1

From Physics to Psychophysics

This first chapter does not present new material, but rather presents some backgrounds for the various BWE topics. The selection of the material to be included in this chapter is mainly motivated by what is considered to be a useful reference if unfamiliar concepts are encountered in later chapters. To keep this backgrounds section concise, it will not always contain all desired information, but references are provided for further reading (as is done throughout the remainder of the book).

The topics covered in this chapter are basics in signal processing in Sec. 1.1, statistics of speech and music in Sec. 1.2, acoustics (mainly concerning loudspeakers) in Sec. 1.3, and auditory perception in Sec. 1.4.

1.1 SIGNAL THEORY

This section reviews some preliminaries for digital signal processing, most notably the concepts of linearity versus non-linearity, and commonly used digital filter structures and some of their properties. An understanding of these concepts is essential to appreciate the algorithms presented in later chapters, but knowledge of more advanced signal processing concepts would be very useful. If necessary, reviews of digital signal processing theory and applications can be found in, for example, Rabiner and Schafer [217], Rabiner and Gold [218], or Oppenheim and Schafer [194]. Van den Enden and Verhoeckx [281] also present a good introduction into digital signal processing. Golub and Van Loan [93] is a good reference for matrix computations, which are extensively used in Chapter 6.

1.1.1 LINEAR AND NON-LINEAR SYSTEMS

Consider a system that transforms an input signal $x(t)$ into an output signal $y(t)$. Assume that this transformation can be described by a function g such that

$$y(t) = g(x(t)). \tag{1.1}$$

Time invariance means that we must have

$$y(t - \tau) = g(x(t - \tau)), \quad \tau \in \mathbb{R}, \tag{1.2}$$

Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design. E. Larsen and R. M. Aarts

that is, the output of a time-shifted input is the time-shifted output. The system is linear iff

$$\begin{aligned} y_1(t) &= g(x_1(t)), \\ y_2(t) &= g(x_2(t)), \\ y_1(t) + y_2(t) &= g(x_1(t) + x_2(t)), \quad (1.3) \\ ay_1(t) &= g(ax_1(t)), \quad a \in \mathbb{R}. \quad (1.4) \end{aligned}$$

Equation 1.3 is called the superposition property, that is, the output of the sum of two signals equals the sum of the outputs. Equation 1.4 is called the homogeneity property, that is, the output of a scaled input equals the scaled output. If any of these two properties (or both) do not hold, the system is non-linear. Vaidyanathan [278] introduces the terminology *homogeneous time-invariant* for a time-invariant system, where Eqn. 1.3 is not true, while Eqn. 1.4 is true; such systems are an important class for BWE algorithms, as we shall see in Chapter 2 and following chapters. Note that these comments and equations are valid for both continuous as well as discrete-time (sampled) systems.

Many mathematical techniques exist for analyzing properties of linear time-invariant (LTI) systems, and some basic ones will be discussed shortly. For non-linear or time-variant systems, such analysis often becomes very complicated or even impossible, which is why one traditionally avoids dealing with such systems (or makes linear approximations of non-linear systems). Nonetheless, non-linear systems can have useful properties that LTI systems do not have. For BWE purposes, an important example is that LTI systems cannot introduce new frequency components into a signal; only the amplitude and/or phase of existing components can be altered. So, if the frequency bandwidth of a signal needs to be extended, the use of a non-linear system is inevitable (and thus desirable). Note that non-linearities in audio applications are often considered as generating undesirable distortion, but controlled use of non-linearities, such as in BWE algorithms, can be beneficial.

1.1.2 CONTINUOUS-TIME LTI (LTC) SYSTEMS

A continuous-time LTI system will be abbreviated as LTC. Beside the input–output function g (Eqn. 1.1), an LTC system can be fully described by its impulse response $h(t)$, which is the output to a delta function $\delta(t)$ input[1] ($h(t) = g(\delta(t))$). Because the system is linear, an arbitrary input signal $x(t)$ can be written as an infinite series of delta functions by

$$x(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)\,\mathrm{d}\tau. \quad (1.5)$$

Using this principle, we can calculate the output signal $y(t)$ as

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)\,\mathrm{d}\tau, \quad (1.6)$$

[1] A delta function is the mathematical concept of a function that is zero everywhere except at $x = 0$, and which has an area of 1, that is, $\int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = 1$.

which is called the convolution integral, compactly written as

$$y(t) = x(t) * h(t). \tag{1.7}$$

An LTC is called stable iff

$$\int_{-\infty}^{\infty} |y(t)| \, \mathrm{d}t < \infty, \tag{1.8}$$

and causal iff

$$h(t) = 0 \quad \text{for } t < 0. \tag{1.9}$$

The Fourier transform of the impulse response $h(t)$ is called the frequency response $H(\omega)$ – like $h(t)$, $H(\omega)$ gives a full description of the LTC system. Here, ω is the *angular* frequency, related to frequency f as $\omega = 2\pi f$. It is convenient to calculate the frequency response (also called frequency spectrum, or simply spectrum) of the output of the system as

$$Y(\omega) = X(\omega)H(\omega). \tag{1.10}$$

We see that convolution in the time domain equals multiplication in the frequency domain. The reverse is also true, and therefore the Fourier transform has the property, which is sometimes called "convolution in one domain is equal to multiplication in the other domain". A slightly more general representation is through the Laplace transform, yielding $X(s)$ as

$$X(s) = \int_{-\infty}^{\infty} x(t) \mathrm{e}^{-st} \, \mathrm{d}t, \tag{1.11}$$

where s is the Laplace variable. For continuous-time, physical frequencies ω lie on the y-axis, thus we can write $s = i\omega$ (note that in practice the i is often dropped when writing $H(i\omega)$). We can also write the LTC system function as

$$H(s) = c \frac{\prod_{i=1}^{N} (s - z_i)}{\prod_{j=1}^{M} (s - p_j)}, \tag{1.12}$$

where c is a constant. The z_i are the N zeros of the system, and p_j are the M poles, either of which can be real or complex. The system will be stable if all poles lie in the left hemifield, that is, $\Re\{p_j\} < 0$.

As an example, a simple AC coupling filter may be considered: a capacitance of C F connecting input and output, and a resistance of R Ω from output to the common terminal. This system transfer function can be written as

$$H(s) = \frac{RCs}{1 + RCs}, \tag{1.13}$$

where the frequency response follows by substituting $s = i\omega$. Comparing Eqn. 1.12 with Eqn. 1.13 shows that $c = RC$, $N = 1$, $z_1 = 0$, $M = 1$, and $p_1 = -\frac{1}{RC}$.

The *magnitude* and *phase* of an LTC system function $H(s)$ are defined as

$$|H(s)| = \left[\Re\{H(s)\}^2 + \Im\{H(s)\}^2\right]^{1/2}, \tag{1.14}$$

$$\angle H(s) = \tan^{-1}\frac{\Im\{H(s)\}}{\Re\{H(s)\}}, \tag{1.15}$$

respectively. From the phase response, we can define the group delay $\tau_{\mathrm{d}}(\omega)$, which can be interpreted as the time delay of a frequency ω between input and output and is given by

$$\tau_{\mathrm{d}}(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\omega}\angle H(i\omega). \tag{1.16}$$

1.1.3 DISCRETE-TIME LTI (LTD) SYSTEMS

In DSP systems, the signals are known only at certain time instants kT_{s}, where $k \in \mathbb{Z}$ and $T_{\mathrm{s}} = 1/f_{\mathrm{s}}$ is the sampling interval, with f_{s} being the sample frequency or sample rate. Thus, these systems are known as discrete-time LTI systems, abbreviated as LTD. According to the *sampling theorem*[2], we can perfectly reconstruct a continuous-time signal $x(t)$ from its sampled version $x(k)$ if f_{s} is at least twice the highest frequency occurring in $x(t)$. To convert $x(k)$ (consisting of appropriately scaled delta function at the sample times) to $x(t)$ (the continuous signal), a filter that has the following impulse response needs to be applied

$$h(t) = \frac{\sin(\pi t/T_{\mathrm{s}})}{\pi t/T_{\mathrm{s}}}. \tag{1.17}$$

In the frequency domain, this corresponds to an ideal low-pass filter ('brick-wall' filter), which only passes frequencies $|f| < f_{\mathrm{s}}/2$. System functions in the discrete-time domain are usually described in the z domain (z being a complex number) as $H(z)$, rather than the s domain. Likewise, the corresponding input and output signals are denoted as $X(z)$ and $Y(z)$ (Jury [138]). $X(z)$ can be obtained from $x(k)$ through the Z-transform

$$X(z) = \sum_{k=-\infty}^{\infty} x(k)z^{-k}. \tag{1.18}$$

Normalized physical frequencies Ω in discrete time lie on the unit circle, thus $z = \mathrm{e}^{-i\Omega}$; therefore $|\Omega| \leq \pi$.

There are various ways to convert a known continuous-time system function $H(s)$ to its discrete-time counterpart $H(z)$, all of which have various advantages and disadvantages. The most widely used method is the bilinear transformation, which relates s and z as

$$s = \frac{2}{T_{\mathrm{s}}}\frac{1-z^{-1}}{1+z^{-1}}. \tag{1.19}$$

[2] The sampling theorem is frequently contributed to Shannon's work in the 1940s, but, at the same time, Kotelnikov worked out similar ideas, and, a few decades before that, Whitaker. Therefore, some texts use the term WKS-sampling theorem (Jerri [135]).

Applying this to the example of Eqn. 1.13, we find

$$H(z) = \frac{k(1 - z^{-1})}{1 + k + (1 - k)z^{-1}}, \tag{1.20}$$

with $k = 2RC/T_\text{s}$, a dimensionless quantity. Like Eqn. 1.12, we can write $H(z)$ in a similar manner, with a zero at $z_1 = 1$ and a pole at $p_1 = \frac{k-1}{k+1}$. Substituting $z = \mathrm{e}^{-i\Omega}$ in Eqn. 1.20 gives the frequency response $H(\mathrm{e}^{-i\Omega})$. Stability for LTD systems requires that all poles lie within the unit circle, that is, $|p_j| < 1$. Magnitude, phase, and group delay of an LTD system are defined analogously as for LTC systems (Eqns. 1.14–1.15).

1.1.4 OTHER PROPERTIES OF LTI SYSTEMS

There are a few other properties of LTI systems that are of interest. We will discuss these using LTD systems, but analogous equations hold for LTC systems.

We have already found that stability requires that all poles must lie within the unit circle, that is, $|p_j| < 1$. Where the locations of zeros z_j are concerned, the system is called minimum phase if all $|z_j| < 1$. This is of interest because a minimum-phase system has a stable inverse. To see this, note that

$$H^{-1}(z) = \left[c \frac{\prod_{i=1}^{N}(z - z_i)}{\prod_{j=1}^{M}(z - p_j)} \right]^{-1} = c^{-1} \frac{\prod_{j=1}^{M}(z - p_j)}{\prod_{i=1}^{N}(z - z_i)}, \tag{1.21}$$

that is, the poles become zeros, and vice versa. We can conclude that a stable minimum-phase system has an inverse, which is also stable and minimum phase. A non-minimum-phase system cannot be inverted[3]. A 'partial inversion' is possible by splitting the system function into a minimum-phase and a non-minimum-phase part and then inverting the minimum-phase part (Neely and Allen [184]).

A linear-phase system $H(z)$ is one for which $\angle H(z) = -a\omega$ $(a > 0)$, that is, the phase is a linear function of frequency. It implies that the group delay is a constant $\tau_\text{d}(\omega) = a$, which is often considered to be beneficial for audio applications. For $H(z)$ to be linear phase, the impulse response $h(k)$ must be either symmetric or anti-symmetric.

1.1.5 DIGITAL FILTERS

The simplest digital filter is $H(z) = z^{-1}$, being a delay of one sample. If one cascades $N + 1$ such delays and sums all the scaled delayed signals, one gets a filter as depicted in Fig. 1.1. Such a filter is called a finite impulse response (FIR) filter, since its impulse response is zero after N time samples. Its system function is written as

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{i=0}^{N} b_i z^{-1}, \tag{1.22}$$

[3] A practical example of a non-minimum-phase system is a room impulse response between a sound source and a receiver (unless they are very close together).

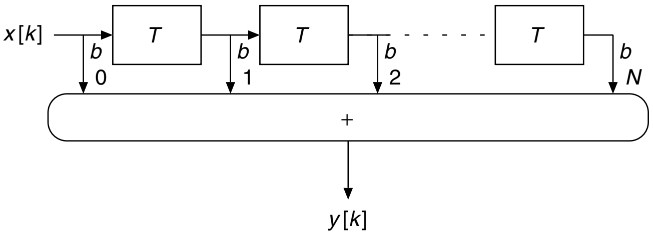


Figure 1.1 A finite impulse response (FIR) filter. The input signal $x(k)$ is filtered, yielding the output signal $y(k)$. The boxes labeled 'T' are one sample (unit) delay elements. The signal at each tap is multiplied with the corresponding coefficients b

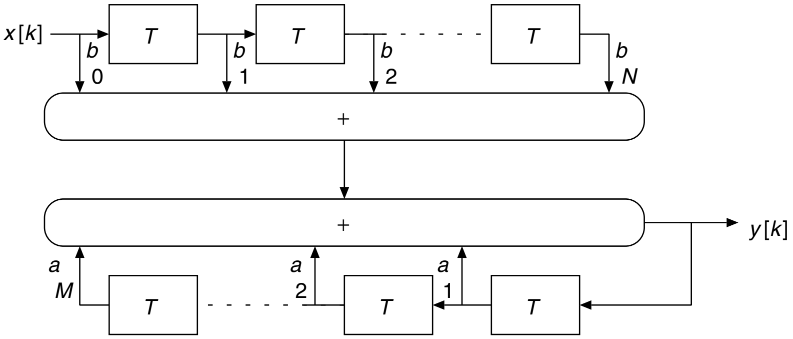


Figure 1.2 An infinite impulse response (IIR) filter in direct form-I structure. The input signal $x(k)$ is filtered by the filter, yielding the output signal $y(k)$. The boxes labeled 'T' are one sample (unit) delay elements. The signal at each forward tap is multiplied with the corresponding coefficients b, while the signal in the recursive (feedback) part is multiplied with the corresponding coefficients a, respectively

and the frequency response can be found by substituting $z = \mathrm{e}^{-i\Omega}$. It is obvious that an FIR filter has only zeros and no poles. Therefore, it is guaranteed to be stable. Note that an FIR filter need not be minimum phase, and therefore a stable inverse is not guaranteed to exist.

By applying feedback to a filter, as shown in Fig. 1.2, a filter with an infinite impulse response (IIR) can be made. Its system function is written as

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{i=0}^{N} b_i z^{-1}}{1 - \sum_{j=1}^{M} a_i z^{-1}}. \tag{1.23}$$

This shows that, in general, an IIR filter has both zeros and poles; therefore, an IIR filter would be unstable if any of the poles lie outside the unit circle. An IIR filter can be both minimum- and non-minimum phase. While the structure of Fig. 1.2 can implement any IIR filter, it is, for practical reasons (like finite word-length calculations), more customary to partition the filter by cascading 'second-order sections' (IIR filters with two delays in both forward and feedback paths), also known as biquads.

FIR and IIR filters have a number of differences, as a result of which each of them has specific applications areas, although overlap exists, of course. The most important differences (for BWE purposes) are:

- *Phase characteristic*: Only FIR filters can be designed to have linear phase; IIR filters can only approximate linear phase at a greatly increased complexity. In Sec. 2.3.3.3, it is discussed why a linear-phase characteristic is beneficial for a particular BWE application.
- *Computational complexity*: For both FIR and IIR filters, the computational complexity is proportional to the number of coefficients used. However, for an FIR filter, the number of coefficients directly represents the length of the impulse response, or, equivalently, the frequency selectivity. An IIR filter has the advantage that very good frequency selectivity is possible using only a small number of coefficients. Therefore, IIR filters are generally much more efficient than FIR filters. For low-frequency BWE applications, filters with narrow passbands are often required, centered on very low frequencies; in such cases, IIR filters are orders of magnitude more efficient than FIR filters.

It will be apparent that the choice of FIR or IIR depends on what features are important for a particular application. Sometimes, as in some BWE applications, both linear phase *and* high frequency selectivity are very desirable. In such cases, IIR filters can be used, but in a special way and at a somewhat increased computational complexity and memory requirement (see e.g. Powell and Chau [213]).

1.2 STATISTICS OF AUDIO SIGNALS

For BWE methods, it is important to know the spectrum of the audio signal. Because these signals are generally not stationary, the spectrum varies from moment to moment, and the spectrogram is useful to visualize this spectro-temporal behaviour of speech and music. First we will consider speech, and then music. For both cases, we will make it plausible that certain bandwidth limitations can be overcome by suitable processing. Some of the material in this chapter is taken from Aarts *et al.* [10].

1.2.1 SPEECH

Speech communication is one of the basic and most essential capabilities of human beings. The speech wave itself conveys linguistic information, the speaker's tone, and the speaker's emotion. Information exchange by speech clearly plays a very significant role in our lives. Therefore, it is important to keep speech communication as transparent as possible, both to be intelligible as well as to be natural. Unfortunately, owing to bandwidth limitation, both aspects can suffer. But, because a lot is known about the statistics of speech, specialized BWE algorithms can be used to restore, to a large extent, missing frequency components, if the available bandwidth is sufficiently large. This is explored in detail in Chapter 6.

