Chapter 4

Sound in space

Federico Avanzini

4.1 Introduction

If we think at the process of sound production in the light of the classic source-medium-receiver model of communication theory, we can say that in Chapters *Sound modeling: signal based approaches* and *Sound modeling: source based approaches* we have studied models for the source of sound signals. We now move a step further and examine the effects of the *medium* in which sound propagates, and the *receiver*, specifically a human receiver with two ears.

One of the most frequently effects produced during sound propagation in a medium is reverberation, which is caused by physical surfaces that partly absorb and partly reflect sound waves in air. We will first examine in Sec. 4.2 the physical and perceptual background of reverberation. The knowledge gained on these aspects will enable us to study some of the most known reverberation algorithms in Sec. 4.3. Finally we will review in Sec. 4.4 more recent approaches to synthetic reverberation, that are based on feedback delay networks and waveguide meshes.

A similar path will be followed in examining the receiver block. We will first examine in Sec. 4.5 how and to what extent a human receiver with two ears can gain information about the incoming direction and distance of an emitted sound, and what are the most relevant perceptual effects involved in *spatial hearing*. Armed with this knowledge we will address in Sec. 4.6 the most popular *3-D sound* processing techniques by which a virtual sound source can be positioned in some point of the space around a listener. We will in particular focus on *binaural techniques*, which assume that two independent sound signals are delivered to the two ears, e.g. through headphones.

4.2 Reverberation: physical and perceptual background

Almost any sound of our everyday life is produced in a reverberant environment, be it the office at work, the living room, or a concert hall. An emitted sound is therefore always accompanied by delayed versions, caused by reflecting surfaces and coming from many different directions. We talk about reverberation when the reflections occur soon after the emitted sound, so that they are not perceived as separate sound events, and instead have the effect of "coloring" the original sound and modifying its spatial characteristics. In this section we first review the physical process of reverberation, then we examine the most perceptually salient characteristics of reverberation. Having knowledge of both these aspects are essential in order to develop algorithms for synthetic reverberation.

4.2.1 Basics of room acoustics

For our purposes a room is a physical enclosure that contains an elastic medium (generally, air) through which acoustic disturbances can be propagated. It also has a boundary (the room walls) that limit the propagation of these acousticdisturbances. In this view a room is simply an acoustic resonator, similar to the string that we have examined in Chapter *Sound modeling: source based approaches*, but with at least two important differences: first, it is a 3-D resonator, because sound can propagates in all spatial directions, and second, its physical dimensions are much larger than typical dimensions of a string in a musical instrument. Put in another way, its physical dimensions are much larger than typical acoustic wavelengths.

4.2.1.1 Sound waves in a closed space

We have analyzed in Chapter *Sound modeling: source based approaches* the D'Alembert equation which describes sound propagation within a perfectly elastic medium. While the 1-D D'Alembert equation can be used to model strings or acoustic tubes, the 3-D equation describes sound propagation in space:

$$\nabla^2 p(\boldsymbol{x},t) = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}(\boldsymbol{x},t), \tag{4.1}$$

where $\boldsymbol{x}$ represents Euclidean coordinates in space and p is the acoustic pressure. The symbol $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ stands for the 3-dimensional Laplacian operator. As opposed to mechanical vibrations in a string or membrane, acoustic vibrations are *longitudinal* rather than transversal, i.e. the air particles are displaced in the same direction of the wave propagation. The constant c has the dimensions m/s of a velocity and indeed is sound velocity in air.

By adding suitable boundary conditions we can gain a description of waves of particle velocity within a three-dimensional enclosure. Let us start with the simplest possible 3-D enclosure, a rectangular room with perfectly smooth and rigid walls. More precisely, we define the domain $\mathcal{D}$ of the problem to be a parallelepiped with edges of length L_x, L_y, L_z:

$$\mathcal{D} = \{\boldsymbol{x} = (x,y,z);\ 0 \le x \le L_x,\ 0 \le y \le L_y,\ 0 \le z \le L_z\} \tag{4.2}$$

Let $\mathcal{B}$ be the boundary of $\mathcal{D}$, i.e. the rigid walls of the parallelepiped. The boundary conditions require the air velocity perpendicular to each wall to be zero on $\mathcal{B}$. Equivalently, if we consider acoustic pressure p then the conditions on the boundary are $\partial p/\partial \boldsymbol{x}(\mathcal{B}) = 0$. Then one can provide an analytical solution of Eq. (4.1) in terms of stationary modes of the kind

$$p(\boldsymbol{x},t) = s(\boldsymbol{x})q(t) \tag{4.3}$$

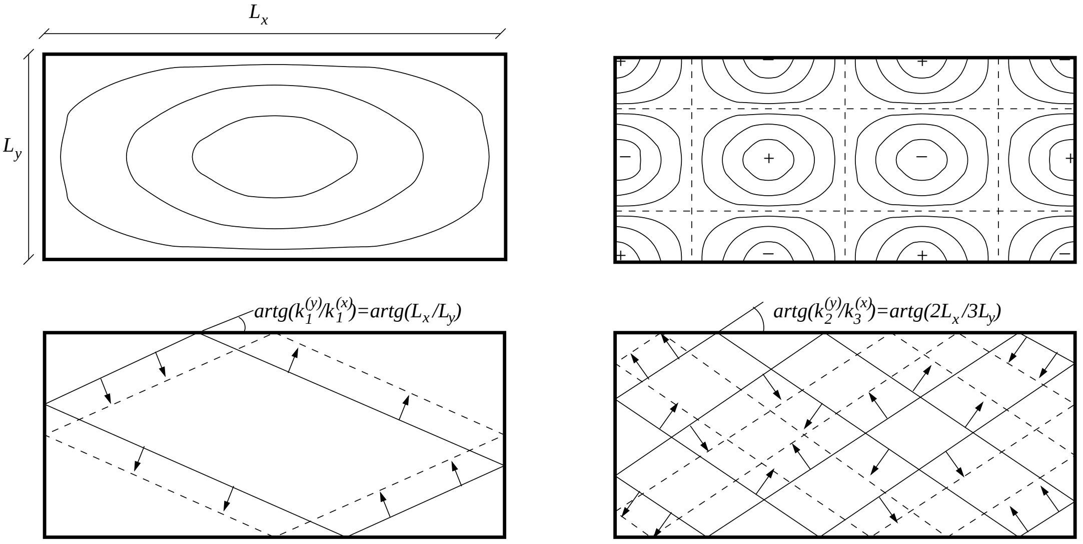

L_x
L_y
+
−
$artg(k_1^{(y)}/k_1^{(x)})=artg(L_x/L_y)$
$artg(k_2^{(y)}/k_3^{(x)})=artg(2L_x/3L_y)$

Figure 4.1: *Plane wave loops* $(1,1,0)$ *and* $(3,2,0)$*, as seen on the* (x,y) *plane.*

Following a line of reasoning analogous to the 1-D case , one can determine the spatial shape $s(\boldsymbol{x})$ of the modes as

$$s_{n,m,l}(\boldsymbol{x}) = \sqrt{\frac{2}{L_x}}\sqrt{\frac{2}{L_y}}\sqrt{\frac{2}{L_z}}\cos\left(k_n^{(x)}x\right)\cos\left(k_m^{(y)}y\right)\cos\left(k_l^{(z)}z\right), \tag{4.4}$$

where we can define the *wavenumbers* $\boldsymbol{k}_{n,m,l}$ as

$$\boldsymbol{k}_{n,m,l} = \left(k_n^{(x)}, k_m^{(y)}, k_l^{(z)}\right) \qquad \text{with} \quad k_n^{(x)} = \frac{n\pi}{L_x},\ k_m^{(y)} = \frac{m\pi}{L_y},\ k_l^{(z)} = \frac{l\pi}{L_z}. \tag{4.5}$$

Analogously to the 1-D case discussed for modal synthesis in *Sound modeling: source based approaches*, these functions are a orthonormal basis for the space $\mathsf{L}^2(\mathcal{D})$. The temporal part is subsequently derived as

$$q_{n,m,l}(t) = \cos\left(\omega_{n,m,l}t + \phi_{n,m,l}\right), \quad \text{with} \quad \omega_{n,m,l} = c\sqrt{\left[k_n^{(x)}\right]^2 + \left[k_m^{(y)}\right]^2 + \left[k_l^{(z)}\right]^2}. \tag{4.6}$$

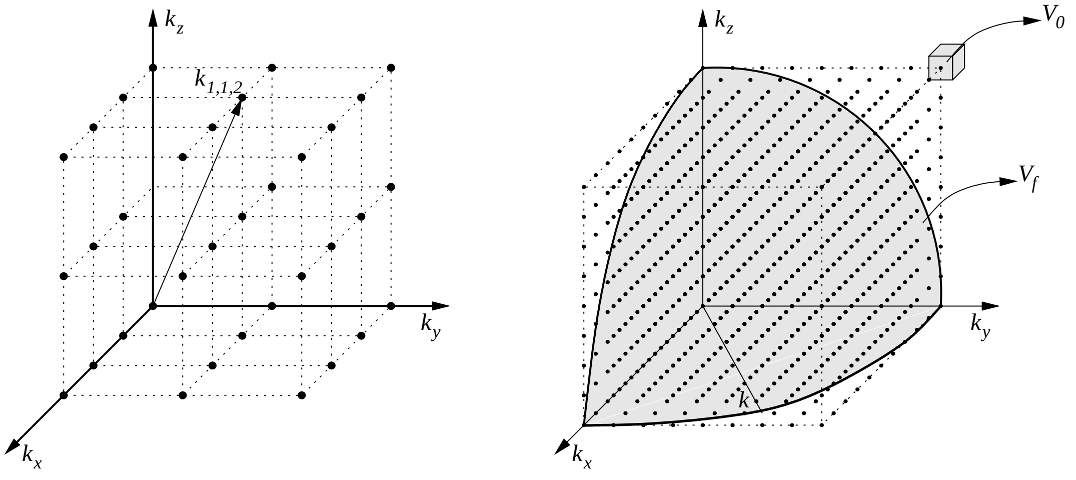
Differently from the 1-D case, and analogously to the 2-D case of the rectangular membrane, the frequencies $\omega_{n,m,l}$ are a non-harmonic series. However each of the three spatial directions (where only one of the three indexes (n,m,l) is varying) is associated to a harmonic subseries. Analogously to the 1-D case a mode (n,m,l) has nodal surfaces, which corresponds to the regions where $s_{n,m,l}(\boldsymbol{x}) = 0$. It is easy to see that these are planes parallel to the walls of the room.

A normal mode $p_{n,m,l}(\boldsymbol{x},t) = s_{n,m,l}(\boldsymbol{x})q_{n,m,l}(t)$ can be written as a superposition of waves traveling in different directions. This can be seen through multiple application of Werner formulas[1], which yelds

$$p_{n,m,l}(\boldsymbol{x},t) = \ldots = \sqrt{\frac{2}{L_x}}\sqrt{\frac{2}{L_y}}\sqrt{\frac{2}{L_z}}\sum\cos\left[\boldsymbol{k}_{n,m,l}^{\pm\pm\pm}\cdot\boldsymbol{x} \pm \left(\omega_{n,m,l}t + \phi_{n,m,l}\right)\right], \tag{4.7}$$

where we have defined $\boldsymbol{k}_{n,m,l}^{\pm\pm\pm} = (\pm k_n^{(x)}, \pm k_m^{(y)}, \pm k_l^{(z)})$, and where the summation has to be extended over the sixteen possible combinations of signs in the argument. This means that for each mode there are eight directions of wave propagation, each one associated to one $\boldsymbol{k}_{n,m,l}^{\pm\pm\pm}$ vector. Figure 4.1 visualizes the wavefronts for the modes $(1,1,0)$ and $(3,2,0)$: these result in plane wave loops having constant length.

[1] $2\cos\alpha\cos\beta = \cos(\alpha-\beta) + \cos(\alpha+\beta)$.



(a)

(b)

Figure 4.2: *Estimation of modal density; (a) distribution of wavenumbers on a regular point lattice, (b) estimation of the amount of wavenumbers contained in a spherical octant of radius k.*

4.2.1.2 Modal density

We now want to derive an estimate of the *modal density*, i.e. the average density of eigenfrequencies on the frequency axis.

From Eq. (4.5) one observes that the allowed values for the wave numbers k are distributed on a regular point lattice in the 3-D space depicted in Fig. 4.2(a). The number N_f of eigenfrequencies in the frequency range from 0 to f equals the number of lattice points contained in the sphere octant of radius $k = c \cdot 2\pi f$ depicted in Fig. 4.2(b). Therefore, $N_f = V_f/V_0$, where V_f is the volume of the sphere octant of radius k and V_0 is the average volume per lattice point. The former is one octave of the sphere volume, $V_f = \pi k^3/6$, while the latter can be estimated as the volume of the cube depicted in Fig. 4.2(b), whose edges have lengths $\pi/L_x, \pi/L_y, \pi/L_z$, respectively (recall Eq. (4.5) for the wavenumbers $\boldsymbol{k}_{n,m,l}$). Therefore $V_0 = \pi^3/V$, where $V = L_xL_yL_z$ is the room volume. One finally obtains

$$N_f = \frac{\pi k^3/6}{\pi^3/V} = \frac{4\pi}{3}V\left(\frac{f}{c}\right)^3. \tag{4.8}$$

The modal density is estimated as the derivative of N_f with respect to frequency:

$$D_f(f) = \frac{dN_f}{df} = \frac{4\pi V}{c^3}f^2 \tag{4.9}$$

In order to gain a quantitative understanding of these equations, let us consider a hypotetical medium-small auditorium with dimensions $(L_x, L_y, L_z) = (35, 20, 14)$ meters, which means $V = 9800$ m^3. From Eq. (4.8) we see that there are approximately 10^9 normal modes with frequencies between 0 and 10 kHz. From Eq. (4.9) we see that at 1 kHz the modal density per Hz is approximately 3500, which means that the average spacing between modes is less than 3×10^{-4} Hz.

4.2.1.3 Sound sources and room impulse responses

Let us now move from the mathematical analysis sketched in the previous sections towards a more realistic situation. First, we assume that a sound source is located within the domain $\mathcal{D}$. The distribution in space of the source is described by a continuous density function $\bar{f}(\boldsymbol{x})$, while the time-domain signal emitted by the source is described by a function $\bar{q}(t)$: this means that $\bar{q}(t) \cdot \bar{f}(\boldsymbol{x}) dV$ is the volume velocity of a volume element dV at time t.

As a second hypothesis, we consider complex, non-ideal, boundary conditions in which walls are not perfectly rigid and instead absorption occurs. This can be restated by assuming that the normal modes have now complex eigenvalues $k_{n,m,l}$:

$$k_{n,m,l} = \omega_{n,m,l}/c + j\delta_{n,m,l}/c, \qquad \delta_{n,m,l} \ll \omega_{n,m,l}. \tag{4.10}$$

We want to find the solution of the wave equation in $\mathcal{D}$ under these two hypotheses. The wave equation in the presence of a sound source can be written as

$$\nabla^2 p(\boldsymbol{x}, t) = \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2}(\boldsymbol{x}, t) - \rho_{air} \bar{f}(\boldsymbol{x}) \frac{d\bar{q}}{dt}(t). \tag{4.11}$$

Since the $f_{n,m,l}$ functions of Eq. (4.4) are a basis for $\mathsf{L}^2(\mathcal{D})$, we can project both the source density function $\bar{f}$ and the solution p of Eq. (4.11) on this basis:

$$\bar{f}(\boldsymbol{x}) = \sum_{n,m,l} \bar{F}_{n,m,l} f_{n,m,l}(\boldsymbol{x}), \qquad P(\boldsymbol{x}, s) = \sum_{n,m,l} P_{n,m,l}(s) f_{n,m,l}(\boldsymbol{x}). \tag{4.12}$$

Note that in the second of the above equations we have implicitly assumed to work in the Laplace domain instead of the time domain. If we can find the unknown coefficients $P_{n,m,l}(s)$ as functions of the known coefficients $\bar{F}_{n,m,l}$, then we have the solution $P(\boldsymbol{x}, s)$ or equivalently $p(\boldsymbol{x}, t)$. If one inserts both series into Eq. (4.11) the result is

$$P_{n,m,l}(s) = s\rho_{air} c^2 Q(s) \frac{\bar{F}_{n,m,l}}{s^2 + c^2 k_{n,m,l}^2}. \tag{4.13}$$

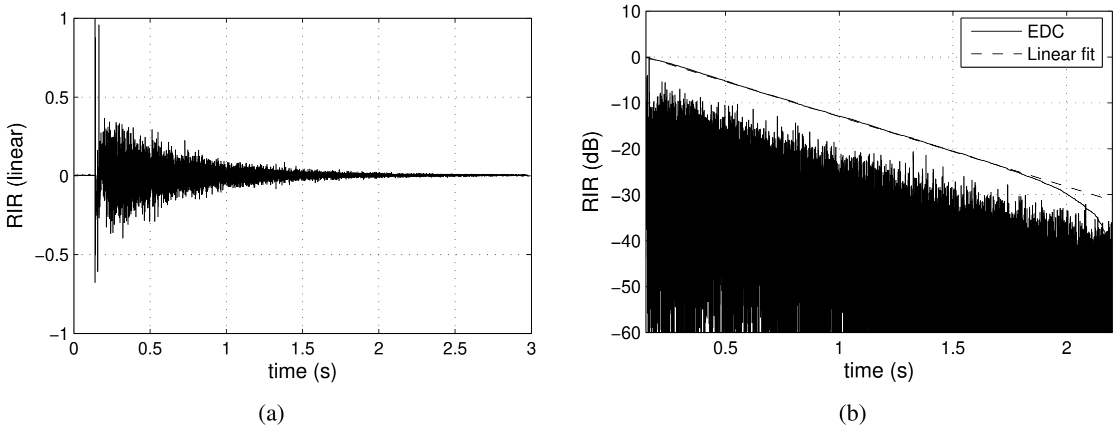
We can find a solution $P(\boldsymbol{x}, s)$ if we consider the special case of a point source located at a certain point $\boldsymbol{x}_0$ of the room and emitting an impulsive sound signal. Under this assumption one has $\bar{f}(\boldsymbol{x}) = \delta_D(\boldsymbol{x} - \boldsymbol{x}_0)$, where the function $\delta_D(\cdot)$ is the Dirac delta. This implies that the coefficients $\bar{F}_{n,m,l}$ are in this case $\bar{F}_{n,m,l} = f_{n,m,l}(\boldsymbol{x}_0)$. Moreover, if the sound source is emitting an impulse $\bar{q}(t) = \delta(t)$, then the corresponding spectrum is $Q(s) = 1$. If one substitutes the coefficients (4.13) into the second of Eqs. (4.12), the result is

$$P(\boldsymbol{x}, s) := H_{x_0,x}(s) = s\rho_{air} c^2 \sum_{n,m,l} \frac{f_{n,m,l}(\boldsymbol{x}) f_{n,m,l}(\boldsymbol{x}_0)}{s^2 + c^2 k_{n,m,l}^2}. \tag{4.14}$$

This is the acoustic pressure generated in $\boldsymbol{x}$ by a point source located at $\boldsymbol{x}_0$ and emitting an impulse. If we take the inverse Laplace transform, $h_{x_0,x}(t) = \mathcal{L}^{-1}\{H_{x_0,x}\}(t)$, this is what we call a *Room Impulse Response (RIR)*, measured at point $\boldsymbol{x}$ after an impulse emitted in $\boldsymbol{x}_0$. Equation (4.14) is telling us that the RIR is a superposition of numerous second-order resonant systems, each with center frequency very close to $\omega_{n,m,l}$ and damping constant very close to $\delta_{n,m,l}$:

$$h_{x_0,x}(t) = \begin{cases} 0 & t < 0 \\ \sum_{n,m,l} A_{n,m,l}(\boldsymbol{x}_0, \boldsymbol{x}) e^{-\delta'_{n,m,l} t} \cos(\omega'_{n,m,l} t + \phi_{n,m,l}) & t \geq 0 \end{cases} \tag{4.15}$$

The function $h_{x_0,x}(t)$ completely describes the room response for a source in $\boldsymbol{x}_0$ and a receiver in $\boldsymbol{x}$: if the emitted sound is not an impulse but a generic signal $\bar{q}(t)$, then the response will be –as usual– the convolution of the signal with the impulse response: $s(t) = [\bar{q} * h_{x_0,x}](t)$.



(a)

(b)

Figure 4.3: *Room Impulse response and reverberation time: (a) RIR of a very reverberant environment (a cathedral); (b) a portion of the same RIR in dB, together with its EDC and a linear fit.*

Now in normal rooms damping constants typically lie between 1 and 20 Hz: this justifies the assumption of very small δ coefficients in Eq. (4.10), and moreover tells that half-widths of these resonant systems are of the order of 1 Hz. If we compare this finding to the modal density estimate given in Eq. (4.9), we see that the average spacing of eigenfrequecies is smaller by several orders of magnitude than half-widths. Therefore each single resonant peak always covers many others and it is practically impossible to excite a single room resonance e.g. with a sinusoidal signal.

4.2.1.4 Reverberation time

From Eq. (4.15) we see that room reverberation adds a decaying tail to a source signal. One of the most important parameters derived from this equation is the *reverberation time* T_r, which is defined as the time required for the sound pressure to decay 60 dB. Clearly T_r is related to the absorption coefficients $\delta_{n,m,l}$. An approximate description of T_r can be derived as follows.

Given a source signal $\bar{q}(t)$ in $\boldsymbol{x_0}$, the resulting room response $s_x(t)$ at a point $\boldsymbol{x}$ will have the form

$$p(\boldsymbol{x},t) = [\bar{q} * h_{x_0,x}](t) = \ldots = \sum_{n,m,l} c_{n,m,l} e^{-\delta'_{n,m,l}t} \cos(\omega'_{n,m,l}t + \psi_{n,m,l}) = \sum_{n,m,l} c_{n,m,l} s_{n,m,l}(t), \tag{4.16}$$

where the $c_{n,m,l}$'s and the $\psi_{n,m,l}$'s will vary depending on the signal $\bar{q}$, and where we have introduced the notation $s_{n,m,l}(t) = e^{-\delta'_{n,m,l}t} \cos(\omega'_{n,m,l}t + \psi_{n,m,l})$ for brevity. The energy density of the response (or actually a quantity proportional to the energy density) is obtained by squaring $s(t)$:

$$w(t) = [s(t)]^2 = \sum_{n,m,l} \sum_{n',m',l'} s_{n,m,l}(t) s_{n',m',l'}(t). \tag{4.17}$$

We can derive an estimate of how $w(t)$ decays by averaging $w(t)$ over time and exploiting the circumstance that the exponential terms vary slowly (as the δ's are small). By averaging the cosine products only, the products with $(n,m,l) \neq (n',m',l')$ cancel on the average, and the products with $(n,m,l) = (n',m',l')$ give a value $1/2$. If one makes the further assumption of nearly uniform damping, i.e. $\delta_{n,m,l} \sim \delta_0$, then we obtain the following result:

$$\langle w(t) \rangle = \sum_{n,m,l} c^2_{n,m,l} e^{-2\delta_{n,m,l}t} \sim e^{-2\delta_0 t} \sum_{n,m,l} c^2_{n,m,l}. \tag{4.18}$$

This equation tells that for uniform damping the energy of the reverberation tail decays exponentially. In particular the reverberation time T_r is in this case derived as

$$-60 = 10\log\left(e^{-2\delta_0 T_r}\right), \qquad \Rightarrow T_r = \frac{6.91}{\delta_0}. \tag{4.19}$$

In general one cannot assume uniform damping, and as a consequence T_r is a function of frequency. However, the reverberation level falls in many practical cases in a fairly exponential fashion and therefore an overall reverberation time T_r can be defined and measured. Figure 4.3(a) shows a RIR measured in a very reverberant environment, a chatedral. Note that, apart from the initial spikes, the overall decay is fairly exponential.

The accuracy with which T_r can be be determined directly from RIR signals is in general severely limited by random fluctuations in the decay curves, which result from mutual beating of normal modes of different frequencies at the moment the excitation signal ceases. Instead T_r is more reliably estimated by looking at another function, the *Energy Decay Curve* (*EDC*, also called the Schroeder integral for historical reasons), defined as follows:

$$EDC(t) = \int_t^{\infty} h^2(\tau)d\tau, \tag{4.20}$$

where $h(t)$, is a RIR. The value $EDC(t)$ provides a measure of the reverberation energy that is left in the RIR at time t. The advantage is that this function has a much more regular behavior than $h(t)$, therefore T_r can be determined by fitting the decay of $EDC(t)$ through linear regression (on a dB scale), and looking at the time needed for this linear fit to drop by 60 dB. Figure 4.3(b) shows this procedure applied to the RIR of Fig. 4.3(a). From the linearly fitted EDC one can see that T_r is in this case close to 4 s, which is a quite large value as one would expect in a cathedral.

M-4.1

Write a function that computes the reverberation time T_r given a signal representing a RIR.

