13

Auditory Modelling

Previous chapters addressed hearing and its capabilities, mainly from an experimental point of view, and a few phenomena were formulated mathematically. Such formulas may be thought to represent a mathematical model of the corresponding phenomenon. However, it is unlikely that a holistic mathematical model or a theory about the complete auditory system can ever be derived due to the enormous complexity of the system. Having said that, simplified mathematical theories are essential for determining causalities and for predicting the perception evoked by a given stimulus, which provides the evident need for experimental analysis and modelling of hearing.

Due to the complexity of the auditory system, computational processing of digitized signals has proven to be the best method to model the functionality of the system. Moreover, computational simulations can be used to design experiments addressing a specific part of the auditory system, which can potentially result in new hypotheses about the physiological functionality. Typically, these computational models are employed to study information processing in the auditory pathway and different input–output relationships.

An even stronger motivation for modelling hearing computationally originates from the engineering point of view, wherein a functional model enables emulating the functionality of hearing in numerous practical applications, especially if the model runs in real time. Such applications include, for example, speech recognition, sound reproduction, spatial audio techniques, hearing aids, and cochlear implants. Mimicking brain functions in the computational domain is also very educating and an inspiring topic: the human brain is a good engineering solution, and reverse engineering it is a good exercise in signal processing.

The history of computational auditory models is relatively short, the first serious attempts being made in the 1960s and the 1970s (Chistovitch, 1974; Dolmazon *et al.*, 1976; Weiss, 1966). However, the number of researchers developing and applying auditory models has increased rapidly since the 1980s, making today's, field rich with a plethora of publications available. An overview of the current status is given by Meddis (2010).

This chapter describes several computational auditory models and their applications. The focus is first placed on the simpler models, moving gradually towards more complex ones.

Communication Acoustics: An Introduction to Speech, Audio, and Psychoacoustics, First Edition.
Ville Pulkki and Matti Karjalainen.

Here, the term *auditory model* is used as a general concept, while the terms *psychoacoustic model* and *perceptual model* are used to refer to models designed to explain results of psychoacoustic experiments without paying specific attention to the physiology.

This chapter briefly overviews existing auditory models classified as follows:

- Simple psychoacoustic models;
- Filter bank models;
- Cochlear models;
- Hair-cell models;
- Models for cognitive processing;
- Models of binaural interaction.

13.1 Simple Psychoacoustic Modelling with DFT

As discussed in previous chapters, the fundamental psychoacoustic theory of hearing describes the peripheral hearing system as a kind of spectral analyser that extracts several perceptual aspects like specific loudness and overall loudness, pitch, duration, sharpness, and roughness, to mention a few, from the ear canal input. Since each of these aspects can be described in a quantitative manner depending on the properties of the input signal, computational models can be designed to extract metrics related to these aspects and, consequently, to emulate the functionality of hearing at least to some extent.

13.1.1 Computation of the Auditory Spectrum through DFT

The most common approach in simple psychoacoustic models is based on the processing involved in loudness perception, as discussed in Section 10.2.5. Here, a discrete Fourier transform (DFT) -based approach is presented, which, instead of actually estimating the loudness, derives an 'auditory spectrum' describing the level of cochlear excitation in dB as a function of frequency on the ERB scale. Such a spectrum can be extracted with the Matlab script listed below.

```
fs=48000; % frequency of sampling
sig=(rand(1,fs/2)-0.5)*10; % 500 ms of white noise
winlen=round(fs/40); % 25 ms time window
[blp,alp]=butter(2,(500 / (fs/2)), 'high'); % high-pass filter
zE=[1:41];                    % utilized ERB channel numbers
fE=228.7*(10.^(zE/ 21.3)-1);  % corresponding frequencies
% gain to implement hump around 4 kHz (ERB 25 +- 7)
hump_coeffs=1+(max(0,7-abs(25-zE))/7)*6;
% approximated spreading function of excitation in ERB scale
spreadfunct=10.^([-80 -60 -40 -20 0 -8 -16 -24 -32 -40 -48 -56
                                                    -64]/10);
sig=filter(blp,alp,sig);%high-pass to simulate LF sensitivity
                                                          loss
a=1; % time position counter

for i=1:winlen/2:(length(sig)-winlen) % loop through the
                                                      signal
```

```
    % window the signal, and take FFT
    SIG=fft(hamming(winlen)' .* sig(i:(i+winlen-1)));
    POWSPECT=SIG.*conj(SIG);     % compute power spectrum
    % scaling linear frequency to ERB
    lowlimit=1; i=1;
    for z=zE(2:end)
        highlimit=round(fE(z)/fs*winlen); % upper FFT-bin for
                                                ERB channel
        % sum the power inside ERB channel
        excitation(i)=sum(POWSPECT(lowlimit:highlimit))*hump_
                                                 coeffs(i);
        lowlimit=highlimit+1; i=i+1; % update counters and
                                                lowlimit
    end
    % implement excitation spreading by convolution with
                                               spreading
    % function and store the excitation patter
    excitpattern(a,:)=conv(excitation,spreadfunct);
    a=a+1; %counter update
end
% computation of auditory spectrum
audspec=10*log10(mean(excitpattern,1)); % avg over time
hearthr=(zE-24).^2/15;  % crude approx. hearing threshold
figure(1); clf;  axes('Position',[0.1 0.1 0.5 0.3])
plot(zE(2:end)-0.5, max(hearthr(2:end), audspec(5:(end-8))),
                                                      '-')
hold on;  plot(zE(1:end), hearthr,'--'); set(gca,'XTick',
                                               [2:4:40]);
xlabel('ERB scale'); ylabel('Auditory spectrum [dB]')
```

The computation of an auditory spectrum begins with the power spectrum computation. The input signal is first divided into, say, 25-ms-long time frames that are then multiplied by a suitable window function, like a Hamming window. Thereafter, a short-term power spectrum is computed for each time frame using the DFT, which is implemented using the fast Fourier transform (FFT). This processing does not reflect the frequency-dependent sensitivity of hearing. In practice, the sensitivity can be emulated at any stage of the computation, but conceptually, it should be emulated at the beginning by filtering the input signal. The Matlab script yields a coarse approximation of the inverse of the equal loudness contour at 60 dB SPL (Figure 9.2) by high-pass filtering the signal at the beginning and then multiplying the short-term power spectra by frequency-dependent weights. As a consequence, both the poor sensitivity of hearing to low frequencies and the increased sensitivity around 4 kHz, resulting from the ear canal resonance, are emulated roughly.

After this, the power spectra on the Hz-scale are converted to the Bark or ERB scale. Such *frequency warping* can be implemented in various ways. For an example, see the explanation of the computation of the mel frequency cepstral coefficients below. The Matlab script above implements the frequency warping by simply summing the power spectrum values within each ERB band.

The next processing step emulates how the excitation evoked by a stimulus spreads to other frequency bands. The simplest emulation approach comprises a convolution with a spreading function, as in Equation (10.6), and such an approach is also exploited in the code above. The function approximates the shape of the simultaneous frequency masking curve at 60 dB SPL. In reality, the shape of such a curve is level dependent, which can be accounted for by selecting the spreading function in a level-dependent manner. This is neglected here, since it would make the structure of the auditory model far more complex. The excitation patterns following the convolution can then be scaled into short-term specific loudness spectra following Equation (10.7). However, the presented model omits this scaling and consequently the resulting auditory spectrum describes the level of the excitation pattern in dB in different frequency bands. In other words, the output provides an estimate of the sound spectrum that is available to hearing mechanisms.

The linear power spectra and auditory spectra for a pure tone, white noise, and two speech signals are shown in Figure 13.1. Overall, the auditory spectrum is seen to differ from the power spectrum due to the use of the ERB scale, the asymmetric spreading of the excitation, and the emulation of the hearing threshold. The pure tone case demonstrates how the excitation spreads to adjacent frequency bands. The white noise case, in turn, visualizes the combined effect of the frequency warping and the emulation of the frequency-dependent sensitivity of hearing on the spectrum. That is, the level of the auditory spectrum increases with frequency, and the boost around 4 kHz resulting from the ear canal resonance is visible as well. In addition, the auditory spectrum differs from the physical one in terms of frequency resolution. The auditory spectra of the speech signals illustrate how the harmonic fine structure in voiced phonemes is visible in the physical spectra but is smoothed in the auditory spectra due to frequency warping. Additionally, the warping increases the visibility of low frequencies in the spectra. However, the formants in the speech signals are clearly also present in the auditory spectra.

Applications of DFT-based auditory models

Several variations of the model above have been designed for different purposes. For instance, the two variations described below have been applied in feature extraction in speech recognition algorithms.

- *Mel frequency cepstral coefficients* (MFCCs) (Davis and Mermelstein, 1980) are commonly used to characterize speech sounds. Figure 13.2 shows how these features are extracted from an input signal that is first processed with a high-pass filter to differentiate the speech waveform. Subsequently, a power spectrum is computed for each windowed time frame. The power spectra are then warped onto the mel scale with a filter bank consisting of a set of M (here 20) triangular-shaped band-pass filters. Thereafter, the logarithm of the filter bank outputs gives the coefficients X_k. Finally, M MFCCs are derived with the discrete cosine transform:

$$c_n = \sum_{k=1}^{20} X_k \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{20}\right], \qquad \text{for } n = 1, 2, \ldots, M. \tag{13.1}$$

  MFCCs have been found to characterize speech sounds efficiently, hence their frequent use in speech recognition algorithms, especially those employing statistical models for the actual identification.