Speech can be voiced or unvoiced (tonal or noise-like), see, for example, Furui [80], Olive *et al.* [191], Rabiner and Schafer [217], and Fig. 6.4. A voiced sound can be modelled as a pulse source, which is repeated at every fundamental period $1/f_{\mathrm{p}}$ (where f_{p}

is the pitch), while an unvoiced sound can be modelled as a white noise generator. The loudness of speech is proportional to the peak amplitudes of the waveform. Articulation can be modelled as a cascade of filters – these filters simulate resonant effects R_i (formants) in the vocal tract, which extends from the vocal cords to the lips, including the nasal cavity. Consequently, any harmonic of the series of pulses with frequency kf_p that happens to lie close to one of the R_i is enhanced. To make the various vowel sounds, a speaker or singer must change these vocal tract resonances by altering the configuration of tongue, jaw, and lips. The distinction between different vowel sounds in Western languages is determined almost entirely by R_1 and R_2, the two lowest resonances, that is, vowels are created by the first few broad peaks on the spectral envelope imposed on the overtone spectrum, by vocal tract resonances. For the vowel sound in 'hood', pronounced by a male speaker, $R_1 \approx 400\,\mathrm{Hz}$ and $R_2 \approx 1000\,\mathrm{Hz}$. In contrast, to produce the vowel in 'had', R_1 and R_2 must be raised to about 600 and 1400 Hz respectively, by opening the mouth wider and pulling the tongue back. For women, the characteristic resonance frequencies are roughly 10% higher. But for both sexes, the pitch frequency f_p in speech and singing is generally well below R_1 for any ordinary vowel sound – except when sopranos are singing very high notes, in which case they raise R_1 towards f_p (Goss Levi [160], Joliveau *et al.* [137]). Finally, radiation of speech sounds can be modelled as arising from a piston sound source attached to an infinite baffle, like a loudspeaker model discussed in Sec. 1.3.2. The range of frequencies for speech is roughly between 100 and 8 kHz (whereas the ordinary telephone channel is limited between 300 and 3400 Hz).

An important parameter for speech is the fundamental frequency or pitch. Furui [80] presents a statistical analysis of temporal variations in the pitch of conversational speech for individual talkers, which indicates that the mean and standard deviation for a female voice are roughly twice those of a male voice. This is shown in Fig. 1.3. The pitch distributed over talkers on a logarithmic frequency scale (not the linear scale of Fig. 1.3) can be approximated by two normal distributions that correspond to the male and female voice, respectively. The mean and standard deviation for a male voice are 125 and 20.5 Hz, respectively, whereas those for a female voice are twice as large. Conversational speech includes discourse as well as pause, and the proportion of actual speech periods relative to the total period is called 'speech ratio'. In conversational speech, the speech ratio for individual talkers is about 1/3 (Furui [80]), and can be used as a feature for speech detection in a speech–music discriminator (Aarts and Toonen Dekkers [6]).

In order to gain insight in the long-term average of speech spectra, six speech fragments of utterances of various speakers of both sexes were measured. Figure 1.4 shows the power spectra of tracks 49–54 of the SQAM disk [255]. To parameterize the spectra in the plot, we derived the following heuristic formula

$$|H(f)| \approx \frac{\left(\frac{f}{f_\mathrm{p}}\right)^6}{1+\left(\frac{f}{f_\mathrm{p}}\right)^6}\,\frac{1}{1+\frac{f}{1000}}\,, \tag{1.24}$$

were f is the frequency and f_p the pitch of the voice. Byrne *et al.* [42] have shown that there is not much difference in the long-term speech spectra of different languages. The first factor in the product of Eqn. 1.24 denotes the high-pass behavior, and the second

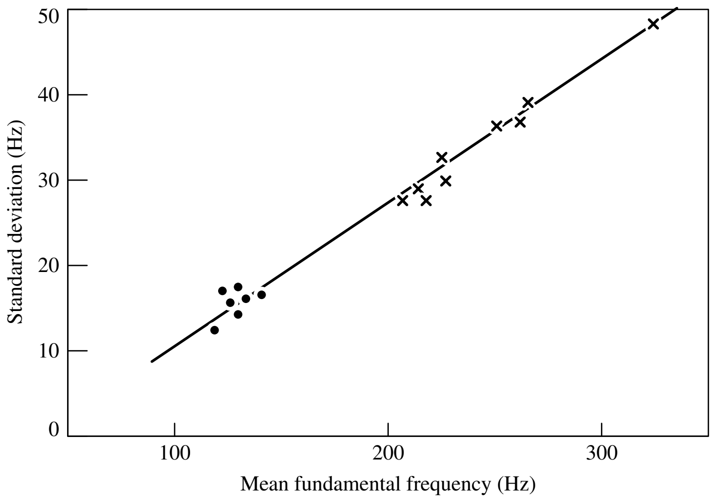


Figure 1.3 Male (•)/female pitch (×), adapted from Furui [80]. The mean and standard deviation for male voice pitch is 125 ± 20.5 Hz and are twice these values for female voice pitch

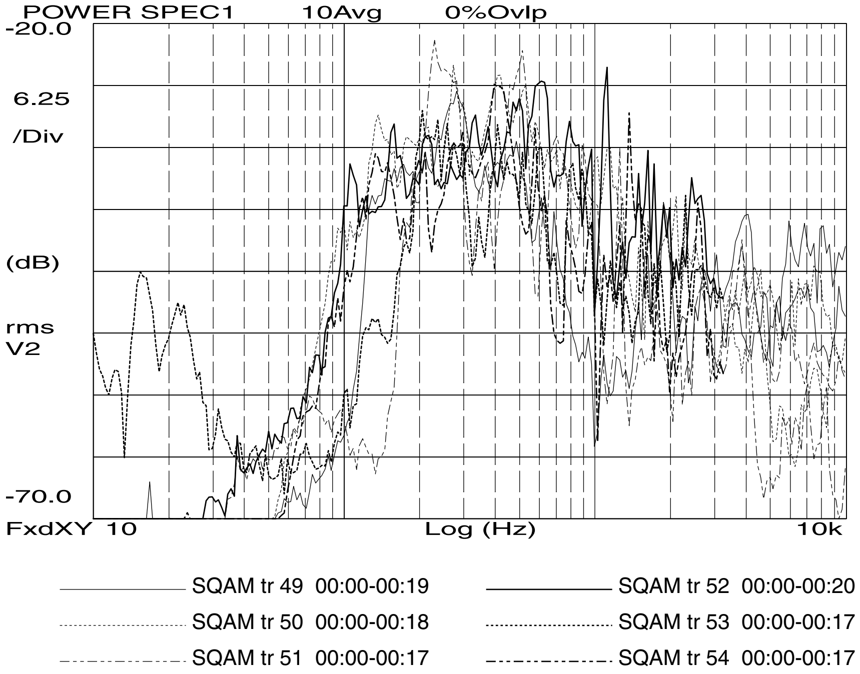


Figure 1.4 Long-term spectrum of speech, for a total of six talkers, both sexes (SQAM disc [255], tracks 49–54)

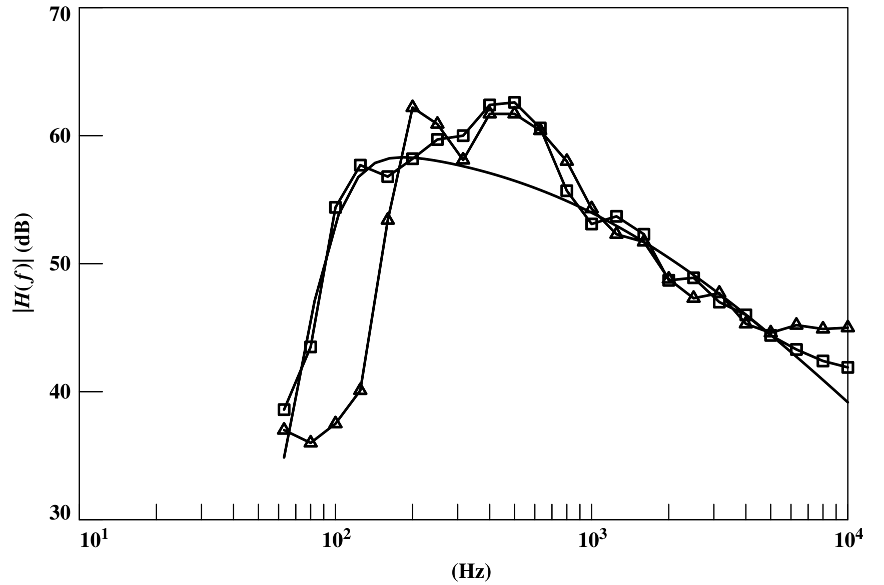


Figure 1.5 Equation 1.24 is plotted (curve without marker) for a male voice (f_P = 120 Hz, offset by 60 dB), together with the data from Byrne *et al.* [42] for males (squares) and females (triangles). Note that the parameterization of Eqn. 1.24 fits the empirical data for the male voice quite well (the female voice would have to be modelled with a higher pitch value)

one the low-pass behaviour. It shows that there is a steep slope at low frequencies, and a rather modest slope at high frequencies, as is depicted in Fig. 1.5. The lowest pitch of the male voice is about 120 Hz (see Fig. 1.3), which corresponds very well with the high-pass frequency in Eqn. 1.24, and Fig. 1.4. The telephone range is 300–3400 Hz, which clearly is not sufficient to pass the high-frequency range, nor the lowest male fundamental frequencies (and also not most female fundamental frequencies). A BWE algorithm that can recreate the fundamental (and possibly the lower harmonics) and the high-frequency band of speech signals should be able to create a more natural sounding speech for telephony.

Because the speech spectrum changes over time, it is instructive to compute spectra at frequent intervals and display the changing spectra. A spectrogram is shown for a particular speech sample in Fig. 1.6 (b); the pitch of the voice is time-varying, as can be seen in (c). A BWE algorithm must be able to follow these pitch changes and change its output frequencies accordingly. Fig. 1.7 shows a waveform and spectrogram (a and b) of the same speech utterance, and an 8-kHz low-pass filtered version thereof, which could occur in perceptual audio coders at very high compression rates; the telephone channel would low-pass filter the signal even more severely at 3.4 kHz. A high-frequency BWE algorithm can resynthesize an octave of high-frequency components, as shown in Fig. 1.7(c); note the similarities and differences with respect to the original spectrogram. High-frequency BWE algorithms are discussed in Chapters 5 (audio) and 6 (speech).

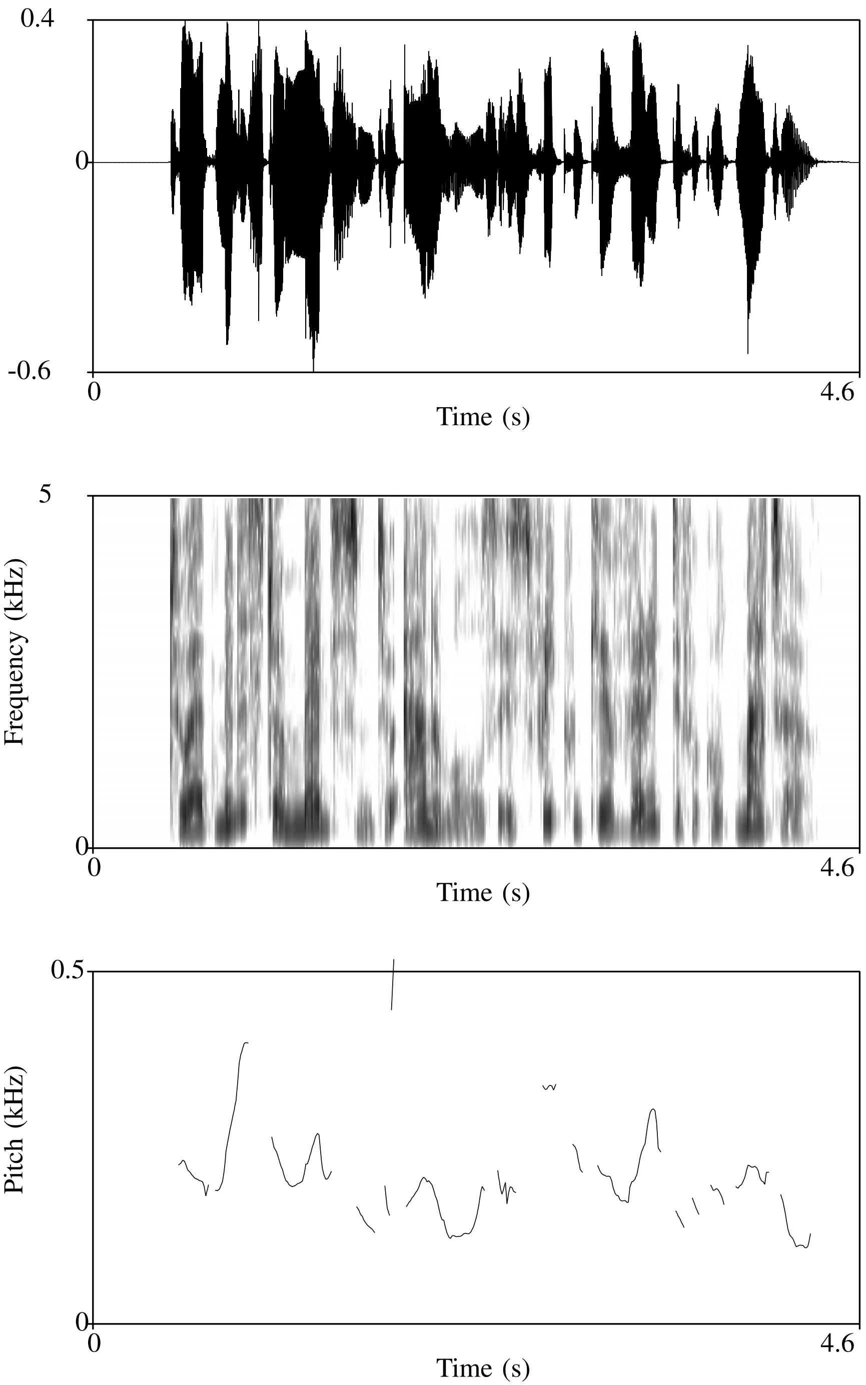
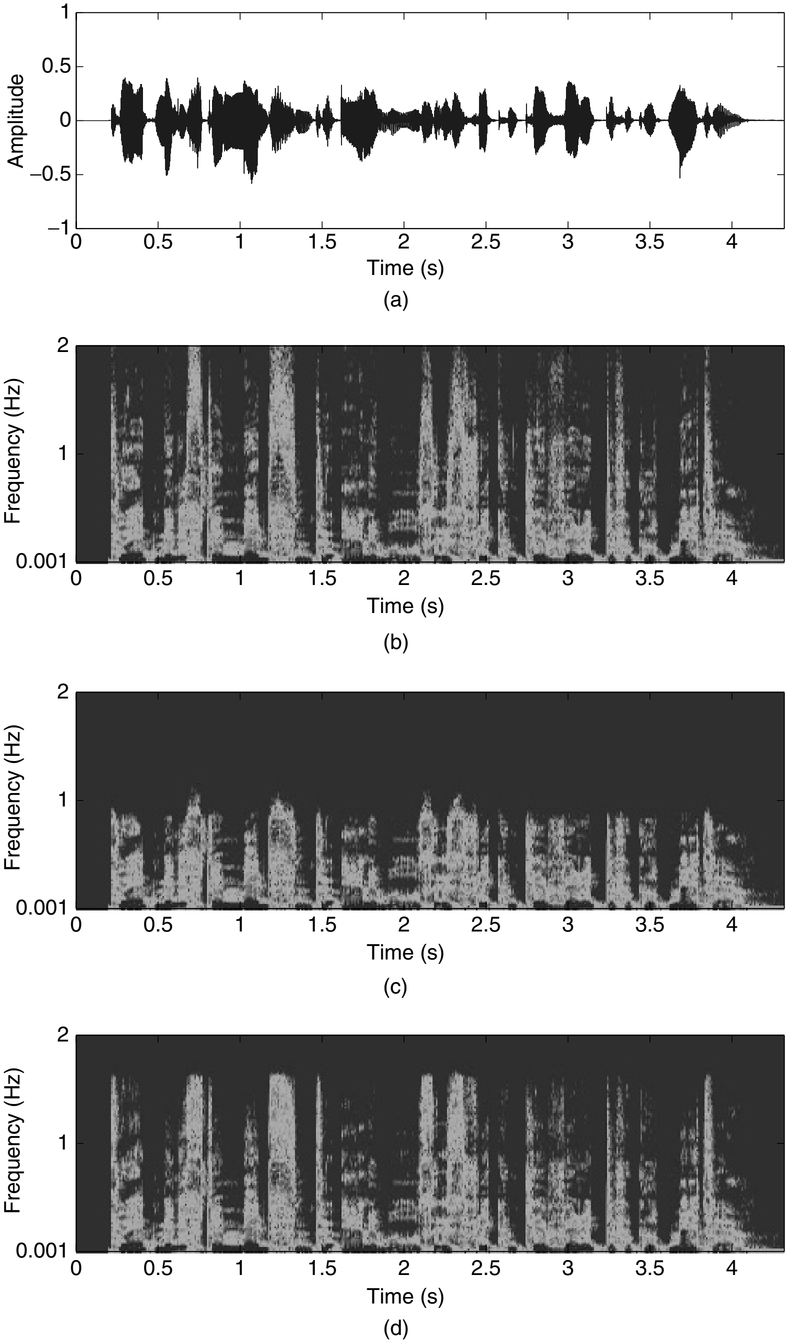


Figure 1.6 For a given speech waveform (a), the spectrogram is displayed in (b) (white−black indicating low−high energy, dB scale). (c) shows the pitch of the voice as determined by a pitch tracker (‘Praat’, Boersma and Weenink [35])



(a)

(b)

(c)

(d)

Figure 1.7 High-frequency BWE for a female voice (waveform in a, spectrogram in b) that has been low-pass filtered at 8 kHz (spectrogram in c). The processed signal (extended to 16 kHz) is displayed as a spectrogram in d

1.2.2 MUSIC

More than 70 years ago, Sivian *et al.* [248] performed a pioneering study of musical spectra using live musicians and – for that time – innovative electronic measurement equipment. Shortly after the introduction of the CD, this study was repeated by Greiner and Eggers [99] by using modern digital equipment and modern source material, at that time CDs. The result of both studies was a series of graphs showing for each instrument or ensemble the spectral amplitude distribution of the performed musical passage. The findings were that, in general, musical spectra have a bandpass characteristic, the exact shape of which is determined by the music and the instrument. As in speech, the fundamental frequency (pitch) is time varying. A complicating factor is that various instruments may be playing together, creating a superposition of several complex tones.

