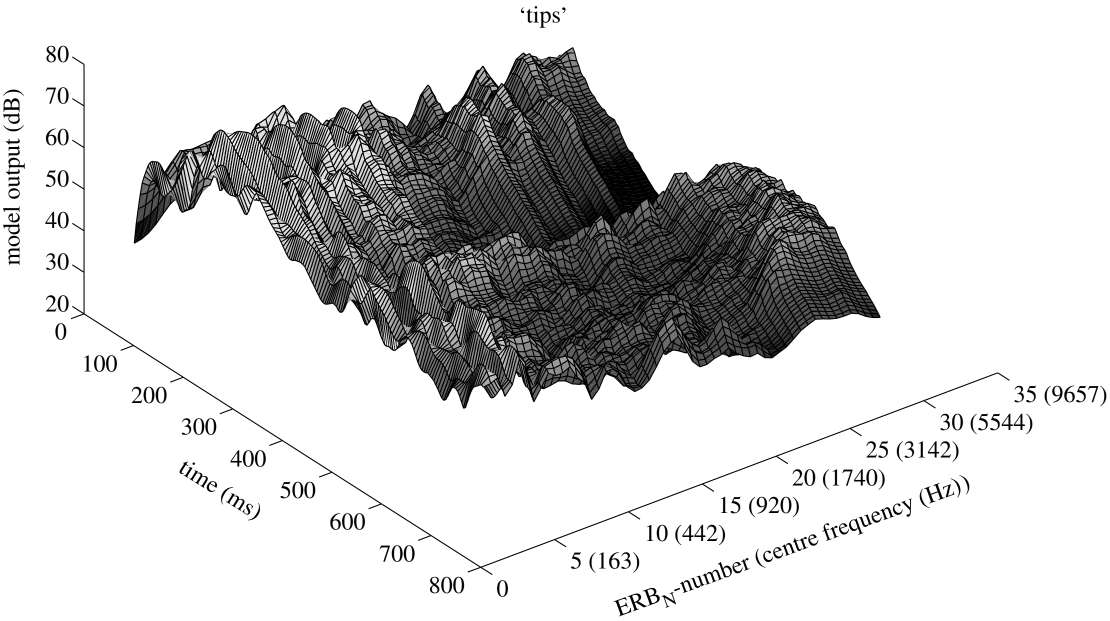

Figure 11. Spectro-temporal excitation pattern (STEP) of the word 'tips'. The figure was produced by Prof. C. J. Plack. Adapted from Moore (2003*c*).

example, if the task is to detect a brief temporal gap in a signal, the decision device might look for a 'dip' in the output of the temporal integrator. If the task is to detect AM of a sound, the device might assess the amount of modulation at the output of the sliding temporal integrator (Viemeister 1979).

7. CALCULATION OF THE INTERNAL REPRESENTATION OF SOUNDS

I describe in the following a method of calculating the internal representation of sounds, including speech, based on processes that are known to occur in the auditory system and taking into account the frequency and temporal resolution of the auditory system. The spectrogram is often regarded as a crude representation of the spectro-temporal analysis that takes place in the auditory system, although this representation is inaccurate in several ways (Moore 2003*a*). I outline below a model that probably gives a better representation, although it is still oversimplified in several respects. The model is based on the assumption that there are certain 'fixed' processes in the peripheral auditory system, which can be modelled as a series of stages including:

(i) fixed filters representing transfer of sound through the outer (Shaw 1974) and middle ear (Aibara *et al.* 2001). The overall transfer function through the outer and middle ear for a frontally incident sound in free field has been estimated by Glasberg & Moore (2002) as shown in their fig. 1. The effect of this transfer function is that low (below 500 Hz) and high frequencies (above 5000 Hz) are attenuated relative to middle frequencies,

(ii) an array of bandpass filters (the auditory filters),

(iii) each auditory filter is followed by nonlinear processes reflecting the compression that occurs on the basilar membrane (Oxenham & Moore 1994; Ruggero *et al.* 1997; see also Young 2008). The compression is weak for very low sound levels (below about 30 dB SPL), and perhaps for very high levels (above 90 dB SPL), but it has a strong influence for mid-range sound levels. Half- or full-wave rectification may also be introduced to mimic the transformation from basilar-membrane vibration to neural activity effected by the inner hair cells (Viemeister 1979), and

(iv) an array of devices (sliding temporal integrators) that 'smooth' the output of each nonlinearity. As described earlier, the smoothing is assumed to reflect a relatively central process, occurring after the auditory nerve.