M-4.1 Solution

```
function [Tr,edc] = revtime(rir,Fs,ti,tf);
%Returns an estimate of Tr and of the edc; rir is a row vector representing
%a RIR, ti and tf are the initial and final times on which edc is computed

rir=rir(round(ti*Fs):round(tf*Fs));  % chunk the RIR on the interval [ti,tf]
edc =  10*log10(fliplr(cumsum(fliplr(rir.^2)))); % compute EDC
edc = edc-max(edc);                              %normalize at 0 dB

%linearly fit edc in the first half and find Tr as
%the instant where the linear fit drops below -60dB
c = polyfit(1:round(length(edc)/2),edc(1:round(length(edc)/2)),1);
Tr = (-60-c(2))/(c(1)*Fs);   % edc=c2 +c1*Fs*t;   edc=-60 => this formula
```

The choice of the initial and final instants is critical: they have to cover the range after the initial impulse and where the decay is almost linear. Moreover we have decided to estimate T_r on the first half of the EDC curve only, in order to avoid the error of the last EDC samples. Of course many techniques exist to choose these parameters automatically, this is just a toy example.

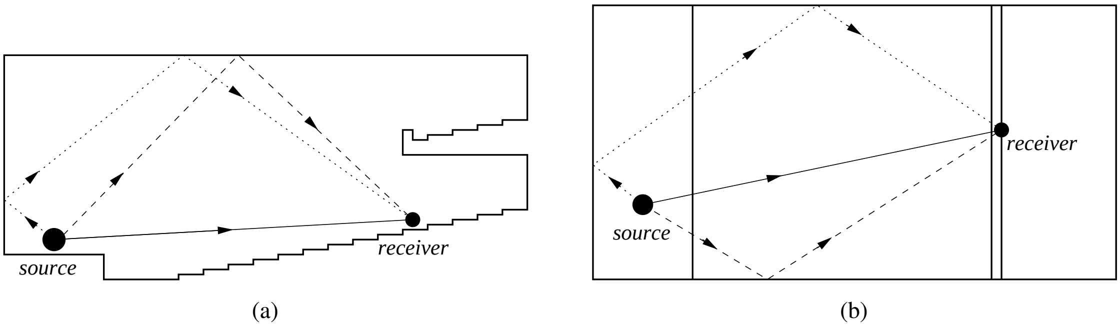


Figure 4.4: *Acoustic rays from a source to a receiver (a) in a vertical room section and (b) in a horizontal room section. Solid lines represent the direct sound, dashed lines represent first-order reflections, dotted lines represent second-order reflections.*

4.2.1.5 Geometrical room acoustics

Few results of practical use are obtained from direct manipulation of the D'Alembert equation, as we did in the previous sections. This is especially true when we consider rooms of arbitrary shapes instead of parallelepipeds: in that case even the computation of a single normal mode can become extremely difficult. An alternative description can be employed if we consider extremely high acoustic frequencies. In this limit, the concept of sound waves can be replaced by the concept of *acoustic rays*. By sound ray, we mean a vanishingly small portion of a spherical wave emitted by a point source in a room. This ray has well-defined direction and velocity of propagation, and conveys a total energy which remains constant (provided that it propagates within an ideal medium with no losses).

This simplified description based on acoustic rays takes the name of *geometrical acoustics* and has strict similarities with geometrical optics, although typical wavelengths and propagation velocities are very different in the two cases. Note that the assumption of extremely high frequencies is practically met in many cases of interest in room acoustics: a frequency of 1 kHz corresponds to a wavelength of approximately 34 cm, which is one or two orders of magnitude smaller than typical linear dimensions of rooms, as well as typical distances traveled by sound waves in a room.

Similarly to an optic ray, an acoustic ray that strikes a plane surface is reflected according to the following principles: *(a)* the reflected ray remains in the plane identified by the incident ray and the normal to the surface, and *(b)* the angles of the incident and reflected rays with the normal are equal. Figure 4.4 shows a room with a non-trivial shape (something like an auditorium), in which we have positioned a sound source and a receiver. All the paths from the source to the receiver can be characterized according to the number of reflections involved. The single source-receiver path with 0 reflections is the *direct sound*, and is followed by a small number of *first-order reflections* that involve one reflection on the room boundary, a larger number of *second-order reflections* that involve two reflections, and so on. In Fig. 4.4 we have drawn two examples of first- and second-order reflections.

Geometrical room acoustics can be used to provide a qualitative description of a RIR. Assume that an ideal impulse shot from a point source reaches a receiver at time $t = 0$. Each reflected ray will then arrive with a certain time delay and also with a certain attenuation, which depends on the path length (absorption in the medium) and on the number of reflections (wall asbsorption). The first reflections are strong and sporadic, but the temporal density of reflections increases rapidly while the average reflection energy decays accordingly. A qualitative reflection diagram is given in Fig. 4.5. Except for the first few isolated reflections, the weaker and denser reflections arriving at later times merge into a unitary percept.

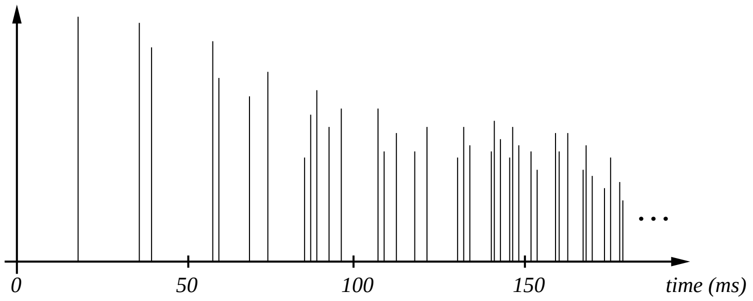


Figure 4.5: *Schematical room response to an ideal impulse: the time axis is relative to the direct sound, which reaches the receiver at* $t = 0$.

This description of room reverberation as a temporal sum of reflected rays is complementary to the view of reverberation as the sum of free decaying normal modes.

We now want to derive an estimate of the temporal structure of reflections. To this end we employ the usual prototype room, i.e. the parallelepiped, and we introduce the concept of *image sources*. If the reflecting surface is a plane the reflection of a sound ray can be simulated by constructing an *image source*. This process is illustrated in Fig. 4.6(a): given a sound source A and a receiver B, the path of a reflected ray r from the wall to B is the same path of the direct ray r' emitted by the *image source* A'. The process can be iterated in order to take into account higher-order reflections, and results in the construction of a grid of image sources that replace the walls altogether.

Now suppose that at time $t = 0$ all the sources emit an impulse. During the time interval from t to $t + dt$, the impulses that reach a receiver in the center of the room are those emitted by image sources whose distance from the receiver lies between cdt and $c(t + dt)$. These sources are located within the spherical shell with radius ct, thickness cdt, and volume $4\pi c^3 t^2 dt$ illustrated in Fig. 4.6(b). Therefore the volume V of an image room is contained $4\pi c^3 t^2 dt/V$ times in the spherical shell, and for t large enough (i.e. for high reflection densities) this number coincides with the number dN_r of image sources contained in the shell. The temporal density of reflections arriving at time t is then

$$D_r(t) = \frac{dN_r}{dt}(t) = 4\pi \frac{c^3 t^2}{V}. \tag{4.21}$$

One could show that this result applies not only to a parallelepiped but to rooms of arbitrary shapes.

4.2.2 Perceptual reverberation parameters

In the previous section we have analyzed reverberation from a purely physical point of view. However in many applications it is important to correlate physical measurements to subjective judgements of acoustical quality, obtained from psychophysical experimentation. This is especially true in the domain of concert hall acoustics, where researchers have tried to isolate the objective parameters that are most relevant in determining the perception of acoustical quality of a hall. Subjective attributes are typically derived from perceptual experiments with musicians and listeners, who answer to detailed interviews, and subsequent comparison of the results with measured objective parameters.

In this section we enter, for the time in this book, the domain of psychoacoustics, and review some of the subjective attributes most commonly used in establishing the acoustical quality of reverberant environments. The literature on this topic is vast and the terminology is not always fully consistent, therefore we try to cluster together similar or equivalent concepts whenever possible.

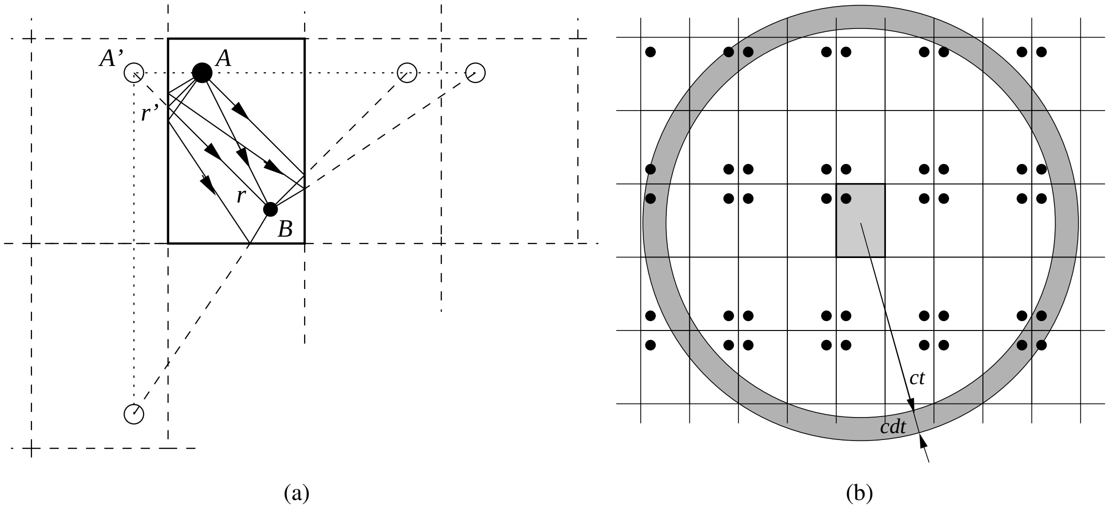


Figure 4.6: *Estimation of temporal reflection density through the image source method; (a) construction of two first-order and two second-order reflections, and (b) estimation of acoustic rays reaching a receiver within the time interval* $(t,\ t+dt)$.

Clearly the perceptual attributes of reverberation are of great importance also for the design of reverberation algorithms. The ultimate goal is to determine an orthogonal set of subjective attributes, using e.g. multidimensional scaling techniques, and provide reverberation algorithms with a set of knobs each of which controls a different perceptual attribute. A fundamental problem with this kind of approach is that the number of perceptual dimensions is not known *a priori*, and moreover it is hard to assign assign relevance to dimensions that are added.

4.2.2.1 Reverberance

We have defined the reverberation time T_r in Sec. 4.2.1 as the time required for the sound pressure to decay 60 dB.[2] This is one of the most important parameters for the perception of the *reverberance*, i.e. the property of the environment of adding fullness and loudness to a dry sound, and of giving the listener a sense of being enveloped by the sound. Some use the term “liveness” to refer to a similar concept, and by contrast call “dead” an environment that is not reverberant.

We have already mentioned that T_r is in general a function of frequency, because absorption in materials is typically higher at higher frequencies. A confirmation of this is given in Figure 4.7, which shows a waterfall representation of a RIR: one can see that each frequency bin decays with a different rate. This dependence of T_r on frequency is also important perceptually. In general the mid-frequency T_r can be considered to be the best measure of the overall reverberant characteristics of a room.

Clearly the audibility of reverberation depends greatly on the sound source. For music or speech, the early portion of the reverberant decay contributes more to the perception of reverberance than does late reverberation, because it is audible during pauses and gaps between notes, syllables, and words. For this reason an *early decay time* (*EDT*) parameter is also used as a complementary measure of reverberance. The EDT is defined as the time required for the sound pressure to decay from 0 to -10 dB, multiplied by a factor of 6 (which merely serves to facilitate comparison with T_r).

[2] An alternative and not completely equivalent definition commonly used in the domain of concert hall acoustics is the following: T_r is the time required for the sound pressure to decay from -5 to -35 dB, multiplied by a factor of 2.

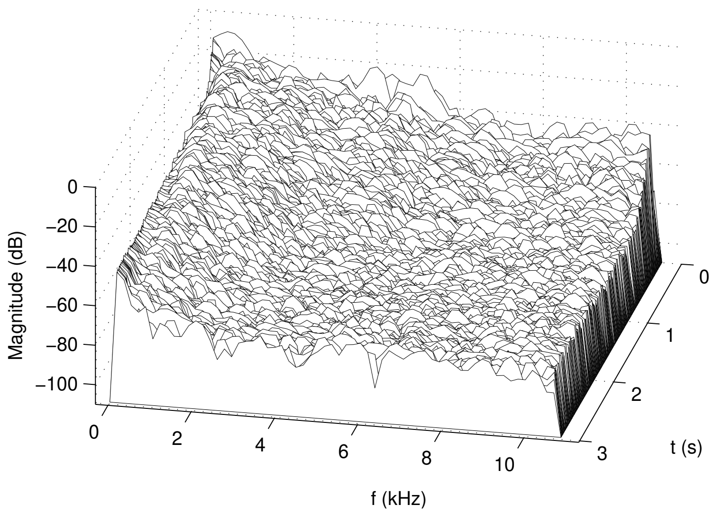


Figure 4.7: *Waterfall representation for the RIR of Fig. 4.3.*

One might wonder what is the “optimal” T_r for a reverberant environment. The answer depends first of all on the source signal: in the case of speech a relatively short T_r is generally preferred, since when listening to speech we generally want to understand what the speaker is saying and thus we need to perceive each element of the sound signal. For music on the contrary, a longer T_r can make a the listening experience more pleasant by masking small imperfections and blending musical sounds. Given this remark, it is not surprising that T_r’s in (good) concert halls are usually in the range 1.8 to 2.2 s, while in opera houses values are usually in the range 0.9 to 1.5 s because the listener has to be able to enjoy the music as well as to understand the text. Note however that T_r’s of renowned opera theaters are more scattered than those of equally renowned concert halls.

4.2.2.2 Early reflections and spatial impression

The subjective attribute of *spatial impression* refers to the sense of a listener of being in close communication with the sound source and sorrounded by the sound. Other terms often found in the literature and referred to similar concepts are spaciousness, envelopment, ambience, apparent source width. Subjective judgements about this property appear to be strongly correlated to the structure of the early reflections of the environment, with two elements being of specific importance.

A first commonly accepted result is that the degree of spatial impression depends on the *initial time-delay gap* t_I, the difference in arrival times between the direct sound and the first reflection. A lack of early reflections (i.e. a long t_I) has the effect of making the sound source perceived as remote and disconnected from the listener, while a short t_I provides the desired sense of envelopment. Some studies suggest that a parameter t_I defined as above becomes useless if the first reflection is much weaker than the following ones.

A second physical correlate of spatial impression is the fraction of lateral energy to the total energy within the early reverberation: a significant amount of *lateral* early reflections, i.e. reflections coming from the sidewalls, provides the listener with the impression of being enveloped by the sound. A quantitative estimate of this property is the so-called *lateral energy fraction* LF, defined as

$$LF_t = \frac{\int_0^t h_{lat}^2(\tau)d\tau}{\int_0^t h^2(\tau)d\tau}, \tag{4.22}$$

where $h(t)$ is the room impulse response measured with an omnidirectional microphone while $h_{lat}(t)$ is the one measured with a dipole microphone (with null axis facing forward this captures lateral energy in the $\pm 20^\circ$ $\pm 90^\circ$ range). A typical integration time is $t = 80$ ms.

The LF_t measure has been superceded by another parameter, the *early interaural cross-correlation coefficient* $IACC_E$. Let us first define the interaural cross-correlation function $IACF(t)$ as

$$IACF(t) = \frac{\int_{t_1}^{t_2} h^{(l)}(\tau)h^{(r)}(\tau+t)d\tau}{\sqrt{\int_{t_1}^{t_2}\left[h^{(l)}(\tau)\right]^2 d\tau \cdot \int_{t_1}^{t_2}\left[h^{(r)}(\tau)\right]^2 d\tau}}, \tag{4.23}$$

where $h^{(l),(r)}(t)$ are the so-called Head-Related Impulse Responses at the entrance of the left and right ear canals, respectively (measured e.g. with a "dummy-head" such as those described later on in Sec. 4.6.1), with the listener facing the sound source. Therefore $IACF(t)$ is a *binaural* attribute of reverberation, while all the parameters previously examined in this section are *monoural* attributes.

The interaural cross-correlation coefficient $IACC$ is the maximum of $IACF(t)$ in a range ±1 ms:

$$IACC = \max_{t\in(-1,1)\cdot 10^{-3}} |\, IACF(t) \,| . \tag{4.24}$$

In particular, if the integration times $t_1 = 0$, $t_2 = 80$ ms are used then the above equations provide the early interaural cross-correlation coefficient $IACC_E$. This is a measure of the similarity of the sound signals arriving at the two ears during the first 80 ms. If the sounds are equal then $IACC_E = 1$, while if they are two independent random signals then $IACC_E = 0$. The $IACC_E$ is a measure of spatial impression because is scales with the fraction of lateral early reflections arriving at the ears: as the number of reflections from outside the median plane increases, the $IACF(t)$ function broadens and consequently $IACC_E$ takes smaller values.

In concert halls initial time-delay gap t_I and the amount of lateral energy are correlated parameters. Measures of t_I in real concert halls show a high correlation of this parameter with the hall width: in a narrow hall it can be shorter than 30 ms, while in a wide hall it can be longer than 50 ms. On the other hand, the hall width is clearly correlated with the fraction of lateral energy arriving at the listener, which will increase as the hall narrows. It is a common finding in the literature of concert hall acoustics that subjective rankings of the acoustic quality of halls scale with their width.

As a final remark, it has to be noted that the perception of spatial impression is largely independent of the reverberation time: halls with similar T_r values but different t_I and $IACC_E$ values sound very different from each other. This finding support the commonly accepted assumption that early reflections and late reverberation play rather separate roles in the perception of reverberation.

4.2.2.3 Clarity

The subjective attribute of *clarity* refers to the "transparency" of a reverberant environment. If the source signal is music, then clarity is associated to the ability of a listener to perceive musical details, while if the source signal is speech then clarity correlates to speech intelligibility. An alternative term which is sometimes found in the literature is that of distinctness.

Single reflections of a reverberant environment are not perceived as individual events, except for exceptional (and generally undesirable) cases. Roughly speaking, early reflections have the effect of making the sound source appear more extended and to increase the apparent loudness of the direct sound. On the contrary, reflections arriving with longer delays are considered to be detrimental for the transmission of information, since they cause different portions of the direct sound signal to merge.

A quantitative measure of clarity is the *clarity index*, or ealy-to-reverberant energy ratio C_t:

$$C_t = 10\log_{10}\left(\frac{\int_0^t h^2(\tau)d\tau}{\int_t^\infty h^2(\tau)d\tau}\right), \tag{4.25}$$

measured in dB. The integration time t is ideally the time instant where late reverberation starts, and is typically selected to be $t = 80$ ms. Thus C_t is a measure of early to late energy ratio.

It is sometimes recommended that $C_t|_{t=0.008}$ for concert halls takes values in the range of -2 to $+1$ dB. Note however that this parameter is not an independent measurable quality, since it correlates to the initial time-delay gap t_I and also on the early decay time EDT. Therefore, subjectively “good” values of the clarity index will also depend on t_I and EDT values. In other words, the subjective attribute of clarity is not orthogonal to reverberance and spaciousness.

Note also that C_t is strongly dependent on the distance between source and listener: the direct sound falls off 6 dBs for each distance doubling, whereas the reverberant level remains approximately constant throughout the room. For this reason, the ratio of direct to reverberated energy is one of the most important cues for the perception of distance, as we will see in Sec. 4.5.

A second objective parameter that relates to the subjective attribute of clarity is the *center time* t_s, defined as the center of gravity time of the sound field:

$$t_s = \frac{\int_0^\infty \tau \cdot h^2(\tau)d\tau}{\int_0^\infty h^2(\tau)d\tau}. \tag{4.26}$$

Obviously a single reflection with a given strength will contribute the more to t_s the longer it is delayed with respect to the direct sound. Therefore high clarity is associated to low values of t_s. It has to be noted however that many studies report a high correlation of t_s with C_t, in the range $50 < t < 80$ ms. Therefore this parameter does not add new information with respect to the clarity index.

4.2.2.4 Other perceptually relevant parameters

In Sec. 4.2.1 we have discussed room acoustics in terms of rays and normal modes, and we have not considered other real-world phenomena. One of the most relevant of these is sound *diffusion* of sound waves: very roughly, diffusion is due to irregularities (at various scales) of reflecting surfaces, that cause scattering of reflected acoustic energy in many directions. This physical concept has a direct perceptual counterpart. If one listen to music in a rectangular hall with perfectly flat sidewalls, the sound takes on an undesirable harsh character. In order to produce the effect of a mellower sound and to increase spaciousness during late reverberation, diffusion should be physically realized at fine and large scales. A commonly accepted measure of diffusion is the *late interaural cross-correlation coefficient* $IACC_L$. This is defined from Eqs. (4.23, 4.24) using integration times $t_1 = 80$ ms and $t_2 = 3$ s, i.e. by estimating the IACF function in the late reverberation portion. Similarly to the $IACC_E$ parameter, $IACC_L$ is a *binaural* attribute of reverberation. It provides a measure of the correlation of the signals at the two ears during late reverberation.

Loudness (or *strength*) is often mentioned as a relevant subjective attribute. Of course the overall loudness depends on the power output of the sound source and not only on the reverberation of the environment. Nonetheless it is useful to introduce a measure of loudness of the environment, which is normalized with respect to the the source power. Such a measure can be used e.g. as a complementary parameter to the clarity index (see Eq. (4.25) above), since high clarity is of no use if the sound cannot be heard at proper loudness. A normalized measure of environmental loudness is given by the following quantity, often called *strength index* G:

$$G = 10\log_{10}\left(\frac{\int_0^\infty h^2(\tau)d\tau}{\int_0^\infty h_0^2(\tau)d\tau}\right), \tag{4.27}$$

where $h(t)$ is as usual the room impulse response and $h_0(t)$ is the response to the same non-directional impulse measured in an anechoic environment at a distance of 10 m. Note however that subjective loudness increases with reverberation time and is affected by the structure of early reflections. Therefore G is not an independent correlate of loudness.

Finally, the most elusive subjective attributes are those related to timbral qualities of a reverberant environment. Roughly speaking, many of the attributes in this family are related to the frequency-dependent shape of the reverberation time. One such attribute is *warmth*, or sometimes *timbre*, which characterizes the musicians' judgement of "richness in bass". This attribute correlates with the variation of the reverberation time in the low- and mid-frequency range: as an example, a quantitative measure of warmth can be the ratio of the average T_r in the range $250 - 500$ Hz to that in the range $500 - 1000$ Hz, or arternatively the slope of a linear interpolation of the EDT function in the range $125 - 2000$ Hz. Other timbre-related attributes are *heaviness* and *liveness*, which roughly relate to low-frequency and high-frequency variations of the reverberation time, respectively.

4.2.2.5 The Energy Decay Relief

A compact representation of the perceptually relevant features of a room impulse response is the so-called *Energy Decay Relief (EDR)* function, which is a time-frequency representation of the reverberation energy. The EDR function is in a way a generalization of the Energy Decay Curve (EDC) discussed previously, and is constructed as follows: given a RIR $h(t)$, this is bandpass filtered into a number N of frequency bands, and the EDC of each of the bandpassed responses $h_i(t)$ $(i = 1 \dots N)$ is computed. The resulting function $EDR(t, \omega)$ can be displayed as a surface in the 3-D space. The section $EDR(0, \omega)$ provides the power gain as a function of frequency. A section $EDR(t, \omega_0)$ shows the energy decay curve for a given frequency bin around ω_0.

M-4.2

Write a function that computes the EDR given a RIR.

M-4.2 Solution

```
function [EDR,F,T] = compute_edr(rir,Fs,frameSizeMS,overlap);
% adapted from http://ccrma.stanford.edu/%7Ejos/vguitar/

% define STFT parameters
minFrameLen = Fs*frameSizeMS/1000; %frameSizeMS is the framelength in ms
frameLen = 2^nextpow2(minFrameLen); % frame length = fft size
frameWindow = hann(frameLen);

%compute spectrogram and energy
[B,F,T] = specgram(rir,frameLen,Fs,frameWindow,round(overlap*frameLen));
[nBins,nFrames] = size(B);
B_energy = B.*conj(B);

%compute EDR (in dB)
EDR = zeros(nBins,nFrames);
for i=1:nBins %compute EDC for each frequency band
    EDR(i,:) = 10*log10(abs(fliplr(cumsum(fliplr(B_energy(i,:))))));
end
EDR = EDR - max(max(EDR)); %normalize at 0 dB
```

The time-frequency EDR function can be parametrized through two functions of frequency only. The first one is $T_r(\omega)$, the frequency-dependent reverberation time. The second one is the *frequency*

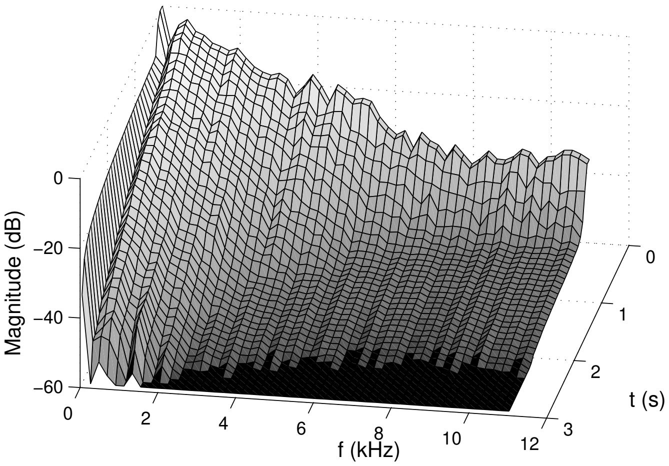


Figure 4.8: *Energy Decay Relief for the RIR of Fig. 4.3, normalized at* 0 *dB and truncated at* -60 *dB.*

response envelope, $G(\omega)$. This latter function is constructed by backward interpolating up to $t = 0$ the exponential decay time. For an ideally diffuse reverberation that decays exponentially, one has the equality $G(\omega) = EDR(0, \omega)$ and G coincides with the power gain of the room. In non-ideal cases, $G(\omega)$ only represent a "conceptual" $EDR(0, \omega)$ of the late reverberation, and the parametrization through $T_r(\omega)$ and $G(\omega)$ is only valid for the late portion of $EDR(t, \omega)$.