An example is shown in Fig. 1.8, where the variable time–frequency characteristic of a 10-s excerpt of music is shown ('One', by Metallica). The waveform is shown in (a), and (b) shows a spectrogram (frequencies 0–140 Hz) of the original signal. The energy extends down to about 40 Hz. By using a low-frequency physical BWE algorithm, we can extend this lower limit to about 20 Hz (c), which requires a subwoofer of excellent quality for correct reproduction. Because the resulting synthetic frequencies have similar spectro-temporal characteristics as the original low frequencies, they will be perceived as an integral part of the signal (lowering the pitch of the bass tones to 20 Hz). Because of the very low frequencies that are now being radiated, it will also add 'feeling' to the music. Low-frequency physical BWE algorithms are discussed in Chapter 3.

Another study (Fielder and Benjamin [70]) was conducted to establish design criteria for the performance of subwoofers to be used for the reproduction of music in homes. The focus on subwoofers was motivated by the fact that low frequencies play an important role in the musical experience. A first conclusion of that study was that recordings with audible bass below 30 Hz are relatively rare. Second, these very low frequencies were generated by pipe organs, synthesizers, or special effects and environmental noise. Other instruments, such as bass guitar, bass viol, tympani, or bass drum, produce relatively little output below 40 Hz, although they may have very high levels at or above that frequency. Fielder and Benjamin [70] gave an example that for an average listening room of 68 m^3, the required acoustic power for reproduction is 0.0316 W (which yields a sound pressure level of 97 dB), which requires a volume displacement of 0.685 l at 20 Hz. This requires an excursion of 13.5 mm for a 10 in. (0.25 m) woofer. These are extraordinary requirements, and very hard to fulfil in practice. An alternative is to use low-frequency psychoacoustic BWE methods, where frequencies that are too low to reproduce are shifted to higher frequencies, in such a way that the pitch percept remains the same. These methods are discussed in Chapter 2. If we consider Fig. 1.8 (c) as the original signal, we could think of such BWE as shifting the frequency band 20–40 Hz to above 40 Hz. The spectrogram of the resulting signal would resemble that of Fig. 1.8(b).

1.3 LOUDSPEAKERS

1.3.1 INTRODUCTION TO ACOUSTICS

BWE methods are closely related to acoustics, particularly acoustics of loudspeakers, so here we will review some basic concepts in this area. Extensive and more general

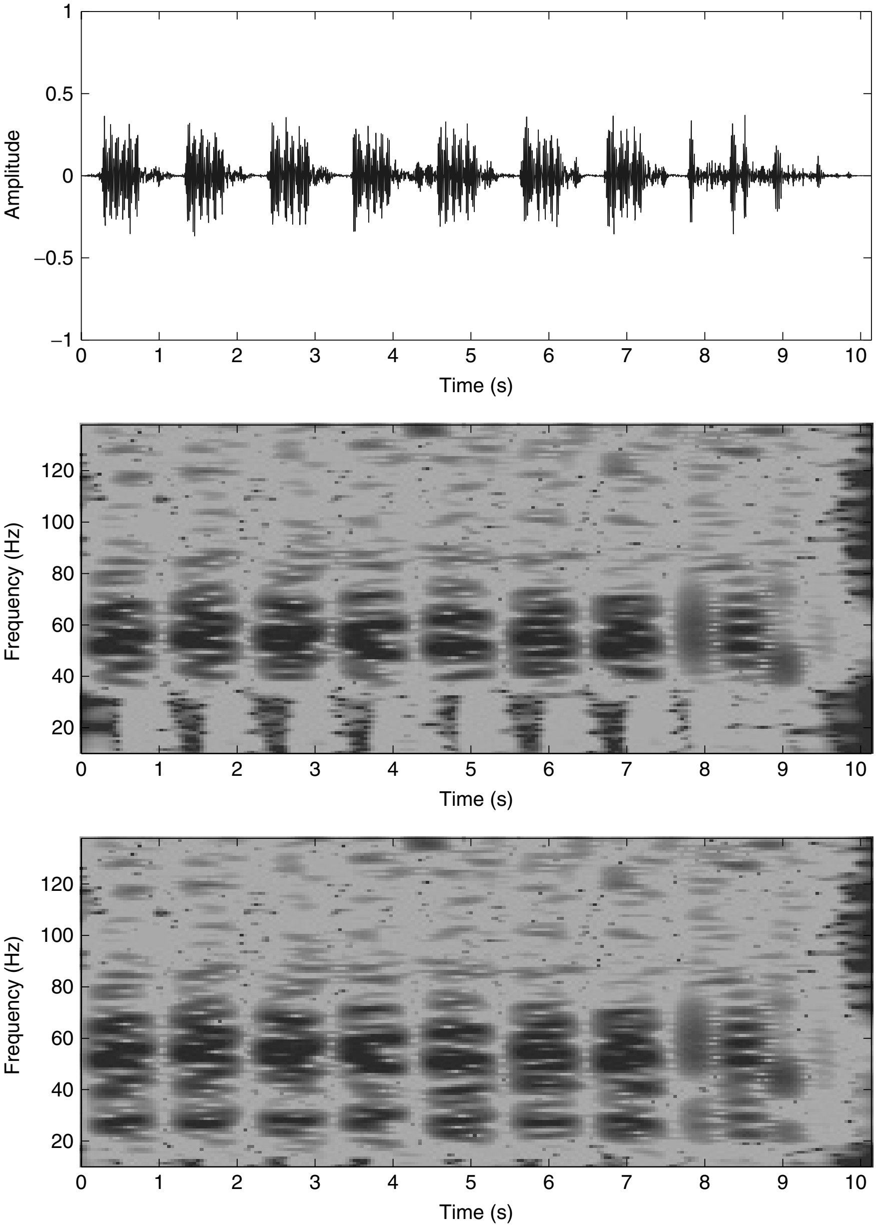


Figure 1.8 A 10-s excerpt of music (a), and its spectrogram for frequencies 0–140 Hz (b), which shows that the lowest frequencies contained in the signal are around 40 Hz. A low-frequency physical BWE algorithm can extend this low-frequency spectrum down to about 20 Hz, as shown in (c). Because the additional frequency components have a proper harmonic relation with the original frequency components, and have a common temporal modulation, they will be perceived as part of the original sound. In this case, the pitch of the bass notes will be lowered to 20 Hz

treatments of acoustics can be found in textbooks such as Kinsler *et al.* [142], Pierce [208], Beranek [28], Morse and Ingard [180]. Acoustics can be defined as the generation, transmission, and reception of energy in the form of vibrational waves in matter. As the atoms or molecules of a fluid or solid are displaced from their normal configurations, an internal elastic restoring force arises. Examples include the tensile force arising when a spring is stretched, the increase in pressure in a compressed fluid, and the transverse restoring force of a stretched wire that is displaced in a direction normal to its length. It is this elastic restoring force, together with the inertia of the system, that enables matter to exhibit oscillatory vibrations, and thereby generate and transmit acoustic waves. Those waves that produce the sensation of sound are of a variety of pressure disturbances that propagate through a compressible fluid.

1.3.1.1 The Wave Equation

The wave equation gives the relation between the spatial ($\mathbf{r}$) and temporal (t) derivates of pressure $p(\mathbf{r}, t)$ as

$$\nabla^2 p(\mathbf{r}, t) = \frac{1}{c^2}\frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} \tag{1.25}$$

where c is the speed of sound, which for air at 293 K is 343 m/s. Equation 1.25 is the linearized, loss-less wave equation for the propagation of sounds in linear inviscid fluids. As a special case of the wave equation, we can consider the one-dimensional case, where the acoustic variables are only a function of one spatial coordinate, say along the x direction. Equation 1.25 then reduces to

$$\frac{\partial^2 p(x, t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p(x, t)}{\partial t^2}. \tag{1.26}$$

The solution of this equation yields two wave fields propagating in $\pm x$ directions, which are called plane (progressive) waves. Sound waves radiated by a loudspeaker are considered to be plane waves in the ‘far field’.

1.3.1.2 Acoustic Impedance

The ratio of acoustic pressure in a medium to the associated particle velocity is called the specific acoustic impedance[4] $z(\mathbf{r})$

$$z(\mathbf{r}) = \frac{p(\mathbf{r})}{u(\mathbf{r})}. \tag{1.27}$$

For plane progressive waves (Eqn. 1.26), this becomes

$$z = \rho_0\, c, \tag{1.28}$$

independent of x, where ρ is the density of the fluid, being 1.21 kg/m^3 for air at 293 K.

[4] Similar to Ohm’s law for electrical circuits.

Table 1.1 Typical sounds and their corresponding SPL values (dB)

Threshold of hearing	0
Whispering	20
Background noise at home	40
Normal talking	60
Noise pollution level	90
Pneumatic drill at 5 m	100
1 m from a loudspeaker at a disco	120
Threshold of pain	140

1.3.1.3 Decibel Scales

Because of the large range of acoustical quantities, it is customary to express values in a logarithmic way. For sound pressure, we define the sound pressure level (SPL) L_{p} in terms of decibel (dB), as

$$L_{\mathrm{p}} = 20 \log(p/p_0), \tag{1.29}$$

where p_0 is a reference level (the log is base 10, as will be used throughout the book), for air $p_0 = 20$ μPa is used. This level is chosen such that it corresponds to the just-noticeable sound pressure level of a 2-kHz sinusoid for an 18-year-old person with normal hearing, see Fig. 1.18 and ISO 226-1987(E) [117]. Table 1.1 lists some typical sounds and their corresponding SPL values. It is convenient to memorize some dB values for the ratio's $\sqrt{2}/2$, 2, 10, and 30 as approximately 3, 6, 20, and 30 dB.

1.3.2 LOUDSPEAKERS

1.3.2.1 Electrodynamic Loudspeakers

Electroacoustic loudspeakers have been around for quite some time. While the first patent for a moving-coil loudspeaker was filed in 1877, by Cuttriss and Redding [55], shortly after Bell's [27] telephone invention, the real impetus to a commercial success was given by Rice and Kellog [223] through their famous paper, so that we can state that the classical electrodynamic loudspeaker as we know now (depicted in Fig. 1.10), is over 80 years old. All practical electroacoustical transducers are limited in their capabilities, owing to their size and excursion possibilities. Among those limitations, there is one in the frequency response, which will be the main topic in the following sections. To study these limitations, we will scrutinize the behaviour of transducers for various parameters. It will appear later that the 'force factor' (Bl) of a loudspeaker plays an important role. To have some qualitative impression regarding the band limitation, various curves are shown in Fig. 1.9. We clearly see that there is a band-pass behaviour of the acoustical power P_{a} (fifth curve in Fig. 1.9), and a high-pass response for the on-axis pressure p (third curve in Fig. 1.9).

First, we will discuss the efficiency of electrodynamic loudspeakers in general, which will be used in a discussion about a special driver with a very low Bl value in Sec. 4.3. This driver can be made very cost efficient, low weight, flat, and with high power

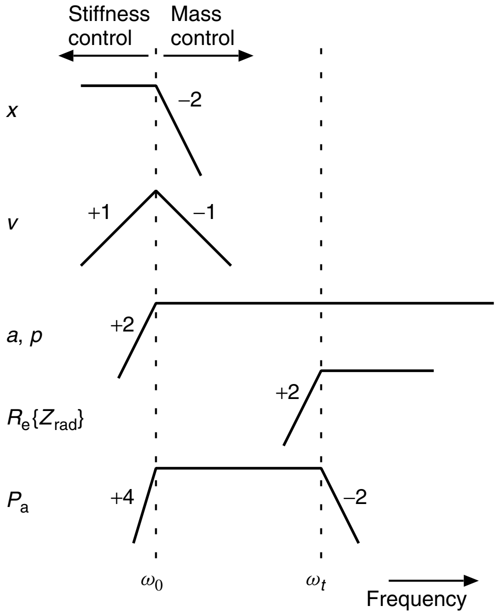


Figure 1.9 The displacement x, velocity v, acceleration a, together with the on-axis pressure p, the real part of the radiation impedance $\Re\{Z_{\text{rad}}\}$, and the acoustical power P_a of a rigid-plane piston in an infinite baffle, driven by a constant force. The numbers denote the slopes of the curves; multiplied by 6, these yield the slope in dB/octave

efficiency. But first (Sec. 1.3.2.3), we show that sound reproduction at low frequencies with small transducers, and at a reasonable efficiency, is very difficult. The reasons for this are that the efficiency is inversely proportional to the moving mass and proportional to the square of the product of cone area and force factor Bl.

1.3.2.2 Construction

An electrodynamic loudspeaker, of the kind depicted in Fig. 1.10, consists of a conical diaphragm, the cone, usually made of paper, being suspended by an outer suspension device, or rim, and an inner suspension device, or spider. The suspension limits the maximum excursion of the cone so that the voice coil remains inside the air gap of the permanent magnet. This limitation can lead to non-linear distortion; see for example, Tannaka *et al.* [264], Olson [193], Klippel [143, 144], Kaizer [140]. The voice coil is attached to the voice coil cylinder, generally made of paper or metal, which is glued to the inner edge of the cone. In most cases, the spider is also attached to this edge. The voice coil is placed in the radial magnetic field of a permanent magnet and is fed with the signal current of the amplifier. For low frequencies, the driver can be modelled in a relatively simple way, as it behaves as a rigid piston. In the next section the electronic behavior of the driver will be described on the basis of a lumped-element model in which the mechanical and acoustical elements can be interpreted in terms of the well-known properties of

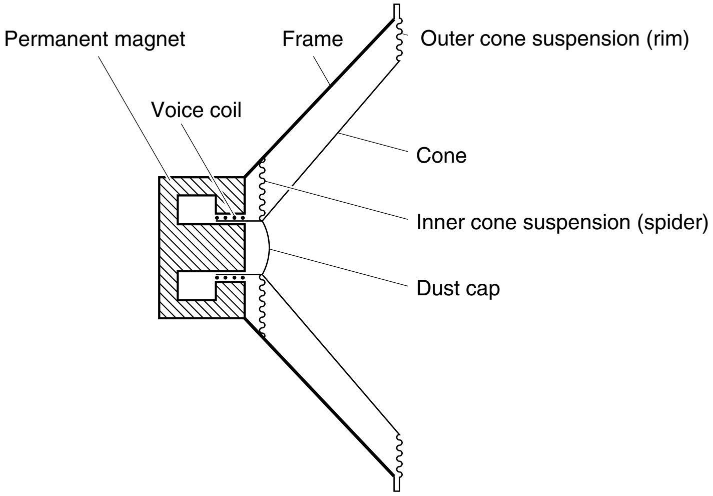


Figure 1.10 Cross section of an electrodynamic cone loudspeaker

their analogous electronic-network counterparts. At higher frequencies (above the cone break-up frequency), deviations from this model occur, as the driver's diaphragm is then no longer rigid. Both transverse and longitudinal waves then appear in the conical shell. These waves are coupled and together they determine the vibration pattern, which has a considerable effect on the sound radiation. Although this is an important issue, it will not be considered here (see e.g. Kaizer [140], Frankort [75], van der Pauw [282]).

An alternative construction method is to have the voice coil stationary, and a moving magnet; this will be discussed in Sec. 4.3.1.

1.3.2.3 Lumped-element Model

For low frequencies, a loudspeaker can be modelled with the aid of some simple elements, allowing the formulation of some approximate analytical expressions for the loudspeaker sound radiation due to an electrical input current, or voltage, which proves to be quite satisfactory for frequencies below the cone break-up frequency. The extreme accelerations experienced by a typical paper cone above about 2 kHz, cause it to flex in a complex pattern. The cone no longer acts as a rigid piston but rather as a collection of vibrating elements.

The forthcoming loudspeaker model will not be extensively derived here, as that has been done elsewhere; see for example, Olson [192], Beranek [28], Borwick [36], Merhaut [173], Thiele [268], Small [252], Clark [51]. We first reiterate briefly the theory for the sealed loudspeaker. In what follows, we use a driver model with a simple acoustic air load. Beranek [28] shows that for a baffled piston this air load is a mass of air equivalent to $0.85a$ in thickness on each side of a piston of radius a. In fact, the air load can exceed this value, since most drivers have a support basket, which obstructs the flow of air from the back of the cone, forcing it to move through smaller openings. This increases the acceleration of this air, augmenting the acoustic load.