In some models, the filtering and the compressive nonlinearity are combined in a single nonlinear filter bank (Irino & Patterson 2001; Lopez-Poveda & Meddis 2001; Zhang *et al.* 2001). Also, the transformation from basilar-membrane vibration to neural activity can be simulated more accurately using a hair-cell model (Sumner *et al.* 2002). However, the basic features of the internal representation can be represented reasonably well using models of the type defined by stages (i)–(iv). The internal representation of a given stimulus can be thought of as a three-dimensional array with centre frequency as one axis (corresponding to the array of auditory filters with different centre frequencies), and time and magnitude as the other axes (corresponding to the output of each temporal integrator plotted as a function of time). The resulting pattern can be called a spectro-temporal excitation pattern (STEP; Moore 1996). An example is shown in figure 11, adapted from Moore (2003*c*). The figure shows the calculated STEP of the word 'tips'. In this figure, the frequency scale has been transformed to an ERB$_N$-number scale, as described earlier. The corresponding frequency is also shown. It should be noted that the STEP does not represent information that is potentially available in the temporal 'fine structure' at the output of each auditory filter.

The role of this fine structure in speech perception is uncertain, and some have suggested that it plays little role (Shannon *et al.* 1995). However, temporal fine structure may play a role in the perception of pitch (Moore *et al.* 2006), in the separation of simultaneous sounds (Moore 2003*a*) and in distinguishing voiced from voiceless sounds.

8. CONCLUDING REMARKS

This paper has reviewed several aspects of auditory perception that are relevant to the perception of speech. These aspects include frequency selectivity, timbre perception, the perception of pitch and temporal analysis. A recurring theme has been the finding that the basic discrimination abilities of the auditory system, measured using simple non-speech stimuli, are very good when considered relative to the acoustic differences that distinguish speech sounds. This partially accounts for the robust nature of speech perception. Indeed, it is remarkable that speech remains reasonably intelligible even under conditions of extreme distortion, such as infinite peak clipping (Licklider & Pollack 1948), time reversal of segments of speech (Saberi & Perrott 1999), representation of speech by three or four sine waves tracking the formant frequencies (Remez *et al.* 1981), or representation of speech by a few amplitude-modulated noise bands (Shannon *et al.* 1995). However, this robustness applies to speech presented in quiet. Speech is often heard under much less ideal conditions. For example, reverberation, background noise and competing talkers may be present. Under these conditions, many of the cues in speech become less discriminable, and some cues may be completely inaudible. Speech perception then becomes much less robust, especially if the functioning of the auditory system is impaired (Moore 2003*b*). The perception of speech when competing sounds are present is reviewed in Darwin (2008).

I thank Tom Baer and Chris Darwin for their helpful comments on an earlier version of this paper.

REFERENCES

Aibara, R., Welsh, J. T., Puria, S. & Goode, R. L. 2001 Human middle-ear sound transfer function and cochlear input impedance. *Hear. Res.* **152**, 100–109. (doi:10.1016/S0378-5955(00)00240-9)

Alcántara, J. I., Moore, B. C. J. & Vickers, D. A. 2000 The relative role of beats and combination tones in determining the shapes of masking patterns at 2 kHz: I. Normal-hearing listeners. *Hear. Res.* **148**, 63–73. (doi:10.1016/S0378-5955(00)00114-3)

ANSI 1994 *ANSI S1.1-1994. American national standard acoustical terminology*. New York, NY: American National Standards Institute.

Bacon, S. P. & Viemeister, N. F. 1985 Temporal modulation transfer functions in normal-hearing and hearing-impaired subjects. *Audiology* **24**, 117–134.

Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L. & Kidd Jr, G. 2005 Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task. *J. Acoust. Soc. Am.* **117**, 292–304. (doi:10.1121/1.1835509)

Buus, S. & Florentine, M. 1985 Gap detection in normal and impaired listeners: the effect of level and frequency. In *Time resolution in auditory systems* (ed. A. Michelsen), pp. 159–179. New York, NY: Springer.

Carrell, T. 1993 The effect of amplitude comodulation on extracting sentences from noise: evidence from a variety of contexts. *J. Acoust. Soc. Am.* **93**, 2327. (doi:10.1121/1.406347)

Carrell, T. D. & Opie, J. M. 1992 The effect of amplitude comodulation on auditory object formation in sentence perception. *Percept. Psychophys.* **52**, 437–445.

Darwin, C. J. 2008 Listening to speech in the presence of other sounds. *Phil. Trans. R. Soc. B* **363**, 1011–1021. (doi:10.1098/rstb.2007.2156)

Dau, T., Kollmeier, B. & Kohlrausch, A. 1997*a* Modeling auditory processing of amplitude modulation I. Detection and masking with narrowband carriers. *J. Acoust. Soc. Am.* **102**, 2892–2905. (doi:10.1121/1.420344)

Dau, T., Kollmeier, B. & Kohlrausch, A. 1997*b* Modeling auditory processing of amplitude modulation II. Spectral and temporal integration. *J. Acoust. Soc. Am.* **102**, 2906–2919. (doi:10.1121/1.420345)

Delgutte, B. 1990 Physiological mechanisms of psychophysical masking: observations from auditory-nerve fibers. *J. Acoust. Soc. Am.* **87**, 791–809. (doi:10.1121/1.398891)

Diehl, R. L. 2008 Acoustic and auditory phonetics: the adaptive design of speech sound systems. *Phil. Trans. R. Soc. B* **363**, 965–978. (doi:10.1098/rstb.2007.2153)

Drullman, R., Festen, J. M. & Plomp, R. 1994*a* Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* **95**, 2670–2680. (doi:10.1121/1.409836)

Drullman, R., Festen, J. M. & Plomp, R. 1994*b* Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* **95**, 1053–1064. (doi:10.1121/1.408467)

Eddins, D. A., Hall, J. W. & Grose, J. H. 1992 Detection of temporal gaps as a function of frequency region and absolute noise bandwidth. *J. Acoust. Soc. Am.* **91**, 1069–1077. (doi:10.1121/1.402633)

Egan, J. P. & Hake, H. W. 1950 On the masking pattern of a simple auditory stimulus. *J. Acoust. Soc. Am.* **22**, 622–630. (doi:10.1121/1.1906661)

Festen, J. M. 1993 Contributions of comodulation masking release and temporal resolution to the speech–reception threshold masked by an interfering voice. *J. Acoust. Soc. Am.* **94**, 1295–1300. (doi:10.1121/1.408156)

Fitzgibbons, P. J. 1983 Temporal gap detection in noise as a function of frequency, bandwidth and level. *J. Acoust. Soc. Am.* **74**, 67–72. (doi:10.1121/1.389619)

Flanagan, J. L. & Saslow, M. G. 1958 Pitch discrimination for synthetic vowels. *J. Acoust. Soc. Am.* **30**, 435–442. (doi:10.1121/1.1909640)

Fletcher, H. 1940 Auditory patterns. *Rev. Mod. Phys.* **12**, 47–65. (doi:10.1103/RevModPhys.12.47)

Fourcin, A. J. & Abberton, E. 1977 Laryngograph studies of vocal-fold vibration. *Phonetica* **34**, 313–315.

Glasberg, B. R. & Moore, B. C. J. 1990 Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* **47**, 103–138. (doi:10.1016/0378-5955(90)90170-T)

Glasberg, B. R. & Moore, B. C. J. 2000 Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *J. Acoust. Soc. Am.* **108**, 2318–2328. (doi:10.1121/1.1315291)

Glasberg, B. R. & Moore, B. C. J. 2002 A model of loudness applicable to time-varying sounds. *J. Audio Eng. Soc.* **50**, 331–342.