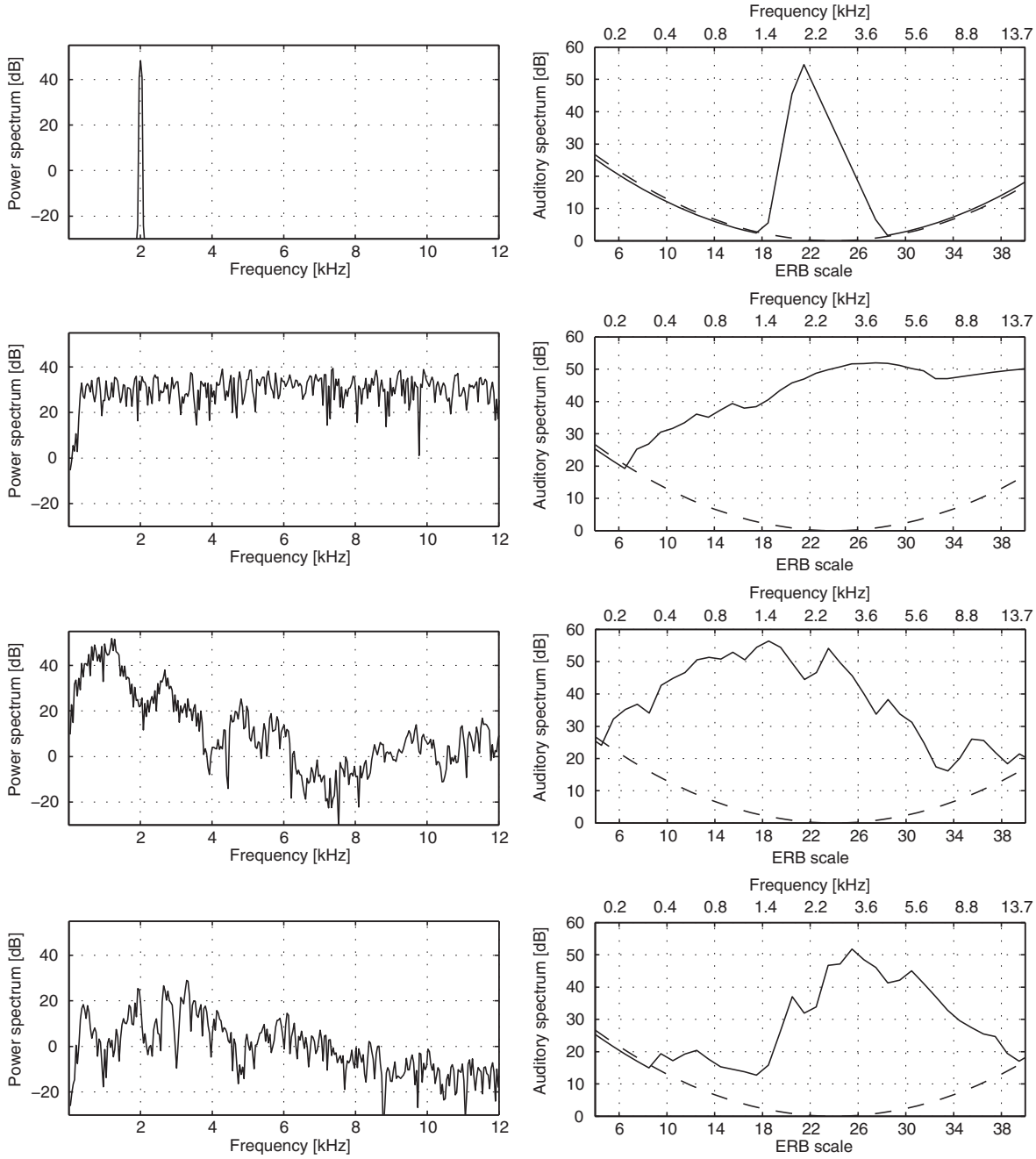

Figure 13.1 Power spectra (left) and auditory spectra (right) for (from top to bottom) a 2-kHz pure tone, white noise, the vowel /a/ and the fricative /s/. The auditory spectra were computed with the Matlab script presented in this section, which provides an implementation of a DFT-based auditory model. The auditory spectra have been computed as averages over several subsequent frames.

- *Perceptual linear prediction* (PLP) (Hermansky, 1990). Conceptually, the PLP coefficients and MFCCs are extracted in a similar manner. In PLP, however, the coefficients are extracted from a specific loudness spectrum. Moreover, the specific loudness spectra are transformed back into autocorrelation functions using the inverse Fourier transform, after which the traditional autocorrelation-based LP algorithm is used to extract the coefficients. The resulting coefficients are able to describe the spectral features of speech signals in a compact and rather speaker-independent manner.

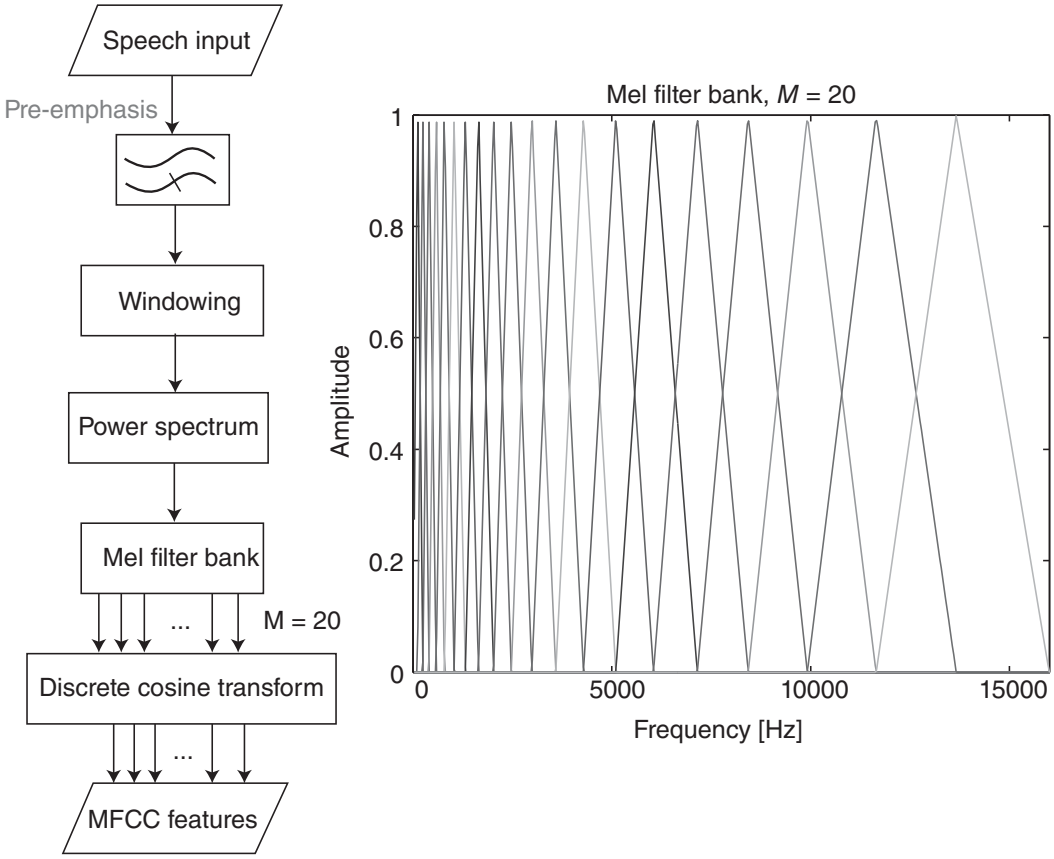

Figure 13.2 The computation of mel cepstral coefficients (MFCCs). The components of the power spectrum are combined with triangular weight functions to give spectral components in the mel frequency scale. The components are further transformed using a cosine transform (Equation (13.1)), which gives out the cepstral coefficients. Courtesy of Marko Takanen.

Similar computations are also performed in many *audio codecs* that map windowed frames of a signal into the frequency domain following the frequency resolution of human hearing. These codecs then perform further quantization or other operations in each time–frequency bin. Such techniques will be described in more detail in Chapter 15.

The technically straightforward and efficient computation of the auditory spectrum described above cannot be used to describe the functionality of hearing accurately, in part due to the following reasons:

- *Temporal resolution*. The length of the time frame and the type of the window function define the temporal resolution of the power spectrum derived with the DFT. However, the time–frequency resolution of hearing does not utilize a time frame of fixed length. The temporal resolution is about 1–2 ms at high frequencies and larger at low frequencies. Such a variation cannot be emulated with the above-mentioned procedure. The filter bank models described below can account for the time–frequency resolution of hearing more accurately.
- *Temporal dynamics*. The DFT-based model presented also ignores a few temporal effects in our hearing resolution. Such effects include, among others, temporal integration (Figure 10.16) and post-masking (Figures 9.10, 9.11, and 9.12). In principle, temporal integration and post-masking can be emulated by processing the time-framed signals, but not very accurately.

- *Level dependency*. If only a simple spreading function is used to simulate the spreading of the excitation to other frequencies, the level dependency cannot be emulated. However, as mentioned above, this can be fixed when a suitable level-dependent function is used.

13.2 Filter Bank Models

The fundamental problem in the above-mentioned simple psychoacoustic models is their inability to emulate the temporal resolution and dynamics of the auditory system. As mentioned previously, this problem originates from the use of fixed-length time frames in the DFT computation, the length of which defines the temporal resolution of such a model. The time–frequency resolution of hearing can be emulated more accurately with filter bank models that process the signals with a set of band-pass filters in the time domain.

Figure 13.3 illustrates how an auditory spectrum can be derived with a model employing a filter bank to emulate the time–frequency resolution. Both Bark and ERB resolutions can be emulated by selecting the filters appropriately. Such an auditory model can also emulate the spreading of the excitation and temporal masking effects.

13.2.1 Modelling the Outer and Middle Ear

The transfer functions of the external and middle ear must be emulated with appropriate filters before processing the signal with a filter bank model. Various approaches can be exploited for this purpose depending on the requirements of the application. The best accuracy in emulating the external ear is achieved using measured HRTFs of an individual subject or of a dummy head. Typically, the middle ear transfer function is considered a band-pass filter with, say, a 6-dB/octave decreasing frequency response at frequencies below 800 Hz as well as above 1.5 kHz, as shown in Figure 7.5.

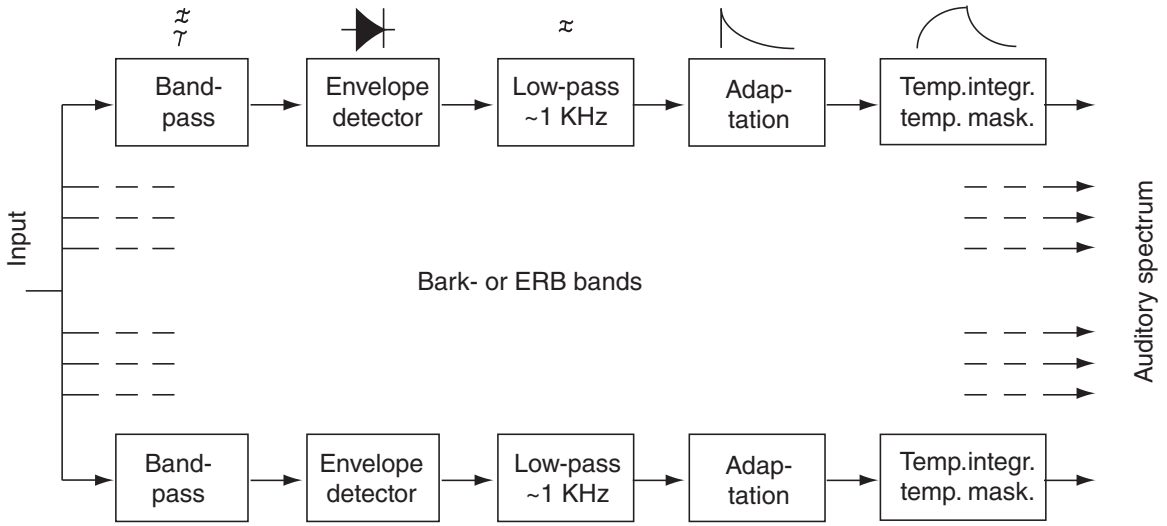

Figure 13.3 An auditory model implemented with a filter bank of multiple band-pass filters. The band-pass filters are followed by signal envelope detection using half-wave rectification and low-pass filtering corresponding to the monaural time resolution. The filters for short-term adaptation and temporal integration for temporal masking then simulate the low-level temporal effects in hearing. Section 13.6 gives two simple Matlab implementations of such auditory models.

13.2.2 Gammatone Filter Bank and Auditory Nerve Responses

The *gammatone filter bank* (Patterson, 1994) is the most commonly used method to emulate the frequency resolution of hearing. The physiological basis for such filters originates from the so-called *reverse correlation technique* measurements (De Boer, 1969) that yield an estimate of the impulse response of the auditory nerve fibre. Since this estimate resembles the shape of a pure tone that has been modulated with a gamma function, the corresponding filter is known as a gammatone filter. In addition, the frequency response of a gammatone filter is very similar to the human auditory filter as estimated by psychoacoustic notched-noise measurements (Glasberg and Moore, 1990). The popularity of the gammatone filter bank is also influenced by its computational efficiency and the relatively simple design. The impulse response of a gammatone filter is given by

$$g(t) = a\, t^{n-1} e^{-2\pi\, b(f_c)\, t} \cos(2\pi f_c t + \phi), \tag{13.2}$$

where a is the peak value of the response; t^{n-1}, which specifies the onset time of the response, together with the exponential term characterizes the bandwidth and decay of the response, f_c is the *characteristic frequency* of the filter, and ϕ is the initial phase of the response. As an example, the impulse response of a gammatone filter and the corresponding magnitude response as well as the magnitude responses of a 32-band gammatone filter bank (100 Hz $\leq f_c \leq$ 10 kHz) are shown in Figure 13.4. Typically, auditory models utilize about 42 bands in the gammatone filter banks, covering a frequency range from about 30 Hz to 18 kHz.