Table 1.2 System parameters of the model of Fig. 1.11

R_e	Electrical resistance of the voice coil
L_e	Inductance of the voice coil
I	Voice coil current
U	Voltage induced in the voice coil
B	Flux density in the air gap
l	Effective length of the voice coil wire
Bl	Force factor
F	Lorentz force acting on the voice coil
V	Velocity of the voice coil
k_t	Total spring constant
m_t	Total moving mass, without air load mass
R_m	Mechanical damping
R_d	Electrical damping = $(Bl)^2/R_\text{e}$
R_t	Total damping = $R_\text{r} + R_\text{m} + R_\text{d}$
Z_rad	Mechanical radiation impedance = $R_\text{r} + jX_\text{r}$

The driver is characterized by a cone or piston of area

$$S = \pi a^2, \tag{1.30}$$

and various other parameters, which will be introduced successively, and are summarized in Table 1.2. The resonance frequency f_0 is given by

$$k_\text{t} = (2\pi f_0)^2 m, \tag{1.31}$$

where m is the total moving mass (which includes the air load), k_t is the total spring constant of the system, including the loudspeaker and possibly its enclosing cabinet of volume V_0. This cabinet exerts a restoring force on the piston with equivalent spring constant

$$k_B = \frac{\gamma P_0 S^2}{V_0} = \frac{\rho c^2 S^2}{V_0}, \tag{1.32}$$

where γ is the ratio of the specific heats (1.4 for air), P_0 is the atmospheric pressure, ρ, the density of the medium, and c, the speed of sound. The current $i(t)$ taken by the driver when driven with a voltage $v(t)$ will be given by equating that voltage to the voice coil resistance R_e and the induced voltage

$$v(t) = i(t)R_\text{e} + Bl\frac{\text{d}x}{\text{d}t} + L_\text{e}\frac{\text{d}i}{\text{d}t} \tag{1.33}$$

where Bl is the force factor (which will be explained later on), x is the piston displacement, and L_e the self-inductance of the voice coil. The term in $\text{d}x/\text{d}t$ is the voltage induced by the driver piston velocity of motion. Using the Laplace transform, Eqn. 1.33 can be written as

$$V(s) = I(s)R_\mathrm{e} + BlsX(s) + L_\mathrm{e}sI(s), \tag{1.34}$$

where capitals are used for the Laplace-transformed variables, and s is the Laplace variable, which can be replaced by $i\omega$ for stationary harmonic signals. The relation between the mechanical forces and the electrical driving force is given by

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + R_\mathrm{m}\frac{\mathrm{d}x}{\mathrm{d}t} + k_\mathrm{t}x = Bli, \tag{1.35}$$

where at the left-hand side, we have the mechanical forces, which are the inertial reaction of the cone with mass m, the mechanical resistance R_m, and the total spring force with total spring constant k_t; at the right-hand side, we have the external electromagnetic Lorentz force $F = Bli$ acting on the voice coil, with B, the flux density in the air gap, i, the voice coil current, and l being the effective length of the voice coil wire. Combining Eqns. 1.34 and 1.35, we get

$$X(s)\left[s^2m + s(R_\mathrm{m} + \frac{(Bl)^2}{L_\mathrm{e}s + R_\mathrm{e}} + k_\mathrm{t}\right] = \frac{BlV(s)}{L_\mathrm{e}s + R_\mathrm{e}}. \tag{1.36}$$

We see that besides the mechanical damping R_m, we also get an electrical damping term $(Bl)^2/(L_\mathrm{e}s + R_\mathrm{e})$, and this term plays an important role. If we ignore the inductance of the loudspeaker, the effect of eddy currents[5] induced in the pole structure (Vanderkooy [283]), and the effect of creep[6], we can write Eqn. 1.36 as the transfer function $H_x(s)$ between voltage and excursion

$$H_x(s) = \frac{X(s)}{V(s)} = \frac{Bl/R_\mathrm{e}}{s^2m + s(R + (Bl)^2/R_\mathrm{e}) + k_\mathrm{t}}. \tag{1.37}$$

We use an infinite baffle to mount the piston, and in the compact-source regime ($a/r \ll c/(\omega a)$) the far-field acoustic pressure $p(t)$ a distance r away becomes

$$p(t) = \rho S(\mathrm{d}^2x/\,\mathrm{d}t^2)/(2\pi r), \tag{1.38}$$

proportional to the volume acceleration of the source (Morse and Ingard [180], Kinsler *et al.* [142]). In the Laplace domain, we have

$$P(s) = s^2\rho SX(s)/(2\pi r). \tag{1.39}$$

[5] Owing to the eddy current losses in the voice coil, the voice coil does not behave as an ideal coil, but it can be modelled very well by means of $L_\mathrm{e} = L_0(1 - j\alpha)$, where α is in the order of magnitude of 0.5.

[6] With a voltage or current step as the input, the displacement would be expected to reach its steady-state value in a fraction of a second, according to the traditional model. The displacement may, however, continue to increase. This phenomenon is called creep. Creep is due the viscoelastic effects (Knudsen and Jensen [145], Flügge [74]) of the spring (spider) and edge of the loudspeaker's suspension.

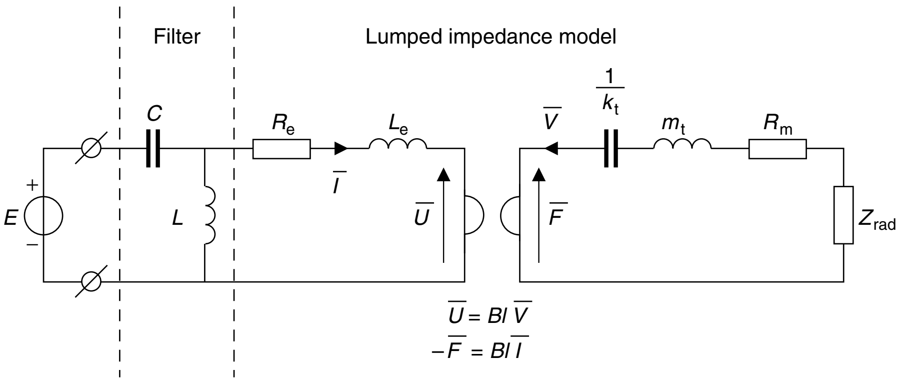


Figure 1.11 Lumped-element model of the impedance-type analogy of an electrodynamic loudspeaker preceded by an LC high-pass crossover filter, which is not part of the actual model. The coupling between the electrical and mechanical parts is represented by a gyrator. The system parameters are given in Table 1.2, from Aarts [4]

Using Eqn. 1.39 and neglecting the self-inductance L_e, we can write Eqn. 1.36 as the transfer function from excursion to pressure

$$H_p(s) = \frac{X(s)}{P(s)} = \frac{s^2 \rho S/(2\pi r) Bl/R_e}{s^2 m + s(R + (Bl)^2/R_e) + k_t}. \tag{1.40}$$

Using Eqns. 1.33 and 1.34, we can make the so-called lumped-element model as shown in Fig. 1.11, which behaves as a simple second-order mass-spring system. We have, for harmonic signals,

$$F = (R_m + i\omega m_t + \frac{k_t}{i\omega} + Z_{rad})V. \tag{1.41}$$

With the aid of Eqn. 1.41 and the properties of the gyrator as shown in Fig. 1.11, the electrical impedance of the loudspeaker (without X_r, which is the imaginary part of the mechanical radiation impedance[7]) can be calculated as follows

$$Z_{in} = R_e + i\omega L_e + \frac{(Bl)^2}{(R_m + R_r) + i\omega m_t + k_t/(i\omega)}. \tag{1.42}$$

Using the following relations

$$\begin{array}{llll} Q_m &= \sqrt{k_t m_t}/R_m, & Q_e &= R_e\sqrt{k_t m_t}/(Bl)^2, \\ Q_r &= \sqrt{k_t m_t}/R_r, & \omega_0 &= \sqrt{k_t/m_t}, \\ \nu &= \omega/\omega_0 - \omega_0/\,\omega, & \tau_e &= L_e/R_e, \\ Q_{mr} &= Q_m Q_r/(Q_m + Q_r), & & \end{array} \tag{1.43}$$

[7] For $\omega \ll \omega_t$ (defined in Eq. 1.49) X_r/ω can be taken into account in m_t.

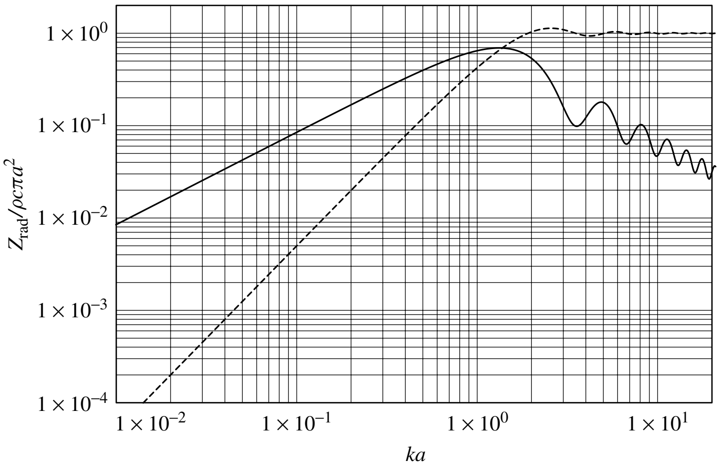


Figure 1.12 Real (dashed line) and imaginary (solid line) parts of the normalized radiation impedance of a rigid disk with a radius a in an infinite baffle

we can write Z_{in} as

$$Z_{\text{in}} = R_{\text{e}}\left[1 + i\omega\tau_{\text{e}} + \frac{Q_{\text{mr}}/Q_{\text{e}}}{1 + iQ_{\text{mr}}\nu}\right], \tag{1.44}$$

if we neglect L_{e}, we get at the resonance frequency ($\nu = 0$) the maximal input impedance

$$Z_{\text{in}}(\omega = \omega_0) = R_{\text{e}}(1 + Q_{\text{mr}}/Q_{\text{e}}) \approx R_{\text{e}} + (Bl)^2/R_{\text{m}}. \tag{1.45}$$

The time-averaged electrical power P_{e} delivered to the driver is then

$$P_{\text{e}} = 0.5|I|^2\Re\{Z_{\text{in}}\} = 0.5|I|^2R_{\text{e}}\left[1 + \frac{Q_{\text{mr}}/Q_{\text{e}}}{1 + Q_{\text{mr}}^2\nu^2}\right]. \tag{1.46}$$

The radiation impedance of a plane-circular rigid piston[8] with a radius a in an infinite baffle can be derived as (Morse and Ingard [180, p. 384])

$$Z_{\text{rad}} = \pi a^2\rho c[1 - 2J_1(2ka)/(2ka) + i2\mathbf{H}_1(2ka)/(2ka)], \tag{1.47}$$

where $\mathbf{H}_1$ is a Struve function (Abramowitz and Stegun [12, 12.1.7]), J_1 is a Bessel function and k is the wave number ω/c. The real and imaginary parts of Z_{rad} are plotted in Fig. 1.12.

[8] The radiation impedance of rigid cones and that of rigid domes is studied in, for example, Suzuki and Tichy [259, 260]. They appeared to be significantly different with respect to rigid pistons for $ka > 1$, revealing that Z_{rad} for convex domes is generally lower than that for pistons and higher than that for concave domes.

Z_{rad} can be approximated as

$$Z_{\mathrm{rad}} \approx \begin{cases} \pi a^2 \rho c[(ka)^2/2 + i8\,ka/(3\pi)], & \omega \ll \omega_{\mathrm{t}} \\ \pi a^2 \rho c[1 + i2/(\pi ka)], & \omega \gg \omega_{\mathrm{t}}, \end{cases} \tag{1.48}$$

where

$$\omega_{\mathrm{t}} = 1.4\,c/a \tag{1.49}$$

is the transition frequency (−3 dB point). A full-range approximation of $\mathbf{H}_1$ is given in Sec. 1.3.3.2 and in Aarts and Janssen [9]. However, for low frequencies, we can either neglect the damping influence of Z_{rad}, or use

$$\Re\{Z_{\mathrm{rad}}\} \approx a^4 f^2/4, \tag{1.50}$$

which follows immediately from Eqn. 1.48. The real part of Z_{rad} is qualitatively depicted in Fig. 1.9, but more precisely in Fig. 1.12. The time-averaged acoustically radiated power can then be calculated as follows

$$P_{\mathrm{a}} = 0.5|V|^2 \Re\{Z_{\mathrm{rad}}\}, \tag{1.51}$$

and with the aid of Eqns. 1.41–1.51 as follows

$$P_{\mathrm{a}} = \frac{0.5(Bl/(R_{\mathrm{m}} + R_{\mathrm{r}}))^2 I^2 R_{\mathrm{r}}}{1 + Q_{\mathrm{mr}}^2 \nu^2}, \tag{1.52}$$

as depicted in Fig. 1.9, which clearly shows the bandwidth limitation similar to a bandpass filter. The acoustic pressure in the far field at distance r and azimuth angle θ is

$$p(r,t) = i\frac{f\rho_0\, V\pi a^2}{r}\left[\frac{J_1(ka\sin\theta)}{ka\sin\theta}\right] \mathrm{e}^{i\omega(t-r/c)}, \tag{1.53}$$

assuming an axis of symmetry at $\theta = 0$ rad, where V is the velocity of the piston (Beranek [28], Kinsler *et al.* [142]), and J_1 is a Bessel function, see Sec. 1.3.3.1. Assuming a velocity profile as depicted in Fig. 1.9, we can calculate the magnitude of the on-axis response ($\theta = 0$). This is also depicted in Fig. 1.9, which shows a ‘flat’ SPL for $\omega \gg \omega_0$. However, owing to the term in square brackets in Eq. 1.53, the off-axis pressure response ($\theta \neq 0$) decreases with increasing ka. This yields an upper frequency limit for the acoustic power, together with the mechanical lower frequency limit ω_0, and is the reason why a practical loudspeaker system needs more than one driver (a multi-way system) to handle the whole audible frequency range.

The power efficiency can be calculated as follows

$$\eta(\nu) = P_{\mathrm{a}}/P_{\mathrm{e}} = [Q_{\mathrm{e}} Q_{\mathrm{r}}(\nu^2 + 1/Q_{\mathrm{mr}}^2) + Q_r/Q_{\mathrm{mr}}]^{-1}. \tag{1.54}$$

In Fig. 4.13, some plots for $\eta(\nu)$ for various drivers are shown. For low frequencies[9], so that $Q_{\mathrm{mr}} \approx Q_{\mathrm{m}}$, the efficiency can be approximated as

$$\eta(\nu) = P_{\mathrm{a}}/P_{\mathrm{e}} \approx [Q_{\mathrm{r}}\{Q_{\mathrm{e}}(\nu^2 + 1/Q_{\mathrm{m}}^2) + 1/Q_{\mathrm{m}}\}]^{-1}. \tag{1.55}$$

A convenient way to relate the sound pressure level L_{p} to the power efficiency η is the following. For a plain wave, we have the relation between sound intensity I and sound pressure p

$$I = \frac{p^2}{\rho c}, \tag{1.56}$$

and the acoustical power is equal to

$$P_{\mathrm{a}} = 2\pi r^2 I \tag{1.57}$$

or

$$P_{\mathrm{a}} = \frac{2\pi r^2 p^2}{\rho c}. \tag{1.58}$$

Using the above relations, we get

$$L_{\mathrm{p}} = 20 \log \left(\sqrt{\frac{P_{\mathrm{a}} \rho c}{2\pi r^2}} / p_0 \right), \tag{1.59}$$

where we assume radiation into one hemifield (solid angle of 2π), that is, we only account for the pressure at one side of the cone, which is mounted in an infinite baffle. For $r = 1$ m, $\rho = 415$, $P_{\mathrm{a}} = 1$ W, and $p_0 = 20\ 10^{-6}$, we get

$$L_{\mathrm{p}} = 112 + \log \eta\,. \tag{1.60}$$

If $\eta = 1$ (in this case $P_{\mathrm{a}} = P_{\mathrm{e}} = 1$ W), we get the maximum attainable L_{p} of 112 dB. Equation 1.60 can also be used to calculate η if L_{p} is known, for example, by measurement.

1.3.3 BESSEL AND STRUVE FUNCTIONS

Bessel and Struve functions occur in many places in physics and quite prominently in acoustics for impedance calculations. The problem of the rigid-piston radiator mounted in an infinite baffle has been studied widely for tutorial as well as for practical reasons, see for example, Greenspan [98], Pierce [208], Kinsler *et al.* [142], Beranek [28], Morse and Ingard [180]. The resulting theory is commonly applied to model a loudspeaker in the audio-frequency range. For a baffled piston, the ratio of the force amplitude to the

[9] It should be noted that Q_{r} depends on ω, but using Eqns. 1.43 and 1.48 for $\omega \ll \omega_{\mathrm{t}}$, we can approximate $Q_{\mathrm{r}} \approx 2c\sqrt{k_{\mathrm{t}} m_{\mathrm{t}}}/(\pi a^4 \rho \omega^2)$.

normal velocity amplitude, which is called the piston mechanical radiation impedance, is given by

$$Z_{\mathrm{m}} = -i\frac{\omega\rho}{2\pi}\int\int\int\int R^{-1}\mathrm{e}^{ikR}\,\mathrm{d}x_{\mathrm{s}}\,\mathrm{d}y_{\mathrm{s}}\,\mathrm{d}x\,\mathrm{d}y. \tag{1.61}$$

Here $R = \sqrt{(x - x_{\mathrm{s}})^2 + (y - y_{\mathrm{s}})^2}$ is the distance between any two points, $(x_{\mathrm{s}},\ y_{\mathrm{s}})$ and (x, y), on the surface of the piston. The integration limits are such that $(x_{\mathrm{s}},\ y_{\mathrm{s}})$ and $(x,\ y)$ are within the area of the piston. The four-fold integral in Eqn. 1.61, known as the Helmholtz integral, was solved by Rayleigh [219, 302] and further elaborated in Pierce [208], with the result

$$Z_{\mathrm{m}} = \rho c\pi a^2[R_1(2ka) - iX_1(2ka)], \tag{1.62}$$

where

$$R_1(2ka) = 1 - \frac{2J_1(2ka)}{2ka} \tag{1.63}$$

and

$$X_1(2ka) = \frac{2\mathbf{H}_1(2ka)}{2ka} \tag{1.64}$$

are the real and imaginary parts of the radiation impedance, respectively. In Eqns. 1.63 and 1.64, J_1 is the first-order Bessel function of the first kind (Abramowitz and Stegun [12, 9.1.21]), and $\mathbf{H}_1(z)$ is the Struve function of the first kind (Abramowitz and Stegun [12, 12.1.6]). Bessel functions are solutions to the homogeneous Bessel equation

$$z^2y'' + zy' + (z^2 - \nu^2)y = 0, \tag{1.65}$$

where a particular kind of solution J_n is discussed in Sec. 1.3.3.1. Struve functions are solutions to the inhomogeneous Bessel equation

$$z^2y'' + zy' + (z^2 - \nu^2)y = \frac{4(z/2)^{\nu+1}}{\sqrt{\pi}\,\Gamma(\nu + 1/2)}, \tag{1.66}$$

and are discussed in Sec. 1.3.3.2; Γ is the gamma function. Also, some useful formulas for Bessel and Struve functions are given and an effective and simple approximation of $\mathbf{H}_1(z)$, which is valid for all z (from Aarts and Janssen [9]).