The EDR is sometimes regarded as a perceptual "signature" of a RIR, meaning with this that a large number of measures of independent perceptual factors can be categorized as energy ratios or energy decay slopes computed in different time-frequency regions of the EDR. Figure 4.8 shows an example of EDR. Note that, in accordance to our predictions, T_r is shorter at higher frequencies.

4.3 Algorithms for synthetic reverberation: the perceptual approach

If a RIR signal is available, the most straightforward approach to synthetic reverberation is to convolve an anechoic input signal with such a RIR. We do not review techniques for impulse response measurement in this chapter, nor we address numerical techinques for convolution. We only observe that *direct convolution*, obtained by storing each sample of the impulse response as a coefficient of an FIR filter whose input is the dry signal, becomes easily impractical if the length of the response exceeds few tenths of a second, as it translates into a FIR filter of order $N \sim 10^4$. But even if we have enough computational resources for direct convolution, or use fast convolution techniques, a real recorder RIR has the disadvantage that it is not easily modified to simulate changes in room attributes.

In order to overcome these limitations, in the second half of the 20th century several engineers and acousticians developed electronic devices, models, and algorithms for synthetic reverberation that are based on a *perceptual approach*, in which efficient filter representations are used, and only the perceptually salient features of reverberation are simulated and controllable.

4.3.1 Late reverberation

We have previously seen that a RIR can be seen as made of two components, early reflections and late reverberation. In this section we discuss perceptual models for late reverberation, and we postpone early reflection modeling to Sec. 4.3.2.

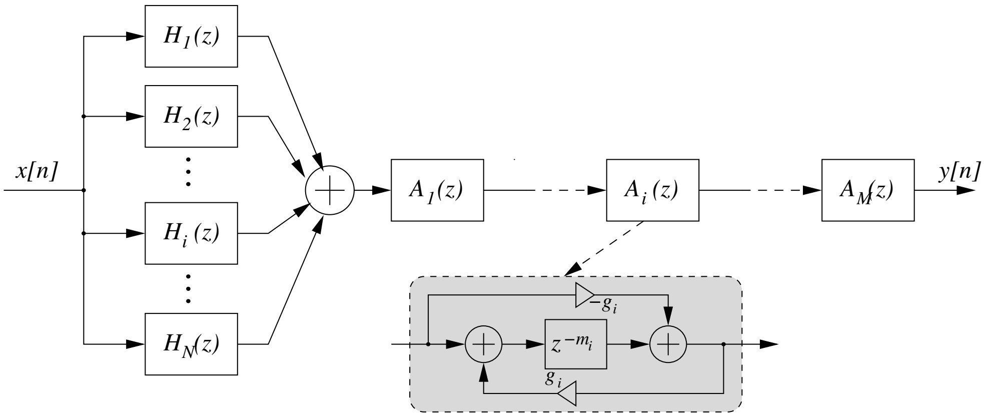


Figure 4.9: *Block scheme of a reverberator based on comb filters (the H_i blocks) and* all-pass comb *filters (the A_i blocks). The internal structure of the A_i filters is shown in the grey box.*

4.3.1.1 Recirculating delays

The two main computational structures that can be used for the inexpensive simulation of complex patterns of echoes associated to late reverberation are the recursive comb filter $H(z)$ (see Karplus-Strong in Ch. *Sound modeling: source based approaches*) and the so-called *all-pass comb filter* $A(z)$:

$$H(z) = \frac{z^{-m}}{1 - gz^{-m}}, \qquad A(z) = \frac{z^{-m} - g}{1 - gz^{-m}}. \tag{4.28}$$

It is easily seen that $A(z)$ is an all-pass structure, since each of the m poles is the reciprocal of one of the m zeros and the amplitude response $|\, A(z)\,|$ is therefore flat. For $m = 1$ the structure reduces to the first-order all-pass filter examined in Ch. *Sound modeling: source based approaches*. The (positive) gain g in $A(z)$ has to be less than unity in order to ensure stability.

Figure 4.9 depicts a reverberator constructed using comb-filters and all-pass comb filters, together with a realization of the all-pass comb (see the grey box). The general idea behind this structure is the following. First, the parallel combination of comb filters generates a frequency response that contains peaks contributed by each comb. In theory we can obtain an arbitrary modal density by using a sufficiently large number N of comb filters. Second, the series combination of all-pass combs that receives the output of the parallel combination of combs has the effect of dramatically increasing the temporal density of reflections, because each echo generated by $A_i(z)$ will create a set of echoes in $A_{i+1}(z)$. Again, an arbitrarily high reflection density can be in principle obtained by using a sufficiently large number M of all-pass combs.

4.3.1.2 Parameter tuning

The choice of a proper set of parameter values is critical in order to obtain convincing results. In the remainder of this section we provide a list of commonly accepted guidelines. The sample delays m_i of the combs should be mutually coprime (or incommensurate), in order to reduce the superimposition of echoes in the impulse response, thus maximizing the modal density and reducing the so-called flutter echoes.

The gains g_i of the combs can be chosen as functions of the sample delays m_i, given a desired reverberation time T_r, as follows: we want to find the number R of loops in the ith comb after which the dB amplitude of an unitary impulse has become -60 dB; the amplitude after R loops is $g^R =$.

Moreover R loops are completed in the time $T_r = Rm_i/F_s$; therefore the following equation holds for the reverberation time of a single comb:

$$\frac{F_s \cdot 20\log_{10}(g_i)}{m_i} = -\frac{60}{T_r} \quad \Rightarrow \quad g_i = 10^{-3\frac{m_i}{F_s T_r}}. \tag{4.29}$$

Note that this choice ensures that the pole moduli $\sqrt[m_i]{g_i} = 10^{-3\frac{1}{F_s T_r}}$ have the same value for all the combs. If this condition was not verified, then the poles with largest moduli would resonate longer and would add an undesired tonal coloration in the late decay.

A quantitative estimate of the modal density provided by the parallel comb structure can be easily obtained. If the m_i's of the combs are mutually coprime, then the modal density D_f (which is number of frequency peaks per Hz) can be estimated as

$$D_f = \sum_{i=1}^{N} \frac{m_i}{F_s} = \frac{N\bar{m}}{F_s}, \tag{4.30}$$

where $\bar{m}$ is the mean sample delay length. Note that this modal density is constant for all frequencies, unlike in real rooms (see Eq. (4.9)). A too low D_f can introduce audible beating between two neighboring modes, especially in response to narrowband signals. In order to avoid this effect, a good rule of thumb is to choose the m_i's such that $D_f \geq T_r$: this ensures that the average beat period is at least equal to the reverberation time.

In a similar way we can estimate quantitatively the temporal reflection density provided by the parallel combination of combs: each filter outputs one echo every m_i/F_s seconds, therefore the combined reflection density (number of reflections per second) is

$$D_r = \sum_{i=1}^{N} \frac{F_s}{m_i} \approx \frac{NF_s}{\bar{m}}, \tag{4.31}$$

where the last approximation only holds when the m_i are similar. Again, the reflection density is constant as a function of time, unlike real rooms (see Eq. (4.21)). A value $D_r = 10^3$ is sometimes considered to be sufficient to sound indistinguishable from diffuse reverberation, although higher values (e.g. $D_r = 10^4$) are preferable.

From the two estimates (4.30) and (4.31) provide an estimate of the number of comb filters needed in order to achieve desired modal and reflection densities:

$$N = \sqrt{D_f D_r}. \tag{4.32}$$

Note however that this estimate does not consider the effect of the cascaded series of all-pass comb filters A_i: as already mentioned, the A_i provide a dramatic increase of the reflection density and allow to a number N of comb filters that is smaller than the one estimated from Eq. (4.32).

M-4.3

Realize the reverberant structure of Fig. 4.9. The reverberator can be tried with $N = 4$, $M = 2$, and the following settings: time delays m_i/F_s ($i = 1 \dots 4$) of the comb filters between 30 and 45 ms, time delays m_i/F_s ($i = 5, 6$) of the all-pass combs between 1.7 and 5 ms, modal density $D_f = 1000$, gains of the all-pass combs $g_i = 0.7$ ($i = 5, 6$). With these settings the structure is known as Schroeder reverberator (see bibliography).

M-4.3 Solution

```
function y = reverb_schroeder(x, Tr, m_H, m_A, g_A);
% x: input signal; m_H: N-dim array of delays (in samples) of combs H_i;
% m_A: M-dim array of delays (in samples) of all-passes A_i;
% g_A: M-dim array of all-pass gains

global Fs;
y = zeros(length(x),1); %output signal updated after each single filter
for i=1:length(m_H)          %parallel comb filtering
    g_H = 10^(-3*m_H(i)/(Fs*Tr));         % gain of ith comb
    num_H = [zeros(1,m_H(i)),1];          % numerator of ith comb
    den_H = [1,zeros(1,m_H(i)-1),-g_H];   % denominator of ith comb
    y = y + filter(num_H, den_H, x);      % update comb parallel
end
for i=1:length(m_A)          %series all-pass filtering
    num_A = [-g_A(i),zeros(1,m_A(i)-1),1]; % numerator of ith all-pass
    den_A = [1,zeros(1,m_A(i)-1),-g_A(i)]; % denominator of ith all-pass
    y = filter(num_A,den_A,y);             % update all-pass series
end
```

4.3.1.3 Low-pass combs

The reverberators discussed above sound reasonably well especially for short reverberation times and low reverberation levels. For different settings however they suffer from a number of problems. First, the reverberation is not dense enough at the beginning, resulting in a “grainy” sound quality (especially with impulsive sounds). Second, late reverberation tends to exhibit an already mentioned “fluttering” effect. Third, especially for long T_r’s a “ringing” effect can be heard, which gives an undesired metallic quality to the reverberation. Fourth, the modal density is not sufficiently large and, as already mentioned, does not increase with frequency. Fifth, the reverberation time T_r does not depend on frequency, unlike in real rooms (see Sec. 4.2.2 and the EDR function there discussed).

A first obvious way of improving the modal density is to increase the number of comb filters in parallel, especially when long reverberation times need to be simulated. A second more substantial improvement amounts to employ, in place of comb filters, a *low-pass comb* filter, where a low-pass filter H_{lp} is inserted in the feedback loop of the comb together with the scalar gain g. The purpose of this modification is to simulate the attenuation effects of higher frequencies, due to air viscosity, heat conduction, and energy losses at reflection. As a result, T_r now decreases at higher frequencies and makes the reverberation sound more realistic. In addition, the response to impulsive sounds is also improved, due to the smoothing effect of the low-pass filtering.

If a simple one-pole low-pass filter H_{lp} is used, then the low-pass comb filter is given as

$$H(z) = \frac{z^{-m}}{1 - H_{lp}(z) z^{-m}}, \quad \text{with } H_{lp}(z) = \frac{g_2}{1 - g_1 z^{-1}}. \tag{4.33}$$

One could verify that in order for $H(z)$ to be stable the condition $\max_z | H_{lp}(z) | = g_2/(1 - g_1) < 1$ must hold. A practical choice is $g_2 = g(1 - g_1)$, with $g < 1$. In this way the overall T_r is still controlled by the parameter g as in Eq (4.29)

Note that we have already introduced the low-pass comb filter for the Karplus-Strong algorithm in Ch. *Sound modeling: source based approaches*, although here we are using a different low-pass filter H_{lp}.

Coefficients of the low-pass combs: g_2 can be determined as a function of the delay length and the desired T_r, as explained in the previous section. The g_1 coefficient can also be related with decay time at a specific frequency or fine tuned by direct experimentation.

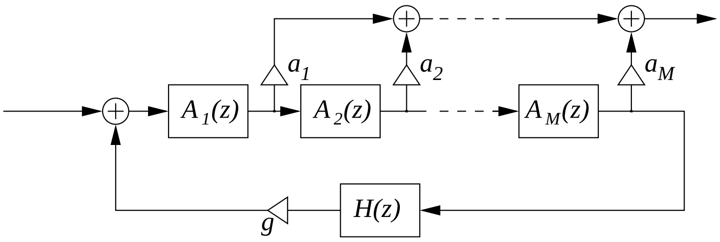


Figure 4.10: *A reverberator constructed with a series connection of all-pass filters and a low-pass filter in feedback.*

M-4.4

Realize the reverberant structure of Fig. 4.9 using low-pass combs of the form (4.33). The reverberator can be tried with $N = 6$, $M = 1$, and with the following settings: time delays m_i/F_s $(i = 1 \dots 6)$ of the combs distributed between 50 and 78 ms, coefficients $g_{1,i}$ of the low-pass filter distributed between 0.40 and 0.48 (at $F_s = 44.1$ kHz), time delay of the all-pass comb $m_7/F_s = 6$ ms, gain of the all-pass comb $g_7 = 0.7$. With these settings the structure is known as Moorer reverberator (see bibliography)

M-4.4 Solution

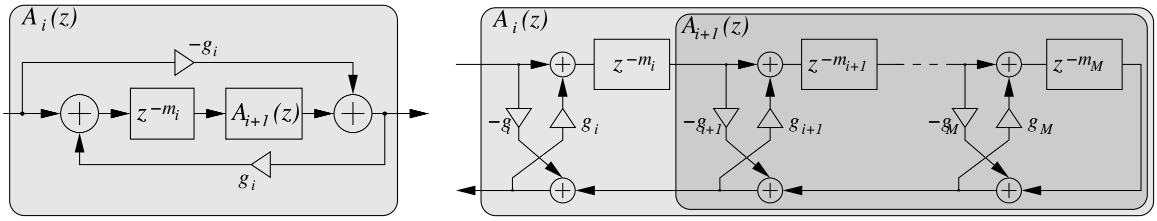
```
function y = reverb_moorer(x,Tr,m_H,g1_H,m_A,g_A)
% x: input signal; m_H: N-dim array of delays (in samples) of combs H_i;
% g1_H: N-dim array of coefficients of the H_lp's; m_A : M-dim array of
% delays (in samples) of all-passes A_i; g_A: M-dim array of all-pass gains

global Fs;
y = zeros(length(x),1); %output signal updated after each single filter
for i=1:length(m_H)          %parallel comb filtering
    g_H = 10^(-3*m_H(i)/(Fs*Tr));          % gain of ith comb
    num_H = [zeros(1,m_H(i)),1,g1_H(i)]; % numerator of ith lowp. comb
    den_H = [1,-g1_H(i),zeros(1,m_H(i)-2),-g_H*(1-g1_H(i))];  % denominator
    y = y + filter(num_H, den_H, x);      % update comb parallel
end
for i=1:length(m_A)          %series all-pass filtering
    num_A = [-g_A(i),zeros(1,m_A(i)-1),1]; % numerator of ith all-pass
    den_A = [1,zeros(1,m_A(i)-1),-g_A(i)]; % denominator of ith all-pass
    y = filter(num_A,den_A,y);             % update all-pass series
end
```

Clearly the filter coefficients `num_H` and `den_H` have been determined by combining Eqs. (4.33).

4.3.1.4 Nested all-pass filters

Despite the improvements provided by this latter reverberator, some problems remain. First, it is not possible to tune the reverberator to a desired $T_r(\omega)$ function. Second, the modal density is still constant with respect to frequency and the ringing quality and the fluttering effect in the reverberation tail remain, although reduced to some extent. In order to overcome this problems some researchers have proposed reverberators with entirely different structures than the one shown in Fig. 4.9. One such structure is shown in Fig. 4.10.

As before, the cascaded all-pass filters $A_i(z)$ provide a high temporal density of reflections, because each echo generated by a filter will create a set of echoes in the following one. In this case however, the



(a) (b)

Figure 4.11: *Nested all-pass filters; (a) generalization of an all-pass structure (see Fig. 4.9), and (b) realizazion by means of a lattice structure.*

output from the last one is recirculated to the series connection through a low-pass filter $H(z)$ and an attenuating gain g. The resulting system is stable, if the condition $|gH(e^{j\omega})| < 1 \; \forall\omega$ is verified.

The low-pass filter $H(z)$ can be interpreted as simulating frequency-dependent absorptive losses, and the gain g provides control over the reverberation time. An important effect of this outer feedback loop is that the characteristic metallic sound of the series all-pass is drastically reduced. Another peculiarity of this structure is that the output is constructed as a linear combination of the all-pass outputs. Since each each tap outputs a different response shape, the coefficients a_i can be adjusted in order to shape the amplitude envelope of the reverberant decay.

A final remark concerns the possibility of generating a reflection density that increases with time, as in real rooms. A structure that achieves this goal is a *nested all-pass filter* $A_1(z)$, which can be defined recursively as follows:

$$\begin{aligned} A_{M+1}(z) = &\quad 1, \\ A_i(z) = &\quad \frac{z^{-m_i} A_{i+1}(z) - g}{1 - g z^{-m_i} A_{i+1}(z)}, \quad \text{for } i = 1 \dots M. \end{aligned} \tag{4.34}$$

Figure 4.11(a) shows that this structure can be seen as a generalization of the all-pass comb, in which part of the delay line has been substituted by an all-pass filter. Figure 4.11(b) explodes this structure into a lattice realization. It is easy to verify that each of the nested filters $A_i(z)$ are all-pass. Moreover, Fig. 4.11(a) shows that each echo generated by the inner all-pass $A_{i+1}(z)$ is recirculated to itself through the outer feedback path of $A_i(z)$: this intuitively explains why this structure provides a reflection density that increases with time.

M-4.5

Realize the reverberant structure of Fig. 4.10 using nested all-pass filters of the form (4.34).

4.3.2 Early reflections

So far we have only examined perceptually-based algorithms for the simulation of late reverberation. In this section we address the simulation of early reflections, which have great importance in the perception of the acoustic space as we have see in Sec. 4.2.2.

4.3.2.1 FIR structures

As previously discussed, the early response of a room is sparsely populated with attenuated impulses. These can be straightforwardly simulated using a direct-form FIR filter that reproduces these impulses explicitly and accurately. For the determination of the filter parameters, a good rule of thumb is to apply to the early reflections delays the same criterion of “mutually-primeness” used before for the comb delays. A better strategy is to derive the parameters from some geometric modeling technique, e.g. the source image method discussed in Sec. 4.2.1.

M-4.6

Write a function that computes a signal containing the first R early reflections

M-4.6 Solution

```
function y = reverb_earlyrefl(x, m_E, a_E);
% x: input signal; m_E: R-dim array of delays (in samples) of early
% reflections; a_E: R-dim array of gains of early reflections

num = zeros(1,max(m_E)+1); %empty FIR numerator
num(m_E+1) = a_E; % populate numerator with early reflection gains
y = filter(num,1,x);
```

The delays in this script have a slightly different meaning than those in Fig. 4.12, since they are not cascaded.

Figure 4.12 shows an example of early reflection modeling, in which the FIR filter simulates the first R reflections and has been realized using a direct form structure. The early reflection filter has to be connected to a late reverberation block: Fig. 4.12(a) and 4.12(b) show two possible connections. In Fig. 4.12(a) the late reverberator receives the delayed input signal, and therefore the FIR response will always occur before the late response in the final output. Figure 4.12(b) shows a more complex coupling between the two blocks. In this case the late reverberator is driven by the output of the FIR filter, with the result of increasing the reflection density in the late reverberation. Moreover, additional control parameters are available: the gain g can be adjusted in order to balance the early/late reverberation ratio, while the delays $D_1,\ D_2$ can be tuned so that the start of the late reverberator output coincides with the last pulse output from the FIR filter, thus avoiding undesired gaps in the overall response.

M-4.7

Realize the reverberator depicted in Fig. 4.12(a), where the early reflection FIR filter has to be coupled to one of the late reverberation structures discussed in the previous sections.

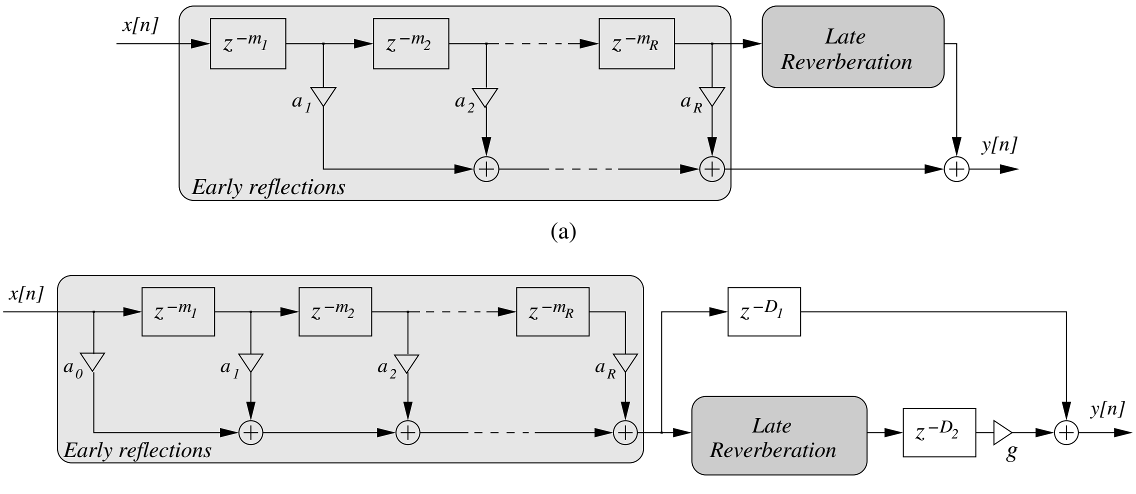
M-4.7 Solution

```
function y = reverb_schroeder_earlyrefl(x, Tr, m_E, a_E, m_H, m_A, g_A);

global Fs;
y_E = [reverb_earlyrefl(x,m_E,a_E); zeros(max(m_E),1)];        %early refl.
y_L = reverb_schroeder([zeros(max(m_E),1); x],Tr,m_H,m_A,g_A); %late rev.
y=y_E+y_L;
```

M-4.8

Realize the reverberator depicted in Fig. 4.12(b), where the early reflection FIR filter has to be coupled to one of the late reverberation structures discussed in the previous sections. Compare the resulting impulse responses with the ones obtained from M-4.7.

M-4.8 Solution



(a)

(b)

Figure 4.12: *Two realizations of a reverberator with early reflections; (a) late reverberation block receiving the delayed input signal, and (b) late reverberation block receiving the output of the early reverberation FIR filter, with additional control parameters D_1, D_2, g. The late reverberation block can be one of the structures examined in the previous sections.*

```
function y = reverb_moorer_earlyrefl(x,Tr,m_E,a_E,m_H,g1_H,m_A,g_A,g_mix);

global Fs;
y_E = reverb_earlyrefl(x,m_E,a_E);              %early refl.
y_L = reverb_moorer(y_E,Tr,m_H,g1_H,m_A,g_A);   %late rev.

delaydiff = max(m_E) - min(m_H);        % diff. in delay between early
if (delaydiff>0)                        % refl. and late rev.
    y = [y_E; zeros(delaydiff,1)] + g_mix*[zeros(delaydiff,1); y_L];
else
    y = [zeros(delaydiff,1); y_E] + g_mix*[y_L; zeros(delaydiff,1)];
end
```

In order to improve the quality of the FIR structure described above, one has to include some form of low-pass filtering that models frequency dependent losses. One possibility is to substitute each of the gains a_i with a low-pass filter, composed by considering the history of reflections for each echo. Early reflections are not perceived as individual events however, therefore it is not necessary to model accurately the spectral content of each single reflection. A cheaper, and often satisfactory, choice is to sum sets of reflections together and and to filter them through the same low-pass.

4.3.2.2 Directional effects

In this brief section we anticipate some concepts that will be addressed in Secs. 4.5 and 4.6, where we will address the topic of rendering the location in space of a sound source.