Although gammatone filters provide a good approximation of the human auditory filters, they suffer from a few shortcomings. They cannot emulate the level-dependent characteristics of the auditory filters. In addition, the impulse response of a gammatone filter has a relatively slow onset, which brings on problems when modelling phenomena involving temporally short sounds, such as the precedence effect.

13.2.3 Level-Dependent Filter Banks

As noted previously, the response of the cochlea shows level-dependent asymmetry in the form of compressive input or output functionalities. Various modelling approaches have been taken to form a filter bank that is able to emulate the suppressive and compressive characteristics of the auditory filters (see, for example, Carney 1993; Irino and Patterson 1997; Meddis *et al.* 2001; and Patterson *et al.* 2003).

An example of these approaches is the dual resonance non-linear (DRNL) filter bank (Meddis *et al.*, 2001). As shown in Figure 13.5, each band of a DRNL filter bank consists of two parallel processing paths, one of which employs a broadly tuned band-pass filter with a linear input–output relation. Additionally, a narrowly-tuned band-pass filter is used in the other processing path so that the gain of the filter compresses the output at higher levels. Both band-pass filters have an asymmetric frequency response, which is achieved by low-pass filtering the outputs of the gammatone filters. A weighted sum is then computed from the outputs of the two processing paths to acquire the DRNL filter bank output for a given frequency band. Moreover, the weights are set so that the outputs of the non-linear and the linear processing paths dominate the filter bank output at low and high levels, respectively. This implements the level dependence of the output.

Unfortunately, the response to a single impulse obtained by adding the outputs of the two processing paths is a double impulse. As a consequence, the output does not retain the temporal fine structure of the input, which may be problematic in some cases.

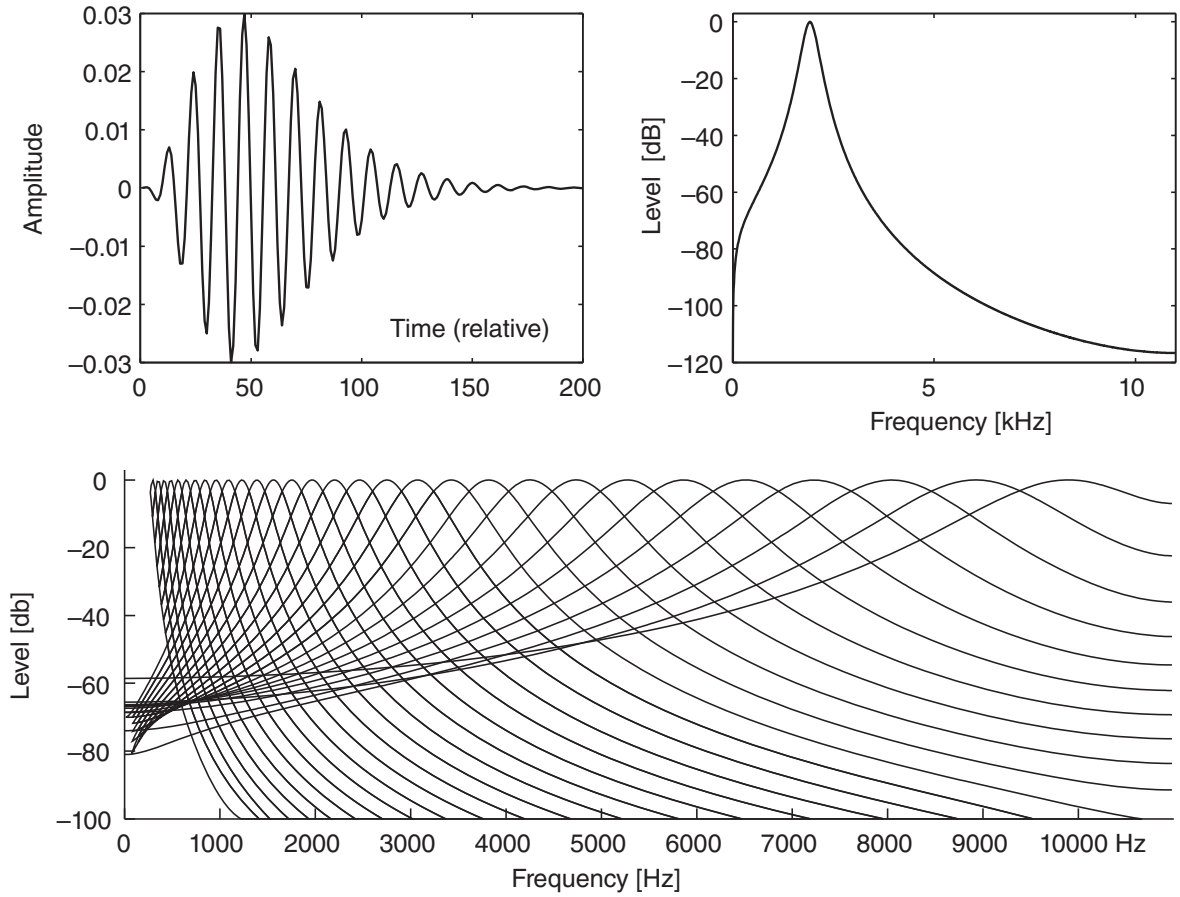

Figure 13.4 The characteristics of a gammatone filter bank: a) the impulse response of an individual filter, b) the corresponding magnitude response, and c) the magnitude responses of the filter bank on a linear frequency scale.

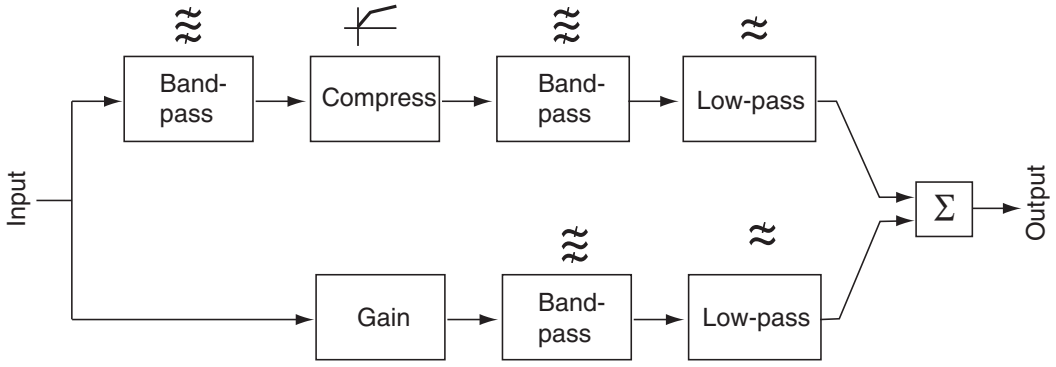

Figure 13.5 A dual-resonance filter, where the lower path employs a broader band-pass filter and the upper path a narrower band-pass filtering with compression. Such filters are used in DRNL filter banks to simulate cochlear processing since they emulate the level-dependent output of the cochlea better than a gammatone filter bank.

An alternative model that simulates the asymmetrical non-linear behaviour of the cochlea is the one by Zhang *et al.* (2001). Their model consists of a filter bank composed of filters that are time-varying, narrowly tuned, linear band-pass filters. Each of these is controlled by a non-linear, broadly tuned control filter. In particular, the output of the control filter sets the instantaneous gain and bandwidth of the corresponding filter, allowing the reproduction of

cochlear non-linearities and phenomena like two-tone suppression. Furthermore, a variant of this model has been implemented by Zilany and Bruce (2006), in which an additional linear filter bank is used to emulate a second pathway of excitation of the inner hair cells. This model also allows the reproduction of the large phase changes in the inner-hair-cell responses at high SPLs. Moreover, it can be used to simulate the functionality of an impaired peripheral auditory system.

13.2.4 Envelope Detection and Temporal Dynamics

The inner hair cells and the auditory nerve fibres transform the mechanical vibrations of the basilar membrane into neural impulses. As noted earlier, the dependency of the rate of impulses on the basilar membrane displacement can be characterized with half-wave rectification (see Figure 7.18). The synchrony between the excitation and the firing rate is lost approximately at frequencies above 1 kHz. This can be modelled with a process that involves temporal integration with a certain temporal window. The 1-kHz limit corresponds to a time constant of 150 μs. Hence, the filter bank model shown in Figure 13.3 typically emulates the neural transduction by processing the half-wave rectified filter-bank outputs with first or second-order low-pass filters.

In the filter bank model of Figure 13.3, the next processing block emulates adaptation with a kind of high-pass filtering that strongly emphasizes the onset of a stationary stimulus. The various adaptation models, described, for instance, by Dau *et al.* (1996), Lyon (1982), and Seneff (1988), can be considered to be based on the idea of *automatic gain control* (AGC) that slowly reduces the amplification as the level of the input increases. Figure 13.6 illustrates an example of how the adaptation can be emulated with a series of feedback loops utilizing different time constants (Dau *et al.*, 1996). Specifically, the divisor elements control the

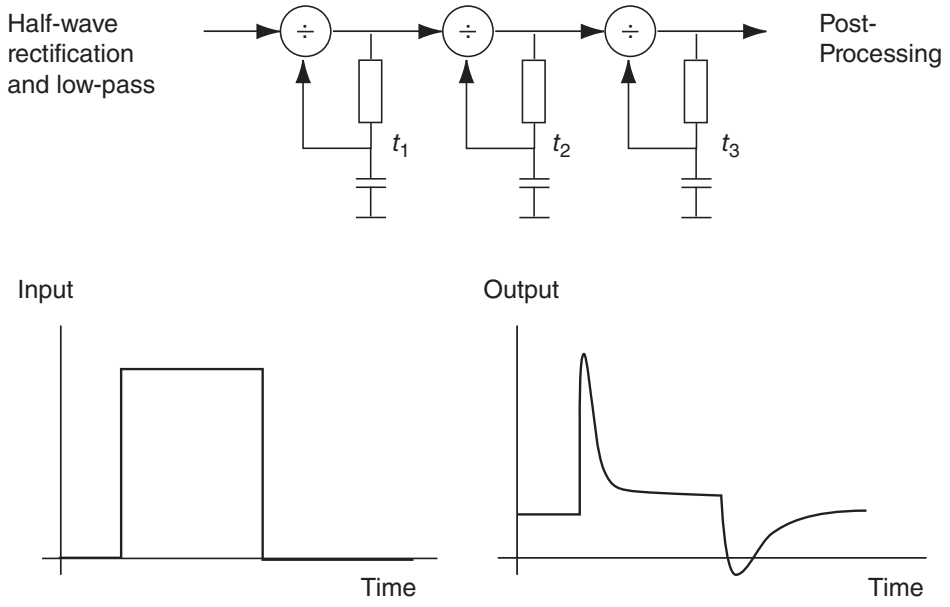

Figure 13.6 A model for adaptation, where the feedback loops act as automatic gain controls via the division operation. The time constants are typically selected to be between 5 ms and 500 ms.

amplification or attenuation by dividing the passing signal by the one coming from the corresponding feedback loop. For a continuous signal, the processing through a series of loops results in a nearly logarithmic output level, while the onsets are emphasized in a similar manner to Figure 3.13.

The last processing block in the model shown in Figure 13.3 utilizes larger time constants to emulate temporal integration and post-masking phenomena. Moreover, the energy of the input signal is low-pass filtered using time constants of 100–200 ms to simulate temporal integration, and post-masking is emulated using the same time constant in a non-linear filter that effectively prolongs the recovery time of the processing block. This block is needed, for example, when modelling dynamic loudness perception. Furthermore, a signal representing the specific loudness as a function of time can be obtained by suitably compressing the output. Optionally, the filter bank model can be designed to have parallel processing paths for temporal integration and adaptation, both of which receive the low-pass filtered envelope signals as input.