1.3.3.1 Bessel Functions $J_n(z)$

The Bessel function of order n can be represented (Abramowitz and Stegun [12, 9.1.21]) by the integral

$$J_n(z) = \frac{1}{\pi}\int_0^{\pi}\cos(z\sin\theta - n\theta)\,\mathrm{d}\theta, \tag{1.67}$$

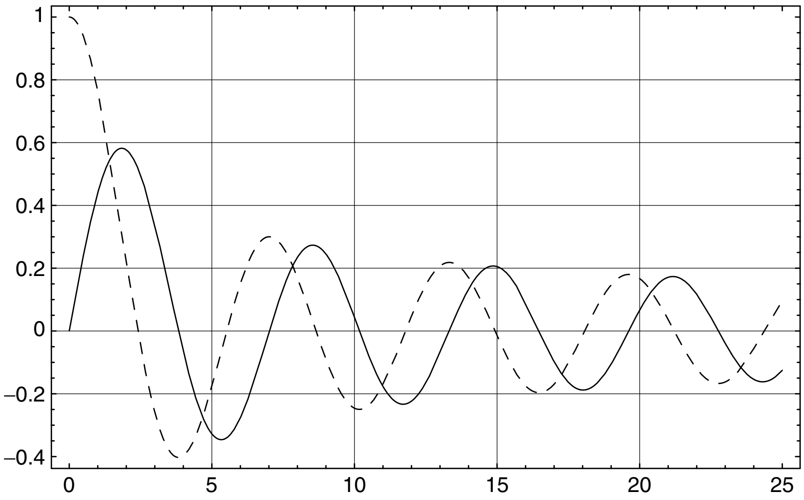


Figure 1.13 Plot of Bessel functions $J_0(x)$ (dashed), and $J_1(x)$ (solid)

and is plotted in Fig. 1.13 for $n = 0, 1$. There is the power series expansion (Abramowitz and Stegun [12, 9.1.10])

$$J_n(z) = \left(\frac{z}{2}\right)^n \sum_{k=0}^{\infty} \frac{\left(\frac{-z^2}{4}\right)^k}{k!\,\Gamma(n+k+1)}, \tag{1.68}$$

which yields

$$J_0(z) = 1 - \frac{\frac{1}{4}z^2}{(1!)^2} + \frac{\left(\frac{1}{4}z^2\right)^2}{(2!)^2} - \frac{\left(\frac{1}{4}z^2\right)^3}{(3!)^2} + \cdots, \tag{1.69}$$

and

$$J_1(z) = \frac{z}{2} - \frac{z^3}{16} + \frac{z^5}{384} - \frac{z^7}{18\,432} + \cdots. \tag{1.70}$$

For the purpose of numerical computation, these series are only useful for small values of z. For small values of z, Eqns. 1.63 and 1.70 yield

$$R_1(ka) \approx \frac{(ka)^2}{2}, \tag{1.71}$$

where we have substituted $ka = z$; this is in agreement with the small ka approximation as can be found in the references given earlier, see also Fig. 1.12. Furthermore, there is the asymptotic result (Abramowitz and Stegun [12, 9.2.1 with $\nu = 1$]), see Fig. 1.13,

$$J_1(z) = \sqrt{\frac{2}{\pi z}}\left(\cos(z - 3\pi/4) + O(1/z)\right), \quad z \to \infty, \tag{1.72}$$

but this is only useful for large values of z. Equation 1.63 and the first term of Equation 1.72 yield for large values of z (again substituting ka)

$$R_1(ka) \approx 1, \tag{1.73}$$

which is in agreement with the large ka approximation, as can be found in the given references as well. The function $J_n(x)$ is tabulated in many books (see e.g. Abramowitz and Stegun [12]), and many approximation formulas exist (see e.g. Abramowitz and Stegun [12, 9.4]). Another method to evaluate J_n is to use the following recurrent relation

$$J_{n-1}(x) + J_{n+1}(x) = \frac{2n}{x} J_n(x), \tag{1.74}$$

provided that $n < x$, otherwise severe accumulation of rounding errors will occur (Abramowitz and Stegun [12, 9.12]). However, $J_n(x)$ is always a decreasing function of n when $n > x$, so the recurrence can always be carried out in the direction of decreasing n. The iteration is started with an arbitrary value zero for J_n, and unity for J_{n-1}. We normalize the results by using the equation

$$J_0(x) + 2J_2(x) + 2J_4(x) + \cdots = 1. \tag{1.75}$$

A heuristic formula to determine the value of m to start the recurrence with $J_{\mathrm{m}} = 1$ and $J_{m-1} = 0$ is ($\lceil \cdot \rfloor$ indicates rounding to the nearest integer)

$$\begin{aligned} m &= \left\lceil \frac{6 + \max(n, p) + \frac{9p}{p+2}}{2} \right\rfloor \\ p &= \frac{3x}{2}. \end{aligned} \tag{1.76}$$

1.3.3.2 The Struve Function $\mathbf{H}_1(z)$

The first-order Struve function $\mathbf{H}_1(z)$ is defined as

$$\mathbf{H}_1(z) = \frac{2z}{\pi} \int_0^1 \sqrt{1 - t^2} \sin zt \, \mathrm{d}t \tag{1.77}$$

and is plotted in Fig. 1.14. There is the power series expansion (Abramowitz and Stegun [12, 12.1.5])

$$\mathbf{H}_1(z) = \frac{2}{\pi}\left[\frac{z^2}{1^2 3} - \frac{z^4}{1^2 3^2 5} + \frac{z^6}{1^2 3^2 5^2 7} - \cdots\right]. \tag{1.78}$$

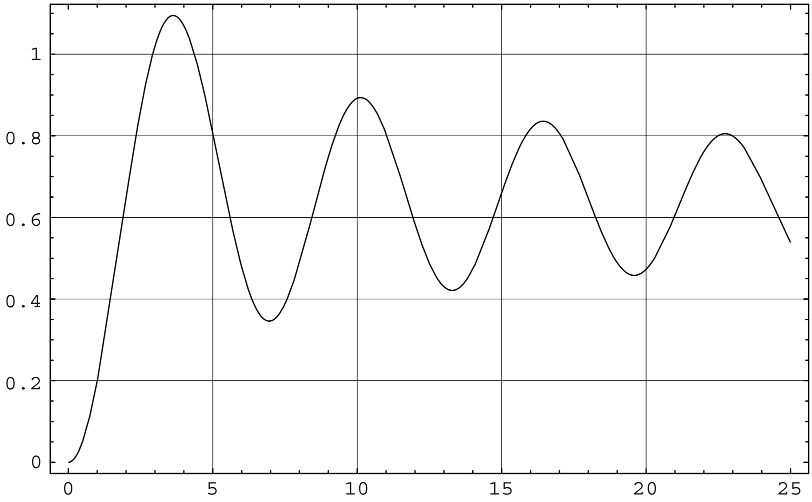


Figure 1.14 Plot of Struve function $\mathbf{H}_1(x)$

For the purpose of numerical computation, this series is only useful for small values of z. Eqns. 1.64 and 1.78 yield, for small values of z (substituting ka for z)

$$X_1(ka) \approx \frac{8\,ka}{3\pi}, \tag{1.79}$$

which is in agreement with the small ka approximation as can be found in the references given earlier, see also Fig. 1.12. Furthermore, there is the asymptotic result (Abramowitz and Stegun [12, 12.1.31, 9.2.2 with $\nu = 1$]).

$$\mathbf{H}_1(z) = \frac{2}{\pi} - \sqrt{\frac{2}{\pi z}}\,(\cos(z - \pi/4) + O(1/z))\,, \quad z \to \infty, \tag{1.80}$$

but this is only useful for large values of z. Eqn. 1.64 and the first term of Eqn. 1.80 yield for large values of ka

$$X_1(ka) \approx \frac{2}{\pi\,ka}, \tag{1.81}$$

which is in agreement with the large ka approximation, as can also be found in the earlier given references. An approximation for all values of ka was developed by Aarts and Janssen [9]. Here, only a limited number of elementary functions is involved:

$$\mathbf{H}_1(z) \approx \frac{2}{\pi} - J_0(z) + \left(\frac{16}{\pi} - 5\right)\frac{\sin z}{z} + \left(12 - \frac{36}{\pi}\right)\frac{1 - \cos z}{z^2}\,. \tag{1.82}$$

The approximation error is small and decently spread out over the whole z-range, vanishes for $z = 0$, and its maximum value is about 0.005. Replacing $\mathbf{H}_1(z)$ in Fig. 1.12 by the approximation in Eqn. 1.82 would result in no visible change. The maximum relative error appears to be less than 1%, equals 0.1% at $z = 0$, and decays to zero for $z \to \infty$.

1.3.3.3 Example

A prime example of the use of the radiation impedance is for the calculation of the radiated acoustic power of a circular piston in an infinite baffle. This is an accurate model for a loudspeaker with radius a mounted in a large cabinet (Beranek [28]). The radiated acoustic power is equal to

$$P_{\mathrm{a}} = 0.5|V|^2\Re\{Z_{\mathrm{m}}\}, \tag{1.83}$$

where V is the velocity of the loudspeaker's cone. The use of the just-obtained approximation for $\mathbf{H}_1$ is to calculate the loudspeaker's electrical input impedance Z_{in}, which is a function of Z_{m} (see Beranek [28]). Using Z_{in}, the time-averaged electrical power delivered to the loudspeaker is calculated as

$$P_{\mathrm{e}} = 0.5|I|^2\Re\{Z_{\mathrm{in}}\}, \tag{1.84}$$

where I is the current fed into the loudspeaker. Finally, the efficiency of a loudspeaker, defined as

$$\eta(ka) = P_{\mathrm{a}}/P_{\mathrm{e}}, \tag{1.85}$$

can be calculated. These techniques are used in Chapter 4 when analyzing the behavior of loudspeakers with special drivers.

1.4 AUDITORY PERCEPTION

This section reviews the basic concepts of the auditory system and auditory perception, insofar as they relate to BWE methods that will be discussed in later chapters. The treatment here is necessarily concise, but there are numerous references provided for further reading, if necessary or desired. Reviews of psychoacoustics can be found in, for example, Moore [177, 178], Yost *et al.* [302]; physiology of the peripheral hearing system is discussed in, for example, Geisler [86].

1.4.1 PHYSICAL CHARACTERISTICS OF THE PERIPHERAL HEARING SYSTEM

The peripheral hearing system consists of outer, middle, and inner ear, see Fig. 1.15. Sound first enters via the pinna, which has an irregular shape that filters impinging sound waves. This feature aids in sound localization, which is not further discussed here (see e.g. Batteau [26], Blauert [34]). Next, sound passes into the auditory canal and on to the eardrum, or tympanic membrane, which transmits vibrations in the air to the three middle ear bones, the ossicles (malleus, incus, and stapes). The stapes connects to the oval window, the entrance to the fluid-filled cochlea. The system from tympanic membrane to oval window serves as an impedance-matching device so that a large portion of sound energy at the frequencies of interest, in the air, is transmitted into the cochlea. Muscles connect the malleus and stapes to the bone of the skull, and contraction of these muscles

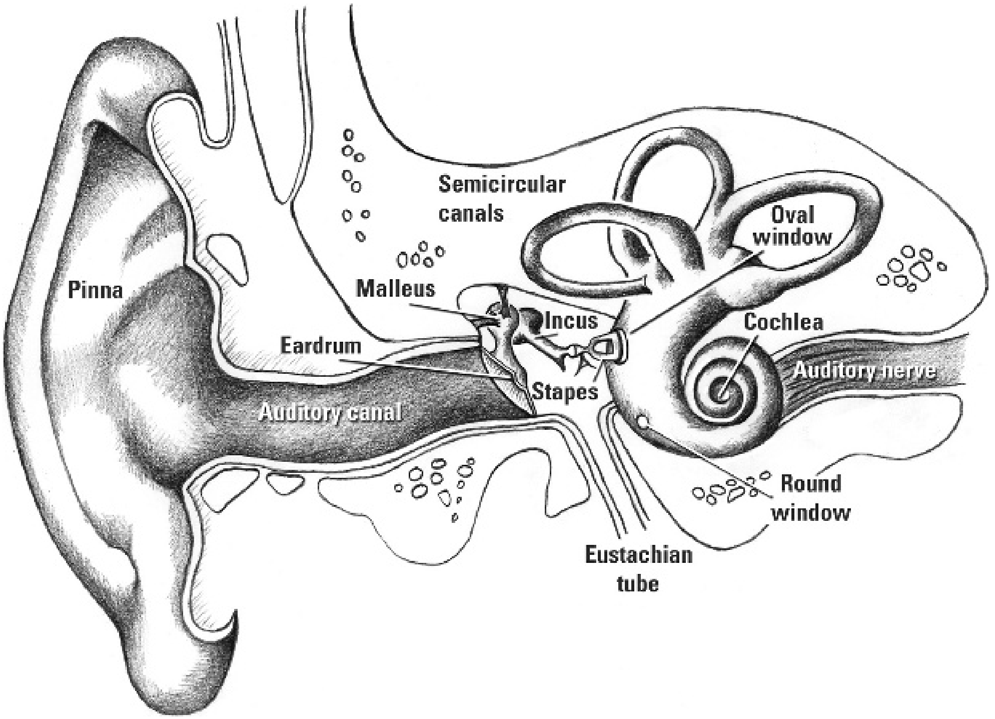


Figure 1.15 Sketch of the peripheral part of the hearing system, showing outer (pinna, auditory canal, eardrum), middle (malleus, incus, stapes), and inner ear (cochlea)

can be used to attenuate high-level sounds (primarily low frequency). Pressure equalization in the middle ear is achieved through the Eustachian tube, which connects the middle ear cavity to the throat.

The cochlea is a spiral-shaped cavity in the bone of the skull, filled with cerebro-spinal fluid; a cross section in shown in Fig. 1.16. The cochlea is wide at the oval window (the base) and narrows towards the other extreme (the apex). It is divided into three parts by the basilar membrane (BM) and Reisner's membrane; from top to bottom are the scala vestibuli, scala media, and scala tympani. The scala vestibuli and scala tympani are connected at the apex. Vibrations of the oval window are transmitted as a travelling wave through the cochlea, and also vibrate the BM. Locations on the BM have a very sharp frequency tuning because the variation in mechanical properties lead to different resonant frequencies at each location; the sharp frequency tuning is also achieved by active processes occurring within the cochlea. Between the BM and the tectorial membrane, in the organ of Corti, are rows of hair cells (outer hair cells and inner hair cells) with attached stereocilia. These oscillate along with the BM motion, which finally leads to a neural response from the inner hair cells. This response is propagated through the auditory nerve onto the cochlear nucleus and subsequent neural processing centres. There is also a descending pathway, from the brain to the outer hair cells, but this is not further discussed here.

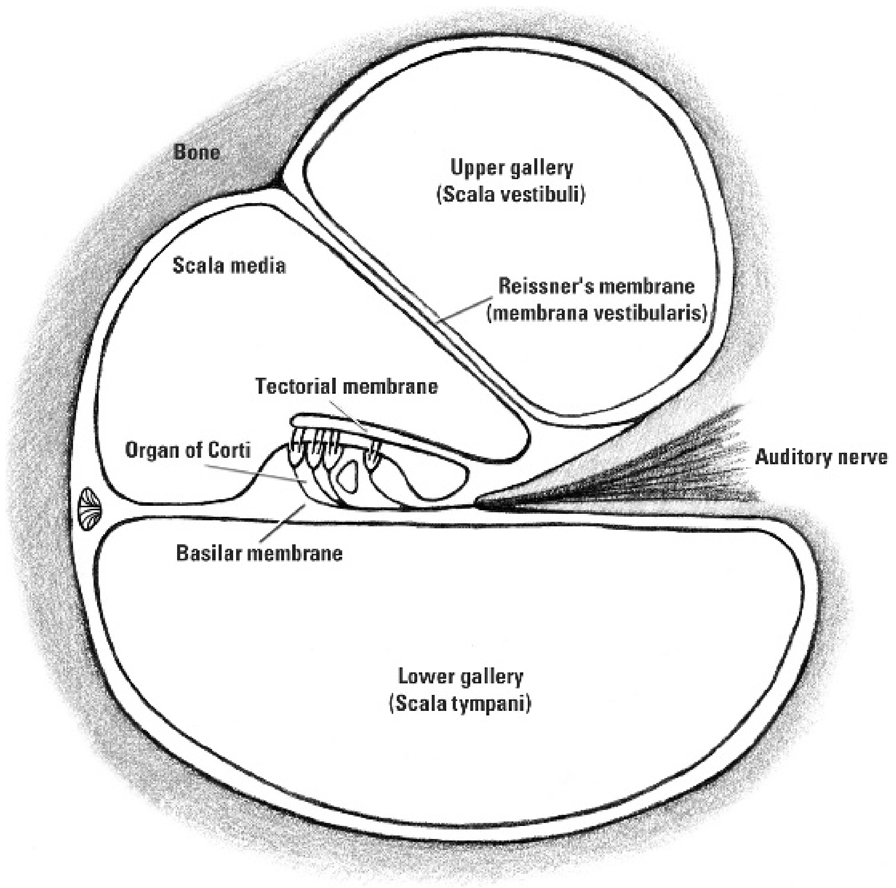


Figure 1.16 Cross section of the cochlea, showing the various membranes and cavities, and the organ of Corti. The hair cells are supported between the organ of Corti and the tectorial membrane, to which they are attached by stereocilia. The movement of the stereocilia, in response to sound entering the ear, causes a neural response, carried to the brain by the auditory nerve

The mechanical properties of the BM vary across the length of the cochlea: the BM is widest and has the lowest compliance at the apex, thereby causing each portion of the BM to respond maximally to a specific frequency (this is somewhat dependent on signal level). High-frequency sound vibrates the BM near the base, and low-frequency sound near the apex. These features were first elucidated by the investigations of the Nobel prize winner, Georg von Békésy [290]. The ordering of frequencies from low to high along the spatial extent of the BM is known as a tonotopic organization. The relation of the position y (distance in cm from the stapes, range approximately from 0–3.5 cm) of the maximum peak amplitude can be well approximated by

$$f = 2.5 \times 10^{4-0.72y}, \tag{1.86}$$

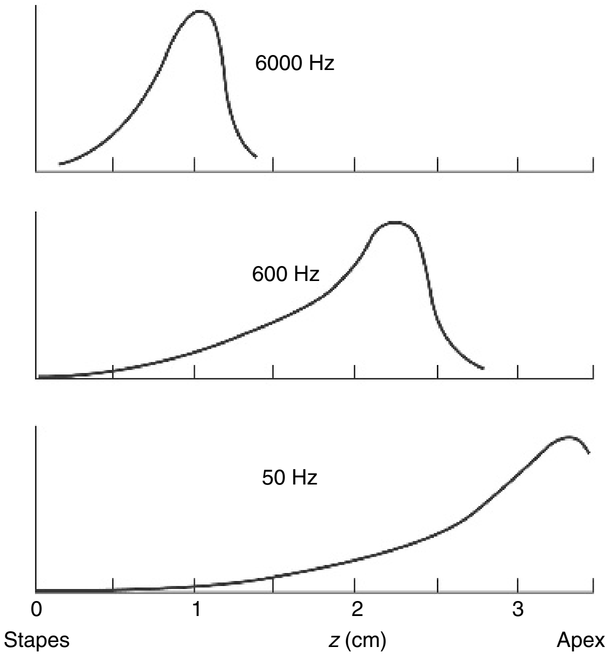


Figure 1.17 Peak displacement amplitude of the BM in response to pure tones of various frequencies

for frequencies f below 200 Hz. Below this frequency, a pattern is produced, which extends all along the basilar membrane, but reaches a maximum before the end of the membrane. Figure 1.17 illustrates for pure tones of various frequencies the peak displacement amplitude along the BM. The response of neurons along the BM generated by the BM motion is qualitatively similar; this neural response is called the excitation pattern.