Effects due to reverberation and spatial perception of sound are related in many respects. On the one hand, reverberation has a relevant role in the perception of the location of a sound source, as we will see in Sec. 4.5. On the other hand, the subjective attribute of *spatial impression* is extremely important in the perception of reverberation, and should be accounted for in any synthetic reverberation algorithm: in

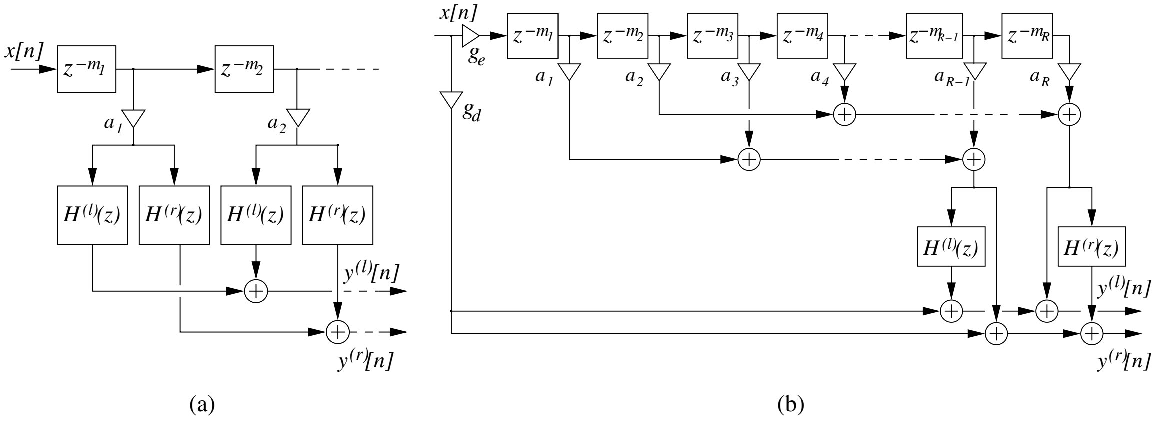


Figure 4.13: *Two stuctures that associate directional filters to early reflections, for binaural reverberation; (a) one directional filter for each reflection, and (b) two directional filters for two sets of reflections.*

Sec. 4.2.2 we have seen that early reflections in particular have a primary role in the formation of spatial impression (see the definition of the early interaural cross-correlation coefficient).

An early reflection reaches the two ears with different intensities and at different times, because of the shadowing effect of the head, the different distance traveled, the filtering properties of the pinna, and so on. For this reason early reveberation is most effective if it is presented *binaurally*, i.e. by taking into account these effects and presenting different early reflections to the two ears (e.g. via headphones). In this case one can associate with each early reflection a directional filter intended to reproduce localization cues. One structure that realizes this idea is shown in Fig. 4.13(a). $H^{(l),(r)}$ are the so-called Head-Related Transfer Functions,[3] that represent the transfer function between the sound source and the entrance of the ear canals. These directional filters are associated to early reflections in a structure analogous to those shown in Fig. 4.12.

Another possibility is to sum sets of early reflections together and process each set with the same directional filter, so that all the reflections in a single set will be rendered with the same spatial location. This approach can still produce a convincing sensation of spatial impression, while being far more efficient. Various realization of this general idea have been proposed. Figure 4.13(b) shows one possible realization: two sets of echoes are formed and each set is processed with the same directional filters. Various degrees of spatial impression can be obtained by playing with the gain g_e, and convincing results are obtained already with $R = 6$ reflections.

As a conclusion to this section it worth mentioning that if no binaural processing is performed the addition of early reflections can in certain cases deteriorate the quality of a reverberator, as they cause tonal coloration of the sound without producing spatial impression.

[3] The transforms of the Head-Related Impulse Responses already introduced in Eq. (4.23).

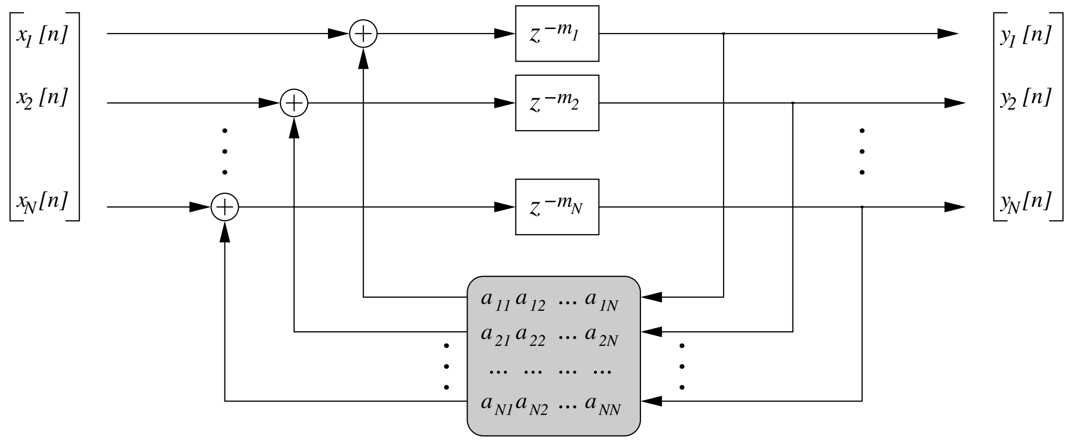

$x_1[n]$
$x_2[n]$
$x_N[n]$
z^{-m_1}
z^{-m_2}
z^{-m_N}
$y_1[n]$
$y_2[n]$
$y_N[n]$
$a_{11}\,a_{12}\,\dots\,a_{1N}$
$a_{21}\,a_{22}\,\dots\,a_{2N}$
$\dots\,\dots\,\dots\,\dots$
$a_{N1}\,a_{N2}\,\dots\,a_{NN}$

Figure 4.14: ...

4.4 Multidimensional reverberation structures

4.4.1 Feedback delay networks

4.4.1.1 A n-D generalization of the recursive comb filter

In the previous section we have seen that the recursive comb filter of Eq. (4.28) has been extensively used as the main building block of perceptual reverberators, as an inexpensive way to generate patterns of resonances. Now the question is: can we generalize the comb structure in order to achieve higher modal densities? The filter structure depicted in Fig. 4.14 provides a first answer. First, it is easily seen to be a vector generalization of the recursive comb filter, as it reduces to a parallel combination of ordinary comb filters when the feedback matrix $\boldsymbol{A} = [a_{ij}]$ is diagonal. Second, and more interesting, it recirculates the output of the ith delay line to the input of the jth delay line, for every non-null element a_{ij}. This observation gives the intuition that when $\boldsymbol{A}$ is non-diagonal this structure is capable of much higher modal densities than a simple parallel of comb filters.

The generalization extend also to stability conditions. While the comb filter of Eq. (4.28) is stable if $|\,g\,| < 1$, the multidimensional structure of Fig. 4.14 is stable if $\|\,\boldsymbol{A}\,\|_2 < 1$, where $\|\cdot\|_2$ is the spectral norm of a matrix.[4] This can be easily verified by applying the conditions for Lyapunov stability, i.e. that the output $\boldsymbol{y}[n]$ decreases in time when the input signal $\boldsymbol{x}$ is zero. Since

$$\|\,\boldsymbol{y}[n]\,\|_2 = \left\|\,\boldsymbol{A}\begin{bmatrix} y_1[n-M_1] \\ \vdots \\ y_N[n-M_N]\end{bmatrix}\right\|_2, \tag{4.35}$$

stability is guaranteed whenever the feedback matrix satisfies

$$\|\,\boldsymbol{A}\boldsymbol{y}\,\|_2 < \|\,\boldsymbol{y}\,\|_2 \qquad \forall\,\boldsymbol{y}. \tag{4.36}$$

In other words, a sufficient condition for stability is that the feedback matrix decreases the L^2 norm of its input vector. Since in general $\|\,\boldsymbol{A}\boldsymbol{y}\,\|_2 < \|\,\boldsymbol{A}\,\|_2 \cdot \|\,\boldsymbol{y}\,\|_2$, we conclude that stability is guaranteed for $\|\,\boldsymbol{A}\,\|_2 < 1$.

[4] The matrix norm corresponding to any vector norm $\|\cdot\|$ may be defined for any matrix $\boldsymbol{A}$ as $\|\,\boldsymbol{A}\,\| = \max_{\boldsymbol{x}\neq 0} \frac{\|\,\boldsymbol{A}\boldsymbol{x}\,\|}{\|\,\boldsymbol{x}\,\|}$. The spectral norm $\|\cdot\|_2$ is the matrix norm induced by the L^2 vector norm.

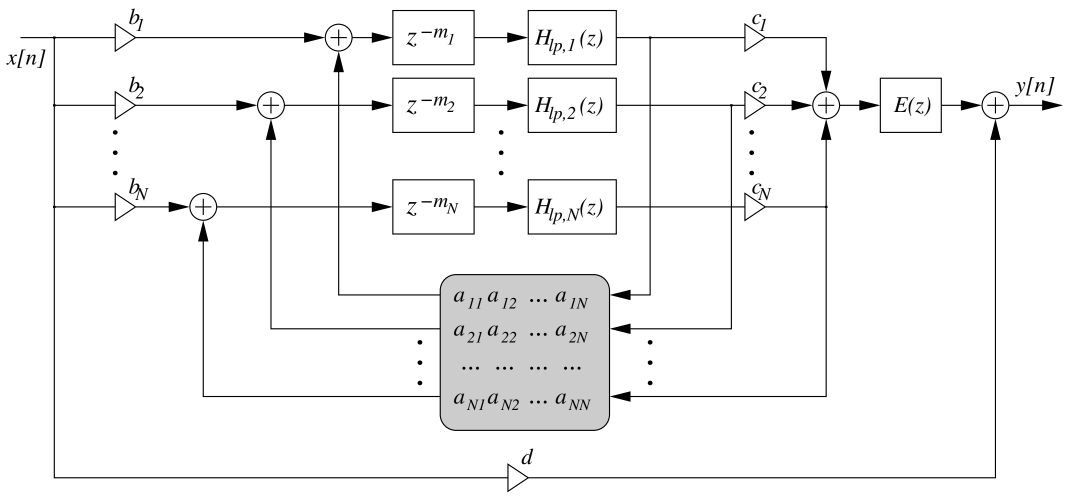


Figure 4.15: *A Feedback Delay Network structure for artificial reverberation.*

A class of matrices that satisfy the stability condition is

$$\boldsymbol{A} = \boldsymbol{\Gamma Q}, \qquad \text{where} \quad \Gamma = \begin{bmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & g_N \end{bmatrix}, \; |\, g_i \,| < 1, \tag{4.37}$$

and where $\boldsymbol{Q}$ is an orthogonal matrix. Recall that *(1)* the spectral norm $\|\, \boldsymbol{A} \,\|_2$ is the square root of the largest eigenvalue of $\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{T}}$, and that *(2)* by definition $\boldsymbol{Q}$ is orthogonal if and only if $\boldsymbol{Q}\boldsymbol{Q}^{\boldsymbol{T}} = \mathbb{I}$. Then $\|\, \boldsymbol{A} \,\|_2 = \|\, \boldsymbol{\Gamma Q} \,\| = \max_i |\, g_i \,|$.

The above analysis justifies the use of the structure of Fig. 4.14 as a multichannel reverberator in which N input signals $\boldsymbol{x}[n]$ (or N replicas of a single input signal $x[n]$) produce N outputs $\boldsymbol{y}[n]$ that are approximately mutually incoherent and thus can be used in a N-channel loudspeaker system to render a diffuse soundfield. A possible choice for the matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = g\frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix}, \quad |\, g \,| < 1, \tag{4.38}$$

which is immediately seen to belong to the class (4.37).

4.4.1.2 A general FDN reverberators

The "vector comb filter" that we have analyzed in the previous section is an example of a class of filter networks, known as *Feedback Delay Networks (FDNs)*. Figure 4.15 shows a more general FDN structure for artificial reverberation, that extends in many ways the one depicted in Fig. 4.14. First, it is a Single-Input, Single-Output structure which uses two $N \times 1$ vectors $\boldsymbol{b} = [b_i]$ and $\boldsymbol{c} = [c_i]$ to split the input into N channels and to combine the N outputs in one channel. Second, low-pass filters $H_{lp,i}(z)$ are cascaded to the delay lines. Third, the final output y is corrected with an additional filter $E(z)$ plus an additive term dx. The transfer function of the system is almost immediately found to be:

$$\frac{Y(z)}{X(z)} = \boldsymbol{c}^{\boldsymbol{T}} \left\{ \left[\mathbb{I} - \boldsymbol{D}(z)\boldsymbol{A}\right]^{-1} \boldsymbol{D}(z) \right\} \boldsymbol{b} \cdot E(z) + d = \boldsymbol{c}^{\boldsymbol{T}} \left[\boldsymbol{D}(z^{-1}) - \boldsymbol{A}\right]^{-1} \boldsymbol{b} \cdot E(z) + d, \tag{4.39}$$

where $\boldsymbol{A} = [a_{ij}]$ is the *feedback matrix* of the system, and

$$\boldsymbol{D}(z) = \begin{bmatrix} z^{-m_1} H_{lp,1}(z) & 0 & \cdots & 0 \\ 0 & z^{-m_2} H_{lp,2}(z) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & z^{-m_N} H_{lp,N}(z) \end{bmatrix}$$

is the *delay matrix* of the system. We shall see that this structure allows to orthogonalize to a great extent the reverberation parameters, as the various blocks can be independently tuned to fit desired values of different reverberation parameters.

M-4.9

Realize the reverberant structure of Fig. 4.15. With the 4×4 matrix given in Eq. (4.38), the structure of Fig. 4.14 is a special case of this.

M-4.9 Solution

```
function y = reverb_fdn(x,Fs,Tr,A,b,c,d,m);
% x: input signal; A: NXN feedback matrix; b: Nx1 array of input weights;
% c: 1XN array of output weights; d: scalar weight for input-to-output contrib.
% m: N-dim array of line delays (in samples)

N=size(A,1); %dimension of the FDN
delaylines = zeros(max(m),N); %create and initialize N delay lines
[num_H,den_H,num_E,den_E] = lossy_fdn(Fs,Tr,m); %initialize lossy components

y = zeros(size(x)); %initialize output signal
for n = 1:length(x)   %audio cycle
    Y=zeros(N,1); % Y is the array of N signals after the lowpass filters
    for i=1:N   Y(i)= filter(num_H(i), den_H(i), delaylines(m(i),i) ); end
    y(n)= filter(num_E, den_E, c*Y) +d*x(n);      %compute output
    linein =b*x(n) + A*Y;      %compute new input to lines
    delaylines = circshift(delaylines,1); %circular shift lines
    delaylines(1,:) = linein;              %write lines
if(mod(n,round(length(x)/20))==0) fprintf(’%d%%\n’,round(n/length(x)*100)); end
end
```

Note that in this case we had to write an audio loop, since an explicit formulation of the rational transfer function (4.39) is not available in general: for this reason this realization is extremely inefficient. Note also that we are using an auxiliary transfer function `lossy_fdn`: see M-4.11 below.

Since an “ideal” late reverberation impulse response should resemble exponentially decaying noise, it is useful to start designing a lossless reverberator (with infinite reverberation time) and work on making it a good noise generator. Once this *lossless prototype* has been designed, one can work on obtaining the desired reverberation time in each frequency band. We associate to the FDN of Fig. 4.15 the lossless prototype of Fig. 4.16.

What does the losslessness requirement imply to the feedback matrix $\boldsymbol{A}$? We know that by definition of losslessness the equality $\int_\omega \left\{ \sum_{i=1}^n \mid Y_i(e^{j\omega}) \mid^2 \right\} d\omega = \int_\omega \left\{ \sum_{i=1}^n \mid X_i(e^{j\omega}) \mid^2 \right\} d\omega$ must hold. Moreover it is a general result that a multidimensional filter is lossless if and only if its frequency response matrix $\boldsymbol{H}(\mathrm{e}^{j\omega})$ is unitary, i.e. $\boldsymbol{H}(\mathrm{e}^{j\omega})\boldsymbol{H}^*(\mathrm{e}^{j\omega}) = \mathbb{I}$ (where * denotes the complex-conjugate transpose as usual). In our case, it is quite straightforward to prove that $\boldsymbol{A}$ being unitary is a sufficient condition for the overall frequency response matrix to be unitary. Moreover the entries a_{ij} have to be real in order for the system to output a real signal $y[n]$, and a unitary matrix with real entries is an orthogonal matrix.

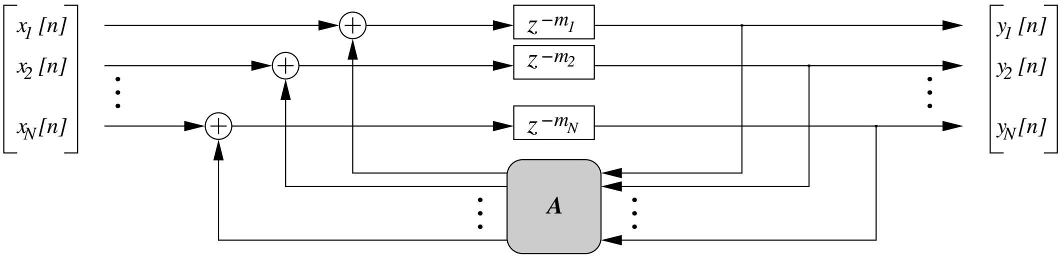


Figure 4.16: *Lossless prototype network associated to the Feedback Delay Network of Fig. 4.15.*

In conclusion, if $\boldsymbol{A}$ is orthogonal then the network of Fig. 4.16 is lossless. Note however that this condition is sufficient but not necessary, thus the system may be lossless even with a non-orthogonal feedback matrix. We will return to this point in Sec. 4.4.2.

4.4.1.3 Designing the lossless prototype

Designing the lossless prototype means choosing the dimension N, the m_i's, and the feedback matrix $\boldsymbol{A}$. Let us start with the dimension N and the delay lengths m_i. Together with the feedback matrix these parameters determine the buildup of reflection density. The criteria that we have examined in Sec. 4.3 (see in particular Eqs. (4.30, 4.31) can be applied also in this case with satisfactory results. Note however that Eqs.Eqs. (4.30, 4.31) are no longer valid here, since, a non-diagonal feedback matrix increases the modal and reflection densities. Therefore in general the parameters have to be chosen on the basis of empirical observations. It is generally noted that $N = 8$ to 16 lines with a total delay $\sum_i m_i/F_s$ of 1 to 2 seconds already produce a response perceptually undistinguishable from white noise.

Let us now consider the lossless feedback matrix $\boldsymbol{A}$. The simplest orthogonal matrix is a diagonal matrix whose diagonal elements (which are the eigenvalues) have unit modulus: as already seen this choice corresponds to a parallel of ordinary comb filters. A more interesting family of orthonormal matrices are *Householder reflection matrices*. A specific Householder matrix is defined given the reference vector $\boldsymbol{u} = [1, \dots, 1]^T$:

$$\boldsymbol{A} = \mathbb{I} - \frac{2}{N}\boldsymbol{u}\boldsymbol{u}^{\boldsymbol{T}}, \qquad \text{then } \boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} x_1 - \frac{2}{N}\sum_i x_i \\ \vdots \\ x_N - \frac{2}{N}\sum_i x_i \end{bmatrix}, \tag{4.40}$$

for any input vector $\boldsymbol{x}$. We will see in Sec. 4.4.2 that $\boldsymbol{u}$ can be interpreted as the specific vector about which an input vector is reflected by the matrix $\boldsymbol{A}$ in an N-dimensional space. A more general formulation may be obtained by replacing the identity matrix in Eq. (4.40) with any $N \times N$ permutation matrix.

The explicit expression for $\boldsymbol{A}\boldsymbol{x}$ in Eq. (4.40) shows that applying a Householder matrix to a vector requires $N - 1$ additions and one multiplication to obtain the term $\frac{2}{N}\sum_i x_i$, plus N additions to subtract this term from $\boldsymbol{x}$. Therefore the matrix-times-vector operation is only $\mathcal{O}(N)$ as opposed to the usual $\mathcal{O}(N^2)$.

Another interesting feature of the Householder feedback matrix is that $\boldsymbol{A}$ does not have null entries for $N \neq 2$. This is a desirable property since it implies that every delay line feeds back to every other delay line, reinforcing the build-up of reflection density. The case $N = 4$ is especially nice, since the matrix entries all have the same magnitude and $\boldsymbol{A}$ is therefore “balanced”. For larger N the diagonal becomes larger than the off-diagonal elements, and $\boldsymbol{A}$ approaches a diagonal matrix as $N \to \infty$. Due to the elegant balance of the $N = 4$ case, a larger ($N = 16$) feedback matrix can be constructed as follows:

$$\boldsymbol{A} = \frac{1}{2}\begin{bmatrix} \boldsymbol{A}_4 & -\boldsymbol{A}_4 & -\boldsymbol{A}_4 & -\boldsymbol{A}_4 \\ -\boldsymbol{A}_4 & \boldsymbol{A}_4 & -\boldsymbol{A}_4 & -\boldsymbol{A}_4 \\ -\boldsymbol{A}_4 & -\boldsymbol{A}_4 & \boldsymbol{A}_4 & -\boldsymbol{A}_4 \\ -\boldsymbol{A}_4 & -\boldsymbol{A}_4 & -\boldsymbol{A}_4 & \boldsymbol{A}_4 \end{bmatrix}, \quad \text{where } \boldsymbol{A}_4 := \frac{1}{2}\begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix} \tag{4.41}$$

is the 4×4 Householder matrix.

Other types of unitary matrices may be used. In particular, unitary feedback matrices can be derived from Hadamard matrices. A Hadamard matrix $\boldsymbol{H}$ is defined as an $N \times N$, $(-1, 1)$-matrix (i.e. a matrix whose elements consist only of the numbers -1 or 1) with the additional property that $\boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}} = N\mathbb{I}$. This means that $\boldsymbol{A} = \boldsymbol{H}/\sqrt{N}$ is an orthogonal matrix whose entries all have the same magnitude $1/\sqrt{N}$. In Sec. 4.4.2 we discuss other classes of feedback matrices.

4.4.1.4 Designing lossy components

So far we have designed the lossless prototype. Now we have to correct it by inserting the low-pass filters $H_{lp,i}$ and the correction filter E. The $H_{lp,i}$'s set the reverberation time from infinity to a finite value, by moving the poles slightly inside the unit circle. More precisely, they can be chosen to tune the reverberator to a desired, frequency-dependent reverberation time $T_r(\omega)$.

The following analysis assumes that the filters $H_{lp,i}$ are all defined as $H_{lp,i}(z) = [G(z)]^{m_i}$. This is conceptually equivalent to substituting each delay z^{-1} in the lines with a "damped delay" $G(z)z^{-1}$, where the factor $G(z)$ represents a *filtering per sample* in the propagation medium. We also make the simplifying hypotheses that *(1)* the response $G(\mathrm{e}^{j\omega})$ is zero-phase and that *(2)* the magnitude $\left| G(\mathrm{e}^{j\omega}) \right|$ is close to 1. Now assume that the lossless prototype has poles $e^{j\omega_i/F_s}$, $i = 1, \ldots N$, then the insertion of the low-pass filters moves the poles to

$$p_i \approx R_i e^{j\omega_i/F_s}, \quad \text{with } R_i = G\left(R_i e^{j\omega_i/F_s}\right) \approx G\left(e^{j\omega_i/F_s}\right), \tag{4.42}$$

where we have exploited our first simplifying hypothesis in assuming that the filters affect the radius of the poles and not their angles, and we have exploited our second simplifying hypothesis in the last approximation for R_i.

We know that the component of the impulse response arising from the ith pole of the system decays like R_i^n, as a function of discrete time n. Therefore the time needed for this response to decay by 60 dB (i.e. $T_r(\omega_i)$) satisfies the relation $20\log_{10}\left(R_i^{T_r(\omega_i)F_s}\right) = -60$ dB. From Eq. (4.42), and recalling that $H_{lp,i} = G^{m_i}$, we conclude that the ideal low-pass filter satisfies the relation

$$20\log_{10}\left| H_{lp,i}\left(e^{j\omega_i/F_s}\right)\right| = -60\frac{m_i}{F_s T_r(\omega_i)}. \tag{4.43}$$

Having been derived in the assumption of zero-phase, this expression disregards the phase response of the $H_{lp,i}$'s, which has the effect of slightly modifying the effective length of the delay m_i. It is usually assumed that in practice this correction has no perceivable effect and can therefore be ignored.

A consequence of incorporating the filters $H_{lp,i}(z)$ into the delay lines is that the energy of each decaying mode of the system response will be affected, i.e. the envelope of the frequency response of the system will no longer be flat. In particular, for exponentially decaying reverberation the envelope is proportional to the reverberation time at all frequencies. The role of the filter $E(z)$ (often referred to as the *tonal correction filter*) is to compensate for this effect: a flat frequency response envelope is restored if the magnitude response of $E(z)$ is inversely proportional to the reverberation time:

$$\left| E\left(e^{j\omega/F_s}\right)\right| \sim \frac{1}{\sqrt{T_r(\omega)}}. \tag{4.44}$$

Having specified ideal filter responses for the $H_{lp,i}$'s and for E, any number of filter-design methods can be used to find low-order filters that reasonably approximate Eqs. (4.43, 4.44). Note that this design effectively decouples the control over reverberation time from the overall reverberator gain.

M-4.10

Write a function that computes filter coefficients for $H_{lp,i}(z)$ and $E(z)$, given a function $T_r(\omega)$ specified on a set of points $\{\omega_k\}$, and given the filter order k.

Since the function $T_r(\omega)$ is typically very smooth and slowly varying with respect to ω, the filters $H_{lp,i}(z)$ can be chosen to have low order. In particular, first-order filters of the form (4.33) can be used:

$$H_{lp,i}(z) = \frac{g_{1,i}}{1 - g_{2,i} z^{-1}}. \tag{4.45}$$

In this case one can use Eq. (4.43) to find the gains (we only report results):

$$g_{2,i} = \frac{\ln(10)}{4} \log_{10}(a_i) \left(1 - \frac{T_r(0)^2}{T_r(\pi F_s)^2} \right), \quad g_{1,i} = a_i(1 - g_{2,i}) \tag{4.46}$$

where $a_i = 10^{-3 \frac{m_i}{F_s T_r(0)}}$ is determined from the desired reverberation time at $\omega = 0$, while $g_{2,i}$ sets the reverberation time at high frequencies.