It should noted that Figure 13.3 shows only an overview of the functional elements, while an actual filter bank model implementation requires detailed design of compatible elements. In addition, some of the elements may be excluded from certain applications. For instance, not all the short time constants are necessary to evaluate loudness, whereas pitch analysis does not benefit from the use of large time constants.

Furthermore, the implementation is not restricted to splitting the processing in the aforementioned manner. For instance, adaptation, temporal integration, and post-masking can all be simulated effectively in a single element (Karjalainen, 1996), as shown in Figure 13.7. The element consists of an envelope detection unit (half-wave rectification and a low-pass filter), a multiplier controlling the amplification of the signal, two parallel low-pass filters utilizing different time constants, and a logarithmic feedback loop connecting the summed outputs of the low-pass filters to the multiplier. The primary output signal is the temporary loudness level (in phons or dBs) that can be transformed into specific loudness following Equation (10.2).

Figure 13.8 illustrates the two outputs of the above-described model (Figure 13.7) for a pure tone signal with a square wave envelope. The auditory nerve response (firing rate) is shown in Figure 13.8b, where the emphasis of the onset and the subsequent adaptation are clearly visible. Figure 13.8c shows the loudness-level output reflecting the temporal integration and post-masking effects. Interestingly, the two output signals that result from the same feedback process can be seen as complementary signals.

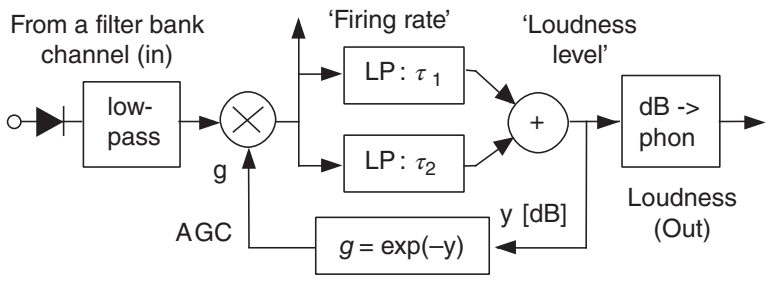

Figure 13.7 A model for adaptation and loudness for filter-bank-based auditory models.

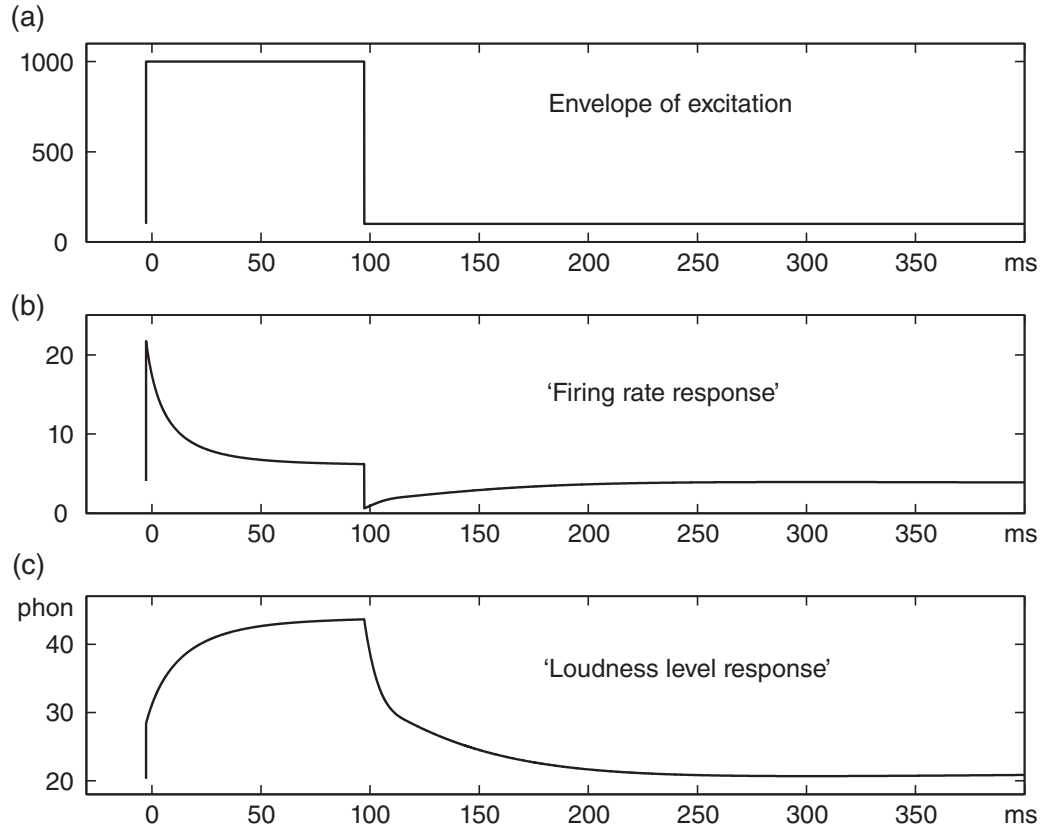

Figure 13.8 Responses computed using the model presented in Figure 13.7: (a) tonal excitation with stepped envelope; (b) fast response (onset response), and (c) slow response (loudness response).

13.3 Cochlear Models

So far, this chapter has presented auditory models that explain the functionalities of hearing, giving less emphasis to the physiological details. In some cases, more accurate modelling of the physiological characteristics is necessary for detailed investigations of the auditory system. This demand has been addressed in several models, many of them aiming to simulate the movement of the basilar membrane inside the cochlea.

13.3.1 Basilar Membrane Models

We saw earlier that the vibration of the stapes that is attached to the oval window generates pressure waves in the fluid inside the cochlea. Since these waves resonate at frequency-specific positions along the basilar membrane, accurate modelling of this phenomenon requires a model consisting of spatially distributed elements. Typically, one-dimensional (1D) travelling wave models are used, but 2D and 3D models may also be used at the expense of increased computational complexity. One option is, for instance, to use a *finite element method* (FEM) to simulate the phenomenon in the frequency domain, but this provides only a rough approximation, assuming the underlying system to be linear and time-invariant. Alternatively, a non-linear time-domain solution may be obtained with a *finite difference method*. However, most of the basilar membrane models see the membrane as a transmission line that can be modelled with electrical equivalent circuits.

Specifically, a transmission-line model represents the basilar membrane as a cascade of coupled mass–spring–damper systems. In the equivalent circuit illustrated in Figure 13.9

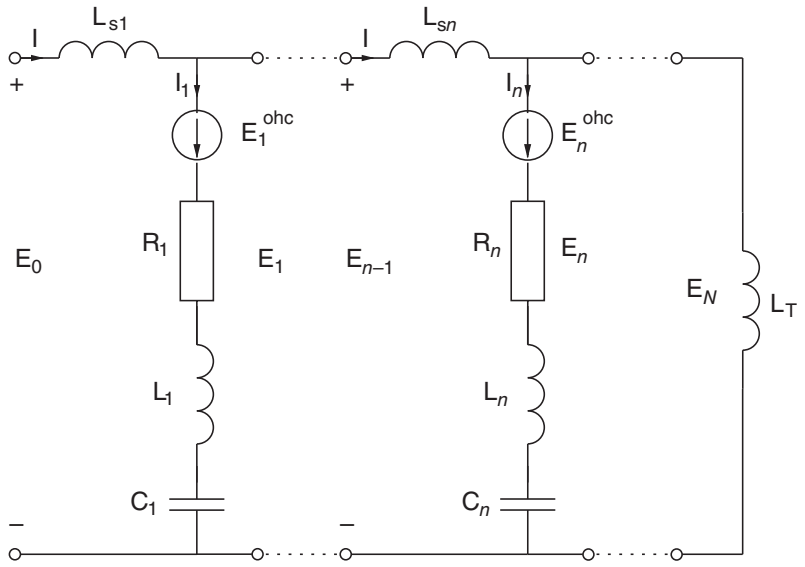

Figure 13.9 Equivalent circuit of the basilar membrane as a transmission line.

(Strube, 1985), the mass and damping characteristics along the basilar membrane are represented by inductors and resistors, respectively, whereas a capacitor is used to represent the energy storage capabilities, such as those of a spring. Using this approach, the vibrations in the membrane are simulated as longitudinal-wave propagation that resonates at a frequency-specific point, abating quickly thereafter. In a digital simulation, the analogue circuit is discretized using so-called *wave-digital filters*, or alternatively numerical methods are used to solve the system of ordinary differential equations that describe the model (Diependaal *et al.*, 1987; Elliott *et al.*, 2007).

Even though the circuit shown in Figure 13.9 is linear and time-invariant, the active role of the cochlear amplifier may be simulated by including negative damping elements (Zweig, 1991) and other non-linear and level-dependent elements (Shera, 2001). For instance, the detectability of amplitude modulation (Figure 10.14) cannot be modelled without accounting for the level dependency of the tuning curves.

13.3.2 Hair-Cell Models

As seen in Section 7.4.1, the bending of inner hair cell stereocilia due to cochlear vibrations modulates the potential difference across the membrane of the cell. This variation in the potentials drives non-deterministically the firing rates of the auditory nerve fibres synaptically connected to the hair cell. The functionality of an auditory fibre has a stochastic nature; one cannot accurately predict when the fibre fires. Therefore, the signal from a single auditory fibre contains somewhat noisy data, and the outputs of large numbers of fibres need to be combined to form the pure and clean sensation that a normal functional auditory system produces. In simple functional models, the combined functionality of inner hair cells and nerve fibres can be emulated deterministically with half-wave rectification and low-pass filtering. However, the stochastic nature of the nerve-fibre firing may be simulated more accurately with a probabilistic model of the inner hair cell and auditory nerve complex. Figure 13.10 depicts the working principle of the model by Meddis (1988), which is the most famous of these models.

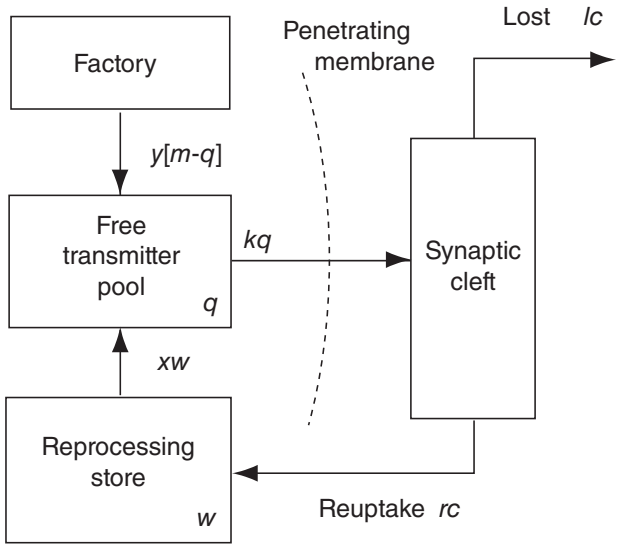

Figure 13.10 Flowchart of the inner-hair-cell model adapted from Meddis (1988).