1.4.2 NON-LINEARITY OF THE BASILAR MEMBRANE RESPONSE

An essential aspect of the cochlear response is that it is non-linear. Therefore, the shape of the graphs in Fig. 1.17 change somewhat, depending on the level of the stimulus. Also, the BM motion (BMM) is strongly compressed at moderate signal levels. This makes it possible for a normal ear to have a useable dynamic range of about 120 dB, while the variation in BMM is much smaller. It is thought that the outer hair cells (OHC) are largely responsible for this non-linear behavior.

An interesting consequence of this non-linearity is that if two tones are presented to the ear at sufficient sound pressure level, distortion products will be generated in the cochlea. These distortion products are also called combination tones (CT), see for example, Goldstein [91], Smoorenburg [253], Zwicker [307]. Assuming that the input frequencies are f_1 and f_2, CTs will appear at frequencies

$$f(n) = (n+1)f_1 - nf_2, \quad f_n > 0. \tag{1.87}$$

These combination tones are audible and, as we shall see in Chapter 2, may serve a useful purpose for certain kinds of BWE applications. The cubic CT ($n = 1$) is usually largest in amplitude, and is relatively independent of frequency, being about 15–20 dB below the level of each component of the two-tone complex (assumed identical). However, this is only true if $f_1/f_2 \approx 1.1$, that is, for closely spaced components. For a ratio $f_1/f_2 = 1.4$, the level of the cubic CT drops to about 40 dB below the level of the components in the two-tone stimulus.

Another distortion product is the difference tone (DT) $f_2 - f_1$ (Goldstein [91]). The DT level does not depend very much on the ratio f_1/f_2, but varies greatly with the level of the two-tone complex. The DT is barely audible at 50 dB and maintains a level of about 50 dB below the level of the components of the two-tone complex. If $f_1 = 2f_2$, the DT and the cubic CT will coincide.

1.4.3 FREQUENCY SELECTIVITY AND AUDITORY FILTERS

It was just shown how the BMM (and correspondingly, the excitation pattern) varies in position, depending on the frequency of a pure tone (Fig. 1.17). When considering a fixed position on the BM, we can use the excitation patterns of pure tones to determine the frequency response of the excitation pattern at that position. One then obtains a filter for each position on the BM; the resulting filters are called auditory filters. These describe how a position on the BM responds to pure tones of various frequencies. Important parameters of these filters are:

- *Characteristic frequency* (CF): This is the frequency (in Hz) of a pure tone, which yields the maximum response, and is also given (approximately) by Eqn. 1.86.
- *Equivalent rectangular bandwidth (*ERB*)*: This is the bandwidth (in Hz) of a rectangular filter having a passband amplitude equal to the maximum response of the auditory filter, that passes the same power of a white noise signal. A small ERB implies a narrow filter, and hence high frequency selectivity. Instead of ERB, sometimes a measure of tuning, Q, is used, to specify frequency selectivity. It is defined as $Q = \mathrm{CF}/\Delta f$, where Δf is some measure of bandwidth, usually the −3-dB bandwidth. Large Q values imply high frequency selectivity.

A relation describing ERB as a function of frequency f is given by (Glasberg and Moore [89])

$$\mathrm{ERB}(f) = 24.7(4.37 \times 10^{-3} f + 1). \tag{1.88}$$

Different experimenters have sometimes found different values for the bandwidth of auditory filters, and there is thus no universally agreed-upon value; the equation given here for auditory filter bandwidth is a widely used version. It is noted, however, that the original measurement of auditory filter bandwidth by Fletcher [72] yielded smaller bandwidths (Fletcher used the concept of ‘critical band’ analogously to the ERB as just described); recent experiments (Shera *et al.* [247]; Oxenham and Shera [195]) using novel techniques for measuring auditory filter bandwidth seem to agree better with Fletcher’s original results than with Eqn. 1.88.

A commonly used model for auditory filter shape is by means of the gammatone filter (Patterson *et al.* [204], Hohman [110]) $g_t(t)$

$$g_t(t) = at^{n-1}e^{-2\pi b \cdot \mathrm{ERB}(f_c)t \cos(2\pi f_c t + \phi)}. \tag{1.89}$$

Here a, b, n, f_c, and ϕ are parameters and $\mathrm{ERB}(f_c)$ is as given in Eqn. 1.88. The gammatone filters are often used in auditory models (e.g. AIM; see Sec. 1.4.8) to simulate the spectral analysis performed by the BM.

The frequency selectivity of the auditory filters is thought to have a large influence on many auditory tasks, such as understanding speech in noise or detecting small timbre differences between sounds. For BWE applications, another interesting aspect of auditory filters is their presumed influence on pitch perception, discussed in Sec. 1.4.5. For the moment, we mention that depending on the ERB of an auditory filter, it may pass one or more harmonics of a complex tone, also depending on the fundamental frequency of that tone. It turns out that roughly up to harmonic number 10, auditory filters pass only one harmonic, that is, these harmonics are *spectrally resolved*. At higher harmonic numbers, the ERB of the auditory filters become wider than the harmonic frequency spacing, and therefore the auditory filters pass two or more harmonics. These harmonics are thus *spectrally unresolved*, and the output of the auditory filter is a superposition of a number of these higher harmonics. Whether a harmonic is resolved or not will have a large influence on the subsequently generated neural response. In the following, we will alternately use the terms harmonic and partial, both referring to one component of a harmonically related complex tone.

1.4.4 LOUDNESS AND MASKING

1.4.4.1 Definitions

Loudness is related to the level, or amplitude, of a sound, but depends in a complicated manner on level and also frequency content. The following definitions are used by the ISO [117]

Definition 4 *Loudness: That attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud. Loudness is expressed in sone, where one sone is the loudness of a sound, whose loudness level is 40 phon.*

Definition 5 *Loudness level: Of a given sound, the sound pressure level of a reference sound, consisting of a sinusoidal plane progressive wave of frequency 1 kHz coming from directly in front of the listener, which is judged by otologically normal persons to be equally loud to the given sound. Loudness level is expressed in phon.*

Definition 6 *Critical bandwidth: The widest frequency band within which the loudness of a band of continuously distributed random noise of constant band sound pressure level is independent of its bandwidth.*

Note that the critical *bandwidth* so defined is intimately related to the ERB of the auditory filters (Eqn. 1.88) and also Fletcher's critical *band.*

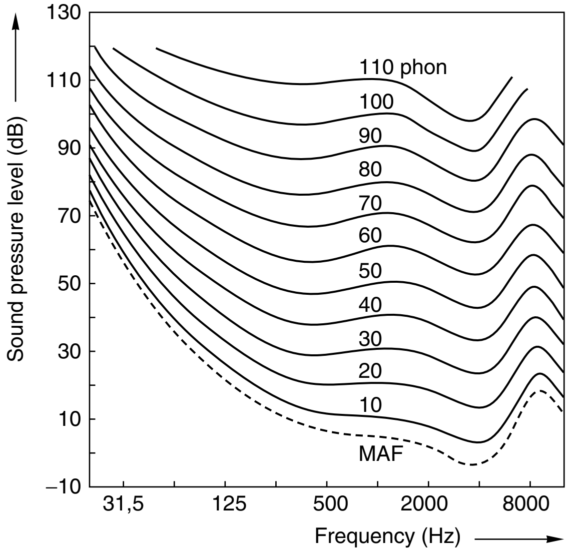


Figure 1.18 Normal equal-loudness level contours for pure tones (binaural free-field listening, frontal incidence. Data from ISO 226-1987(E) [117, Fig. 1]. MAF indicates minimum audible field, the just-noticeable sound pressure level

The threshold of audibility is the minimum perceptible free-field listening intensity level of a tone that can be detected at each frequency over the entire range of the ear. The average threshold of audibility for the normal ear is shown as the curve labeled MAF (minimum audible field) in Fig. 1.18; the other curves show the equal-loudness contours at various phon levels. The frequency of maximum sensitivity is near 4 kHz. Below this frequency, the threshold rises to about 70 dB. For high frequencies, the threshold rises rapidly, which also strongly depends on age, as is shown in Fig. 1.19. This has some implication for high-frequency BWE methods, as will be discussed in Chapters 6 and 5. Elderly listeners might, on average, not benefit as much (or at all) from processing strategies that increase high-frequency content of audio signals. The curves are also level dependent, especially at low frequencies, where they are compressed; this is illustrated in Fig. 1.20, which shows the normalized difference between the 80-phon contour and the 20-, 40-, 60-, and 100-phon contours. The compression of the equal-loudness contours at low frequencies implies that small changes in SPL lead to large changes in loudness.

1.4.4.2 Scaling of Loudness

Several experimenters have made contributions to the scaling of loudness. The earliest published work seems to be that credited to Richardson and Ross [224], who required an observer to rate one of two tones of different intensities, which he heard as a certain multiple or fraction of the other. Since then, various methods of evaluating loudness of complex sounds from objective spectrum analysis have been proposed. The earliest attempt to use algebraic models in psychophysical measurement is probably that of Fletcher and

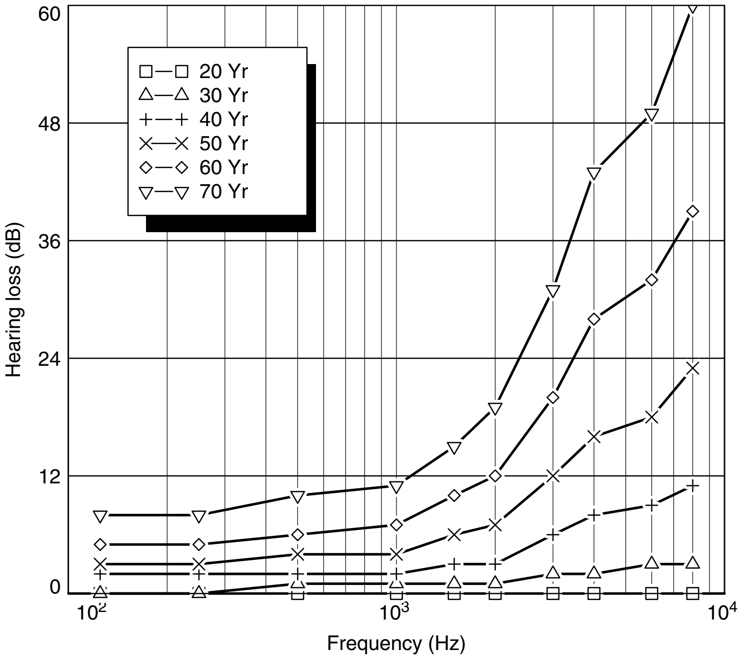


Figure 1.19 Hearing loss (with respect to threshold of hearing) for a group of normal males of various ages. For each age, the 50th percentile is shown. Data from ISO 7029-1984(E) [118]

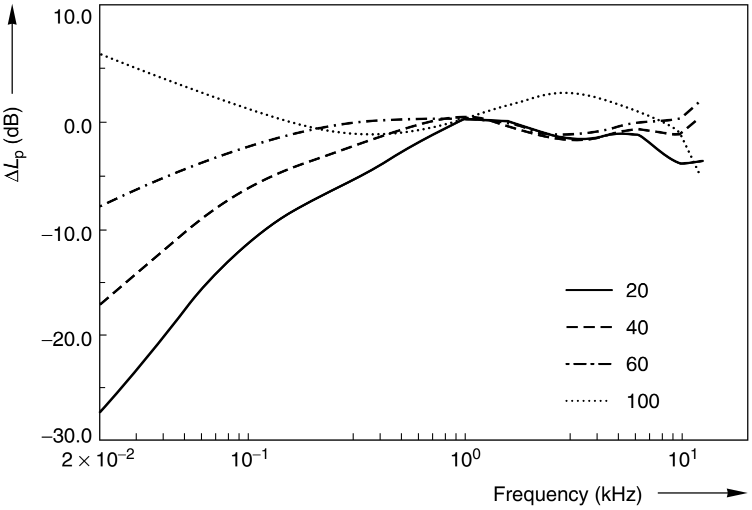


Figure 1.20 Differences between the 80-phon contour and the 20-, 40-, 60-, and 100-phon contours, respectively. The differences have been normalized to 0 dB at 1 kHz. From Aarts [4]

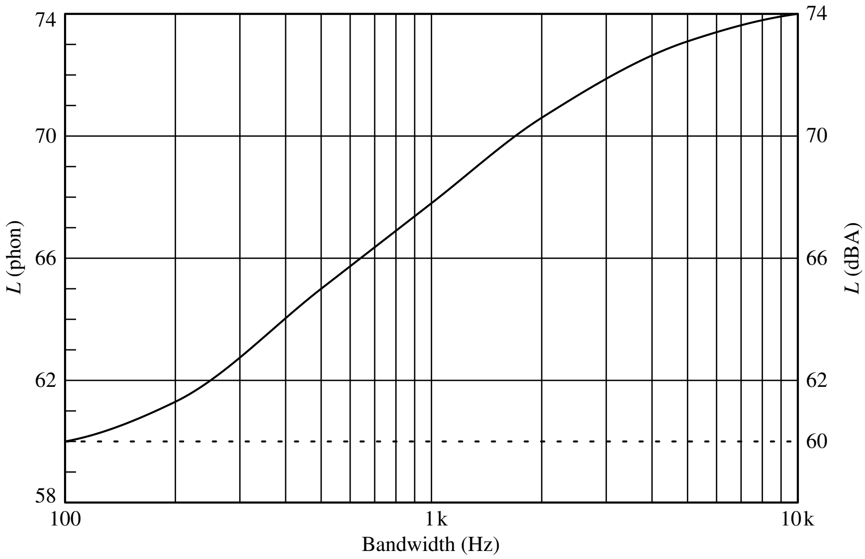


Figure 1.21 Increasing loudness (solid curve) versus bandwidth of white noise, while the dB(A) Level (dashed line) was kept constant. From Aarts [4]

Munson [73]. However, there is still much interest in this subject, and nowadays there are two standardized procedures for calculating loudness levels. The first is based on a method developed by Stevens [256], hereafter referred to as method 532A and the second one, 532B, by Zwicker [305, 310]. A-weighting is a widely used method, traditionally applied in sound level meters to measure the loudness of signals or to determine the annoyance of noise. It is based on an early 40-phon contour and is a rough approximation of a frequency weighting of the human auditory system. However, considerable differences are ascertained between subjective loudness ratings and the A-weighted measurements. For example, the loudness of noise increases when the dB(A) level is kept constant and the bandwidth of the noise is increased. As depicted in Fig. 1.21, with increasing bandwidth the loudness has increased from 60 to 74 phon (solid curve), while the dB(A) level (dashed line) was kept constant. The effect has been studied by Brittain [39], and is a striking example that for wideband signals the A-weighted method is generally too simple. As another example, consider the loudness of a tone of 200 Hz, 2000 Hz, or both combined, as in Fig. 1.22 (a, b, and c, respectively). Each part shows four dB values: the top value is that computed by a loudness model (ISO532B, to be discussed hereafter), the second and third by A- and B-weighting respectively, and the last value is the acoustic SPL. B-weighting is obtained by approximating the inverse of the 80-phon contour. Note that the A-weighting underestimates the perceived loudness for the 200-Hz tone and also for the combination. The B-weighting works better for the 200-Hz tone. Neither weighting procedure works for the combination, though.

The sone scale The loudness level is expressed in phon. However, loudness values expressed on this scale do not immediately suggest the actual magnitude of the sensation.

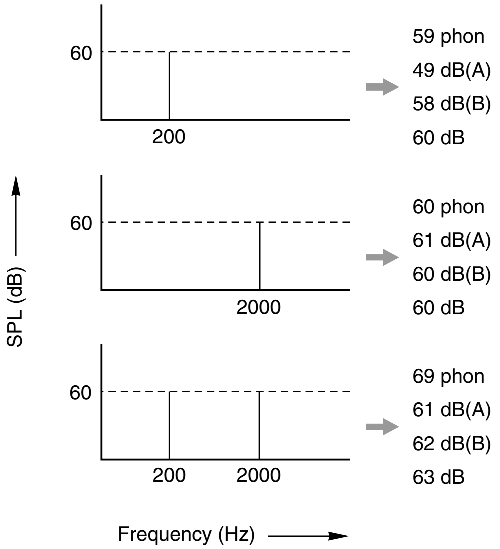


Figure 1.22 Levels of tones of 200 Hz (a), 2000 Hz (b), and the two tones simultaneously (c). The phon value is computed with a loudness model; dB(A) and dB(B) represent A- and B-weighted levels respectively. The lowest value in each part gives the acoustic SPL. From Aarts [4]

Therefore the sone scale, which is the numerical assignment of the strength of a sound, has been established. It has been obtained through subjective magnitude estimation using listeners with normal hearing. As a result of numerous experiments (Scharf [231]), the following expression has evolved to calculate the loudness S of a 1-kHz tone in sone:

$$S = 0.01 \times (p - p_0)^{0.6} \tag{1.90}$$

where $p_0 = 45$ μPa approximates the effective threshold of audibility and p is the sound pressure in μPa. For values $p \gg p_0$, Eqn. 1.90 can be approximated by the well-known expression

$$S = 2^{(P-40)/10} \tag{1.91}$$

or

$$P = 40 + 10 \log_2 S \tag{1.92}$$

where P is the loudness in phon.