If first-order low-pass filters of the form (4.45) are used, then one can use a correction filter which is also first-order and is determined as follows (we only report results):

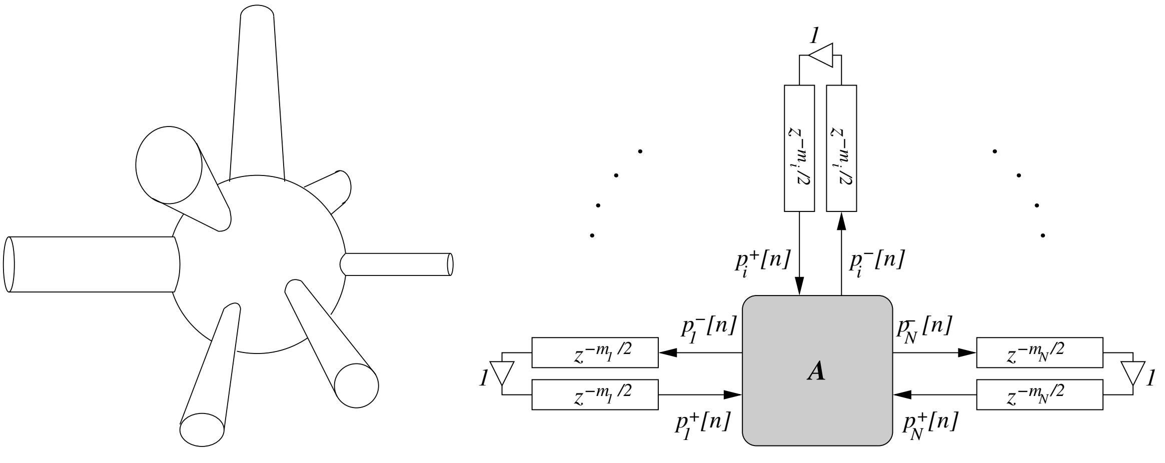
$$E(z) = \frac{1 - bz^{-1}}{1 - b}, \quad \text{with } b = \frac{1 - \frac{T_r(\pi F_s)}{T_r(0)}}{1 + \frac{T_r(\pi F_s)}{T_r(0)}}. \tag{4.47}$$

M-4.11

Write a function that computes filter coefficients for $H_{lp,i}(z)$ and $E(z)$ in the first-order case described above, given a function $T_r(\omega)$ specified on a set of points $\{\omega_k\}$.

M-4.11 Solution

```
function [num_H,den_H,num_E,den_E] = lossy_fdn(Fs,Tr,m);
%computes (first-order) lowpass filters H and correction filter E for a fdn,
%given an array of frequency-dependent Tr and the array m of fdn delays

N=length(m);
for i=1:N
    a=10^( -3*m(i)/(Fs*Tr(1)) );
    g2=(log(10)/4)*log10(a)*(1- ( Tr(1)^2/Tr(length(Tr))^2 ) );
    num_H(i,:)=[a*(1-g2), 0];
    den_H(i,:)=[1, -g2];
end
b=(1-(Tr(length(Tr))/Tr(1)))/(1+(Tr(length(Tr))/Tr(1)));
num_E=[1, -b];    den_E=[1-b, 0];
```



(a) (b)

Figure 4.17: *DWN reverberator*

4.4.2 Digital waveguide networks

4.4.2.1 The link between FDNs and DWNs

In Eq. (4.40) we have introduced a specific Householder reflection matrix, constructed from the reference vector $\boldsymbol{u} = [1, \dots, 1]^T$. In fact a Householder matrix can be constructed given any reference vector $\boldsymbol{u}$. We now want to provide a geometric interpretation of this family of matrices.

Consider the *projection matrix* $\boldsymbol{P}_u$, which orthogonally projects any vector $\boldsymbol{x}$ onto the vector $\boldsymbol{u}$:

$$\boldsymbol{P}_u = \frac{\boldsymbol{u}\,\boldsymbol{u}^T}{\boldsymbol{u}^T\,\boldsymbol{u}} = \frac{\boldsymbol{u}\,\boldsymbol{u}^T}{\|\,\boldsymbol{u}\,\|^2}, \qquad \text{then} \qquad \boldsymbol{x}_u := \boldsymbol{P}_u\,\boldsymbol{x} = \boldsymbol{u}\,\frac{\langle \boldsymbol{u}, \boldsymbol{x}\rangle}{\|\,\boldsymbol{u}\,\|^2} \tag{4.48}$$

is the orthogonal projection of $\boldsymbol{x}$ onto $\boldsymbol{u}$. Now consider the vector $\boldsymbol{x}_u^{\perp} := (\mathbb{I} - \boldsymbol{P}_u)\boldsymbol{x}$: this is the projection of $\boldsymbol{x}$ onto the hyperplane orthogonal to $\boldsymbol{u}$, since it is easily verified that $\boldsymbol{x}_u^{\perp} \perp \boldsymbol{x}_u$ and that $\boldsymbol{x}_u^{\perp} + \boldsymbol{x}_u = \boldsymbol{x}$.

Finally consider the vector $\boldsymbol{y}$ obtained by *reflecting* $\boldsymbol{x}$ about $\boldsymbol{u}$. Elementary geometrical considerations allow to conclude that this vector is the difference between $\boldsymbol{x}_u$ and $\boldsymbol{x}_u^{\perp}$:

$$\boldsymbol{y} = \boldsymbol{x}_u - \boldsymbol{x}_u^{\perp} = \boldsymbol{P}_u\boldsymbol{x} - (\mathbb{I} - \boldsymbol{P}_u)\boldsymbol{x} = (2\boldsymbol{P}_u - \mathbb{I})\boldsymbol{x}. \tag{4.49}$$

The matrix $(2\boldsymbol{P}_u - \mathbb{I})$ is a Householder matrix as defined in Eq. (4.40), except for a sign. Therefore we conclude that given a reference vector $\boldsymbol{u}$ the corresponding Householder matrix reflects any vector $\boldsymbol{x}$ about $\boldsymbol{u}$.

Having undestood the meaning of Householder matrices, we now construct a digital waveguide network (DWN) that is equivalent to the FDN lossless prototypes considered in the previous section. We start by considering the physical resonator depicted in Fig. 4.17(a). It is composed by N acoustic bores connected in parallel. In Chapter *Sound modeling: source based approaches* we have derived the $N \times N$ *scattering matrix* $\boldsymbol{A}$ that relates the incoming pressure waves $\boldsymbol{p}^+$ to the outgoing pressure waves $\boldsymbol{p}^-$. In this section we reconsider that matrix when the pressure waves in the ith bore are defined as

$$p_i^+ = \sqrt{\Gamma_i}\frac{p_i + Z_i u_i}{2}, \qquad p_i^- = \sqrt{\Gamma_i}\frac{p_i - Z_i u_i}{2}, \tag{4.50}$$

where Z_i and $\Gamma_i = 1/Z_i$ are the wave impedance and admittance of the ith bore. These are often referred to as *normalized* waves, and differ from our previous definition of wave variables uniquely for the scaling factor $\sqrt{\Gamma_i}$. It is straightforward to see that normalized pressure waves are scattered as $\boldsymbol{p}^- = \boldsymbol{A}\boldsymbol{p}^+$, where

$$\boldsymbol{A} = \begin{bmatrix} \frac{2\Gamma_1}{\Gamma_J} - 1, & \frac{2\sqrt{\Gamma_1\Gamma_2}}{\Gamma_J}, & \dots & \frac{2\sqrt{\Gamma_1\Gamma_N}}{\Gamma_J} \\ \frac{2\sqrt{\Gamma_2\Gamma_1}}{\Gamma_J}, & \frac{2\Gamma_2}{\Gamma_J} - 1, & \dots & \frac{2\sqrt{\Gamma_2\Gamma_N}}{\Gamma_J} \\ \vdots & & \ddots & \vdots \\ \frac{2\sqrt{\Gamma_N\Gamma_1}}{\Gamma_J}, & \frac{2\sqrt{\Gamma_N\Gamma_2}}{\Gamma_J}, & \dots & \frac{2\Gamma_N}{\Gamma_J} - 1 \end{bmatrix}, \qquad \text{where} \quad \Gamma_J = \sum_{l=1}^{N} \Gamma_l. \tag{4.51}$$

This normalized scattering matrix is immediately recognized as a Householder matrix:

$$\boldsymbol{A} = \frac{2}{\|\boldsymbol{\Gamma}\|}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T - \mathbb{I}, \qquad \text{with} \quad \Gamma := \left[\sqrt{\Gamma_1}, \sqrt{\Gamma_2}, \dots, \sqrt{\Gamma_N}\right], \tag{4.52}$$

so we have this interesting geometrical interpretation: scattering of normalized pressure waves corresponds to a reflection around the vector $\boldsymbol{\Gamma}$.

If the acoustic bores are lossless and with ideal closed terminations, and if the length (in samples) of the ith bore is $m_i/2$, then the physical resonator of Fig. 4.17(a) can be modeled with the digital waveguide network given in Fig. 4.17(b). Now compare this scheme with the lossless FDN of Fig. 4.16: apart from the input signals $x_i[n]$, the two schemes implement the same computational structure. The incoming pressure waves $p_i^+[n]$ correspond to the output signals $y_i[n]$, and the outgoing pressure waves $p_i^-[n]$ correspond to the feedback signals generated by the feedback matrix.

4.4.2.2 General lossless scattering matrices

Showing the equivalence between DWNs and FDNs is more than a mere intellectual exercise: we can now design an entire new class of lossless FDN prototypes, in which the feedback matrix $\boldsymbol{A}$ is given by Eq. (4.51) and have a straightforward physical interpretation.

Note that the matrix in Eq. (4.51) is still orthogonal (it is easy to verify that $\boldsymbol{A}\boldsymbol{A}^T = \mathbb{I}$. We can push the generalization further by generalizing our definition of losslessness, and consequently define new classes of lossless feedback matrices that are neither physical nor orthogonal. Consider a Hermitian, positive-definite $N \times N$ matrix $\boldsymbol{\Gamma}$ (we use this notation because we interpret $\boldsymbol{\Gamma}$ as a generalized junction admittance). This matrix induces a norm $\|\cdot\|_\Gamma$, defined as follows: $\|\boldsymbol{x}\|_\Gamma := \boldsymbol{x}^T\boldsymbol{\Gamma}\boldsymbol{x}$ for any real valued N-dimensional vector $\boldsymbol{x}$. We can then define a waveguide scattering matrix $\boldsymbol{A}$ to be "lossless" if the scattering preserve the norm, i.e. the equality $\|\boldsymbol{p}^+\|_\Gamma = \|\boldsymbol{p}^-\|_\Gamma$ holds. This condition is clearly equivalent to the condition

$$\boldsymbol{A}^T\boldsymbol{\Gamma}\boldsymbol{A} = \boldsymbol{\Gamma} \tag{4.53}$$

for the scattering matrix $\boldsymbol{A}$. In the case $\boldsymbol{\Gamma} = \mathbb{I}$, the norm $\|\cdot\|_\Gamma$ is the euclidean norm and the above equation reduces to the condition of $\boldsymbol{A}$ being orthogonal. In the general case $\boldsymbol{\Gamma} \neq \mathbb{I}$ it can be shown that Eq. (4.53) holds if and only if $\boldsymbol{A}$ has eigenvalues with modulus 1 and N linearly independent eigenvectors. We do not provide a proof of this characterization: intuitively it means that when such a feedback matrix is used in a lossless FDN prototype the system poles all have unit modulus and thus the system response consists of non-decaying eigenmodes.

Clearly orthogonal matrices are lossless in this sense, since they have unitary eigenvalues and pairwise orthogonal eigenvectors. Another class of matrices that satisfy this condition are triangular matrices: designing a triangular matrix with unitary eigenvalues is straightforward since we know from linear algebra that they lie on the diagonal. Additional care is required in order to ensure that the triangular matrix possesses N independent eigenvectors.

4.4.2.3 Waveguide meshes

So far we have seen DWNs in analogy with FDNs. In this section we discuss a new multidimensional waveguide structure, named *waveguide mesh*, that can be used to physically simulate resonating enclosures. What follows is only a quick and qualitative introduction to the subject, the interested reader can refer to the bibliography.

Consider again the N-D D'Alembert equation (4.1). Similarly to what we have done in the 1-D case (Chapter *Sound modeling: source based approaches*), we can simulate the traveling wave solution by using delay lines. In this case the delay lines are arranged in a mesh, that represents waves propagating in the x, y, z directions. At each node of the mesh continuity constraints must be satisfied, namely the pressure waves in each direction must provide the same pressure value.[5] This means that at each node of the mesh the incoming pressure waves are scattered by a matrix identical to the matrix $\boldsymbol{A}$ given in Chapter *Sound modeling: source based approaches*, in which all the incoming branches share the same impedance:

$$\boldsymbol{A} = \begin{bmatrix} \frac{2}{N}-1 & \frac{2}{N} & \dots & \frac{2}{N} \\ \\ \frac{2}{N} & \frac{2}{N}-1 & \dots & \frac{2}{N} \\ \vdots & & \ddots & \vdots \\ \frac{2}{N} & \frac{2}{N} & \dots & \frac{2}{N}-1 \end{bmatrix}. \tag{4.54}$$

In order to clarify this idea, let us examine the 2-D case shown in Fig. 4.18. The outgoing pressure waves at each node are computed as $\boldsymbol{p}^- = \boldsymbol{A}\boldsymbol{p}^+$, i.e.

$$p_i^-[n] = p_J[n] - p_i^+[n] \quad (i = 1 \dots 4) \quad \text{where } p_J[n] = \frac{\sum_{i=1}^4 p_i^+[n]}{2} \tag{4.55}$$

is the junction pressure. It can be shown that this rectangular waveguide mesh is equivalent to a finite-difference numerical solution of the the 2-D D'alembert equation, in which the pressure at a certain node is expressed in terms of the pressures at its neighboring nodes one sample earlier, and itself two samples earlier.

The rectangular layout depicted in Fig. 4.18 is not the only possible one: other geometries may be used for assembling the mesh, like triangular, hexagonal, and so on. The choice of the geometry has a major influence on the *dispersion* error in the mesh, i.e the error in propagation speed as a function of frequency and direction along the mesh. It can be shown that the triangular waveguide mesh is the simplest 2-D mesh geometry with the least dispersion variation as a function of direction of propagation. In other words, the triangular mesh is closer to isotropic than all other known elementary geometries. Isotropy can be obtained also through interpolation, i.e. by using non integer propagation delays, but computational costs are higher. As far as frequency dispersion is concerned, frequency-warping methods can be used to minimize it in the mesh.

The waveguide meshes analyzed so far simulates lossless propagation in an infinite medium. In order to model something similar to a real resonating enclosure we must add losses and boundary conditions into the structure. The techniques discussed in Chapter *Sound modeling: source based approaches* to simulate lossless in 1-D wave propagation can be extended to the waveguide mesh: the basic idea is once again that wave propagation during one sampling interval (in time) is associated with linear filtering by $G(z)$. The problem of modeling mesh boundaries is particularly important in the context of artificial reverberation: in order to obtain high temporal reflection densities, maximally *diffusing* boundaries have to be modeled.

As efficient solutions are found to deal with the above mentioned problems, 3-D waveguide meshes are being more and more used for the simulation of acoustic spaces.

[5] In this section we are using waveguide meshes to simulate resonating enclosures and thus we work with pressure waves and consider parallel junctions. Waveguide meshes can also be used to simulate mechanical resonators, e.g membranes, and in that case it is natural to choose velocity waves and to consider series junctions at mesh nodes.

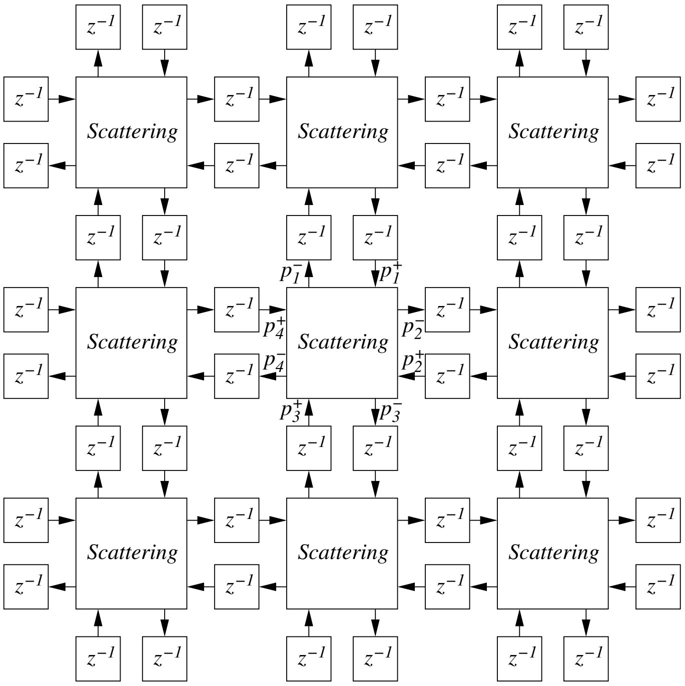


Figure 4.18: *2D rectilinear digital waveguide mesh.*

4.5 Spatial hearing

In the previous sections we have learned how a sound produced by an acoustic source is affected by the surrounding environment. So far we have assumed that the receiver is a point in the space, which is reasonable e.g. for a omnidirectional microphone. We now want to study a different type of receiver, i.e. a human receiver with two ears and one head in between.

Throughout the next sections our assumption will be that the two acoustic pressure signals at the two eardrums contain all the information that is used by a human listener to elaborate his/her auditory perception. In other words we assume that if different acoustic events (e.g. different sounds, or different sound/listener positions in the environment, etc.) produce the same pair of acoustic pressures at the eardrums, they will be perceived by a human listener as the same acoustic event.[6] In particular these signal will provide the listener with *spatial information*, about the location of the sound source relative to the listener.

With this assumption, our goal is to understand and simulate how sound is transformed in his path to the eardrum by neighboring parts of the body (such as head and shoulders), by the pinna (the visible portion of the outer ear), and by the ear canal (the meatus that leads to the eardrum).

4.5.1 The sound field at the eardrum

Spatial attributes of the sound field are coded into temporal and spectral attributes of the acoustic pressure at the eardrum, via the filtering effect of three main elements: head, external ear, and torso/shoulders.

[6] In fact this is not entirely correct. Sound reaches our ears also through bone conduction. Moreover auditory perception interacts with other types of information (e.g., conflicting visual cues) and is affected by adaptation and expectations.

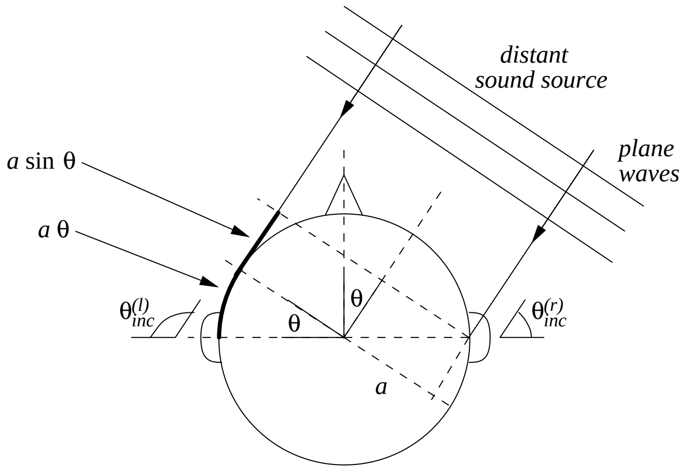


Figure 4.19: *Estimate of ITD in the case of a distance sound source (plane waves) and spherical head.*

4.5.1.1 Head

Our ears are not isolated objects in space. They are located, at the same height, on opposite sides of an acoustically rigid object: the head. This acts as an obstacle to the free propagation of sound and has two main effects: *(1)* it introduces an *interaural time difference (ITD)*, because a sound wave has to travel an extra distance in order to reach the farthest ear, and *(2)* it introduces an *interaural level difference (ILD)* because the farthest ear is acoustically "shadowed" by the presence of the head.

An approximate yet quite accurate description of the ITD can be derived using a few simplifying assumptions, in particular by considering the case of "distant" sound sources and a spherical head: this situation is depicted in Fig. 4.19. The first assumption implies that the sound waves that strike the head are plane waves. Then the extra-distance Δx needed for a sound ray to reach the farthest ear is estimated from elementary geometrical considerations, as shown in Fig. 4.19, and the ITD is simply $\Delta x/c$. Therefore

$$\text{ITD} \sim \frac{a}{c}(\theta + \sin\theta), \tag{4.56}$$

where a is the head radius and θ is the *azimuth* angle that defines the direction of the incoming sound on the horizontal plane. This formula shows that the ITD is zero when the source is directly ahead ($\theta = 0$), and is a maximum of $a/c(\pi/2 + 1)$ when the source is off to one side ($\theta = \pi/2$). This represents an ITD of more than 0.6 ms for a head radius $a = 8.5$ cm, which is a realistic value.

While it is acceptable to approximate the ITD as a frequency independent parameter, as we did in Eq. (4.56), the ILD is highly frequency dependent: at low frequencies (i.e., for wavelengths that are long relative to the head diameter) there is hardly any difference in sound pressure at the two ears, while at high frequencies differences become very significant. Again, the ILD can be studied in the case of an ideal spherical head of radius a, with a point sound source located at a distance $r > a$ from the center of the sphere. It is customary to use the normalized variables $\mu = \omega a/c$ (normalized frequency) and $\rho = r/a$ (normalized distance). If we consider a point on the sphere, then the diffraction of an acoustic wave by the sphere seen on the chosen point is expressed with the transfer function

$$H_{\text{sphere}}(\rho, \theta_{inc}, \mu) = -\frac{\rho}{\mu} e^{-i\mu\rho} \sum_{m=0}^{+\infty} (2m+1) P_m(\cos\theta_{inc}) \frac{h_m(\mu\rho)}{h'_m(\mu)}, \tag{4.57}$$

where P_m and h_m are the mth order Legendre polynomial and spherical Hankel function, respectively,

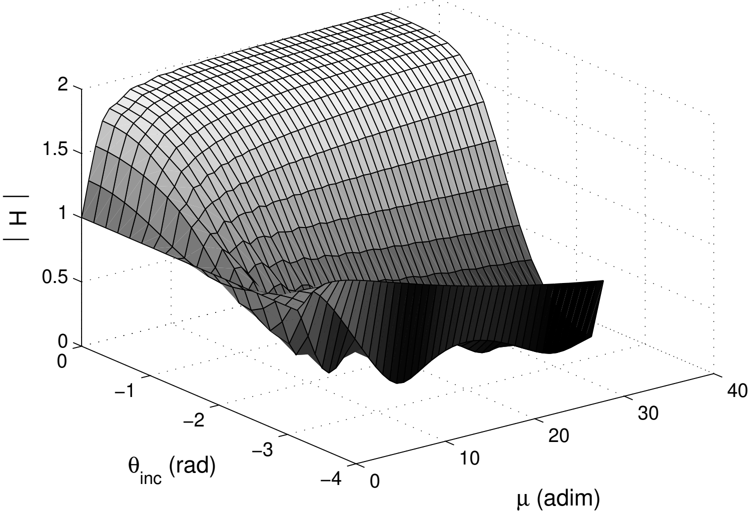


Figure 4.20: *Magnitude response* $\left| H_{sphere}(\infty, \theta_{inc}), \mu \right|$ *of a sphere for an infinitely distant source.*

and θ_{inc} is the angle of incidence, i.e. the angle between the ray from the center of the sphere to the source and the ray to the measurement point on the surface of the sphere.[7] Normal incidence corresponds to $\theta_{inc} = 0$, while the sphere point opposite to the source is at $\theta_{inc} = \pi$.

It is known that the Hankel function $h_m(x)$ admits an asymptotic approximation as the argument x goes to infinity. By exploiting this approximation one can study the behavior of the transfer function $H_{\text{sphere}}(\infty, \theta_{inc}, \mu)$ as the distance r between the source and the sphere becomes arbitrarily large. The approximate solution $\left| H_{\text{sphere}}(\infty, \theta_{inc}), \mu \right|$ is plotted in Fig. 4.20.

At low frequencies the transfer function is not directionally dependent and the magnitude $\left| H_{\text{sphere}} \right|$ is essentially unity for any angle θ_{inc}. When μ exceeds 1 the dependence on θ_{inc} becomes noticeable. The response increases around the front of the sphere, and in particular exhibits a 6 dB boost at high frequencies near the front of the sphere ($\left| H_{\text{sphere}}(\infty, 0, \infty) \right| = 2$), consistently with the requirement that in this limit the solution must reduce to that of a plane wave normally incident on a rigid plane surface. $\left| H_{\text{sphere}} \right|$ is approximately flat when θ_{inc} is around 100 degrees, and progressively decreases around the back of the sphere. Note however that the minimum response does not occur at the very back ($\theta_{inc} = \pi$). Instead, this point exhibits a so-called "bright spot" effect, which is due to the fact that all the waves propagating around the sphere arrive at that point in phase. At very high frequencies the bright-spot lobe becomes extremely narrow, and the back of the sphere is effectively in a sound shadow. Finally, note that interference effects caused by waves propagating in various directions around the sphere introduce ripples in the response that are quite prominent on the shadowed side.