The left-most blocks in the figure represent an inner hair cell that is synaptically connected to the auditory nerve fibre. The model assumes that the probability of a neural impulse $p(t)$ is linearly dependent on the amount of transmittal material $c(t)$ in the *synaptic cleft* between the hair cell and the nerve fibre:

$$p(t) = h\,c(t)\,\mathrm{d}t, \tag{13.3}$$

where h is a parameter of the model and $\mathrm{d}t$ corresponds to the computational sampling rate. Furthermore, the release of the transmittal material is designed to depend on the *permeability* of the hair-cell membrane $k(t)$, which is modulated by the amplitude of the excitatory stimulus $x(t)$:

$$k(t) = g\,\mathrm{d}t\,\frac{x(t)+A}{x(t)+A+B},\ \text{when } x(t)-A>0,\ \text{and} \tag{13.4a}$$

$$k(t) = 0,\ \text{when } x(t)-A<0. \tag{13.4b}$$

Here, g, A, and B are parameters of the model. Hence, the amount of released transmittal material at a time instant t corresponds to $k(t)\,q(t)\,\mathrm{d}t$, where $q(t)$ denotes the amount of transmittal material in the transmitter pool next to the membrane. The majority of the transmittal material returns from the cleft to the hair cell at the rate $r\,c(t)$, while some of the material is lost in the cleft and from the system with the speed $l\,c(t)$. This loss introduces an adaptation to the firing rate. Moreover, the returned transmittal material first spends some time in the reprocessing store before entering the transmitter pool again. The speed of transmittal material entering the pool corresponds to $x\,w(t)$, where $w(t)$ denotes the amount of transmittal material in the store. Additionally, the hair cell contains a factory producing the transmittal material at the speed $y\{m-q(t)\}$, where $m=1$.

In practice, the functionality of the model can be characterized with three differential equations:

$$\mathrm{d}q/\mathrm{d}t = y\{m - q(t)\} + x\,w(t) - k(t)\,q(t) \tag{13.5}$$

$$\mathrm{d}c/\mathrm{d}t = k(t)\,q(t) + l\,c(t) - r\,c(t) \tag{13.6}$$

$$\mathrm{d}w/\mathrm{d}t = r\,c(t) - x\,w(t). \tag{13.7}$$

The model has been shown to yield accurate simulations of physiological responses.

13.4 Modelling of Higher-Level Systemic Properties

The above-mentioned auditory models are related to a relatively low level of neural processing. It is useful and necessary to simulate the functionality of hearing at higher levels to understand the functionality of the auditory system in detail. The functional models may either focus on a specific phenomenon or aim to describe the bigger picture of information processing. Unfortunately, there is a lack of precise knowledge and experimental data regarding phenomena requiring higher-level cognitive processing, and the models are based on high-level assumptions of neurophysiology and subsequent testing of the models against psychoacoustic data.

The following parts of this section describe a few functional models for higher-level processing. Some of them also have a limited physiological basis, but, in general, they are hypothetical models.

13.4.1 Analysis of Pitch and Periodicity

The existence of the two alternative theories for pitch perception is also reflected in the auditory models, that are based on either spectral (place theory) or periodicity (temporal theory) analysis (Plack *et al.*, 2005). Spectral analysis of pitch assumes that the auditory system can extract frequency information with a high resolution which is then analysed at the neural level by a central processor. Alternatively, the time-domain models bolster the idea that several pitch perception phenomena can be explained with simple, low-level time-domain processing based on periodicity. The models are generally based on the idea that the auditory system extracts pitch with neural processing resembling the computation of an autocorrelation function (Licklider, 1951, 1959; Meddis and Hewitt, 1991, 1992; Meddis and O'Mard, 1997).

Figure 13.11 shows the general concept for autocorrelation-based pitch analysis. The signals originating from the filter bank are first processed by a hair-cell model consisting of a half-wave rectifier and a subsequent low-pass filter. A separate autocorrelation function (ACF) is then computed for each signal in order to detect the periodicities in the different sub-bands. The ACF presentation is often called a *correlogram*. Thereafter, the separate ACFs are summed to obtain the *summary autocorrelation function*, characterizing the periodicities in the original stimulus.

The Matlab code in Section 13.6.1 produces the outputs plotted in Figure 13.12, demonstrating the effects of the different processing steps. Figure 13.12a shows the waveform of the vowel signal, and the outputs of the filter bank are shown in Figure 13.12b. The latter plot is often referred to as a *cochleogram*. In addition, the signals following the hair-cell processing are shown in Figure 13.12c, while the correlogram and the summary autocorrelation function are illustrated in Figure 13.12d. The last of the graphs shows a clear peak at 9 ms that corresponds to the fundamental frequency of 110 Hz of the vowel input.

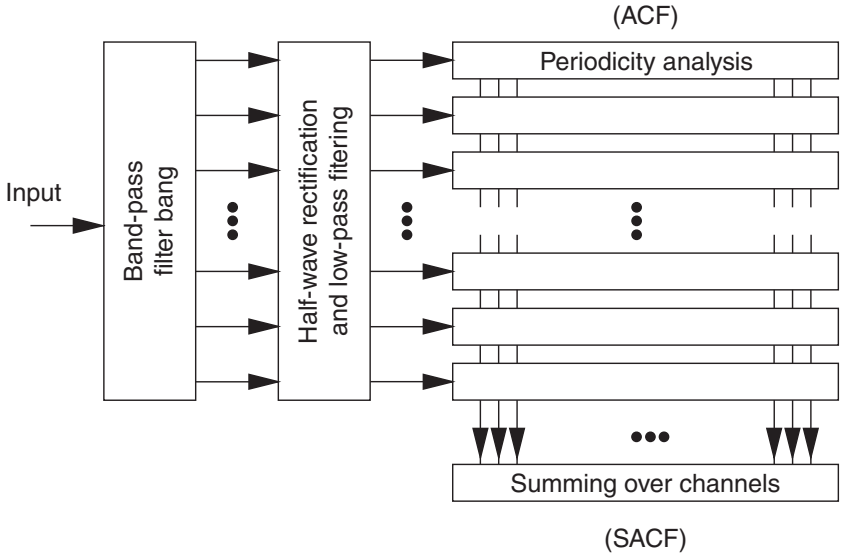

Figure 13.11 The general concept of pitch analysis with an autocorrelation-based auditory model. An autocorrelation function (ACF) is computed in the periodicity analysis, and the results are summed to form the summary autocorrelation function (SACF).

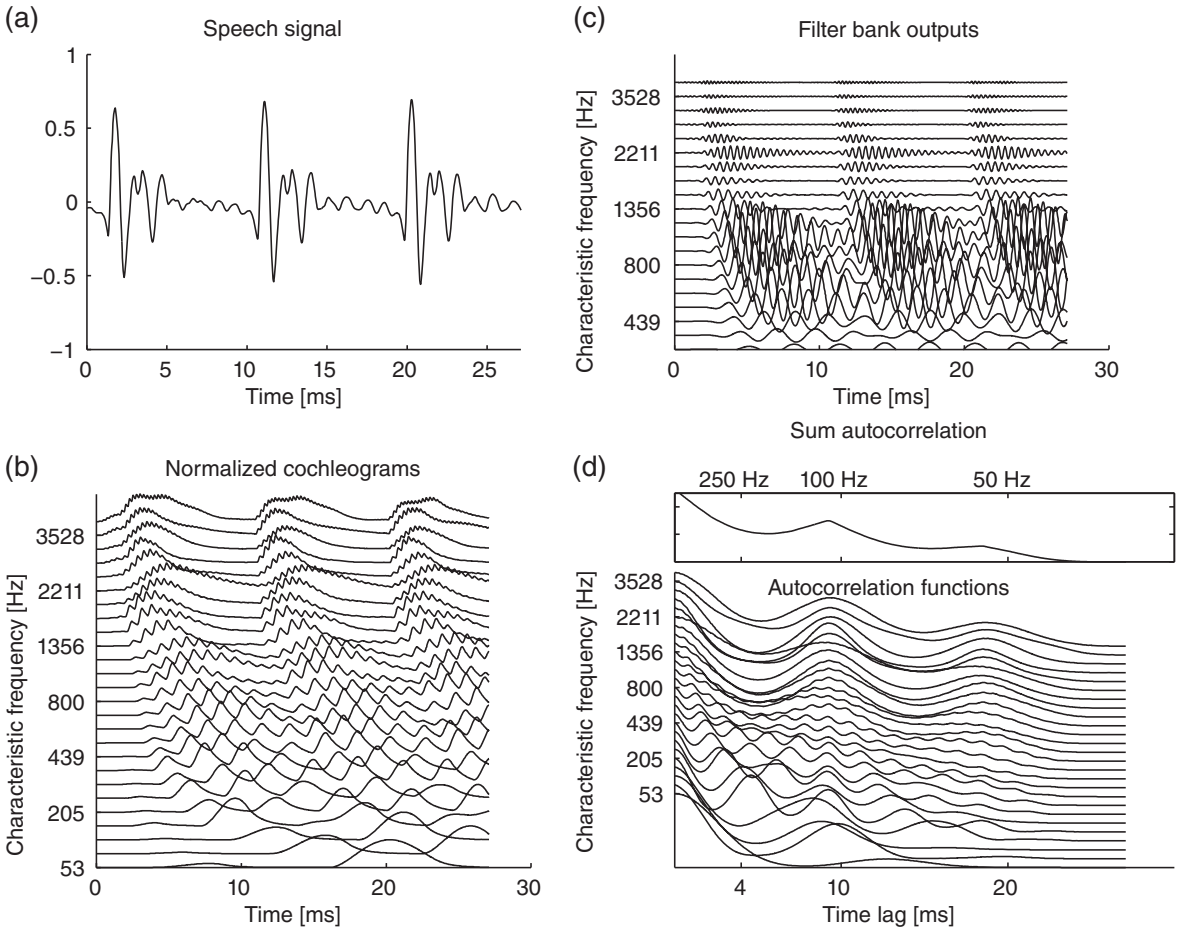

Figure 13.12 Auditory pitch analysis for a 27-ms segment of the vowel /a/: (a) the time-domain signal, (b) the cochleogram, (c) the hair-cell model output with the highest peaks normalized to unity, and (d) the normalized autocorrelation functions and the summary autocorrelation function.

Autocorrelation-based analysis has proven to be able to explain several phenomena of pitch perception (Plack *et al.*, 2005). In addition, such models can be applied to segregate different sound sources, particularly to segregate concurrent vowel sounds from each other when the sounds differ in terms of the fundamental frequency (Meddis and Hewitt, 1992).

13.4.2 Modelling of Loudness Perception

Section 10.2 introduced a simple loudness model that accurately estimates the loudness perception evoked by many relatively simple signals. Such models have been found to be unable to derive accurate loudness estimates for spectrally complex and time-variant signals. This shortcoming has motivated the design of more advanced loudness models (Florentine *et al.*, 2005), of which two approaches are briefly touched upon next.

The first one (Zwicker, 1977) processes the signal in the time domain and produces a continuous signal representing the loudness as a function of time. This approach opens up the possibility of predicting the loudness perception evoked by a time-variant signal based on the peak values in the model output. The second model (Glasberg and Moore, 2002) divides the signal into overlapping time frames and extracts short-term loudness values from each time frame. The model also emulates the temporal integration of loudness between adjacent time frames, and consequently, the model provides accurate estimates for time-variant signals as well, although this estimate is derived by simply averaging across the short-term loudness values. Despite the improved accuracy in predicting perceived loudness of complex signals, the different models cannot yet fully explain loudness perception. The objective audio and speech quality methods discussed in Sections 17.5.2 and 17.8.1 can also be seen as models that estimate the specific loudness depending on time, and the interested reader might find the references in those sections worth exploring.

13.5 Models of Spatial Hearing

As discussed in Chapter 12, human spatial hearing capabilities are based on the binaural and monaural analysis of the ear canal signals. Spatial hearing is able to localize sources with good accuracy, although the reflections and reverberations of the room may corrupt the directional cues in the ear canal signals. This remarkable ability has, for decades, inspired researchers to model spatial hearing. A plethora of binaural and monaural models of spatial hearing have been proposed (Blauert, 1996, 2013; Colburn, 1996; Stern and Trahiotis, 1995), and some of them are discussed next.

13.5.1 Delay-Network-Based Models of Binaural Hearing

The majority of the binaural processing algorithms are based on the coincidence detection model proposed by Jeffress (1948). The model suggests that certain neurons in the brain are narrowly tuned to specific ITDs between the ear canal signals. As illustrated in Figure 13.13, the model consists of an array of coincidence-detector neurons receiving excitatory signals from both ears, and delay lines are used to represent axons connecting the neuron to the cochlear nuclei of the left and right ears. The highest activity is then received from the coincidence-detector neuron where the propagation delay in the inputs effectively cancels out the ITD between the left and right ear inputs. This probably also facilitates channels sensitive to specific ITDs, as suggested by (Fastl and Zwicker, 2007).