ISO532A and 532B The ISO532A method is equal to the Mark VI version as described in Stevens [256]. However, Stevens refined the method, resulting in the Mark VII version

[257], which is not standardized. Here, the 532A method is discussed briefly. The SPL of each one-third octave band is converted into a loudness index using a table based on subjective measurements. The total loudness in sone S is then calculated by means of the equation

$$S = S_{\mathrm{m}} + F\left(\sum S_i - S_{\mathrm{m}}\right) \tag{1.93}$$

where S_{m} is the greatest of the loudness indices and $\sum S_i$ is the sum of the loudness indices of all the bands. For one-third octave bands the value of F is 0.15, for one-half octave bands it is 0.2, and for octave bands it is 0.3.

An early version of the ISO532B method is described in Zwicker [305] and later it has been refined, see for example, Zwicker and Feldtkeller [311], Paulus and Zwicker [205], and Zwicker [308]. The essential steps of the procedure are as follows. The sound spectrum measured in one-third octave bands is converted to bands with bandwidth roughly equal to critical bands. Each critical band is subdivided into bands of 0.1 Bark (which is an alternate measure of auditory filter bandwidth) wide. The SPL in each critical band is converted, by means of a table, into a loudness index for each of its sub-bands. In order to incorporate masking effects, contributions are also made to higher bands. The total loudness is finally calculated by integrating the loudness indices over all the sub-bands, resulting in the loudness in sone. The total loudness may be converted into loudness level in phon using Eqn. 1.92 (of course this can also be done for loudness as computed using method 532A).

Zwicker's method is elegant because of its compatibility with the accepted models of the human ear, whereas Stevens' method is based on a heuristic approach. Zwicker's procedure tends to give values systematically larger than Stevens'.

Time-varying loudness model The main drawback of both Stevens' and Zwicker's loudness models is that they are, in principle, only valid for stationary signals. This would seriously limit their applicability, but fortunately both models seem to correlate quite well with subjective judgements, even for realistic time-varying signals, see Sec. 1.4.4.4. Nonetheless, Glasberg and Moore [90] devised a method to predict loudness for time-varying sounds, building on the earlier models. The time-varying model was designed to predict known subjective data for stationary sounds, amplitude-modulated sounds, and short-duration (<100 ms) sounds. Broadly speaking, Glasberg and Moore's model is similar to Zwicker's, but loudness is temporally integrated to account for the time-varying nature of the signals. Specifically, the momentary excitation pattern generated by the sound at a specific time is used to compute the excitation pattern, and from this the 'instantaneous' loudness. This quantity is not consciously observable, but might correspond to total activity in the auditory nerve, for example. The instantaneous loudness is then 'smoothed' to obtain the short-term loudness, with a relatively fast attack and slower decay time. The short-term loudness is observable, for example, as would be perceivable for a 10-Hz amplitude-modulated signal. The short-term loudness is smoothed again, with larger time constants, to obtain the 'long-term' loudness. The long-term loudness corresponds to the overall loudness percept of the signal.

The model seems appropriate for use with audio signals, which are always time varying. The main limitation the authors mention is the fact that the relative phases of harmonics are not taken into account; the crest factor (peak-to-rms ratio) of waveforms on the BM can differ substantially for complex tones with identical power spectra but different phases, which may lead to loudness differences. This might be of some importance for BWE applications, where in some cases the harmonic structure of signals is modified. However, in practice, relative phases of harmonics are unpredictable, because of the randomizing effect of room reflections (unless the distance to the speaker is very small such that reflections are small compared with direct sound, or with headphone presentation).

1.4.4.3 Sensitivity to Changes in Intensity

Sensitivity to changes in intensity can be measured in several different ways, which mostly give similar trends. Results are usually expressed as the smallest detectable increment, or just-noticeable difference (JND), ΔI of a sound with intensity I, expressed as $\Delta L = 10\log([I + \Delta I]/I)$. For wideband noise, $\Delta L \approx 0.5$–$1.0\,\text{dB}$ over most of the dynamic range of the auditory system. Thus, $\Delta I/I \approx 0.13$–0.25; this ratio is called the *Weber fraction*. A constant Weber fraction implies that sensitivity to changes in the stimulus is proportional to the magnitude of the stimulus; this property is known as Weber's law. Weber's law does not hold for pure tones, where it has been found that sensitivity increases with increasing intensity. Much work has been done to explain how intensity variations are coded by the auditory system, and how to explain the intensity versus loudness curves for various signals, but no definite theory exists as yet; Moore [178] presents a review. Allen and Neely [19] present a model that does account for the intensity JND of pure tones and wideband noise, on the basis of the assumption that the intensity JND is related to the variance of an internal loudness variable.

1.4.4.4 Loudness Issues for Listening Tests

Although the BWE algorithms to be discussed later can be analyzed in objective ways, the ultimate quality test is of course through subjective experiments. For this, not only is the quality of the algorithms important but, perhaps, also the quality of the loudspeaker. The perceived sound quality of a loudspeaker and its relation to its various physical properties have been a subject of discussion and research for a long time, see for example, Toole [272, 273, 274, 275], Gabrielsson and Lindström [82], Tannaka and Koshikawa [263]. In this regard, it is important that reproduction levels are chosen appropriately for the various signals tested, in particular, if different loudspeakers are used in the same test. Although normally one would prefer to use the same experimental hardware throughout one listening test, there are situations where this is not desirable. For example, to evaluate certain low-frequency enhancement algorithms (discussed in Chapter 2), one might want to subjectively compare a processed signal reproduced on a flat-panel loudspeaker with an unprocessed signal on a high-quality electrodynamic loudspeaker. Loudness matching across loudspeakers is especially important as it is well known that a higher reproduction level, or loudness level, of a loudspeaker can lead to a higher appreciation score than that of another one of the same quality, or even the same loudspeaker. The importance of equal-loudness levels of the sounds being compared is shown by a striking investigation of

Illényi and Korpássy [116]. They found that the rank order of the loudspeakers, according to the subjective quality judgements, was in good agreement with the rank order obtained by the corresponding calculated loudness.

In Aarts [4], it was found that the ISO 532B method was the most suitable of the two ISO methods to adjust interloudness levels of loudspeakers, while the simple B-weighting gave the most satisfactory results of all the tested methods (both ISO and A–D weightings). The widely used A-weighting gave poor results, though (see related comments in Sec. 1.4.4.2). It was also found that loudness levels were hardly influenced by the choice of the repertoire, more specifically that a varied repertoire, on average, sounds equally loud, as was computed for pink noise. This considerably facilitates the computation of appropriate loudness levels if multiple loudspeakers are used.

1.4.4.5 Masking

Masking is defined by the American Standards Association as [20]

Definition 7 *The process, and amount, by which the threshold of audibility for one sound is raised by the presence of another (masking) sound. The unit customarily used is the decibel.*

Masking and frequency selectivity are intimately related; it has been known for a long time that a sound is masked most easily by another sound that has similar frequency components (Wegel and Lane [295]). In fact, Fletcher [72] assumed, in his studies of the critical bandwidth, that masking is only possible if the masker and masked signal (maskee) fall within the same critical band, even though it was known that masking is possible at greater frequency separations.

For BWE applications, masking may be of interest to consider the audibility of distortion components, which are generated by some of the algorithms, which is a form of tone-on-tone masking. This kind of masking is known as energetic, simultaneous masking. Energetic masking refers to the fact that the detection threshold is determined by the power spectra of masker and maskee (power spectrum model of masking); masking that cannot be explained by the power spectrum model is informational masking. It is thought that this involves higher-level (attentional) processes. Simultaneous masking refers to the fact that masker and maskee occur at the same time; masking is also possible if the masker precedes the maskee (forward masking), or if the maskee precedes the masker (backward masking). Both informational and non-simultaneous masking do not seem very relevant for BWE applications.

Masking effects have not generally been studied in relation to BWE methods; it might have some use in connection with audibility of distortion components that some of the algorithms generate. The energy of these unwanted components can be analyzed and compared with respect to the energy of desired frequency components, and quantified as a kind of ‘signal-to-noise’ ratio. This is then used to assess the performance of various algorithms; see for example, Sec. 2.3.2.1, and following sections, for such analysis. However, this ‘signal-to-noise’ ratio is a purely physical description that does not factor in any perceptual effects, such as masking. This implies that conclusions thus reached are to be considered with some caution. A better modelling of BWE performance could be achieved if masking effects were also considered.

1.4.5 PITCH

1.4.5.1 Factors Influencing Pitch

According to the American Standards Association [20] pitch is defined as follows

Definition 8 *Pitch: that attribute of an auditory sensation in terms of which sounds may be ordered on a scale extending from low to high.*

According to Moore [177], there are various ways how the pitch of a pure tone depends on its frequency. One can obtain a pitch–frequency relation by various methods, the classical result being the mel scale. It has an arbitrary pitch reference of 1000 mel at a frequency of 1000 Hz. A tone that sounds, on average, twice as high receives a value of 2000 mel, whereas a tone that sounds only half as high has a pitch of 500 mel. Although the mel scale suggests that the pitch of a pure tone is simply determined by its frequency, the perceived pitch also depends on some other factors, one of them being the intensity. If one measures for a group of subjects how, on average, the pitch of a pure tone changes with the tone's intensity, one typically finds that (1) for tones below 1000 Hz the pitch decreases with increasing intensity (about 15%), (2) for tones between 1000 and 2000 Hz the pitch remains rather constant, and (3) for tones above 2000 Hz the pitch rises with increasing intensity (about 20%). This effect varies considerably between listeners and also depends on the duration of the tone. Hartmann [105] found that the pitch of short-duration tones ($\approx$100 ms) with decaying envelopes is higher than the frequency of the tone. The upward pitch shift seems to increase with decreasing frequency, being 2.6% at a frequency of 412 Hz (the lowest frequency used by Hartmann). The shift at even lower frequencies could be considerably higher, although there is no data to support this hypothesis.

For a complex tone, consisting of more than one frequency component, the situation is more complicated. Pitch should then be measured by psychophysical experiments. A pitch that is produced by a set of frequency components, rather than by a single sinusoid, is called a residue. Even if in a harmonic complex the fundamental frequency is missing, it will still be perceived as a residue pitch, which in this case is sometimes called virtual pitch, because the frequency corresponding to the pitch is absent. There is a vast literature on pitch perception and residue pitch. Some of the earlier systematic investigations are described in Bilsen and Ritsma [33], de Boer [58], Houtsma and Goldstein [113], and Schouten [239, 240]). The fact that low-order harmonics need not be physically present to evoke a pitch percept at the fundamental[10] is an attractive option to enhance low-pitched sounds reproduced by (small) loudspeakers; in Chapter 2, we shall see how this can be exploited. One factor that remains unclear is the strength of the residue pitch at very low frequencies ($<$100 or 200 Hz); most investigators have looked at higher frequencies. Ritsma [225, 226] has studied the existence region above 200 Hz.

Repetition pitch Some noise-like sounds do evoke pitch sensations. An example was described by Huygens [115] in the seventeenth century. He noticed that the noise of a water fountain, reflected by marble stairs, produced a distinct musical pitch equal to that of an organ pipe whose length matched the depth of the stairs. He essentially discovered

[10] In fact, the residue pitch can be heard even if there is a masking noise present in the frequency region of the fundamental such that it would normally be masked (Licklider [161]).

that when one or more consistently delayed images of a sound interfere with the original sound, one hears a pitch that corresponds to the inverse of the delay. The original sound can be noise, music, speech, or just about any other sound. Because the frequency response characteristic of such a time-delay system has a periodic comb-like structure, this process is often referred to as comb filtering (Boff *et al.* [232], Bilsen [32]).

Special cases However, the clearest pitch sensations are evoked by sounds that are periodic or, equivalently, sounds that have line spectra of harmonically related frequencies. Most string and wind instruments produce near-periodic sounds and are therefore very efficient in conveying pitch information. Other instruments, such as bells or chimes produce line spectra with inharmonically related frequencies that evoke the ambiguous sensations, characteristic of these instruments. Still other instruments, such as the snare drum or cymbals, produce sounds with continuous spectra that evoke a sensation of noise without any pitch. Accordingly, these are instruments used for rhythmic rather than for melodic or harmonic purposes. Other pitch phenomena are edge pitch (Small and Daniloff [251], Kohlrausch and Houtsma [146]) – which refers to a weak pitch sensation evoked by low-pass or high-pass filtering noise with a sufficiently sharp spectral edge; adaptation pitch (Zwicker [306]) is heard when one is exposed to wideband noise with a spectral notch of about half an octave. A weak tonal afterimage is heard when the tone is abruptly switched off.

1.4.5.2 Sensitivity to Changes in Frequency

Sensitivity to changes in stimulus frequency is remarkably high for pure tones of moderate frequency. Around 500–1000 Hz, the *difference limen for frequency* (DLF) is about 0.2–0.3% (Sek and Moore [244]); this means that two pure tones differing in frequency by the DLF will be correctly discriminated 75% of the time (or some other threshold). At low and high frequencies the DLF increases, and above 4–5 kHz the DLF exceeds 1%. Also, for short-duration tones (<100 ms) the DLF increases, being roughly 5 times larger than the values quoted previously for tones of 6.25 ms duration (Moore [176]).

Sensitivity to a modulation in frequency, the frequency modulation difference limen (FMDL), is less frequency dependent. The FMDL is about 0.5–1% over most of the audible frequency range. The FMDL seems the more suitable value to use when considering if frequency differences will be perceived in continuous musical or speech sounds.

One can also study the detectability of deviations from a perfect harmonic relationship between partials in a complex tone, see for example, Moore *et al.* [179]. In Le Goff *et al.* [158], experiments were performed to study the effect of mistuning only the fundamental component in a harmonic complex tone. Subjects had to distinguish the complex tone with its lowest harmonic at the fundamental frequency from a complex with the lowest harmonic shifted in frequency. Thresholds were determined for fundamental frequencies of 60 and 100 Hz. Complexes had either a flat spectrum, or components were generated with a spectral slope of −5 or −10 dB/octave. Conditions with the second or both the second and third harmonic omitted from the complex were also included. In additional conditions, the complex was slowly amplitude modulated and/or presented with a simultaneous distracting sound. The results showed a range of detectability from better than 0.5 to 7%, depending on the various conditions. Presenting the sound with a spectral slope strongly lowers thresholds. Adding a distractor or applying amplitude modulation, both lead to higher thresholds. The aim of this study was to investigate the perceptual consequences of a special frequency mapping technique, that is used to exploit the high efficiency of a loudspeaker design (trading high efficiency at the resonance frequency for decreased efficiency at higher frequencies), described in Sec. 4.3. The more complex stimuli used in the subjective experiment were thought to mimic signals that would typically be reproduced through this special driver.

1.4.5.3 Pitch Theories

A number of theories have been developed to explain the pitch of complex tones, which led to what are now known as the place and periodicity theories of pitch perception. Among these are the early theories of Ohm [190] and Seebeck [243]. Others, including Helmholtz [108], assumed that if the fundamental frequency of a harmonic stimulus was absent, non-linear distortion in the middle ear could recreate that fundamental as a difference tone.

Place theories of pitch perception Briefly, the place theory of pitch perception assumes that the cochlea performs a spectral analysis of the sound and maps different frequency components along the BM (tonotopic organization, see Sec. 1.4.1); then the various locations where spectrally resolved partials are detected are used to derive a pitch percept. This implies that tones without resolved partials have no definite pitch. Several models for such pitch extraction have been put forward, some of which are:

- *Closest matching subharmonic of lowest partial (Walliser [293])*: The first step is to determine the stimulus envelope repetition rate; this corresponds to the frequency difference of the partials. Next, a subharmonic of the lowest partial is determined that is closest to the frequency found in the first step.
- *Most frequently occurring subharmonic*: Terhardt [266] elaborated upon Walliser's model, and assumed that each partial elicits a number of subharmonics that are pitch 'candidates'. For a complex tone, there will be one subharmonic that occurs most frequently, which determines the perceived pitch. Ambiguous pitch percepts are possible if several candidate subharmonics have roughly equal number of occurrences or if subharmonics cluster around multiple values.
- *Optimum processor (Goldstein [92])*: The estimated frequency values of the partials are fed into a kind of pattern recognition device, which first estimates the corresponding harmonic numbers and then derives the best-matching subharmonic to fit the observed partials; that subharmonic will be the perceived pitch. This model explains ambiguous pitch percepts by assuming that errors can be made in the estimation of harmonic number.

It is noted that in all of these models the use of place information is not an absolute requirement. The frequencies of the partials could be derived through timing information.