4.5.1.2 The external ear

The external ear consists of the *pinna* and the *ear canal* until the eardrum. Beyond the eardrum are the middle ear and the internal ear. For the purpose of this chapter we are interested in the external ear only. In Chapter *Auditory based processing* we will study the middle and internal ear.

The pinna, schematically depicted in Fig. 4.21(a), has a characteristic "bas-relief" form with features

[7] We are using a different notation with respect to the azimuth angle θ used previously, in order to avoid confusion. Given a 2-D reference system like that in Fig. 4.19, the transfer functions (4.57) at the right and left ear will use the angles $\theta_{inc}^{(r)} = \pi/2 - \theta$ and $\theta_{inc}^{(l)} = \pi/2 + \theta$, respectively.

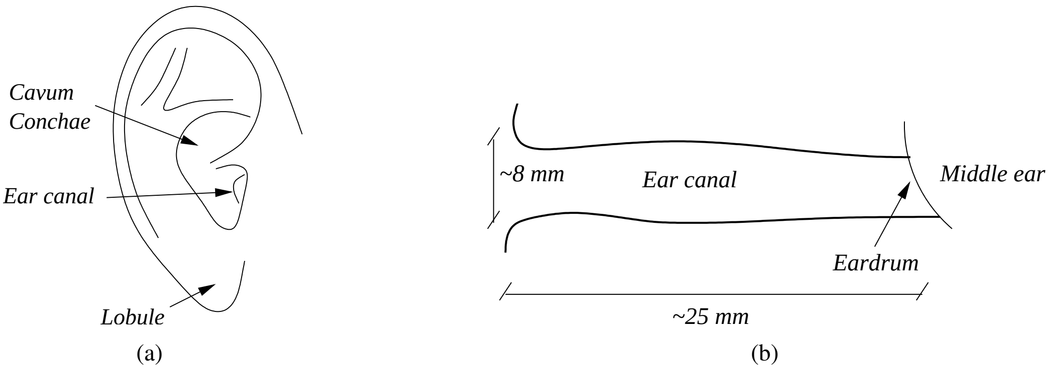


Figure 4.21: *External ear: (a) pinna, and (b) ear canal.*

that differ greatly from one individual to another (just look at people's ears). The pinna is connected to the ear canal, depicted in Fig. 4.21(b). It can be approximately described as a tube of constant width, with walls of high acoustic impedance. At the end opposite to the pinna, the ear canal is terminated by the eardrum diaphragm.

At a first approximation the acoustic behaviour of the ear canal is easily understood: it behaves like a one-dimensional resonator. On the other hand the pinna has much more complex effects, as it basically acts like an acoustic antenna. Its resonant cavities amplify some frequencies, and its geometry leads to interference effects that attenuate other frequencies. Moreover, its frequency response is directionally dependent. Acoustically it acts like a filter whose transfer function depends in general on the distance and direction of the sound source relative to the ear. Like for any other resonator, we can interpret these filtering effect either by looking at reflections of sound rays or in the frequency domain.

First approach: external ear as a sound reflector. Figure 4.22(a) shows two different directions of arrival. In each case there are two paths from the source to the ear canal –a direct path and a longer path following a reflection from the pinna. At moderately low frequencies, the pinna essentially collects additional sound energy, and the signals from the two paths arrive in phase. However, at high frequencies, the delayed signal is out of phase with the direct signal, and destructive interference occurs. The greatest interference occurs when the difference in path length is a half wavelength: this produces a “pinna notch”. Since the pinna is a more effective reflector for sounds coming from the front than for sounds from above, the resulting notch is much more pronounced for sources in front than for sources above. In addition, the path length difference changes with elevation.

However reflection models are suspect whenever the dimensions of the reflecting surfaces are comparable to (or even smaller than) the acoustic wavelengths under exam. At the very least, the reflection coefficients should be frequency dependent. A more thorough approach is modal analysis of the external ear resonator, through measurements of frequency responses using an imitation pinna and a ear canal with high impedance termination. Such measurements give results like those depicted in Fig. 4.22(b). First resonance is that of a open-closed tube $\sim 33\%$ longer then the ear canal: the pinna acts as a prolongation of the ear canal with an aperture effect. Second resonance is a resonance of the *cavum concha* alone: the pressure distribution is similar to what would be obtained if the canal were plugged. The higher resonances instead are again associated to longitudinal standing waves: these are not very widely spaced and are quite dependent on the individual, therefore it can combine in a single broad peak of the magnitude response.

The synthetic conclusion of this section is then that the pinna and the ear canal form a systems of acoustic resonators, whose resonances are excited to different extents depending on the direction and

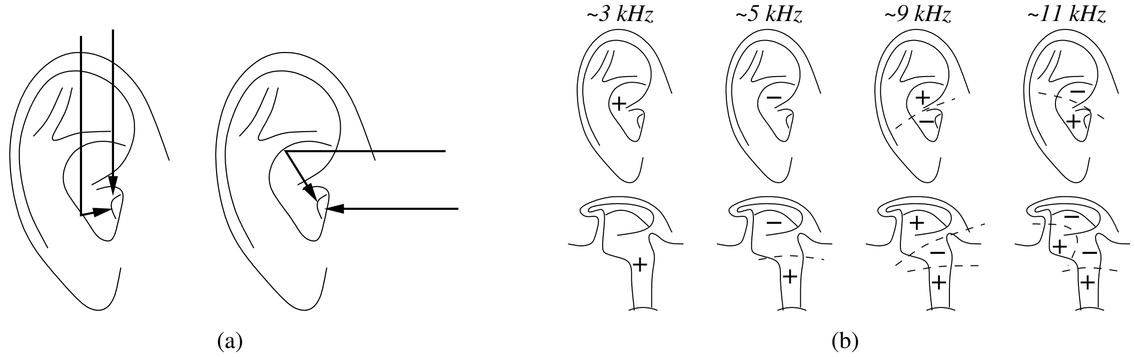


Figure 4.22: *Effects of pinna: (a) direction-dependent reflections, and (b) resonances.*

distance of the sound source.

4.5.1.3 Torso and shoulders

In the discussion up to now we have not considered a third element that, together with the head and the external ear, contributes to the shaping of the sound field at the eardrum: the torso. Torso and shoulders affect incident sound waves in two main respects. First, they provide additional reflections that sum up with the direct sound. Second, they have a shadowing effect for sound rays coming from below.

The geometry of the torso is quite complicated. However a simplified description can be derived by considering an ellipsoidal torso below a spherical head. These kind of approximate descriptions are sometimes called “snowman models”, for obvious reasons. Figure 4.23(a) depicts a snowman model and shows the main effects of the ellipsoidal torso on the sound field at the ear.

Reflections: Fig. 4.23(a). If we measured the impulse response at the right ear for the sound source locations depicted in Fig. 4.23(a) we would see that the initial pulse is followed by a series of subsequent pulses, whose delays increase and then decrease with elevation. These additional pulses are caused by reflections on the torso.

We could exploit the simplified geometry of the snowman model to compute analitycally the delay of the reflected rays, given the model parameters and the sound source position. However some important remarks can already be made from a qualitative analysis. First, the delay between the direct sound and the reflected ray does not vary much if the sound source moves on a circumference in the horizontal plane (especially if its radius is large compared to the head radius). Second, the delay varies considerably if the sound source moves vertically, and in particular the reflected pulses are maximally delayed for sound source locations right above the listener. If we consider that the distance from the ear canal to the shoulder is roughly 16 cm, then a reflected ray from a source right above the subject has to travel an extra distance of approximately 32 cm, which corresponds to a delay of almost 1 ms.

In the frequency domain the torso reflections act as a comb filter, introducing periodic notches in the spectrum. The frequencies at which the notches occur are inversely related to the delays, and thus produce a pattern that varies with the elevation of the source. The lowest notch frequency corresponds to the longest delay. Delays longer than a sixth of a millisecond will produce one or more notches below 3 kHz, which is approximately the lowest frequency where pinna effects start to be noticeable.

Modeling the effects of the torso as specular reflections means accounting for only a part of the story. First, reflection is a high frequency concept. Second, and perhaps more important, as the source descends in elevation, a point of grazing incidence is reached, below which torso reflections disappear

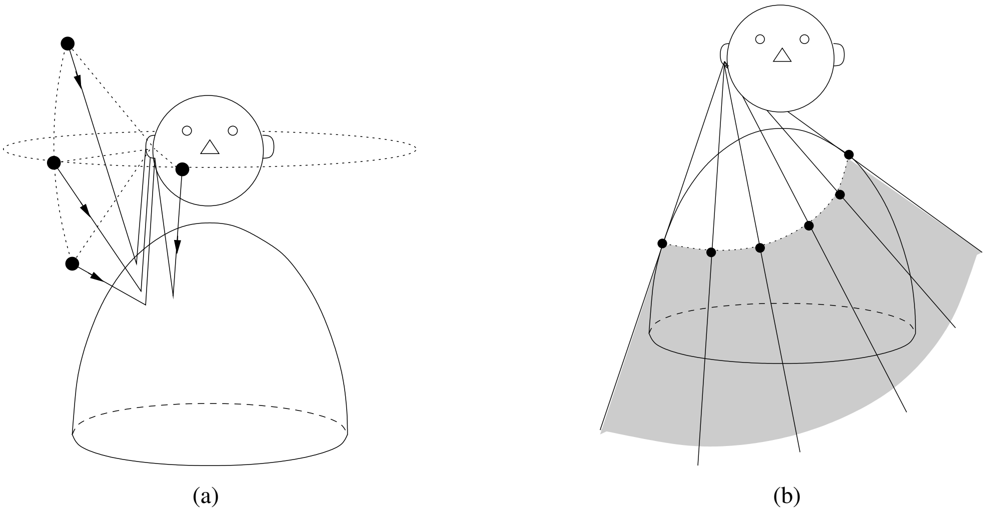


Figure 4.23: *Effects of torso: (a) reflections, and (b) shadowing.*

and *torso shadowing* emerges. As shown in Fig. 4.23(b), rays drawn from the ear to points of tangency around the upper torso define a torso-shadow cone. Clearly, the specular reflection model does not apply within the torso shadow cone. Instead, diffraction and scattering produce a qualitatively different behavior, characterized by a stronger attenuation for high frequencies (i.e. for wavelength comparable to or smaller than the size of the torso).

Although the acoustic effects of torso and shoulders are not as strong as those introduced by the pinna, they are important because they appear at lower frequencies, where typical sound signals have most of their energy and where the response of the pinna is essentially flat. In terms of frequency ranges the effects provided by the torso are therefore complementary to those provided by the pinna.

4.5.1.4 Head-related transfer functions

In the preceding sections we have investigated the influence of hear, torso and external ear on the sound field at the eardrum. All the effects that we have examined are linear, which means that *(1)* they can be described by means of transfer functions, and *(2)* they combine additively. Therefore the sound pressure produced by an arbitrary sound source at the eardrum is uniquely determined by the impulse response from the source to the eardrum. This is called *Head-Related Impulse Response (HRIR)*, and its Fourier transform is called *Head Related Transfer Function (HRTF)*. The HRTF captures all of the physical effects that we have examined separately in the previous sections.

The HRTF is a function of three spatial coordinates and frequency. Given the approximately spherical shape of the head, it is customary to use the spherical coordinates depicted in Fig. 4.24, which use slightly different notations and conventions with respect to more traditional definition (see our definition of spherical coordinates in Chapter *Sound modeling: source based approaches*). Specifically, in this context the vertical and horizontal angular coordinates *azimuth* and *elevation* are noted as θ and ϕ, respectively, while the radial coordinate is named *range* and noted as r. Moreover, two different spherical coordinate systems are used in the literature. Figure 4.24(a) show the most popular one, sometimes called *vertical polar* coordinate system: in this system the azimuth is measured as the angle from the yz plane to a vertical plane containing the source and the z azis, and the elevation is measured as the angle up from the xy plane. With this choice, surfaces of constant azimuth are planes through the z axis, and surfaces

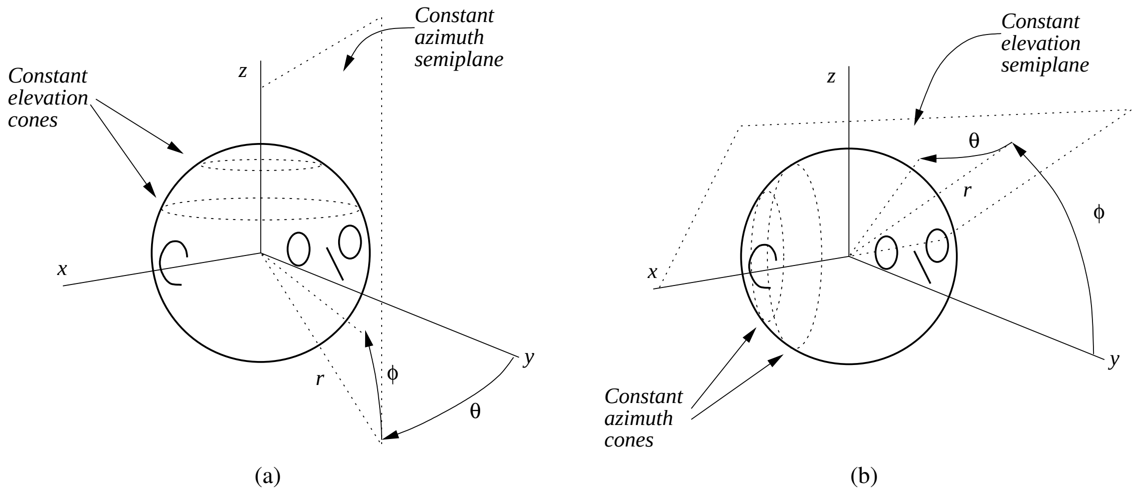


Figure 4.24: *Spherical coordinate systems used in the definition of HRTFs: (a) vertical-polar coordinate system, and (b) interaural-polar coordinate system.*

of constant elevation are cones concentric about the z axis.

In alternative the so-called *interaural-polar* coordinate system, shown in Fig. 4.24(b), is sometimes used. In this case the elevation is measured as the angle from the xy plane to a plane containing the source and the x axis, and the azimuth is then measured as the angle from the yz plane. With this choice, surfaces of constant elevation are planes through the x axis, and surfaces of constant azimuth are cones concentric with the x axis. One advantage of this system is that it makes it significantly simpler to express interaural differences at all elevations (in particular the constant-azimuth cones are the loci of points that share equals ILD and ITD values for a spherical head).

In the remainder of this chapter we will specify, when necessary, whether we are using the vertical-polar or the interaural-polar coordinate system. In any case we will indicate the HRTFs as $H^{(l),(r)}(r, \theta, \phi, \omega)$, where superscripts $(l), (r)$ indicate the HRTF at the left and right ear, respectively. When $r \rightarrow +\infty$ (which in practice means $r > 1$ m, a condition that is met in most applications), the source is said to be in the *far field*. In this case we will use the notation $H^{(l),(r)}(\theta, \phi, \omega)$. Finally, in the hypotesis of a perfectly symmetrical geometry will will simply write $H(\theta, \phi, \omega)$, with $H^{(r)}(\theta, \phi, \omega) = H(\theta, \phi, \omega)$ and $H^{(l)}(\theta, \phi, \omega) = H(-\theta, \phi, \omega)$.

We formally define the HRTF at one ear as the frequency-dependent ratio between the sound pressure level (SPL) $\Phi^{(l),(r)}(\theta, \phi, \omega)$ at the corresponding eardrum and the free-field SPL at the center of the head $\Phi_f(\omega)$ as if the listener were absent:

$$H^{(l)}(\theta, \phi, \omega) = \frac{\Phi^{(l)}(\theta, \phi, \omega)}{\Phi_f(\omega)}, \qquad H^{(r)}(\theta, \phi, \omega) = \frac{\Phi^{(r)}(\theta, \phi, \omega)}{\Phi_f(\omega)}. \tag{4.58}$$

Figures 4.25(a) and 4.25(b) show two examples of HRTFs (magnitude response only): all the effects examined in this section combine to form a surprisingly complicated function of θ and ϕ.

4.5.2 Perception of sound source location

This is complicate matter. Many competing and interfering effects can influence auditory perception of sound source location. In this section we provide a brief summary, but we warn the reader to be cautios when dealing with this matter and always to be aware of limitations and simplifying hypoteses.

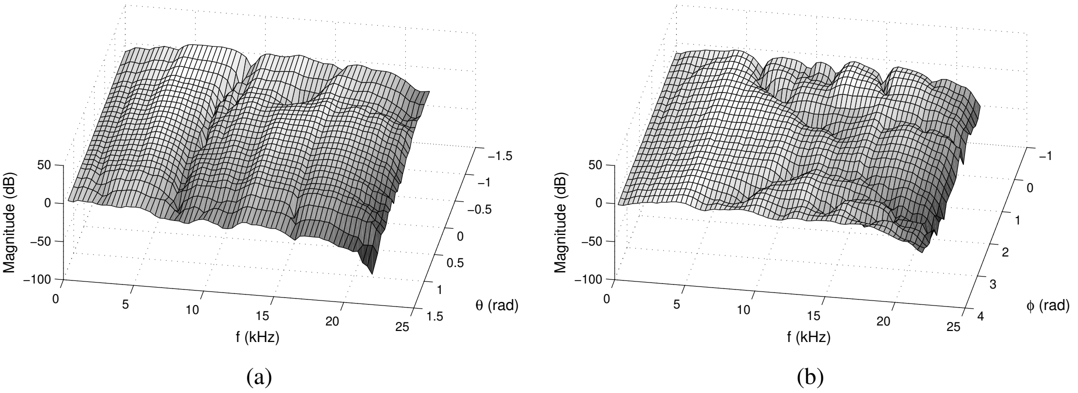


Figure 4.25: *Example of magnitude of HRTFs (a) in the* xy *plane (*$\theta \in [-\pi/2, \pi/2]$*,* $\phi = 0$*) and (b) in the* yz *plane (*$\theta = 0$*,* $\phi \in [-\pi/4, \pi]$*). Interaural polar coordinates are used.*

4.5.2.1 Azimuth perception

The horizontal placement of the ears maximizes differences for sound events occurring around the listener, rather than from below or above, enabling audition of sound sources at the terrain level and outside the visual field of view. The ITD and the ILD are considered to be the key parameters for azimuth perception, in what is sometimes referred to as the *Duplex Theory* of localization.

For the sake of clarity, consider a sine wave reaching the left and right ear. At low frequencies the ITD shifts the waveform a fraction of a cycle, which is easily detected: see Fig 4.26(a). Qualitatively one can say that if the half wavelength is larger than the size of the head, then it is possible for the auditory system to detect the phase of these waveforms unambiguously, and the ITD cue can function. On the other hand, at high frequencies there is ambiguity in the ITD, since there can be several cycles of shift: see Fig 4.26(b). Qualitatively, we can consider the critical point to be the point where the half wavelength becomes shorter than the head size: for shorter wavelengths, the phase information in relation to relative time of arrival at the ears can no longer convey which is the leading wavefront. The critical point in frequency is usually assumed to be a value around 1.5 kHz.

If we now look at the ILD the situation is reversed. As we have seen in Sec. 4.5.1 (see in particular Fig. 4.20), at low frequencies the head transfer function is essentially flat and therefore there is little ILD information. On the other hand, at high frequencies the ILD is more marked and can become very large. For this reason the Duplex Theory asserts that the ILD and the ITD are complementary cues to azimuth perception, and that taken together they provide azimuth perception throughout the audible frequency range.

This is not completely true, though. In fact timing information can be exploited for azimuth perception also in the high frequency range because the timing differences in amplitude envelopes are detected. Again, for the sake of clarity consider a sine wave that is modulated in amplitude as in Fig. 4.26(c). Then an ITD envelope cue, sometimes referred to as *Interaural Envelope Difference (IED)* can be exploited, based on the hearing system's extraction of the timing differences from the transients of amplitude envelopes, rather than from the timing of the waveform within the envelope. This is demonstrated by the so-called Franssen Effect: if a sine wave is suddenly turned on and a high-pass-filtered version is sent to a loudspeaker "A" while a low-pass filtered version is sent to a loudspeaker "B", most listeners will localize the sound at A. This is true even if the frequency of the sine wave is sufficiently low that in steady state most of the energy is coming from B.

The information provided by ITD and ILD can be ambiguous. If we assume the spherical geometry of

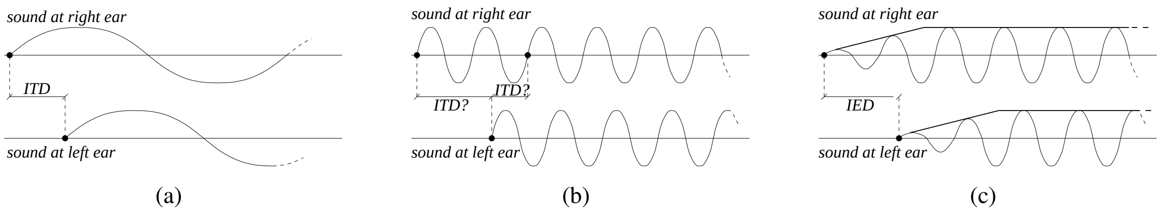


Figure 4.26: *Time differences at the ears; (a) non ambiguous ITD, (b) ambiguous ITD, and (c) IED.*

Fig. 4.19, a sound source located in front of the listener at a certain θ, and a second one located at the rear, at $\pi - \theta$, provide identical ITD and ILD values. In reality ITD and ILD will not be exactly identical at θ and $\pi - \theta$ because *(1)* human heads are not spherical, *(2)* there are asymmetries and other facial features, and (3) ears are not positioned as in Fig. 4.24 but lie below and behind the x axis. Nonetheless the values will be very similar, and *front-back confusion* is in fact often observed experimentally: listeners operate *reversals* in azimuth judgements, erroneously locating sources at the rear instead of at the front, or viceversa. The former reversal occurs more often than the latter. Some argue that this asymmetry may originate from a sort of ancestral "survival mechanism", according to which if something (a predator?) can be heard but not seen then it must be at the rear (danger!).

The Duplex Theory essentially works in anechoic conditions. But in everyday conditions reverberation can severely degrade especially ITD information. As we know, in a typical room reflections begin to arrive a few milliseconds after the direct sound. Below a certain sound frequency, the first reflections reach the ear before one oscillation period is completed. Before the auditory system estimates the frequency of the incoming sound wave, and consequently infers the ITD, the number of reflections at the ear has increased exponentially and the auditory system is not able to estimate the ITD. Therefore sounds that possess energy in the low-frequency range only (indicatively below 250 Hz) are essentially impossible to localize in a reverberant environment.[8] Instead the IED is used, because the starting transient provides unambiguous localization information, while the steady-state signal is very difficult to localize. In conclusion we can state –with some risk of oversimplification– that high-frequency energy only is important for localization in reverberant environments.

4.5.2.2 Lateralization and externalization

In Sec. 4.6 we will see that the simplest systems for spatial sound rendering are based on manipulation of the interaural cues examined above, and on headphone-based auditory display. These systems can be used in applications where only two-dimensional localization –in the horizontal plane– is required.

In this context, the term *lateralization* is typically used to indicate a special case of localization, where the spatial percept is heard inside the head, mostly along the interaural axis (the x of Fig. 4.24), and the means of producing the percept involves manipulation of ITD and/or ILD over headphones. Lateralization illustrates a fundamental example of virtual, as opposed to actual, sound source position. When identical monaural sounds are delivered from stereo headphones, the listener does not hear two distinct sounds coming from the transducers, and instead perceives a single virtual sound source which appears to be positioned at the center of the head. As ITD and ILD are increased, the perceived position of the virtual sound source will start to shift toward one ear, along an imaginary line. Once a critical value of the ITD or the ILD is reached, the perceived sound source will stop moving along the interaural axis and will be located at one of the ears. This effect is sometimes termed *inside-the-head localization*

[8] This is why surround systems use many small loudspeakers for high frequencies and one subwoofer for low frequencies.

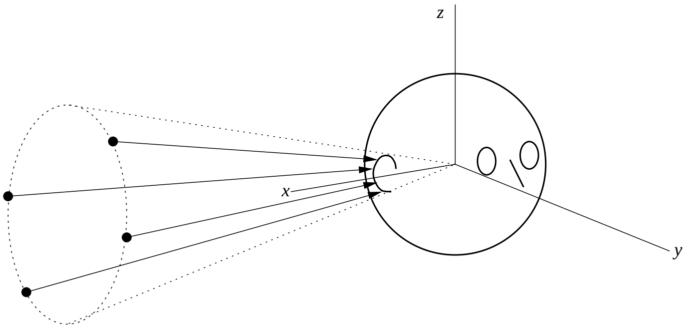


Figure 4.27: *Cone of confusion.*

(IHL). Having knowledge of this effect is important since headphone playback is otherwise superior to loudspeakers for transmitting virtual acoustic imagery in three dimensions.

Achieving *externalization* of the sound (i.e. in removing the IHL effect) is in many respects the "sacred graal" of headphone-based spatial audio systems. It is not completely clear what additional cues are most effective in producing sound externalization. However it has been observed by many that externalization increases as the stimulation approximates more closely a stimulation that is natural and that especially reverberation, either natural or artificial, can enhance dramatically externalization. In general, IHL is not an inevitable consequence of headphone listening, simply because externalized sounds can be heard through headphones in many instances.