Glasberg, B. R., Moore, B. C. J., Patterson, R. D. & Nimmo-Smith, I. 1984 Dynamic range and asymmetry of the auditory filter. *J. Acoust. Soc. Am.* **76**, 419–427. (doi:10.1121/1.391584)

Gockel, H., Carlyon, R. P. & Plack, C. J. 2005 Dominance region for pitch: effects of duration and dichotic presentation. *J. Acoust. Soc. Am.* **117**, 1326–1336. (doi:10.1121/1.1853111)

Green, D. M. 1973 Temporal acuity as a function of frequency. *J. Acoust. Soc. Am.* **54**, 373–379. (doi:10.1121/1.1913587)

Green, D. M. 1988 *Profile analysis*. Oxford, UK: Oxford University Press.

Grose, J. H. & Hall, J. W. 1992 Comodulation masking release for speech stimuli. *J. Acoust. Soc. Am.* **91**, 1042–1050. (doi:10.1121/1.402630)

Hall, J. W., Haggard, M. P. & Fernandes, M. A. 1984 Detection in noise by spectro-temporal pattern analysis. *J. Acoust. Soc. Am.* **76**, 50–56. (doi:10.1121/1.391005)

Hall, J. W., Grose, J. H. & Mendoza, L. 1995 Across-channel processes in masking. In *Hearing* (ed. B. C. J. Moore), pp. 243–266. San Diego, CA: Academic Press.

Helmholtz, H. L. F. 1863 *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Braunschweig, Germany: F. Vieweg.

Hermes, D. J. & van Gestel, J. C. 1991 The frequency scale of speech intonation. *J. Acoust. Soc. Am.* **90**, 97–102. (doi:10.1121/1.402397)

Hoekstra, A. & Ritsma, R. J. 1977 Perceptive hearing loss and frequency selectivity. In *Psychophysics and physiology of hearing* (eds E. F. Evans & J. P. Wilson), pp. 263–271. London, UK: Academic.

Houtsma, A. J. M. & Smurzynski, J. 1990 Pitch identification and discrimination for complex tones with many harmonics. *J. Acoust. Soc. Am.* **87**, 304–310. (doi:10.1121/1.399297)

Irino, T. & Patterson, R. D. 2001 A compressive gammachirp auditory filter for both physiological and psychophysical data. *J. Acoust. Soc. Am.* **109**, 2008–2022. (doi:10.1121/1.1367253)

Johnson-Davies, D. & Patterson, R. D. 1979 Psychophysical tuning curves: restricting the listening band to the signal region. *J. Acoust. Soc. Am.* **65**, 765–770. (doi:10.1121/1.382490)

Kay, R. H. & Mathews, D. R. 1972 On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. *J. Physiol.* **225**, 657–677.

Klatt, D. H. 1973 Discrimination of fundamental frequency contours in speech: implications for models of pitch perception. *J. Acoust. Soc. Am.* **53**, 8–16. (doi:10.1121/1.1913333)

Kluender, K. R., Coady, J. A. & Kiefte, M. 2003 Sensitivity to change in perception of speech. *Speech Commun.* **41**, 59–69. (doi:10.1016/S0167-6393(02)00093-6)

Kluk, K. & Moore, B. C. J. 2004 Factors affecting psychophysical tuning curves for normally hearing subjects. *Hear. Res.* **194**, 118–134. (doi:10.1016/j.heares.2004.04.012)

Kohlrausch, A., Fassel, R. & Dau, T. 2000 The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. *J. Acoust. Soc. Am.* **108**, 723–734. (doi:10.1121/1.429605)

Kortekaas, R. W. & Kohlrausch, A. 1999 Psychoacoustical evaluation of PSOLA II. Double-formant stimuli and the role of vocal perturbation. *J. Acoust. Soc. Am.* **105**, 522–535. (doi:10.1121/1.424588)

Langner, G. & Schreiner, C. E. 1988 Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J. Neurophysiol.* **60**, 1799–1822.