These models predict pitch values not only for 'normal' complex tones, but also for complex tones where each partial is shifted in frequency by a fixed amount, creating an inharmonic complex that does not have a 'proper' fundamental frequency. Although

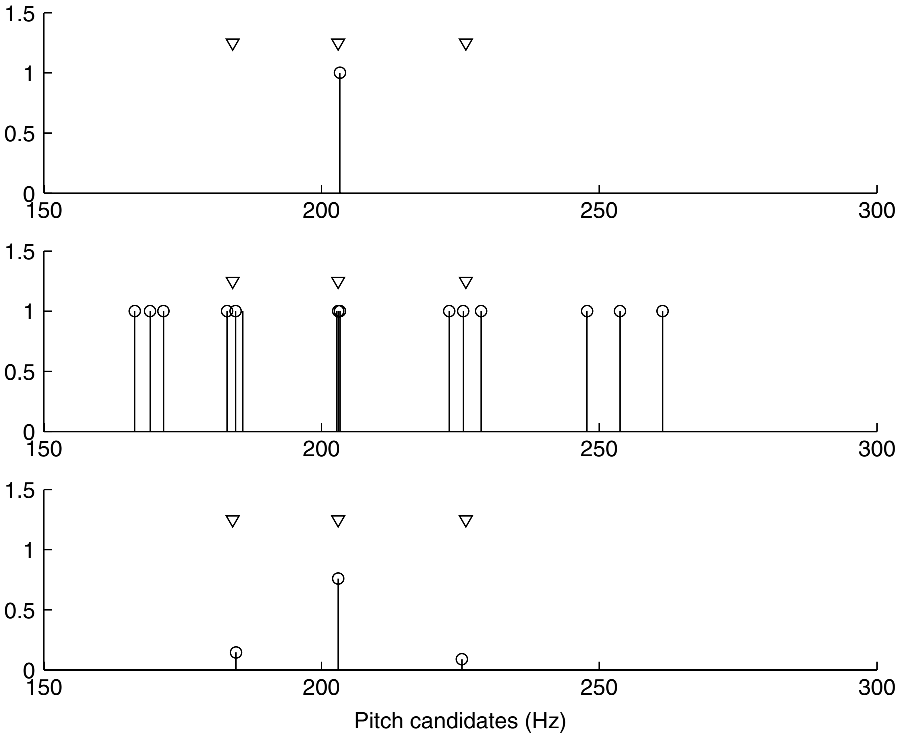


Figure 1.23 Predicted pitch percepts according to models of Walliser (a), Terhardt (b), and Goldstein (c), given a complex tone with partials at 1830, 2030, and 2230 Hz. The three arrows displayed in each part indicate frequencies at which subjects report pitch values; the pitch at 203 Hz is most strong, but percepts also occur at 184 and 226 Hz. Walliser's model gives only a single pitch prediction; Terhardt's computes subharmonics of the partials and predicts pitches for frequencies where subharmonics (nearly) coincide; Golstein's finds best-matching subharmonics for the observed partials. The amplitudes of the three predicted pitches in Goldstein's model are scaled proportionally to the inverse minimum mean-square error of that pitch value with reference to the observed partials; the sum of amplitudes equals 1, such that individual amplitudes can be interpreted as probabilities

such a shift does not alter the frequency spacing of the partials, the perceived pitch does shift. Schouten [240] found that a complex consisting of frequencies 1830, 2030, and 2230 Hz elicits a pitch of about 203 Hz; additionally, pitch matches around 184 and 226 Hz are also obtained. We investigate what pitch predictions the three pitch models will yield, with illustration thereof in Fig. 1.23. Fig. 1.23 (a) shows that the prediction by Walliser's model correctly finds a pitch of 203 Hz; it fails, however, to predict the alternate pitch values. Note that in all three parts the three arrows at frequencies 184, 203, and 226 Hz indicate the subjectively obtained pitch matches. Terhardt's model, shown in Fig. 1.23 (b), finds the most likely pitch at 203 Hz, because the subharmonics cluster most closely around that value. Pitch matches at 184 and 226 Hz are also correctly obtained, but with less likelihood, as the clustering is not as tight around these values. Fig. 1.23 (c) shows predictions from Goldstein's model. In this case, the amplitudes of the pitch matches indicate the probability of subjectively finding that particular pitch on any given occasion. For this, we have used an ad hoc metric that relates the probability of a pitch match to the inverse of the minimum mean-square error for that pitch match, given the observed harmonics. Goldstein's model also correctly matches the 203-Hz pitch, by assuming the harmonic numbers to be 9, 10, and 11; if the harmonic numbers are overestimated by 1, the pitch match at 184 Hz is predicted, and if the harmonic numbers are underestimated by 1, the pitch match at 226 Hz is predicted. Note that the pitch ambiguity that is observed in this example is primarily due to the absence of lower harmonics, which, if present, would give far less ambiguous predictions in either Terhardt's or Goldtein's model. This is also subjectively observed, and, in fact, it is believed that low-order harmonics (number $\approx 3-6$) are dominant with respect to pitch determination, even if the fundamental is physically present (Plomp [209], Ritsma [227]).

Periodicity theories of pitch perception In the periodicity, or temporal, theory the locations of the BM that are excited are not important (although the tonotopic organization of the cochlea *per se* is not disputed); rather, periodicities in neural activity are used to derive pitch information. Periodicities in neural activity are caused by the fact that neural response occurs preferentially during a specific phase of the BMM waveform (phase locking). The original temporal theory is mainly due to Schouten [239], who devised an ingenious theory that combined peripheral frequency analysis and central periodicity detection. According to his theory, the lower components of a harmonic complex are spectrally resolved in the cochlea (see Sec. 1.4.3) and each map into their own pitch. The higher components, which are not resolved, create a periodic interference pattern that reflects the periodicity of the waveform. This periodicity is detected by higher neural centers and maps into a sensation of (fundamental) pitch. This gave rise to the term residue pitch, because, according to Schouten, it results from the residue of spectral components that the cochlea fails to resolve (Boff *et al.* [232]). The actual pitch that is assigned to the sound is that pitch to which attention is mainly drawn; for complex tones this is generally the residue pitch. Note that the temporal theory can also account for ambiguous pitch of inharmonically related partials, as in the example of Fig. 1.23.

Numerous experiments have shown support for both theories. It appears that neither theory alone can account for all conditions, and as such it seems likely that both place and timing information can and are used for pitch perception. Moore [177] presents a qualitative model that incorporates both place and timing information and can account for all experimental data.

1.4.6 TIMBRE

Timbre is defined by the American Standards Association [20] as follows.

Definition 9 *Timbre: that attribute of an auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.*

This is, as has also been noted by others, a better description of what timbre *is not* than what it *is*. It is problematic to define exactly what timbre is because it does not appear to be a one-dimensional quantity. Timbre is known to depend on short-term power spectrum (or more properly, the excitation pattern), amplitude envelope (in particular, attack and decay time), and phase spectrum. Plomp and Steeneken [210] carried out a clever experiment to investigate the relative influence of power spectrum versus phase spectrum on timbre. They used signals $s(t)$ of the form

$$s(t) = \sum_{n=1}^{N} a_n \sin(2\pi f_0 t) + b_n \cos(2\pi f_0 t), \tag{1.94}$$

and found (using triadic comparisons and multidimensional scaling) that the largest timbral difference occurred between signals where all a_n (or b_n) were zero and signals where the a_n and b_n were alternately zero. They proceeded to investigate what changes in power spectrum would yield the same magnitude of timbral difference as was found for the phase effect. The a_n and b_n were chosen to decay by a fixed amount in dB per octave, varied between −4.5 and −7.5 dB, in 0.5-dB steps. The maximum influence of phase spectrum as discussed above could be matched by a change in the power spectrum slope of 2 dB, at $f_0 = 146.2\,\text{Hz}$. The influence of phase spectrum decreases for increasing f_0, being matched by a 0.7-dB change in the power spectrum slope for $f_0 = 584.4\,\text{Hz}$. Also, it appeared that phase spectrum and power spectrum influences were independent of one another. In conclusion, this experiment showed that phase spectrum does influence timbre, but less so than the power spectrum. In practice, when listening to signals over loudspeakers, the relative phases of frequency components become randomized because of room reflections, which means that the signal's original phase spectrum will be modified. It thus appears that it is not practical to try to control the phase spectrum, unless listening occurs via headphones. In BWE algorithms, timbre can thus be adjusted by modifying harmonic amplitudes, for example, through filtering.

Timbre of a sound is usually qualitatively described using several descriptors. One such descriptor that seems to be well linked to an objective parameter of the signal is 'brightness'; it is related to the relative amount of high frequencies versus low frequencies contained in the signal. Brightness can be quantatively described by the spectral centroid C_S, or power spectral center of gravity, as

$$C_S = \frac{\int f 10 \log S^2(f)\, \mathrm{d}f}{\int 10 \log S^2(f)\, \mathrm{d}f}, \tag{1.95}$$

given a signal with power spectrum[11] $S^2(f)$. If, for example,, the relative amplitudes of the harmonics of a complex tone are modified (as most of the BWE algorithms do), C_S will change, while the pitch remains the same. Equation 1.95 is only a first-order approximation of what the spectral centre of gravity is for the internal representation of

[11] It is more proper to use a compressed power spectrum (e.g. by taking the cube root), which corresponds better to the BMM. For simplicity, we keep using $S(f)$ in the remainder of the book. It should be noted that if spectral centroid is quantatively used for analysis purposes, a perceptually much more accurate version must be employed.

the signal; refinements can be made that incorporate knowledge of auditory processing, but that is beyond the scope of this chapter.

The temporal envelope, in particular, attack and decay time, has a large influence on timbre, but as BWE algorithms do not greatly modify the temporal structure of signals, we do not need to consider this in depth.

1.4.7 AUDITORY SCENE ANALYSIS

Auditory scene analysis (ASA) is a relatively new area of interest, but has experienced increasing attention since Bregman's [38] seminal book. Bregman defines ASA as the study of how the auditory system uses sensory information to form a mental representation of the world around us. The task of forming such a representation can be subdivided into many sub-tasks: how many sound sources are present, which frequency components belong together, what is the relative positioning of the sound sources? The relevance of such questions for BWE is that BWE algorithms typically take one part of an audio signal, process it, and add it back to the other part. After the final addition of signals, the auditory system should perceive the result as deriving from one sound source. This is not an academic problem; for instance, a short time delay between the onsets of two frequency bands of a signal can be enough to perceptually separate the two bands. We will introduce some terminology here and discuss a few of the relevant ASA principles that are thought to be important for BWE applications, following Bregman [38].

Bregman uses the following concepts:

- *(Auditory) stream*: Perceptual grouping of the parts of the neural spectrogram that go together; the perceptual unit that represents a single happening. A stream can consist of more than one sound, for example, a soprano singing with a piano accompaniment.
- *Grouping*: Formation of a stream from separate sensory elements. Grouping can occur across time (sequential integration) and across frequency (simultaneous integration). The opposite of grouping is segregation, where two or more streams are formed from the sensory elements.
- *Belongingness*: A sensory element has to belong to a stream.
- *Exclusive allocation*: A sensory element can only belong to one stream; it can not be used in more than one description at the same time.
- *Closure*: Perceived continuity of a stream even if the sensory elements are interrupted. For closure to occur, it is necessary that the interruption is 'plausible'. In hearing, a plausible interruption could be masking noise, whereas a mere silent gap would not constitute a plausible interruption, and thus would not yield a continuous stream.

As explained above, BWE algorithms could potentially create two streams if an audio signal is not processed properly; in that case, the processed part appears segregated from the unprocessed part. Because this concerns the grouping of simultaneous frequency components, we will investigate the factors influencing simultaneous integration and how they might apply to BWE processing. Sequential integration is of less concern, as BWE processing does not alter the temporal structure of the processed signal, and thus should not influence grouping along the temporal dimension.

Factors influencing simultaneous integration are:

- *Correspondence over time*: If a complex spectrum consists, in part, of a simpler spectrum that was present previously, the simple spectrum would appear to continue. The 'difference' spectrum would appear as a new tone ('old-plus-new' heuristic).
- *Harmonicity*: Components with a harmonic frequency relationship are grouped, and assigned a fundamental pitch.
- *Common fate*: Components that have correlated changes in frequency (frequency modulation, FM) or amplitude (amplitude modulation, AM) are grouped. Naturally produced speech and music sounds often exhibit micromodulation (slight modulations on small time scales) as well as modulations on a larger time scale. An extreme example of amplitude modulation is an onset or offset. The common AM during an onset is an especially strong grouping cue.
- *Spatial direction*: Components arriving from the same direction are grouped; in contrast, components arriving from different directions are segregated. This cue can be quite strong, although it can be acoustically ambiguous, which is why it breaks down easily if in conflict with any of the other factors mentioned earlier.

The tendency for simultaneous frequency components to group depends on how many of the factors mentioned here favor grouping versus segregation. The various factors reinforce each other, and grouping (or segregation) will be strongest if many or all of the factors are in 'agreement'.

For BWE applications, harmonicity and common fate principles seem of greatest importance. Music and speech signals abound with harmonic signals (although noise-like signals also occur), and this harmonic structure should be maintained as much as possible. Some of the BWE algorithms actively produce harmonics of input signal components, which promotes grouping. Because the production of harmonics occurs through non-linear processing, inharmonic (distortion) components can, in some cases, also be generated. Depending on the relative energy of harmonic versus distortion products, the distortion will be audible. The distortion will covary in amplitude with the harmonic components, and may therefore be grouped into the same stream. Of course, perceptible distortion should be avoided as much as possible, regardless of grouping or segregation. Another important result is that the processed signals should not be excessively delayed with respect to the unprocessed signal: the common fate principle implies that the delayed onset of the processed signal could segregate it from the unprocessed signal. Zera and Green [304] found that in some cases, delays on the order of milliseconds can lead to discriminable changes in perception (though this does not necessarily imply segregation). This will be discussed in more detail in Sec. 2.3.3.3, and solutions will be given to avoid such potential problems.

1.4.8 PERCEPTUAL MODELLING – AUDITORY IMAGE MODEL

The auditory image model (AIM) can be used to visualize internal representations ('auditory images') of sounds, and is described elaborately in Patterson *et al.* [203, 202]. Both functional and physiological modules can be selected that model the use of fine-grain

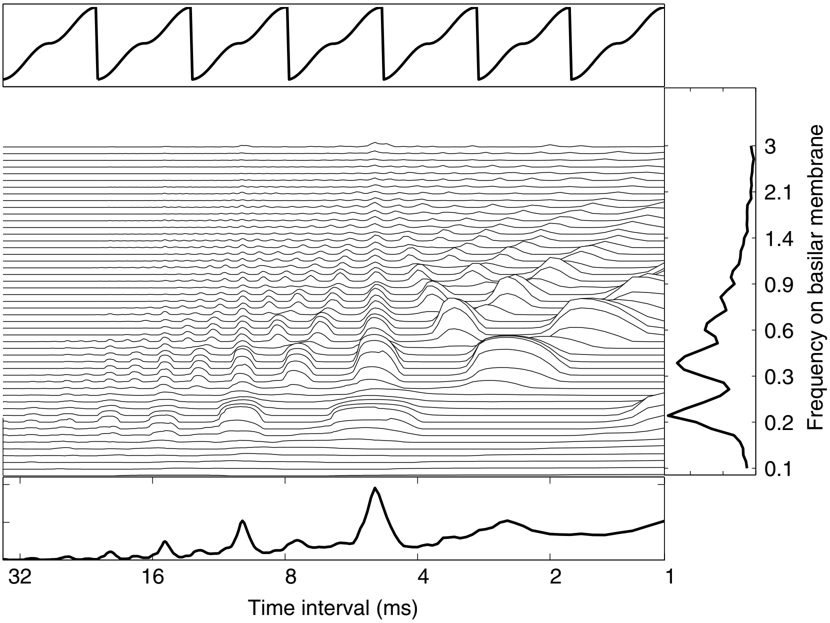


Figure 1.24 AIM calculation for a complex tone with $f_0 = 200\,\text{Hz}$ and a decaying harmonic spectrum (waveform shown in upper panel). The main panel displays the auditory image, as a function of time lag (ordinate) and frequency (abscissa), and is explained in the text. The lower panel is obtained by collapsing the auditory image over frequency, giving a lag-domain representation. The largest peak occurring at a lag of 5 ms indicates the perceived pitch at 200 Hz. The right panel is obtained by collapsing the auditory image over the lags, giving a frequency-domain representation. This clearly shows the harmonic structure of the signal, and also shows that low harmonics are better resolved than higher harmonics

timing information by the auditory system. In the following chapters, we will use AIM as a tool for predicting pitch percepts of complex tones.

The MatLab version of AIM was used (AIM-MAT); in all calculations, the functional (instead of physiological) modules of the package were used. The following processing stages are included:

- *Middle ear filtering*: The filtering described in Glasberg and Moore [90] is used.
- *Spectral analysis*: Here the response on the BM is computed in terms of BMM. The gammatone auditory filters (Eqn. 1.89 in Sec. 1.4.3) are used for this purpose.
- *Neural encoding*: First, there is a global compression of the BMM to allow a large dynamic range of sounds to be processed, as also occurs in the auditory system. Second, there is fast-acting expansion of the larger peaks in the compressed BMM response, which serves to enhance these presumably more important peaks. The last stage is a two-dimensional (over time and frequency) adaptive thresholding that serves to sharpen the resulting neural activity pattern (NAP).

- *Time-interval stabilization*: Because periodic sounds give rise to static, rather than oscillating, percepts, it is assumed that a temporal integration is performed on the NAP. To preserve the fine structure present in the timing of the NAP, a so-called 'strobed' temporal integration occurs. The integration commences when a peak in the NAP is encountered, after which the NAP input to the integrator decays in time.

The result of these processing steps is the auditory image, an example of which is shown in Fig. 1.24, where a complex tone with a 200-Hz fundamental was used. The auditory image can be collapsed over the time lags, to obtain an internal *spectral* representation of the signal (right panel of Fig. 1.24). Alternatively, a collapse over frequency provides a *temporal* representation of the signal (lower panel). Owing to the nature of the strobed temporal integration, this temporal representation is interpreted as a *lag* with respect to the last strobe: repetitions in the NAP show up as peaks in the auditory image at the appropriate lag values. Therefore, a signal with a temporal periodicity will exhibit peaks in the auditory image, which occur at a lag value equal to the periodicity interval; thereby predicting a pitch percept for the given signal at a frequency, which is the inverse of the lag. In Fig. 1.24, we find that the largest peak in the lag domain occurs at 5 ms, as we would expect on the basis of the 200-Hz fundamental frequency. We interpret the pitch strength as corresponding to the width of the peak, with wider peaks corresponding to the weaker pitch. If multiple peaks occur, the likelihood of pitch matches corresponds to the relative height of peaks. On the other hand, multiple peaks could indicate that the given signal segregates into two (or more) streams, each with their own pitch. As AIM does not incorporate auditory scene analysis principles, it does not predict which of these two possibilities applies to the given signal.