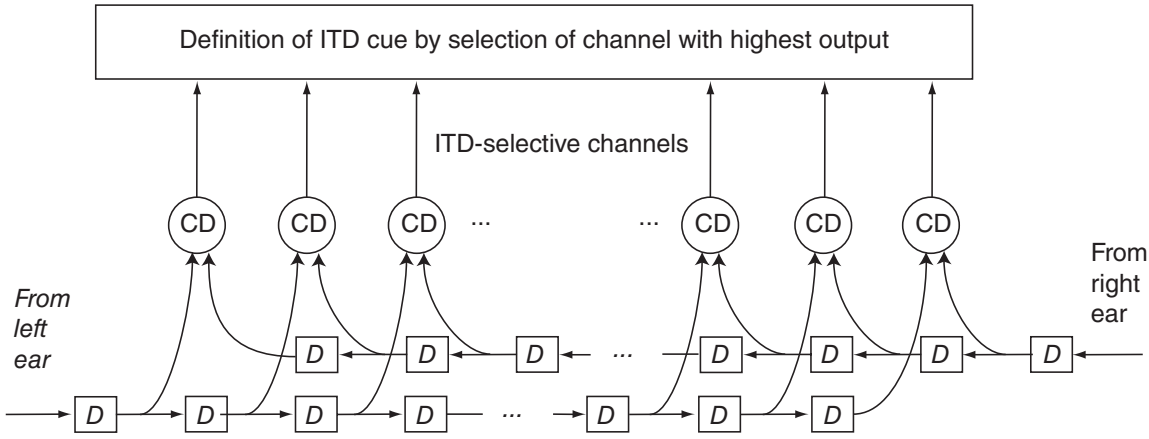

Figure 13.13 A schematic illustration of the coincidence detection model proposed by Jeffress (1948), where delay lines represent axons connecting the ear canal inputs to the coincidence-detector neuron (CD). Here, D is the unit delay. The outputs of the CD-neurons are then thought to be compared with each other, and the highest outputs are thought to define the ITD cue(s).

Such processing can be elegantly emulated by computing the normalized interaural cross-correlation (IACC) (Sayers and Cherry, 1957)

$$\gamma(t,\tau) = \frac{\int_t^{t+\Delta t} x_\mathrm{l}(T-\tau/2)x_\mathrm{r}(T+\tau/2)\,\mathrm{d}T}{\sqrt{\int_t^{t+\Delta t} x_\mathrm{l}^2(T)\,\mathrm{d}T + \int_t^{t+\Delta t} x_\mathrm{r}^2(T)\,\mathrm{d}T}}, \tag{13.8}$$

where t is time, τ is the interaural delay, Δt denotes the length of the integration window, and x_l and x_r are the signals from the left and right ears, respectively. An estimate of the ITD is then obtained as the interaural delay corresponding to the maximum of the IACC function. The output of the IACC computation may also be used to visualize the auditory scene as a cross-correlogram-type binaural activity map (Shackleton *et al.*, 1992). The Matlab script listed in Section 13.6.2 demonstrates how such a map can be extracted for a binaural input signal.

We will next discuss the output of the IACC computation, which is shown in Figure 13.14. Figure 13.12c shows the signals used as input in the computation, to make this discussion more comprehensible. The normalized cross-correlation function is plotted for each frequency channel. The plotting shows that the functions are tuned to a certain time lag, which corresponds to the ITD between the ear canal signals. In this case, the source was in the direction 30° azimuth, and the maximum value of the IACC depends relatively strongly on the frequency content in the interval between 0.2 ms and 0.5 ms. The value from the theoretical broadband curve shown in Figure 12.9 matches with the IACC-based estimate at 700 Hz. The time lags corresponding to the maxima of the IACC functions are plotted in the lower panel in Figure 13.14. This is the ITD function often seen in the literature, which is thought to represent the ITD cue accessible to the higher levels in processing (Blauert, 2013).

The IACC function is normalized with the power of signals, which means that it attains the value one only when the ear canal signals differ by the amount of the ITD and the ILD in the natural range, otherwise it gets a lower, non-negative value. The maximum value can thus be used to estimate the *interaural coherence* (Faller and Merimaa, 2004).

The IACC humps are broader at low frequencies than at high frequencies. This is because a constant change in time lag corresponds to a smaller change in phase at low frequencies, and

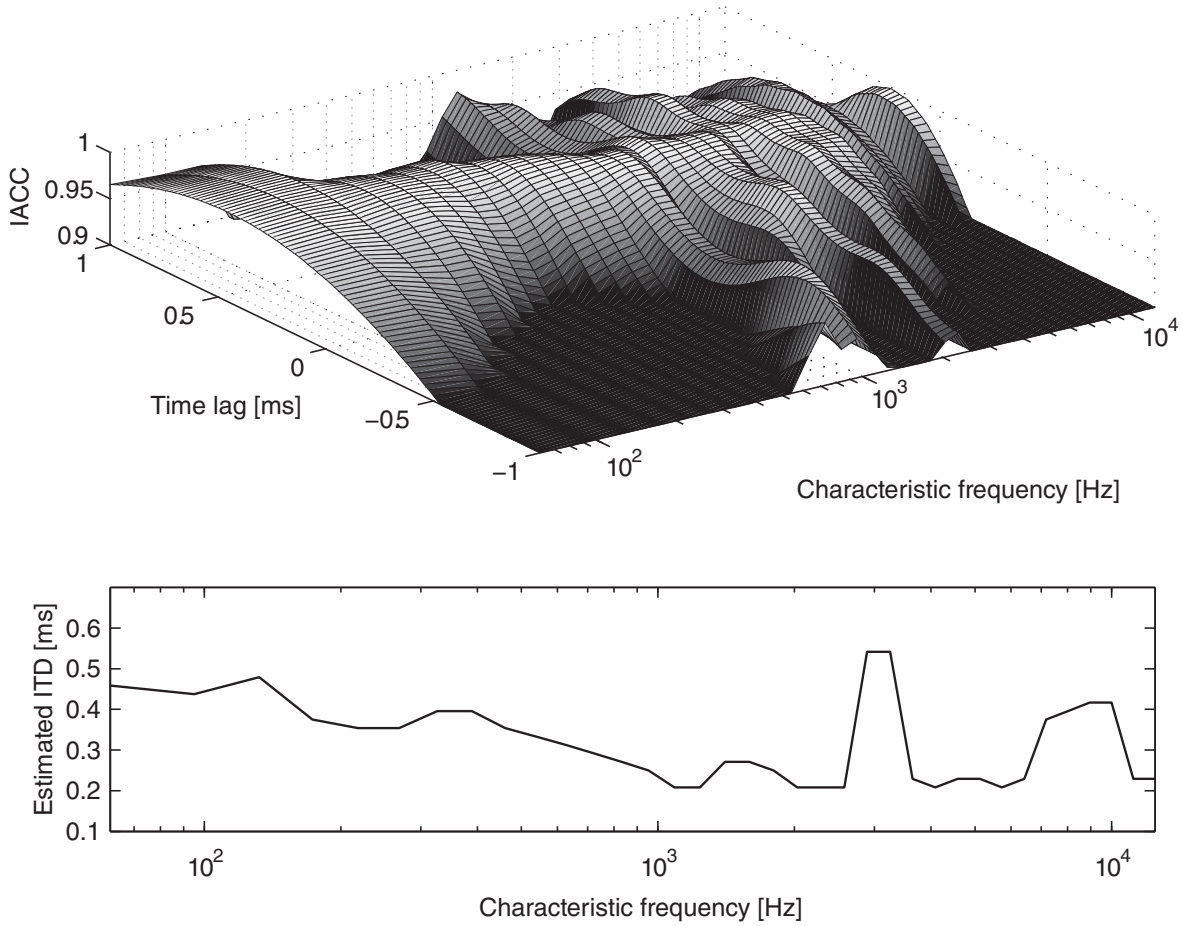

Figure 13.14 A cross-correlogram-type binaural activity map for a scenario where a single speech source at the azimuthal angle of 30° is simulated with HRTFs measured from a real subject. The bottom-most graph shows the estimated ITD in different frequency bands.

a larger change at higher frequencies. Near 1 kHz, sidelobes are seen in the functions, which occur because the period of the centre frequency of auditory bands is shorter than 1 ms, and the IACC analysis also finds high correlation in the signals when the time lags equal the delay or the phase advances by 2π. The effect of frequency on temporal spacing of peaks is clearly shown in Figure 13.12c.

At higher frequencies, the IACC no longer shows side bands, since the temporal details of the signals are mostly lost, which is also evident in Figure 13.12c. Instead, only the temporal envelope is preserved, and since, in this case, the signal was a short excerpt of the vowel /a/, a rather strong temporal structure remains. The cross-correlation between the left and right signals then results in only a single hump in the IACC functions. Such a clear unitary hump is not always found in the simulation results. If the input had been, for example, a high-frequency sinusoidal tone, the high-frequency IACC functions would be completely flat as a function of the time lag.

A number of extensions have been proposed to the coincidence detection model, such as the one presented by Breebaart *et al.* (2001). In their model, the delay lines are connected to a chain of attenuators and each coincidence detector of the original model (see Figure 13.13) is replaced by two excitation–inhibition cells, one receiving the excitation from the left ear and

inhibition from the right ear, and the other with opposite connections. Effectively, they extend the coincidence detection model to account also for ILD sensitivity. For a binaural input signal, the model outputs an activity map with local minima around the positions corresponding to the ITD and ILD values, and the depths of the troughs depend on the interaural coherence between the ear canal signals.

Such correlation- or coherence-based models have also been used to explain the sensitivity of humans to the coherence between ear canal signals (Bernstein and Trahiotis, 1996). Additionally, in real rooms, when the coherence between ear canal signals varies temporally depending on the source signal and on the room response, it has been suggested that listeners utilize directional cues only when the interaural coherence value is larger than a threshold value (Faller and Merimaa, 2004).

13.5.2 Equalization-Cancellation and ILD Models

The above-mentioned delay-line models analyse the coherence of the ear canal signals to extract directional information. Alternative modelling concepts have also been exploited, some of them being based on subtracting the signals from each other. The equalization-cancellation model (Durlach, 1963) was designed to account for binaural signal detection in the presence of masking noise, and no attempts were made to emulate processing in the auditory pathway. In this model, the left and right inputs are first filtered with a set of band-pass filters so that the narrowband target can be more easily separated from the masker. Thereafter, the masker signal components are equalized in the two ears by adjusting the ITD and ILD values, and the ear canal signals are subtracted from each other, which ideally eliminates the masker from the signal.

A straightforward method to estimate the ILD cue accessible to the auditory system comprises the computation of the level difference between the ear canal signals (Blauert, 1996). Typically, separate ILD estimates are derived for each auditory channel and for each time frame of the signal to maximize accuracy. First, the signal levels in each auditory channel are measured within each time frame, after which the values obtained for the left and right ear signals are compared to each other to derive estimates of the ILD in each time frame and for each auditory channel. Thus, such an ILD estimation can be interpreted as an equalization cancellation model without the equalization phase.

The selection of the length of the time frame opens up possibilities to broaden the model performance for different applications. The output also becomes sensitive to interaural coherence when time frames as short as 5–10 ms are used in the ILD estimation. The ILD values fluctuate randomly in a diffuse sound field (Goupell and Hartmann, 2007; Pulkki and Hirvonen, 2009) and therefore also provide a potential cue for humans to sense spatial attributes related to reverberation.