4.5.2.3 Elevation perception

While the relevant cues for the localization of a sound source in the horizontal plane are relatively well understood, things become more complicated when we consider non-null elevations.

Figure 4.27 shows that sound sources located anywhere on a conical surface extending out from the ear of a spherical head produce identical values of ITD and ILD. These surfaces are often referred to as *cones of confusion*, and extend the concept of front/back confusion that we have examined above. Of course this situation is only theoretical: in reality ITD and ILD will never be completely identical on the cone of Fig. 4.27, because of the facial features and asymmetries already mentioned. Nonetheless, when ITD and ILD cues are maximally similar between two locations, a potential for confusion between the positions exists in the absence of other spatial cues.

The directional effects of the pinnae can disambiguate this confusion, and are considered to be particularly important for vertical localization. The role of the pinnae in improving vertical localization can be evaluated experimentally e.g. by comparing judgments made under normal conditions to a condition where the pinnae are bypassed or occluded. In fact vertical localization can be achieved even when one ear is completely occluded. This evidence supports the idea that the spectral cues provided by the pinnae work mainly monaurally.

There are many theories about the role of pinnae spectral cues. Very roughly, all of them suggest that a major cue for elevation involves movement of spectral notches and/or peaks, that change as a function of source and listener orientation. A way of appreciating the pinnae spectral cues is to examine the case of sound sources along the yz plane of the listener: note that this is the locus of the points where not only IID and ITD are null, but also spectral differences between the left and right HRTFs are null as long as the left and right pinnae are identical. If we look back at Fig. 4.25(b), we can notice a moving spectral notch that is thought to be important for elevation perception.

In general it is difficult without extensive psychoacoustic evaluation to ascertain how importantly these changes function as spatial cues. In particular, it is unclear if localization cues are derived from a particular spectral feature such as a peak or a notch, or from the overall spectral shape. Also, it is generally considered that a sound source has to contain substantial energy in the high-frequency range for accurate judgment of elevation, because the pinna has limited dimensions in space and wavelengths longer than the size of the pinna are not affected (see also Fig. 4.22(a)). One could roughly state that the pinnae have a relatively little effect below 3 kHz.

While the role of the pinna in vertical localization has been extensively studied, the role of the torso is less well understood. We have seen in Sec. 4.5.1 that the torso disturbs incident sound waves at frequencies lower than those affected by the pinna. However, the effects of the torso are relatively weak, and experiments to establish the perceptual importance of low-frequency cues have produced mixed results.

4.5.2.4 Distance perception

It is an unanimous claim that auditory estimation of azimuth is more accurate that elevation estimation, and that distance estimation is the most difficult task. Accordingly, the cues for azimuth are quite well understood, those for elevation are less well understood, and those for distance are least well understood. Distance perception involves a process of integrating multiple cues, any of which can be rendered ineffective by the summed result of other potential cues.

In the absence of other information, *intensity* is the primary distance cue used by listeners, who learn from experience to correlate the physical displacement of sound sources with corresponding increases or reductions in intensity. Under anechoic conditions, sound intensity reduction with increasing distance is predicted by the inverse square law: an omnidirectional sound source's intensity will fall approximately 6 dB for each distance doubling (see also our discussion of the clarity index parameter in Sec. 4.2.2). However this law is not well motivated perceptually: it expresses the ratio of a sound source's intensity to a reference level, whereas the *perceived* magnitude of intensity is called *loudness*. Thus a mapping where the relative estimation of doubled distance follows "half-loudness" rather than "half-intensity" seems preferable: the two scales are different.[9]

Loudness (or intensity) increments can only operate effectively as distance cues in the absence of other information, in particular reverberation. When reverberation is present the overall loudness at a listener's ear does not change much for very close and very distant sources: the distance-dependent scaling applies only to the direct sound whereas the *reflected energy* remains approximately constant. The change in the proportion of reflected to direct energy, the so-called *R/D ratio*, seems to function as a stronger cue for distance than intensity scaling. In particular a sensation of changing distance can occur if the overall loudness remains constant but the R/D ratio is altered. Note however that in some cases the possible R/D ratio variation can be limited by the size of the particular environmental context, causing the cue to be less robust (e.g. in a small, acoustically treated room, the ratio would vary between smaller limits than in a large room like a gymnasium).

Estimation of distance with anechoic stimuli is usually worse than in experiments with "optimal" reverberation conditions. Many experimental results show an overall underestimation of the apparent distance of a sound source in an anechoic environment, which may be explained by the absence of reverberation. It can be said that reverberation provides the "spatiality" that allows listeners to move from the domain of loudness inferences to the domain of distance inferences, i.e. from an analytic listening attitude to an *everyday listening* attitude.

Distance perception is also affected by expectation or *familiarity* with the sound source. If the sound is completely synthetic (e.g., pulsed white noise), then a listener will typically focus on parametric

[9] We will return on the concept of loudness in Chapter *Auditory based processing*.

changes in loudness and R/D ratio (in this case loudness probably plays a more important role than reverberation effects). On the other hand, if the sound source is cognitively associated with a typical distance range, that range will be more easily perceived than unexpected or unfamiliar distances. This is especially true for speech: as an example, it is easier to simulate a whispering voice 20 cm away from your ear than it is to simulate an unnaturally loud whisper 10 m away.

Spectral effects can also affect distance perception, although to a lesser extent than the cues discussed above. Atmospheric conditions and air absorption play a role: with increasing distance, higher frequencies of a complex sound are increasingly attenuated by air humidity and temperature. There is little experimental evidence this cue is actually used by listeners in forming the distance of an auditory event, although some experimental results suggest that, in the absence of other cues, a low-frequency emphasis applied to a stimulus would be interpreted as "more distant" compared to an untreated stimulus. A second spectral effect is produced in the so-called *near field*, i.e. for distances less than approximately 1 m. Within this range it is not possible to assume the sound wavefronts to be planar, and the effect of their curvature must be taken into account. As the source approaches, emphasis is added to lower frequencies. This phenomenon corresponds to the "darkening" of tone color that occurs as a sound source is moved very close to one's ear.

Note that all the cues discussed above are essentially monoural cues. An open question is whether binaural listening improves the perception of distance. This could indeed be the case again in the near-field limit. The spherical head model shows that in this limit both the ILD and the ITD at low frequencies are emphasized, especially for very lateralized sound sources ($\theta \sim \pm\pi/2$). This effect is sometimes termed *auditory parallax*, and has been interpreted by some to mean that the accuracy of estimation of a sound from the side should be improved when compared to distance perception on the median plane. There are numerous discrepancies in the literature, however, and the question of the influence binaural cues to distance perception is still unresolved.

4.5.2.5 Dynamic cues

So far we have examined sound source perception in the implicit assumption of static conditions, i.e. with both listener and source not moving. However in everyday perception we use also *dynamic* cues in addition to static ones to reinforce localization. These arise from active, sometimes unconscious, motions of listeners, who change their position relative to the source. When we hear a sound that we want to localize, we move e.g. in order to minimize the interaural differences, using our head as a sort of "pointer". Animals use movable pinnae for the same purpose (think of a cat).

When moving their head, listeners apparently integrate some combination of the changes in ITD, ILD, and movement of spectral notches and peaks that occur with head movement over time, and subsequently use this information to improve localization ability. Perhaps the most clear example is represented by front/back confusions, which are common in static listening tests (see our discussion about cones of confusion), and instead disappear when listeners are allowed to turn their heads during a localization task: a listener who is trying to localize a source at, e.g. $\theta = 30^\circ$, $\phi = 0^\circ$ will probably attempt to center the auditory image by moving his head to the right. If the sound becomes increasingly centered –i.e. interaural differences are minimized– consequently to head motion, then it must be in the front. If instead it becomes increasingly lateralized –i.e., the sound becomes louder and arrives sooner at the right ear relative to the left– then it must be to the rear.

Dynamic cues are important also for externalization. IHL, which can be experienced with headphone reproduction as discussed previously, is less likely to occur when head movement is allowed, probably for the same reason that front/back confusion is avoided: dynamic cues arising from head motion are used to disambiguate locations, while static conditions can potentially lead to judgments at a "default" position inside or at the edge of the head. A very undesirable situation is when the sound scene is presented through headphones without traking of head/body motion, *and* the listener can move: in this case dynamic cues are absent and the scene rotates together with the user, creating discomfort and preventing externalization. When visual cues are supplied however, e.g. one can move in a fully immersive virtual environment and can see the virtual sound source, it is quite likely that the combination of vestibular and visual cues will enable externalization. In fact externalization can occur even when listening to a television with a single earpiece: this is because vision is more reliable than audition in spatial location, and therefore our brain "trusts" visual rather than auditory feedback (the general mechanism underlying this phenomenon is known as "visual capture").

Finally, active listener motion provides cues for distance perception. One is the motion-induced rate of change in intensity the so-called *acoustic* τ,[10] by which a listener who moves e.g. towards the sound can infer distance information. A one is the so-called *motion parallax*, which indicates the rate of change in angular direction resulting from listener translation: for a very close source, a small shift of the head causes a large change in angular direction, while for a very distant source the change is almost null irrespective of the amount of shift. The rate of change of ITD, ILD, and spectral notches/peaks will therefore be affected by the distance. This dynamic cue is in many respects similar to its visual counterpart (a large, distant sphere and a small, near sphere look the same, but if we move the different changes in perspective reveal the different distances).

4.6 Algorithms for 3-D sound rendering

Before examining processing algorithms for 3-D sound rendering we have to understand that the techniques to be developed depend on the type of system that is going to be used: the type of the effectors (e.g loudspeakers vs. headphones), as well as their number and geometric arrangement (e.g. stereo systems vs. 5.1 surround systems, etc.).

Stereo is the simplest system involving "spatial" sound. In order to place a sound to the left or to the right, its signal is sent to the corresponding loudspeaker. If the same signal if sent to both speakers, the speakers are wired "in phase", and the listener is approximately equidistant from the speakers, then the listener will perceive a "phantom source" located midway between the two loudspeakers. By crossfading the signal from one speaker to the other, one can create the impression of the source moving continuously between the two louspeaker positions. With this technique however the perceived source will never move outside the line segment between the two speakers.

Multichannel systems are the next step in complexity. The idea is to have a separate channel for every desired direction, possibly including above and below. Commercial home-theater systems are based on this idea. In typically reverberant environments, one can exploit the limitations of our perception (see in Sec. 4.5.2 our discussion about azimuth perception in reverberant environments) and use small loudspeakers everywhere, except for one large speaker (the "subwoofer") that provides the nondirectional, low-frequency content.

Headphone-based systems have some disadvantages compared to loudspeakers: headphones are invasive and can be uncomfortable to wear for long periods of time; they have non-flat frequency responses that can compromise spatialization effects; they tend to provide the impression of too close sources, and do not compensate for listener motion unless a tracking system is used. On the other hand they have two main advantages: first, they eliminate reverberation of the listening space; second, and more important, they allow to deliver distinct signals to each ear, which greatly simplifies the design of 3-D sound rendering techniques. On the contrary loudspeaker based systems suffer from "cross-talk", i.e. the sound emitted by one loudspeaker will be always heard by both ears. If one ignores the effects of the listening

[10] This name comes from studies in visual perception, where the *optical* τ specifies the time-to-contact estimated by a subject in relative motion with respect to a target.

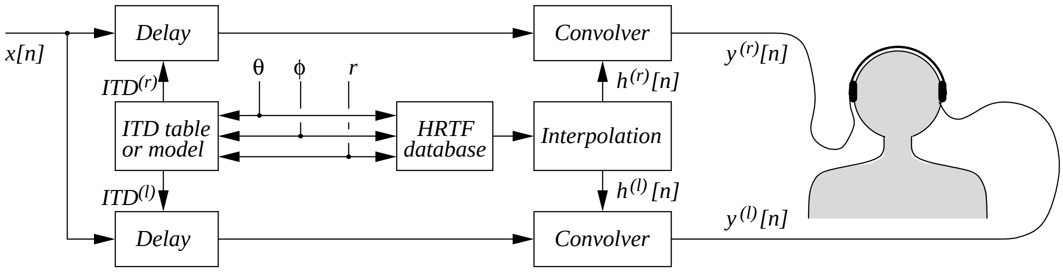


Figure 4.28: *Block scheme of a headphone 3-D audio rendering system based on HRTFs.*

environment, headphone listening conditions can be roughly approximated from stereo loudspeakers using *cross-talk cancellation* techniques, which try to pre-process the stereo signals in such a way that the sound emitted from one loudspeaker is cancelled at the opposite ear. Using these techniques the phantom source can be placed significantly outside of the line segment between the two loudspeakers and in particular elevation effects can be produced. The main problem is that the result will depend on where the listener is relative to the speakers: cross-talk cancellation is obtained only near the so-called "sweet spot", a specific listener location assumed by the system.

In this section we will focus on techniques for headphone-based systems. We will implicitly assume that a single (point) sound source is rendered in space: if multiple sound sources have to be rendered, then each one has to be processed with a different replica of the rendering scheme, with consequent increases in the computational costs.

4.6.1 HRTF-based rendering

The general idea in HRTF-based 3-D audio systems is to use measured HRIRs and HRTFs. Given an anechoic signal and a desired virtual sound source position (θ, ϕ), a left and right signals are synthesized by *(1)* delaying the anechoic signal by appropriate amount, in order to introduce the desired ITD, and *(2)* convolving the anechoic signal with the corresponding left and right head-related impulse responses. A synthetic block scheme is given in Fig. 4.28. In the remainder of this section we summarize the main steps involved in the development of a HRTF-based 3-D audio system, including HRTF measurement and processing, approximation through synthetic HRTFs, and interpolation.

4.6.1.1 Measuring HRTFs and ITDs

The typical setting for HRTF measurement is the following: an anechoic chamber, a set of speakers mounted on a geodesic sphere (with a radius of at least one meter in order to avoid near-field effects), at fixed intervals in azimuth and elevation. The listener is at the center of the sphere, with microphones placed in each ear. HRIRs are measured by playing an analytic signal and recording the responses produced at the ears, for each desired virtual position.[11] Listener and speakers do not need to be moved, facilitating the collection of measurements. Microphone placing is an issue: it can be placed at the entrance of a plugged ear canal, or near the eardrum to account for the response of the ear canal.

Measured HRTFs can be analyzed in order to estimate ITD values and derive a table to be subsequently used in the rendering stage (see the first processing block in Fig. 4.28). ITD estimation can be performed through various methods, including cross-correlation methods (where ITD is computed as the offset in cross-correlation maxima of $h^{(l)}$ and $h^{(r)}$), leading-edge methods (where the time difference

[11] There is a plethora of sophisticated techniques for Impulse Response estimation, which we do not discuss here.

of the start of the impulses is estimated), and so on. Some approaches allow to derive a frequency-dependent ITD, while in other cases frequency-independent estimates are derived. Alternatively, theoretical ITD models can also be used instead of empirically estimated values. We have already examined a frequency-independent ITD model, given in Eq. (4.56). Other models exist, that introduce frequency dependence or even elevation dependence of ITD.

In most 3-D sound applications one typically wants to use a single set of HRTFs for every user. One approach might be to use the features of a person who has "desirable" HRTFs, based on some criteria. A set of HRTFs from a good localizer could be used if the criterion were localization performance. An alternative approach is to construct *generalized HRTFs*, that represent the common features of a number of individuals. Binaural impulse responses from many individuals can be "spectrally averaged" in the Fourier domain. However this can cause the resultant HRTF to have diminished spectral features with respect to individual ones. In the extreme case, one person has a 20 dB notch at 8 kHz, and another has a 20 dB peak – the average is no spectral feature at all.

Generalized HRTFs can also be obtained through the use of so-called "dummy heads", which are mannequins constructed from averaged anthropometric measures and represent standardized heads with average pinnae and torso. The most widely used one is probably the *KEMAR* (Knowles Electronics Manikin for Auditory Research), although many others are commercially available. Measurements with dummy heads are easier, since they are often part of integrated measurement and analysis systems. The low frequency response of the microphones built into the head will be better than that of probe mics, and the results will be more replicable. Moreover, 3-D sound systems based on dummy head HRTFs will be closely matched to recordings made by the same binaural head, allowing compatibility between the two different types of processing. One dummy head might sound more natural to a particular set of users than another, depending on the microphones, the technique used for simulating the ear canal, the head's dimensions, and so on. The head size (and correspondingly, its diffraction effects and overall ITD) is a major component in the suitability of one dummy head versus another.

4.6.1.2 Post-processing of measured HRTFs

Measured HRTFs undergo a series of processing steps. A typical procedure is post-equalization of HRTFs to eliminate potential spectral nonlinearities originated from the loudspeaker, the measuring microphone, and the headphones used for playback. As an example, probe microphones are usually small and are especially inefficient at low (< 400 Hz) frequencies, making high-pass filtering or "bass boosting" a fairly common HRTF post-equalization procedure. A frequency curve approximating the ear canal filter, usually derived from some standard equalization, can be applied if it was not part of the impulse response measurement. Since this filter is independent on the angle of incidence, it needs to be compensated only once. For most applications, the listener's own ear canal resonance will be present during headphone listening; this requires removal of the ear canal resonance that may have been present in the original measurement, to avoid a "double resonance".

One more post-processing procedure is often applied to reduce redundancy in HRTF data. Spectral features that are common to raw HRTFs at all locations do not contain important directional cues, and do not need to be encoded in each single HRTF. Therefore a so-called *Common Transfer Function (CTF)* is often estimated, by computing the mean log-magnitude of the HRTFs measured at several spatial locations. The CTF will include the direction-independent spectral features shared by all HRTFs (e.g., the ear canal filter). It will also include systematic measurement artifacts, if any. During postprocessing, the CTF can be removed from the raw HRTFs to yield the *Directional Transfer Function (DTF)*. The DTF is a function of $\theta,\ \phi$, and is the quantity that contains spectral cues responsible for spatial hearing. Let $C(\omega)$ be the known CTF and $D^{(l),(r)}(\theta,\phi,\omega)$ be the unknown left and right ear DTFs respectively.

Then $D^{(l),(r)}$ are estimated from $H^{(l),(r)}$ and C with the equality

$$H^{(l),(r)}(\theta, \phi, \omega) = C(\omega) D^{(l),(r)}(\theta, \phi, \omega). \tag{4.59}$$

The CTF captures the overall structure and dynamic range of the HRTFs, allowing each DTF to operate over a smaller dynamic range. This division allows us to vary a smaller parameter set (corresponding to only the DTF) to achieve space-varying HRTF approximations. Many of the algorithms described in the next sections can be applied either to the "raw" HRTFs or to the DTFs.

M-4.12

Write a script that computes the Common Transfer Function $C(\omega)$ and the Directional Transfer Functions $D(\theta_k, \phi_k, \omega)$ given a set of HRTFs $H(\theta_k, \phi_k, \omega)$ measured on M directions θ_k, ϕ_k ($k = 1 \ldots M$).

A third post-processing procedure is minimum-phase reconstruction of the HRTF filters. Recall that a minimum-phase reconstruction of a filter is a filter with the same magnitude response of the original one, in which all zeros and poles are inside the unit circle. Minimum-phase filters have many benefits in terms of realization, coefficient interpolation, and so on. Various studies show that this processing step does not have any perceptual consequences, provided that ITD is introduced before convolution with the minimum-phased reconstructed HRTF (as in Fig. 4.28), as detailed phase information is not perceptually relevant.

Having acquired HRTF magnitude responses, one can design *synthetic HRTFs*, low-order filters that approximate the original HRTFs in a perceptually motivated way while providing significant computational advantages. In fact direct use of measured HRTFs requires a convolution with long FIR filters: assuming a duration of $\sim$ 10 ms for a measured HRIR (reported durations vary across studies), the corresponding HRIR filter length is ~ 440 samples for $F_s = 44.1$ kHz. Despite the ever increasing computational power at our disposal, such filter sizes can make it difficult to synthesize complex acoustic environments in real time, particularly when multiple sound sources and reverberant environments have to be rendered.

Developing perceptually appropriate low-order representations of the HRTFs may also provide insight into sound localization mechanisms and into the usefulness of various cues embodied in the HRTF, which is incompletely understood. Moreover, such representations can be used to improve our understanding of the physical mechanisms that produce certain features in the HRTF.

We can schematically synthetic HRTF design techniques into two families. In *pole-zero models* the problem is viewed as one of filter design, which has several classical solutions. One drawback is that the coefficients are usually complicated functions of azimuth and elevation, and have to be tabulated, which hinders the usefulness of the model. *Series expansions* let one represent the HRTF as a weighted sum of simpler basis functions. While this is useful for inspecting the data, the run-time complexity of such models can limit their usefulness. In both cases, the original HRTFs can be further processed prior to the design of the synthetic HRTFs. Usually some form of *auditory smoothing* is used, that performs a non-uniform frequency-dependent smoothing of the responses based on psychoacoustic models. This produces more regular magnitude responses without any relevant perceptual consequences, and filter approximation is easier. We pospone the description of auditory smoothing techniques to Chapter *Auditory based processing*. In the next sections we discuss both pole-zero and series expansion approaches to synthetic HRTFs.

4.6.1.3 Synthetic HRTFs: pole-zero models

Given a direction (θ, ϕ), a *pole-zero model* (or an *ARMA* model)[12] approximates the corresponding HRTF, $H(z)$, with the rational transfer function

$$\tilde{H}(z) = \frac{b_0 + \sum_{k=1}^{q} b_k z^{-k}}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{B(z)}{A(z)}. \tag{4.60}$$

For brevity, here and in the following we omit in the notation any dependence on (θ, ϕ): in particular the coefficient vectors $\boldsymbol{b} = \{b_k\}, \boldsymbol{a} = \{a_k\}$ will depend on θ, ϕ.

In the particular case $p = 0$, Eq. (4.60) is an *all-zero* (FIR) model: in this case the most straightforward approximation consists in windowing the impulse response h to a shorter length. This approach can be since further refined to account for frequency-domain weighting that models the non-uniform frequency resolution of the ear.[13] Various studies report of synthetic all-zero HRTF models obtained with this approach, with filter orders between 20 and 64.

In the particular case $q = 0$, Eq. (4.60) is an *all-pole* model: we have already seen in Chapter *Sound modeling: signal based approaches* that linear prediction can be used in this case to estimate the coefficients $\{a_k\}$ that allow $\tilde{H}$ to best approximate H.

In the general case $q, p \geq 1$, traditional digital filter design techniques still state the problem as one of minimizing the difference between $\tilde{H}$ and the target response H, which is typically known on a set of L "design frequencies" $\{\omega_k\}_{k=1}^{L}$ (e.g. $\omega_k = 2k\pi/LF_s$ if the ω_k are evenly distributed along the frequency axis). Usually this difference is expressed as a weighted error function $\mathcal{E}$ given by

$$\mathcal{E}(\omega_d) = W\left(e^{j\omega_d}\right)\left[H\left(e^{j\omega_d}\right) - \tilde{H}\left(e^{j\omega_d}\right)\right], \tag{4.61}$$

where W is some positive weighing function specified in the design. Moreover the error is usually estimated with regard to the magnitude response while the phase response is disregarded since, as already mentioned, the effect of the ITD is rendered separately and minimum phase transfer functions are used (see Fig. 4.28). Commonly used error functions are based on the L^{p} norm of the function $\mathcal{E}$. The most straightforward choice is the L^2 norm, which is generally known as the *Least-Squares Error* and corresponds to the energy of the difference signal:

$$E_{\text{LS}}\{\mathcal{E}\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\,\mathcal{E}(\omega_d)\,|^2\, d\omega_d \sim \frac{1}{L}\sum_{k=1}^{L}\left[\mathcal{E}(\omega_k)\right]^2. \tag{4.62}$$

Minimizing the error $E_{\text{LS}}\{\mathcal{E}\}$ means finding the coefficient vectors $\boldsymbol{b}, \boldsymbol{a}$ for which the gradient of $E_{\text{LS}}\{\mathcal{E}\}$ is null, that is solving the set of equations

$$\nabla_a E_{\text{LS}}\{\mathcal{E}\} = \nabla_b E_{\text{LS}}\{\mathcal{E}\} = \boldsymbol{0}, \tag{4.63}$$

where the notation $\nabla_x E_{\text{LS}}$ stands for the gradient of E_{LS} with respect to the vector $\boldsymbol{x}$. One of the main advantages of the least squares formulation is that the error function has a global minimum, since it is quadratic. We do not enter into the mathematics involved in writing and solving these equations and refer the reader to the literature on linear Least-Squares Error estimation.

M-4.13

Write a function that computes a LS pole-zero approximation of a target impulse response (representing e.g. a HRIR), given the orders p and q.

[12] See linear prediction in Chapter *Sound modeling: signal based approaches*.