Leshowitz, B. 1971 Measurement of the two-click threshold. *J. Acoust. Soc. Am.* **49**, 426–466.

Licklider, J. C. & Pollack, I. 1948 Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech. *J. Acoust. Soc. Am.* **20**, 42–52. (doi:10.1121/1.1906346)

Lopez-Poveda, E. A. & Meddis, R. 2001 A human nonlinear cochlear filterbank. *J. Acoust. Soc. Am.* **110**, 3107–3118. (doi:10.1121/1.1416197)

Moore, B. C. J. 1986 Parallels between frequency selectivity measured psychophysically and in cochlear mechanics. *Scand. Audiol.* **25**(Suppl.), 139–152.

Moore, B. C. J. 1992 Across-channel processes in auditory masking. *J. Acoust. Soc. Jpn (E)* **13**, 25–37.

Moore, B. C. J. 1996 Masking in the human auditory system. In *Collected papers on digital audio bit-rate reduction* (eds N. Gilchrist & C. Grewin), pp. 9–19. New York, NY: Audio Engineering Society.

Moore, B. C. J. 1998 *Cochlear hearing loss*. London, UK: Whurr Publishers Ltd.

Moore, B. C. J. 2003*a* *An introduction to the psychology of hearing*, 5th edn. San Diego, CA: Academic Press.

Moore, B. C. J. 2003*b* Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms. *Speech Commun.* **41**, 81–91. (doi:10.1016/S0167-6393(02)00095-X)

Moore, B. C. J. 2003*c* Temporal integration and context effects in hearing. *J. Phonet.* **31**, 563–574. (doi:10.1016/S0095-4470(03)00011-1)

Moore, B. C. J. & Carlyon, R. P. 2005 Perception of pitch by people with cochlear hearing loss and by cochlear implant users. In *Pitch perception* (eds C. J. Plack, A. J. Oxenham, R. R. Fay & A. N. Popper), pp. 234–277. New York, NY: Springer.

Moore, B. C. J. & Glasberg, B. R. 1983*a* Masking patterns of synthetic vowels in simultaneous and forward masking. *J. Acoust. Soc. Am.* **73**, 906–917. (doi:10.1121/1.389015)

Moore, B. C. J. & Glasberg, B. R. 1983*b* Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* **74**, 750–753. (doi:10.1121/1.389861)

Moore, B. C. J. & Glasberg, B. R. 1987 Formulae describing frequency selectivity as a function of frequency and level and their use in calculating excitation patterns. *Hear. Res.* **28**, 209–225. (doi:10.1016/0378-5955(87)90050-5)

Moore, B. C. J. & Glasberg, B. R. 1988 Effects of the relative phase of the components on the pitch discrimination of complex tones by subjects with unilateral and bilateral cochlear impairments. In *Basic issues in hearing* (eds H. Duifhuis, H. Wit & J. Horst), pp. 421–430. London, UK: Academic Press.

Moore, B. C. J. & Glasberg, B. R. 1990 Frequency discrimination of complex tones with overlapping and non-overlapping harmonics. *J. Acoust. Soc. Am.* **87**, 2163–2177. (doi:10.1121/1.399184)

Moore, B. C. J. & Ohgushi, K. 1993 Audibility of partials in inharmonic complex tones. *J. Acoust. Soc. Am.* **93**, 452–461. (doi:10.1121/1.405625)

Moore, B. C. J. & Peters, R. W. 1992 Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity. *J. Acoust. Soc. Am.* **91**, 2881–2893. (doi:10.1121/1.402925)

Moore, B. C. J. & Shailer, M. J. 1992 Modulation discrimination interference and auditory grouping. *Phil. Trans. R. Soc. B* **336**, 339–346. (doi:10.1098/rstb.1992.0067)

Moore, B. C. J., Glasberg, B. R. & Shailer, M. J. 1984 Frequency and intensity difference limens for harmonics within complex tones. *J. Acoust. Soc. Am.* **75**, 550–561. (doi:10.1121/1.390527)