13.5.3 Count-Comparison Models

Another group of binaural hearing models based on the count-comparison principle (van Bergeijk, 1962; von Békésy, 1930; 1960), which proposes that the nuclei in the two hemispheres encode the spatial direction of sound at the rate of their output (see Figure 13.15), and the spatial location is then indicated by the relative activation rates of the nuclei in the two hemispheres (Stecker *et al.*, 2005). Actually, the above-mentioned ILD models already follow the count-comparison principle, and now it is argued that the ITD is extracted similarly to the ILD. It has also been shown that the ‘comparison’ phase is not needed if the outputs of the

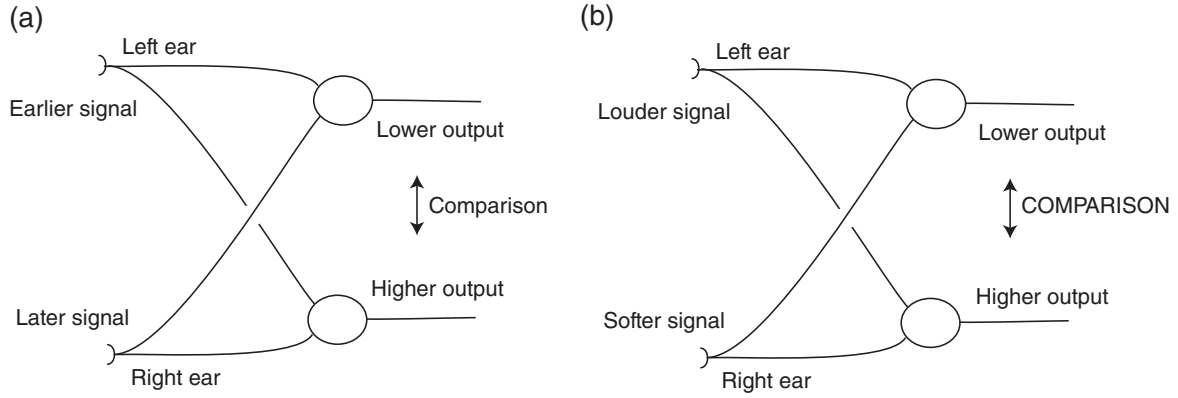

Figure 13.15 The count comparison principle for (a) ITD extraction (b) ILD extraction.

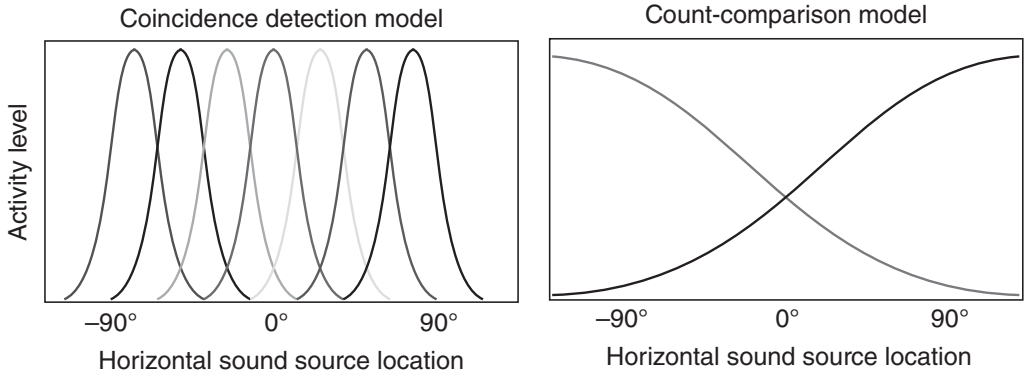

Figure 13.16 The schematic representation of the effect of a horizontal sound source location on the activity levels of different receptive fields in the brain as determined by the coincidence detection and count-comparison models.

nuclei models are self-normalized using their input signals (Pulkki and Hirvonen, 2009). Such an approach yields left- or right-coordinate outputs that depend only on the direction of the sound event and not on the sound pressure level or any other attribute.

One of the principal differences between the count-comparison and coincidence detection models lies in the nature of the output. A count-comparison model outputs a left or right coordinate, whereas each coincidence-detector neuron in a Jeffress-type model provides its own output. These neurons may be thought to be most sensitive to specific left or right directions. In other words, Jeffress-type models assume that receptive fields in the brain are narrowly tuned to specific horizontal directions, while the count-comparison principle bolsters the idea of two wide receptive fields spanning an entire hemifield, as illustrated in Figure 13.16. The left or right coordinate that a count-comparison model outputs cannot be used as such to visualize the surrounding auditory scene as an elegant binaural activity map, as in Figure 13.14. This issue was addressed in a study by Takanen *et al.* (2014), where the binaural cues in the ear canal signals were extracted following the count-comparison principle, and the resulting directional cues were utilized to steer the spectral content of the signals to specific locations on a topographically organized binaural activity map. Thus, the auditory scene is visualized similarly to the cross-correlogram of the Jeffress-type model output.

13.5.4 Models of Localization in the Median Plane

As seen in Section 12.3.4, the localization of a sound source in the median plane relies on spectral cues. Generically, a machine-learning approach can be used to simulate human performance in localizing the elevation of a sound source. In particular, an accurate prediction can be achieved using artificial neural networks trained with large data sets from a single subject (Jin *et al.*, 2000). Recently, a functional model that predicts listeners' performance in such localization tasks has been proposed by Baumgartner *et al.* (2013). This model computes the probability of perceiving a sound coming from a certain elevation angle by comparing the internal representation of the sound as processed by the peripheral auditory system with an internal representation of the HRTF.

13.6 Matlab Examples

This section shows two examples of filter-bank-based auditory models. The models rely on the implementation of a gammatone filter bank in the auditory modelling toolbox (AMT) that can be downloaded from `amtoolbox.sourceforge.net`.

13.6.1 Filter-Bank Model with Autocorrelation-Based Pitch Analysis

This example is explained in Section 13.4, and the plots of its outputs are shown in Figure 13.12.

```
%code for autocorrelation-based pitch analysis in auditory
                                                    models
clc;close all;clear;

%some parameters for the auditory model
fLow = 50;    % the lowest characteristic frequency of the
                                                 filter bank
fHigh = 4000; % and the highest
fCut = 1000;  % cut-off frequency of the low-pass filter

%load a speech file, the example contains 27 ms of /a/ vowel
[sample,fs] = wavread('kaksi.wav',[5900 7200]);

sampleLen = length(sample);
%create of a gammatone filter bank using a command from the
                                                   auditory
%modelling toolbox (http://amtoolbox.sourceforge.net)
[b,a] = gammatone(erbspacebw(fLow,fHigh),fs,'complex');

%processing the signal through the filter bank
filterOut = real(ufilterbankz(b,a,sample));

%emulation of the neural transduction with half-wave
                                    rectification and
```

```
%low-pass filtering of the filter bank output
rectified = filterOut.*(filterOut>0);
%a first-order IIR filter is used as the low-pass filter
beta = exp(-fCut/fs);
outSig = filter(1-beta,[1 -beta],rectified);

for freqInd=1:size(outSig,2) % autocorrelation for each band
    auCorr(:,freqInd)= xcorr(outSig(:,freqInd),'coeff')';
end

%plotting of the figures
auCorr=auCorr((size(auCorr,1)+1)/2:end,:);
lags=[[0:(size(auCorr,1)-1)]/fs*1000];
fcs = erbspacebw(fLow,fHigh);
h=figure;
%plot the input signal
g(1) = subplot('position',[0.13 0.6438 0.3326 0.3012]);hold on;
plot((1:sampleLen)./fs*1000,sample,'k');
xlabel('Time [ms]');title('a) speech signal');
set(gca,'xlim',[0 sampleLen/fs*1000]);

%plot the filter bank outputs
g(2) = subplot('position',[0.5803 0.6438 0.3326 0.3012]);hold
                                                          on;
for freqInd=size(outSig,2):-1:1
    plot((1:sampleLen)./fs*1000,(freqInd-1)/30+filterOut
                                        (:,freqInd),'k');
end
set(gca,'YTick',(0:4:(size(outSig,2)-1))/30);
set(gca,'YTickLabel',round(fcs(1:4:end)));
axis([0 30 0.2 0.9])
xlabel('Time [ms]');ylabel('Characteristic frequency [Hz]');
title('b) filter bank outputs');

%plot the cochleograms
g(3) = subplot('position',[0.13 0.115 0.333 0.3812]);  hold on;
for freqInd=size(outSig,2):-1:1
    plot((1:sampleLen)./fs*1000,(freqInd-1)/2+outSig
            (:,freqInd)/max(outSig(:,freqInd)),'k');
end
set(gca,'YTick',(0:4:(size(outSig,2)-1))/2);
set(gca,'YTickLabel',round(fcs(1:4:end)));
xlabel('Time [ms]');ylabel('Characteristic frequency [Hz]');
title('c) normalized cochleograms');
axis([0 30 0 13.5])

%plot the normalized autocorrelation functions
```

```
g(4) = subplot('position',[0.5803 0.1150 0.3826 0.3012]);hold
                                                          on;
for freqInd=size(outSig,2):-1:1
    plot(lags,(freqInd-1)*0.12+auCorr(:,freqInd),'k');
end

axis([0 30 0 4])
set(gca,'YTick',(0:4:(size(outSig,2)-1))*0.12+1);
set(gca,'YTickLabel',round(fcs(1:4:end)));
set(gca,'xTick',floor(1./[250 100 50]*1000));
ylabel('Characteristic frequency [Hz]');
xlabel('Time lag [ms]');
text(15,3.8,'d) autocorrelation functions',
          'HorizontalAlignment','center');

%plot the sum autocorrelation
g(5) = subplot('position',[0.5803 0.427 0.3826 0.07]);
plot(lags,sum(auCorr'),'k');
set(gca,'xaxisLocation','top')
set(gca,'YTickLabel',[]);
axis([0 30 0 25])
set(gca,'xTick',floor(1./[250 100 50]*1000));
set(gca,'xTickLabel',{'250 Hz','100 Hz','50 Hz'});
title('Sum autocorrelation');
```

13.6.2 Binaural Filter-Bank Model with Cross-Correlation-Based ITD Analysis

This example is explained in Section 13.5.1, and the plots of its outputs are shown in Figure 13.14.

```
%function for IACC-based binaural activity map computation

clear;clc;close all; fs=48000;
% the data files can be obtained from the web siteof this book
load hrir30.mat % a HRTF

%some parameters for auditory model
fLow = 50;%the lowest characteristic frequency of the filter
                                                        bank
fHigh = 13000;%and the highest
fCut = 1000; % cut-off frequency of the low-pass filter
maxLag= floor(0.001*fs); % IACC-values are computed within
                                                  -1...1 ms

%a 29-ms-long sample of speech (vowel /a/) is used as the
                                            source material
```

```
[stim] = wavread('kaksi.wav',[5000 6400]);

% convolve stimulus with HRIRs of left and right ear
insig = [conv(stim,hrir30(:,1)) conv(stim,hrir30(:,2))];

%create of a gammatone filter bank using a command from the
                                                    auditory
%modelling toolbox (http://amtoolbox.sourceforge.net)

cfs = erbspacebw(fLow,fHigh);%characteristic frequencies of
                                              the filter bank
[b,a] = gammatone(cfs,fs,'complex');

%processing the signal through the filter bank
filterOut.left = 2*real(ufilterbankz(b,a,insig(:,1)));
filterOut.right = 2*real(ufilterbankz(b,a,insig(:,2)));

%emulation of the neural transduction with half-wave
                                    rectification and
%low-pass filtering of the filter bank output
rectified.left = filterOut.left.*(filterOut.left>0);
rectified.right = filterOut.right.*(filterOut.right>0);

%a first-order IIR filter is used as the low-pass filter
beta = exp(-fCut/fs);
outSig.left = filter(1-beta,[1 -beta],rectified.left);
outSig.right = filter(1-beta,[1 -beta],rectified.right);

%compute interaural cross-correlation at each frequency band
iaccFuncts = zeros(2*maxLag+1,length(cfs));
lagValues = (-maxLag:maxLag)./fs;
for freqInd=1:length(cfs)
    iaccFuncts(:,freqInd) = xcorr(outSig.left(:,freqInd),...
        outSig.right(:,freqInd),maxLag,'coeff');
end

%compute the ITD estimate at different frequency bands based
                                                  on the maxima
%of the IACC functions
[temp,lag] = max(iaccFuncts);
itdEst = lagValues(lag);

%plotting of figures
h=figure;
%for visualization purposes, only the IACC-values above 0.9
                                                   are plotted
iaccFuncts(iaccFuncts<=0.9)=0.9;
```

```
g(1) = subplot('Position',[0.1300 0.4838 0.7750 0.4412]);
surf(round(cfs),lagValues*1000,iaccFuncts);
set(gca,'XScale', 'log','xTick',[100 1000 10000]);
xlabel('Characteristic frequency [Hz]');ylabel('Time lag
                                                    [ms]');
zlabel('IACC');
xlim(round([min(cfs) max(cfs)]));
view(-35,70);colormap(colormap('gray'));
g(2) = subplot('Position',[0.1300 0.1100 0.7750 0.2412]);
semilogx(round(cfs),itdEst*1000,'k');ylabel('Estimated ITD
                                                      [ms]');
xlabel('Characteristic frequency [Hz]');
axis([round([min(cfs) max(cfs)]) 0.1 0.7])
```