Moore, B. C. J., Glasberg, B. R. & Peters, R. W. 1985 Relative dominance of individual partials in determining the pitch of complex tones. *J. Acoust. Soc. Am.* **77**, 1853–1860. (doi:10.1121/1.391936)

Moore, B. C. J., Glasberg, B. R., Plack, C. J. & Biswas, A. K. 1988 The shape of the ear's temporal window. *J. Acoust. Soc. Am.* **83**, 1102–1116. (doi:10.1121/1.396055)

Moore, B. C. J., Glasberg, B. R., Gaunt, T. & Child, T. 1991 Across-channel masking of changes in modulation depth for amplitude-and frequency-modulated signals. *Q. J. Exp. Psychol.* **43A**, 327–347.

Moore, B. C. J., Peters, R. W. & Glasberg, B. R. 1993 Detection of temporal gaps in sinusoids: effects of frequency and level. *J. Acoust. Soc. Am.* **93**, 1563–1570. (doi:10.1121/1.406815)

Moore, B. C. J., Alcántara, J. I. & Dau, T. 1998 Masking patterns for sinusoidal and narrowband noise maskers. *J. Acoust. Soc. Am.* **104**, 1023–1038. (doi:10.1121/1.423321)

Moore, B. C. J., Glasberg, B. R., Flanagan, H. J. & Adams, J. 2006 Frequency discrimination of complex tones; assessing the role of component resolvability and temporal fine structure. *J. Acoust. Soc. Am.* **119**, 480–490. (doi:10.1121/1.2139070)

O'Loughlin, B. J. & Moore, B. C. J. 1981*a* Improving psychoacoustical tuning curves. *Hear. Res.* **5**, 343–346. (doi:10.1016/0378-5955(81)90057-5)

O'Loughlin, B. J. & Moore, B. C. J. 1981*b* Off-frequency listening: effects on psychoacoustical tuning curves obtained in simultaneous and forward masking. *J. Acoust. Soc. Am.* **69**, 1119–1125. (doi:10.1121/1.385691)

Oxenham, A. J. & Moore, B. C. J. 1994 Modeling the additivity of nonsimultaneous masking. *Hear. Res.* **80**, 105–118. (doi:10.1016/0378-5955(94)90014-0)

Patterson, R. D. 1976 Auditory filter shapes derived with noise stimuli. *J. Acoust. Soc. Am.* **59**, 640–654. (doi:10.1121/1.380914)

Patterson, R. D. 1987 A pulse ribbon model of monaural phase perception. *J. Acoust. Soc. Am.* **82**, 1560–1586. (doi:10.1121/1.395146)

Patterson, R. D. & Moore, B. C. J. 1986 Auditory filters and excitation patterns as representations of frequency resolution. In *Frequency selectivity in hearing* (ed. B. C. J. Moore), pp. 123–177. London, UK: Academic.

Patterson, R. D. & Nimmo-Smith, I. 1980 Off-frequency listening and auditory filter asymmetry. *J. Acoust. Soc. Am.* **67**, 229–245. (doi:10.1121/1.383732)

Patterson, R. D. & Wightman, F. L. 1976 Residue pitch as a function of component spacing. *J. Acoust. Soc. Am.* **59**, 1450–1459. (doi:10.1121/1.381034)

Pierrehumbert, J. 1979 The perception of fundamental frequency declination. *J. Acoust. Soc. Am.* **66**, 363–368. (doi:10.1121/1.383670)

Plack, C. J. & Moore, B. C. J. 1990 Temporal window shape as a function of frequency and level. *J. Acoust. Soc. Am.* **87**, 2178–2187. (doi:10.1121/1.399185)

Plomp, R. 1964*a* The ear as a frequency analyzer. *J. Acoust. Soc. Am.* **36**, 1628–1636. (doi:10.1121/1.1919256)

Plomp, R. 1964*b* The rate of decay of auditory sensation. *J. Acoust. Soc. Am.* **36**, 277–282. (doi:10.1121/1.1918946)

Plomp, R. 1967 Pitch of complex tones. *J. Acoust. Soc. Am.* **41**, 1526–1533. (doi:10.1121/1.1910515)

Plomp, R. 1970 Timbre as a multidimensional attribute of complex tones. In *Frequency analysis and periodicity detection in hearing* (eds R. Plomp & G. F. Smoorenburg), Leiden, The Netherlands: Sijthoff.

Plomp, R. 1976 *Aspects of tone sensation*. London, UK: Academic Press.

Plomp, R. & Steeneken, H. J. M. 1969 Effect of phase on the timbre of complex tones. *J. Acoust. Soc. Am.* **46**, 409–421. (doi:10.1121/1.1911705)

Pollack, I. 1968 Periodicity discrimination for auditory pulse trains. *J. Acoust. Soc. Am.* **43**, 1113–1119. (doi:10.1121/1.1910946)

Remez, R. E., Rubin, P. E., Pisoni, D. B. & Carrell, T. D. 1981 Speech perception without traditional speech cues. *Science* **212**, 947–950. (doi:10.1126/science.7233191)

Ritsma, R. J. 1967 Frequencies dominant in the perception of the pitch of complex sounds. *J. Acoust. Soc. Am.* **42**, 191–198. (doi:10.1121/1.1910550)

Ronken, D. 1970 Monaural detection of a phase difference between clicks. *J. Acoust. Soc. Am.* **47**, 1091–1099. (doi:10.1121/1.1912010)

Rosen, S. & Fourcin, A. 1986 Frequency selectivity and the perception of speech. In *Frequency selectivity in hearing* (ed. B. C. J. Moore), pp. 373–487. London, UK: Academic.

Rosen, S., Baker, R. J. & Darling, A. 1998 Auditory filter nonlinearity at 2 kHz in normal hearing listeners. *J. Acoust. Soc. Am.* **103**, 2539–2550. (doi:10.1121/1.422775)

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S. & Robles, L. 1997 Basilar-membrane responses to tones at the base of the chinchilla cochlea. *J. Acoust. Soc. Am.* **101**, 2151–2163. (doi:10.1121/1.418265)

Saberi, K. & Perrott, D. R. 1999 Cognitive restoration of reversed speech. *Nature* **398**, 760. (doi:10.1038/19652)

Sachs, M. B. & Kiang, N. Y. S. 1968 Two-tone inhibition in auditory nerve fibers. *J. Acoust. Soc. Am.* **43**, 1120–1128. (doi:10.1121/1.1910947)

Schouten, J. F. 1968 The perception of timbre. In *Sixth Int. Conf. on Acoustics*, vol. 1, Tokyo, GP-6-2.

Schouten, J. F. 1970 The residue revisited. In *Frequency analysis and periodicity detection in hearing* (eds R. Plomp & G. F. Smoorenburg), pp. 41–54. Leiden, The Netherlands: Sijthoff.

Schreiner, C. E. & Urbas, J. V. 1986 Representation of amplitude modulation in the auditory cortex of the cat I. The anterior auditory field (AAF). *Hear. Res.* **21**, 227–241. (doi:10.1016/0378-5955(86)90221-2)

Shackleton, T. M. & Carlyon, R. P. 1994 The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *J. Acoust. Soc. Am.* **95**, 3529–3540. (doi:10.1121/1.409970)

Shailer, M. J. & Moore, B. C. J. 1983 Gap detection as a function of frequency, bandwidth and level. *J. Acoust. Soc. Am.* **74**, 467–473. (doi:10.1121/1.389812)

Shailer, M. J. & Moore, B. C. J. 1987 Gap detection and the auditory filter: phase effects using sinusoidal stimuli. *J. Acoust. Soc. Am.* **81**, 1110–1117. (doi:10.1121/1.394631)

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. 1995 Speech recognition with primarily temporal cues. *Science* **270**, 303–304. (doi:10.1126/science.270.5234.303)

Shaw, E. A. G. 1974 Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *J. Acoust. Soc. Am.* **56**, 1848–1861. (doi:10.1121/1.1903522)

Steeneken, H. J. M. & Houtgast, T. 1980 A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.* **69**, 318–326. (doi:10.1121/1.384464)

Stone, M. A. & Moore, B. C. J. 2004 Side effects of fast-acting dynamic range compression that affect intelligibility in a competing speech task. *J. Acoust. Soc. Am.* **116**, 2311–2323. (doi:10.1121/1.1784447)

Summerfield, A. Q. & Assmann, P. 1987 Auditory enhancement in speech perception. In *The psychophysics of speech perception* (ed. M. E. H. Schouten), pp. 140–150. Dordrecht, The Netherlands: Martinus Nijhoff.