Summary

This chapter reviewed different computational models of the auditory system, covering a wide range of modelling principles that aim to emulate the processing occurring within different regions of the auditory pathway in varying detail. The models based on windowing the ear canal signals and performing DFT-based processing can be used to explain the basic properties of auditory frequency resolution. Although DFT-based modelling has some drawbacks in temporal accuracy, it is interesting in the scope of this book, as many perceptual audio-coding methods are based on similar processing. The auditory models that model the cochlea using filter banks are more precise, and the best accuracy of peripheral modelling is obtained with transmission-line models. Unfortunately, the computational complexity increases drastically when the accuracy of modelling is increased. The models for binaural interaction are based either on sound being encoded into multiple direction-dependent channels, or on directional cue computation for each time–frequency position. The models for pitch and loudness perception succeed in some simple scenarios, however, none of the models can explain perception accurately in all cases.

Further Reading

The main directions of research are covered in the books by Blauert (2013) and Meddis (2010). Many models are also available publicly as part of computational toolboxes, like Majdak and Søndergaard (2013), and Slaney (1998). The auditory models have also found applications in speech and audio techniques, and the remaining chapters of this book review their use in many applications.

References

Baumgartner, R., Majdak, P., and Laback, B. (2013) Assessment of sagittal-plane sound localization performance in spatial-audio applications. In Blauert J. (ed.) *The Technology of Binaural Listening*. Springer, pp. 93–119.

Bernstein, L.R. and Trahiotis, C. (1996) The normalized correlation: Accounting for binaural detection across center frequency. *J. Acoust. Soc. Am.*, **100**(6), 3774–3784.

Blauert, J. (1997) *Spatial Hearing – Psychophysics of Human Sound Localization*. MIT Press.

Blauert, J. (2013) *The Technology of Binaural Listening*. Springer.

Breebaart, J., van de Par, S., and Kohlrausch, A. (2001) Binaural processing model based on contralateral inhibition. I. Model structure. *J. Acoust. Soc. Am.*, **110**(2), 1074–1088.

Carney, L. (1993) A model for the responses of low-frequency auditory-nerve fibers in cat. *J. Acoust. Soc. Am.*, **93**(1), 401–417.

Chistovich, I.A., Granstrem, M.P., Kozhevnikov, V.A., Lesogor, L.W., Shupljakov, V.S., Taljasin, P.A., and Tjulkov, W.A. (1974) A functional model of signal processing in the peripheral auditory system. *Acustica*, **31**(6), 349–354.

Colburn, H.S. (1996) Computational models of binaural processing. In Hawkins, H.L., McMullen, T.A., Popper, A.N., and Fay, R.R. (eds) *Auditory Computation*. Springer, pp. 332–400.

Dau, T., Püschel, D., and Kohlraush, A. (1996) A quantitative model of the "effective" signal processing in the auditory system. I. model structure, II. simulations and measurements. *J. Acoust. Soc. Am.*, **99**, 3615–3631.

Davis, S. and Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. ASSP*, **ASSP-28**, 357–366.

De Boer, E. (1969) Encoding of frequency information in the discharge pattern of auditory nerve fibers. *Int. J. Audiol.*, **8**(4), 547–556.

Diependaal, R.J., Duifhuis, H., Hoogstraten, H., and Viergever, M.A. (1987) Numerical methods for solving one-dimensional cochlear models in the time domain. *J. Acoust. Soc. Am.*, **82**(5), 1655–1666.

Dolmazon, J.M., Bastet, L., and Shupljakov, V.S. (1976) A functional model of the peripheral auditory system in speech processing. *Proc. of IEEE ICASSP'77*, pp. 261–264.

Durlach, N.I. (1963) Equalization and cancellation theory of binaural masking-level differences. *J. Acoust. Soc. Am.*, **35**(8), 1206–1218.

Elliott, S.J., Ku, E.M., and Lineton, B. (2007) A state space model for cochlear mechanics. *J. Acoust. Soc. Am.*, **122**(5), 2759–2771.

Faller, C. and Merimaa, J. (2004) Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, **116**(5), 3075–3089.

Fastl, H. and Zwicker, E. (2007) *Psychoacoustics – Facts and Models*. Springer.

Florentine, M., Popper, A., and Fay, R.R. (eds) (2005) *Loudness*, volume 37. Springer.

Glasberg, B.R. and Moore, B.C.J. (1990) Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, **47**(1–2), 103–138.

Glasberg, B.R. and Moore, B.C.J. (2002) A model of loudness applicable to time-varying sounds. *J. Audio Eng. Soc.*, **50**(5), 331–342.

Goupell, M.J. and Hartmann, W.M. (2007) Interaural fluctuations and the detection of interaural incoherence. III. Narrowband experiments and binaural models. *J. Acoust. Soc. Am.*, **122**, 1029–1045.

Hermansky, H. (1990) Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.*, **87**(4), 1738–1752.

Irino, T. and Patterson, R. (1997) A time-domain, level-dependent auditory filter: The gammachirp. *J. Audio Eng. Soc. Am.*, **101**(1), 412–419.

Jeffress, L.A. (1948) A place theory of sound localization. *J. Comp. Physiol. Psychol.*, **41**(1), 35–39.

Jin, C., Schenkel, M., and Carlile, S. (2000) Neural system identification model of human sound localization. *J. Acoust. Soc. Am.*, **108**(3), 1215–1235.

Karjalainen, M. (1996) A binaural auditory model for sound quality measurements and spatial hearing studies. *Proc. of IEEE ICASSP'96*, pp. 985–988.

Licklider, J. (1951) A duplex theory of pitch perception. *Experientia*, **7**, 128–133.

Licklider, J.C.R. (1959) *Three Auditory Theories*. McGraw-Hill, pp. 41–144.

Lopez-Poveda, E., Fay, R.R., and Popper, A.N. (2010) *Computational Models of the Auditory System*, volume 35. Springer.

Lyon, R.F. (1982) A computational model of filtering, detection and compression in the cochlea. *Proc. of IEEE ICASSP'82*, pp. 1282–1285.

Majdak, P. and Søndergaard, P. (2013) The auditory modeling toolbox `http://amtoolbox.sourceforge.net`.

Meddis, R. (1988) Simulation of auditory neural transduction: Further studies. *J. Acoust. Soc. Am.*, **83**, 1056–1063.

Meddis, R. and Hewitt, M.J. (1991) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J. Acoust. Soc. Am.*, **89**(6), 2866–2882.

Meddis, R. and Hewitt, M.J. (1992) Modelling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, **91**(1), 233–245.

Meddis, R. and O'Mard, L. (1997) A unitary model of pitch perception. *J. Acoust. Soc. Am.*, **102**(3), 1811–1820.

Meddis, R., O'Mard, L.P., and Lopez-Poveda, E.A. (2001) A computational algorithm for computing nonlinear auditory frequency selectivity. *J. Acoust. Soc. Am.*, **109**(6), 2852–2861.

Patterson, R.D. (1994) The sound of a sinusoid: Spectral models. *J. Acoust. Soc. Am.*, **96**(3), 1409–1418.

Patterson, R.D., Unoki, M., and Irino, T. (2003) Extending the domain of center frequencies for the compressive gammachirp auditory filter. *J. Acoust. Soc. Am.*, **114**(3), 1529–1542.

Plack, C.J., Oxenham, A.J., Fay, R.R., and Popper, A.N. (2005) *Pitch: Neural Coding and Perception*, volume 24. Springer.

Pulkki, V. and Hirvonen, T. (2009) Functional count-comparison model for binaural decoding. *Acta Acustica United with Acustica*, **95**, 883–900.

Sayers, B.M. and Cherry, E.C. (1957) Mechanism of binaural fusion in the hearing of speech. *J. Acoust. Soc. Am.*, **29**(9), 973–987.

Seneff, S. (1990) A joint synchrony/mean-rate model of auditory speech processing. In Waibel A. and Lee K-F, Readings in Speech recognition. Morgan-Kaufmann. pp. 101–113.

Shackleton, T.M., Meddis, R., and Hewitt, M.J. (1992) Across frequency integration in a model of lateralization. *J. Acoust. Soc. Am.*, **91**(4), 2276–2279.

Shera, C.A. (2001) Intensity-invariance of fine time structure in basilar-membrane click responses: Implications for cochlear mechanics. *J. Acoust. Soc. Am.*, **110**(1), 332–348.

Slaney, M. (1998) Auditory toolbox `https://engineering.purdue.edu/~malcolm/interval/1998-010/`.

Stecker, G.C., Harrington, I.A., and Middlebrooks, J.C. (2005) Location coding by opponent neural populations in the auditory cortex. *PLoS Biol*, **3**(3), 520–528.

Stern, R.M. and Trahiotis, C. (1995) Models of binaural interaction. *Handbook of Perception and Cognition*, **6**, 347–386.

Strube, H.W. (1985) A computationally efficient basilar-membrane model. *Acustica*, **58**(4), 207–214.

Takanen, M., Santala, O., and Pulkki, V. (2014) Visualization of functional count-comparison-based binaural auditory model output. *Hear. Res.*, **309**, 147–163.

van Bergeijk, W.A. (1962) Variation on a theme of Békésy: A Model of Binaural Interaction. *J. Acoust. Soc. Am.*, **34**(8), 1431–1437.

von Békésy, G. (1930) Zur theorie des Hörens. Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkeungleighheit der beiderseitigen Schalleinwirkungen. *Physik. Zeitschr.* pp. 824–835, 857–868.

von Békésy, G. (1960) *Experiments in Hearing*. McGraw-Hill and Acoustical Society of America.

Weiss, T.F. (1966) A model of the peripheral auditory system. *Kybernetik*, **3**(4), 153–175.

Zhang, X., Heinz, M.G., Bruce, I.C., and 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**(2), 648–670.

Zilany, M.S. and Bruce, I.C. (2006) Modelling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *J. Acoust. Soc. Am.*, **120**(3), 1446–1466.

Zweig, G. (1991) Finding the impedance of the organ of Corti. *J. Acoust. Soc. Am.*, **89**(3), 1229–1254.

Zwicker, E. (1977) Procedure for calculating loudness of temporally variable sounds. *J. Acoust. Soc. Am.*, **62**(3), 675–682.