Summerfield, A. Q., Sidwell, A. S. & Nelson, T. 1987 Auditory enhancement of changes in spectral amplitude. *J. Acoust. Soc. Am.* **81**, 700–708. (doi:10.1121/1.394838)

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P. & Meddis, R. 2002 A revised model of the inner-hair cell and auditory-nerve complex. *J. Acoust. Soc. Am.* **111**, 2178–2188. (doi:10.1121/1.1453451)

't Hart, J. 1981 Differential sensitivity to pitch distance, particularly in speech. *J. Acoust. Soc. Am.* **69**, 811–821. (doi:10.1121/1.385592)

Verhey, J. L., Dau, T. & Kollmeier, B. 1999 Within-channel cues in comodulation masking release (CMR): experiments and model predictions using a modulation-filterbank model. *J. Acoust. Soc. Am.* **106**, 2733–2745. (doi:10.1121/1.428101)

Viemeister, N. F. 1979 Temporal modulation transfer functions based on modulation thresholds. *J. Acoust. Soc. Am.* **66**, 1364–1380. (doi:10.1121/1.383531)

Vogten, L. L. 1978 Low-level pure-tone masking: a comparison of "tuning curves" obtained with simultaneous and forward masking. *J. Acoust. Soc. Am.* **63**, 1520–1527. (doi:10.1121/1.381846)

von Bismarck, G. 1974 Sharpness as an attribute of the timbre of steady sounds. *Acustica* **30**, 159–172.

Watkins, A. J. 1991 Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *J. Acoust. Soc. Am.* **90**, 2942–2955. (doi:10.1121/1.401769)

Watkins, A. J. & Makin, S. J. 1996*a* Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *J. Acoust. Soc. Am.* **99**, 3749–3757. (doi:10.1121/1.414981)

Watkins, A. J. & Makin, S. J. 1996*b* Some effects of filtered contexts on the perception of vowels and fricatives. *J. Acoust. Soc. Am.* **99**, 588–594. (doi:10.1121/1.414515)

Yost, W. A. & Sheft, S. 1989 Across-critical-band processing of amplitude-modulated tones. *J. Acoust. Soc. Am.* **85**, 848–857. (doi:10.1121/1.397556)

Yost, W. A., Sheft, S. & Opie, J. 1989 Modulation interference in detection and discrimination of amplitude modulation. *J. Acoust. Soc. Am.* **86**, 2138–2147. (doi:10.1121/1.398474)

Young, E. D. 2008 Neural representation of spectral and temporal information in speech. *Phil. Trans. R. Soc. B* **363**, 923–945. (doi:10.1098/rstb.2007.2151)

Yumoto, E., Gould, W. J. & Baer, T. 1982 Harmonics-to-noise ratio as an index of the degree of hoarseness. *J. Acoust. Soc. Am.* **71**, 1544–1549. (doi:10.1121/1.387808)

Zhang, X., Heinz, M. G., Bruce, I. C. & Carney, L. H. 2001 A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *J. Acoust. Soc. Am.* **109**, 648–670. (doi:10.1121/1.1336503)

Zwicker, E. 1964 'Negative afterimage' in hearing. *J. Acoust. Soc. Am.* **36**, 2413–2415. (doi:10.1121/1.1919373)

Zwicker, E. & Fastl, H. 1999 *Psychoacoustics–facts and models*, 2nd edn. Berlin, Germany: Springer.

Zwicker, E. & Terhardt, E. 1980 Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.* **68**, 1523–1525. (doi:10.1121/1.385079)