[13] See Chapter *Auditory based processing*

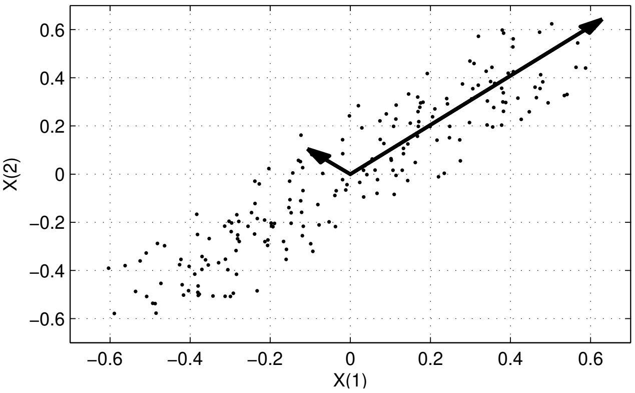


Figure 4.29: *Example of principal component analysis: a two-dimensional data set with 0 mean, and the two basis vectors (principal axes) extracted using PCA.*

Another choice is to minimize the L^{∞} norm of the difference function:

$$E_{\infty}\{\mathcal{E}\} = \max_{-\pi<\omega_d<+\pi} |\,\mathcal{E}(\omega_d)\,|\,. \tag{4.64}$$

This is often referred to as Chebyshev or minimax criterion. Since this norm tries to minimize the maximum error, it should be able to provide good approximation of peaks and valleys of the HRTFs, which are relevant for localization as we know. On the other hand one drawback is that this error surface may not always be convex and thus may lead to unstable or locally optimal results.

As already mentioned for the all-zero case, for our particular filter design problem it is desirable to account for frequency-domain weighing that models the non-uniform frequency resolution of the ear. In this sense, error metrics that utilize absolute LS error on a linear scale are not the best choice, whereas an error criteria based on the difference in log magnitude might be perceptually more appropriate. Since both spectral peaks and spectral notches provide relevant information about the sound source location, minimizing the error on a log scale ensures that the solution is not biased toward peaks relative to notches. An example of such a perceptually motivated error criterion is

$$E_{\log}\{H,\tilde{H}\} = \frac{1}{L}\sum_{k=1}^{L}\left(\ln|\,H(\omega_k)\,| - \ln\left|\,\tilde{H}(\omega_k)\,\right|\right)^2, \tag{4.65}$$

A drawback of this kind of error functions is their minimization is a nonlinear problem, whose solution can be found only with iterative numerical solvers. Another way to construct a perceptually motivated error criterion is to choose the weighing function W in order to model auditory frequency resolution.

4.6.1.4 Synthetic HRTFs: series expansions

Based on the notions given in Sec. 4.5.1, one can argue on a physical basis that HRTFs should be completely determined by a relatively small number of physical parameters: average head radius, maximum pinna diameter, etc. This suggests that the intrinsic dimensionality of the HRTFs might be small, and that their complexity primarily reflects the fact that we are not viewing them correctly.

Among the statistical procedures used to provide a “simpler” representation of a set of correlated measures, a powerful and popular one is *principal component analysis (PCA)*, also known as Karhunen-Loève transformation. The central idea of PCA is to reduce the dimensionality of a large dataset while retaining as much as possible of the variation present in the data. A small set of *basis vectors* is derived, and these are used to compute the *principal components*, i.e. the sets of weights that reflect the relative contributions of each basis vector to the original data.

To start, assume we wish to represent M N-dimensional column vectors $\boldsymbol{x}_1 \dots \boldsymbol{x}_M$ with a 1-dimensional projection (a line) through their mean. The vector will then be represented as

$$\boldsymbol{x}_k \sim \boldsymbol{m} + a_k \boldsymbol{e} \quad k = 1, \dots M, \tag{4.66}$$

where $\boldsymbol{e}$ is a unit vector in the direction of the line, and a_k is a constant coefficient that estimates the distance of $\boldsymbol{x}_k$ from the sample mean $\boldsymbol{m} = 1/M \sum_{k=1}^{M} \boldsymbol{x}_k$. The optimal coefficient a_k can be obtained by minimizing the "squared error criterion function"

$$E(a_1 \dots, a_k, \boldsymbol{e}) = \sum_{k=1}^{M} \| (\boldsymbol{m} - a_k \boldsymbol{e}) - \boldsymbol{x}_k \|^2 . \tag{4.67}$$

For a given direction $\boldsymbol{e}$, the optimal coefficients are clearly $a_k = \boldsymbol{e}^T(\boldsymbol{x}_k - \boldsymbol{m})$, i.e. they are obtained by projecting the data vectors onto the line $\boldsymbol{e}$ that passes through the sample mean. The question is now: what is the optimal direction $\boldsymbol{e}$? By exploiting the expression written above for the optimal a_k's, the error E can be rewritten after some straightforward algebra as

$$E(a_1 \dots, a_k, \boldsymbol{e}) = \sum_{k=1}^{M} \| (\boldsymbol{m} - a_k \boldsymbol{e}) - \boldsymbol{x}_k \|^2 = \dots = -\boldsymbol{e}^T \boldsymbol{S} \boldsymbol{e} + \sum_{k=1}^{M} \| \boldsymbol{x}_k - \boldsymbol{m} \|^2 , \tag{4.68}$$

where $\boldsymbol{S} = \sum_{k=1}^{M} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^T$ is the $N \times N$ *scattering matrix* of the data (which coincides with the covariance matrix except for a multiplying factor $1/(N-1)$). Therefore minimizing E means maximizing the function $f(\boldsymbol{e}) = \boldsymbol{e}^T \boldsymbol{S} \boldsymbol{e}$, with the constraint $\| \boldsymbol{e} \| = 1$. This can be done using Lagrange multipliers.[14] For our PCA problem we have $L(\boldsymbol{e}, \lambda) = \boldsymbol{e}^T \boldsymbol{S} \boldsymbol{e} - \lambda(1 - \boldsymbol{e}^T \boldsymbol{e})$, and $\nabla_e L(\boldsymbol{e}, \lambda) = 2\boldsymbol{S}\boldsymbol{e} - 2\lambda \boldsymbol{e}$. In conclusion the points $\boldsymbol{e}$ that maximize $f(\boldsymbol{e})$ are those for which

$$\boldsymbol{S}\boldsymbol{e} = \lambda \boldsymbol{e}, \tag{4.69}$$

i.e. are the eigenvectors of $\boldsymbol{S}$ for the eigenvalue λ. The single "best" line that represents the data is found by picking the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{S}$ so to ensure that $\boldsymbol{e}^T \boldsymbol{S} \boldsymbol{e} = \lambda$ is maximized. This can be readily extended to larger dimensions. If we wish to represent the $\boldsymbol{x}_k$'s on a q-dimensional hyperplane through the sample mean, written as

$$\boldsymbol{x}_k \sim \boldsymbol{m} + \sum_{i=1}^{q} a_{k,i} \boldsymbol{e}_i, \tag{4.70}$$

then we project the data onto the q eigenvectors of $\boldsymbol{S}$ corresponding to the q largest eigenvalues. If we choose to use all eigenvectors, that is $q = M$ in Eq. (4.70), we will get the original data back (with no dimensionality reduction). From a geometrical standpoint, eigenvectors of $\boldsymbol{S}$ represent the *principal axes* along which the data exhibit largest variance. The weight coefficients $a_{k,i}$ are called the *principal components*. Moreover the basis vectors are derived in such a way that the first one captures the majority of common variation present in the data and that the remaining vectors reflect decreasing common variation and increasing unique variation. The number q of principal axes required to provide an adequate representation of the data is largely a function of the amount of redundancy or correlation present in the data. The greater the redundancy, the smaller the number q needed.

[14] Recall that in order to find the extremum of a function $f(\boldsymbol{x})$ subject to a constraint $g(\boldsymbol{x}) = 0$, one can construct the Lagrange function $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$ and look for a zero of the gradient $\nabla_x L(\boldsymbol{x}, \lambda)$.

Now suppose we have measured directional transfer functions $D(\theta_k, \phi_k, \omega_j)$, on M directions θ_k, ϕ_k ($k = 1 \dots M$) and on N frequency points ω_j ($j = 1 \dots N$). We can apply PCA to the particular set of M N-dimensional vectors $\boldsymbol{x}_k$ constructed as $x_{k,j} = \log |\, D(\theta_k, \phi_k, \omega_j)\, |$, i.e. we work on the log magnitudes of the DTFs (as already remarked, approximation of log-magnitudes is perceptually more appropriate than approximation of linear magnitudes). The result is a set of q basis vectors $\boldsymbol{e}_i$ (where $e_{i,j} = e_i(\omega_j)$), such that for the kth direction (θ_k, ϕ_k) the DTF can be approximated as

$$\log |\, D(\theta_k, \phi_k, \omega_j)\, | \sim \sum_{i=1}^{q} a_i(\theta_k, \phi_k) e_i(\omega_j). \tag{4.71}$$

Studies on the evaluation of this procedure have shown that the first five basis functions ($q = 5$) can accurately represent the magnitudes of the DTF set, and listening tests have shown a high correlation between responses to the synthesized and measured conditions. Moreover the dependence on space and frequency have been decoupled in Eq. (4.71), with consequent computational advantages.

M-4.14

Write a function that computes the first q principal axes $\boldsymbol{e}_i$ ($i = 1 \dots q$) and components $a_{i,k}$ for a set of DTFs $D(\theta_k, \phi_k, \omega_j)$ measured on M directions θ_k, ϕ_k ($k = 1 \dots M$) and on N frequency points ω_j ($j = 1 \dots N$).

4.6.1.5 Interpolation

HRTF measurements can only be made a finite set of locations, and when a sound source at an intermediate location must be rendered, the HRTF must be *interpolated.* If interpolation is not applied (e.g.. if a nearest neighbor approach is used) audible artifacts like clicks and noise are generated in the sound spectrum when the source position changes.

A straightforward way to perform interpolation directly on the HRIR samples is the bilinear method, which simply consists of computing the response at a given point (θ, ϕ) as a weighted mean of the measured responses associated with the four nearest points. More precisely, if the corresponding set of HRIRs has been measured over a spherical grid with steps θ_{grid} and ϕ_{grid}, the estimate $\hat{h}$ of the HRIR at an arbitrary point (θ, ϕ) can be obtained as (see Fig. 4.30(a))

$$\hat{h}[n] = (1 - c_\theta)(1 - c_\phi) h_1[n] + c_\theta (1 - c_\phi) h_2[n] + c_\theta c_\phi h_3[n] + (1 - c_\theta) c_\phi h_4[n], \tag{4.72}$$

where $h_k[n]$ ($k = 1, \dots, 4$) are the HRIRs associated with the four nearest points to the desired position. The parameters c_θ and c_ϕ are computed as

$$c_\theta = \frac{\theta \mod \theta_{\text{grid}}}{\theta_{\text{grid}}}, \qquad c_\phi = \frac{\phi \mod \phi_{\text{grid}}}{\phi_{\text{grid}}}. \tag{4.73}$$

Several refinements can be applied to this simple technique, in order to improve efficiency. In particular, reduced-order HRIR such as those described earlier in this section can be used. Also, interpolation can be performed using only three grid points (those which form a triangle around the desired position). However, since some HRTF features arise due to coherent addition or cancelation of reflected and diffracted waves, interpolation may not preserve these features and produce perceptually poor results. Moreover, the interpolating filters are required to be minimum-phase: if this requirement is not satisfied, severe comb-filtering effects in the frequency domain can be produced when the phase delays of the interpolating filters vary considerably. Also, to capture fine details of the HRTF the sampling must be fine enough, i.e. satisfy a spatial Nyquist criterion. Interpolation can be performed in the frequency domain as well (i.e. estimate the DFT of $\hat{h}$ by interpolating the DFTs of the h_k's). Besides linear approaches, geometric and spline interpolation can be used as well.

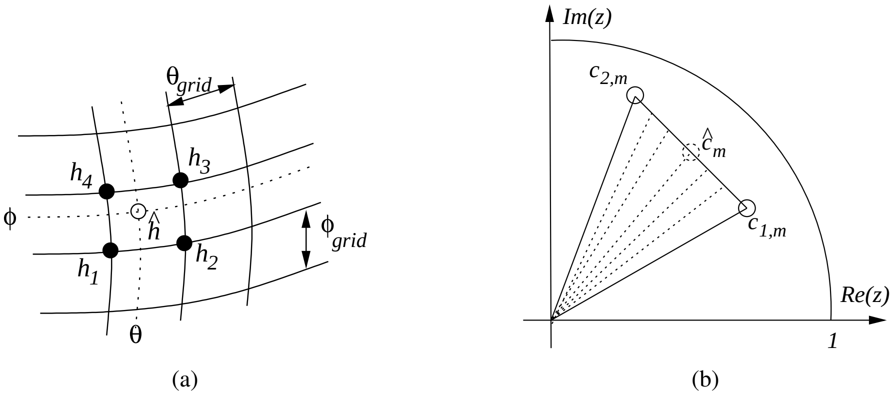


Figure 4.30: *Approaches to HRTF interpolation; (a) bilinear interpolation of the HRIRs, and (b) interpolation of zeros for pole-zero synthetic HRTFs*

M-4.15

Realize bilinear interpolation.

If synthetic HRTFs in the form of pole-zero filters are being used, interpolation can be performed on the poles and the zeros themselves. The case of an all-zero filter is relatively straightforward. Suppose that we want to interpolate between two transfer functions $H_k(z)$ ($k = 1, 2$) of the form

$$H_k(z) = 1 + \sum_{m=1}^{q} b_{k,m} z^{-m} = \prod_{k=0}^{q} (1 - c_{k,m} z^{-1}), \qquad k = 1, 2, \tag{4.74}$$

where $b_{k,0} = 1$ without loss of generality, and where we are assuming that the zeros of both filters are sorted according to their phases. Then an interpolated filter $\hat{H}(z) = \prod_{m=0}^{q}(1 - \hat{c}_m z^{-1})$ can be obtained by *(1)* pairing the zeros according to angular proximity, and *(2)* computing the interpolated zeros $\hat{c}_m = (1-\rho)c_{1,m} + \rho c_{2,m}$ ($m = 1, \ldots, q$). Note that if the H_k are minimum-phase the interpolated filter is also minimum-phase (see also Fig. 4.30(b))

If we use pole-zero synthetic HRTFs of the form (4.60) with $p > 0$, then interpolation becomes more complicated. One can still use convex combinations of pole and zero values from neighbouring DTF approximations (note in particular that linear combination of stable poles is guaranteed to be stable). However a naive realization of this approach can result in erratic and occasionally large errors of the interpolated filters. In order to achieve regularity in the interpolation, more refined algorithms are needed that provide pairing and ordering on the entire HRTF database.

Synthetic HRTFs based on PCA expansions, in the form (4.71) are well suited for interpolation, since the dependence on frequency is decoupled from the dependence on spatial variables: therefore only the spatially-dependent coefficients $a_i(\theta, \phi)$ need to be interpolated while the frequency-dependent basis-vectors are not involved in the interpolation process. Functional representations of the a_i's can be obtained through standard techniques such as spline interpolation.

In general, reconstruction of the underlying continuous coefficient functions from the samples obtained is an inherently ill-posed problem, because the samples do not uniquely define the functions in the absence of additional assumptions, and because the samples can be corrupted by the presence of noise. Some form of *smoothness constraints* must be used, so that a small change in θ, ϕ induces a small change in the coefficients.

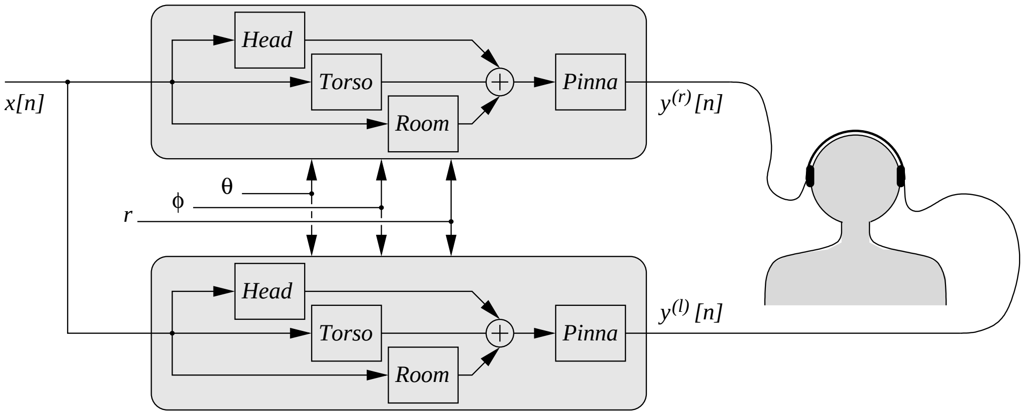


Figure 4.31: *Block scheme of a headphone 3-D audio rendering system based on a structural model.*

4.6.2 Structural models

As opposed to the HRTF-based rendering approach discussed above, the structural approach presented in this section is based on the modeling of the separate effects of the torso, head, and pinna, which combine to form the head related transfer function.

The HRTF is then modeled as a combination of filter blocks, each accounting for the contribution of one anatomical structure. The parameters of each block can in principle be related to anthropometric measures (e.g. the interaural distance, or the diameter of the cavum conchae), with the advantage that a generic structural HRTF model can be adapted to a specific listener and can account for posture-related effects. Another advantage is that room effects can be incorporated into the rendering scheme, specifically early reflections can be processed through the pinna model.

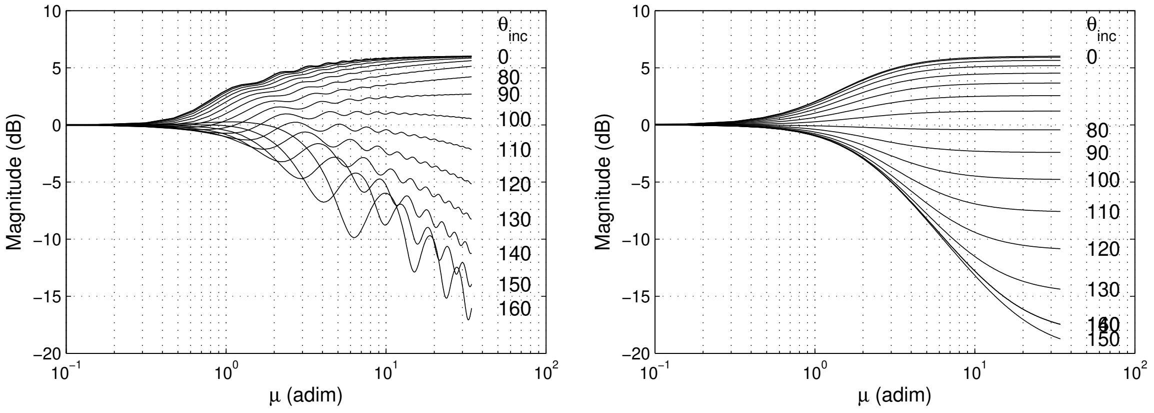
It is clear that separating the effects of various anatomical structures into perfectly independent filter structures is a heuristic approximation that disregards interactions due to waves scattered from one structure to another. However research in this direction has shown that structural models are able to provide good approximations of real HRTFs.

A synthetic block scheme of a generic structural model is given in Fig. 4.31. In the remainder of this section we describe modeling approaches for each of the three main components depicted in the figure, namely head, torso, and pinna. Room effects can be also accounted for in this structure: in particular early reflections can be convolved with a pinna model, depending on their incoming direction (see also our discussion on directional effects with early reflection modeling in Sec. 4.3.2).

4.6.2.1 Head models

In Sec. 4.5.1 we have analyzed the effects of the head on the sound field at the eardrum by approximating the head with a sphere. We have seen that given a sphere of radius a, a point sound source at a distance $r > a$ from the center of the sphere, and a point on the sphere, then the diffraction of an acoustic wave by the sphere seen on the chosen point can expressed with a transfer function $H_{\text{sphere}}(\rho, \theta_{inc}, \mu)$ (where we are using the normalized frequency $\mu = \omega a/c$ and the normalized distance $\rho = r/a$, and θ_{inc} is the angle of incidence). We have also studied this tranfer function in the limit of $\rho \to +\infty$.

In this limit the response H_{sphere} can be approximated with a parametric filter $\tilde{H}_{\text{sphere}}(\theta_{inc}, \mu)$, whose parameters depend on θ_{inc} only. In fact already a first order filter can provide reasonable results, if



(a) (b)

Figure 4.32: *Spherical head model; (a) ideal response of Eq. (4.57) for $\rho \to +\infty$, (b) approximated response with the first-order filter of Eq. (4.75) with $\alpha_{min} = 0.1$ and $\theta_{min} = 170°$.*

properly parametrized. Some authors have proposed the following form:

$$\tilde{H}_{\text{sphere}}(\theta_{inc}, \mu) = \frac{1 + \frac{j}{2}\mu \cdot \alpha(\theta_{inc})}{1 + \frac{j}{2}\mu}, \qquad 0 \le \alpha(\theta_{inc}) \le 2. \tag{4.75}$$

The idea behind this equation is that the θ_{inc}-dependent parameter α controls the location of the zero in the numerator: for $\alpha = 2$ the filter gives a 6 dB boost at high frequencies (which corresponds to the behavior of H_{sphere} for $\theta_{inc} = 0$), while for $\alpha < 1$ there is a low pass effect. Moreover, in order for $\tilde{H}_{\text{sphere}}$ to match the behavior of H_{sphere} at values $\theta_{inc} \neq 0$, the parameter α must depend in a nonlinear way on θ_{inc}. A possible choice is

$$\alpha(\theta_{inc}) = \left(1 + \frac{\alpha_{min}}{2}\right) + \left(1 - \frac{\alpha_{min}}{2}\right)\cos\left(\frac{\theta_{inc}}{\theta_{min}}\right) \tag{4.76}$$

where values of the auxiliary parameters $\alpha_{min},\ \theta_{min}$ can be chosen in order to tune the dependence of α on θ_{inc}. The result can be seen in Fig. 4.32.

The filter $\tilde{H}_{\text{sphere}}$ can already produce fairly convincing azimuth effects, even though it only matches the gross magnitude characteristics of the spectrum. In order to enhance its effectiveness, an all-pass section has to be cascaded to it, to account for the interaural time difference: we can implement this additional block as a fractional delay filter[15] $F_{ITD}(\theta_{inc}, z)$, so that the complete head models is $\tilde{H}_{\text{sphere}}(\theta_{inc}, z) \cdot F_{ITD}(\theta_{inc}, z)$. The ITD values used to parametrize the filter F_{ITD} can be values derived from measured HRTFs, or values derived from theoretical ITD models. We have already discussed this point at the beginning of Sec. 4.6.1.

M-4.16

Write a function that relizes the first-order filter (4.75).

It is clear that a sphere provides only a first approximation to a human head. Better approximation can be already obtained by introducing two simple refinements. First, one can use a non-spherical shape:

[15] Recall that we have discussed fractional delay filters in Chapter *Sound modeling: source based approaches*.

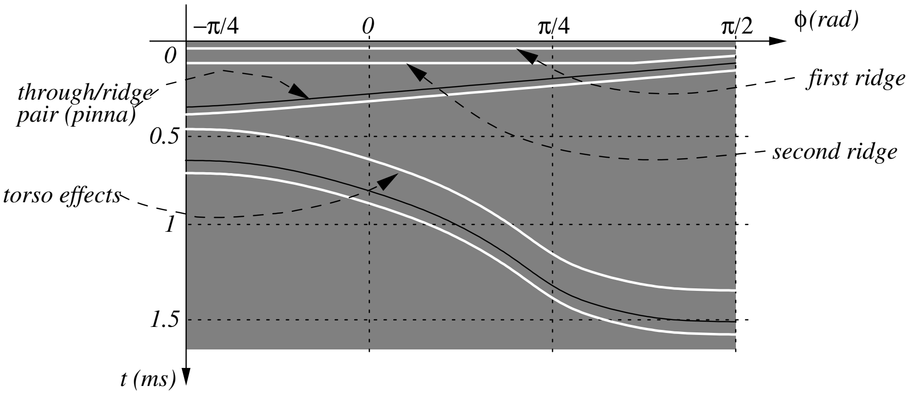


Figure 4.33: *A schematic representation of the major features of the HRIR in the median plane ($\theta = 0$) for a human subject. White and black lines indicate ridges and throughs in the response, respectively.*

an ellipsoid is an obvious choice. Second, one can note that the ears are not positioned across a diameter, but are displaced behind and below the center of the head. As already remarked in Sec. 4.5.2, these two anatomical details have the consequence that the ITD is a function of elevation as well as azimuth. In fact analysis on measured HRTFs shows that for a fixed value of θ and varying values of ϕ,[16] the ITD can vary by almost 20% of its maximum value, with noticeable perceptual effects.

4.6.2.2 Modeling torso and pinna reflections

Based on what we have said in the previous sections, we can assume that the main effects of torso and pinna that need to be accounted for are reflections. This means that both torso and pinna will be modeled as FIR comb filters, in which each reflection determines a comb series in the spectrum. We should be aware however that reflection is a short-wavelength or high-frequency concept, and modeling the effects of torso and pinna by specular reflections is only a first approximation.

In order to realize a model for the torso and the pinna, everything reduces down to estimating reflection delays and their dependence on θ and ϕ, either through analysis of measured HRIRs/HRTFs, or through numerical simulations. As remarked by many authors, a general trend can be observed in measured HRTFs. A schematic representation is given in Fig. 4.33, where only elevations in the range $[-\pi/4, \pi/2]$ have been considered: for values $\phi < -\pi/4$ head shadowing effects start to appear, while HRIR features (end especially pinna related features) are less clear for $\phi > \pi/2$.

The initial ridge due to the direct impulse is followed by a sequence of ridges and troughs. A second ridge occurs roughly 50μs after the initial ridge and varies only slightly with elevation. It is followed by a very prominent trough and ridge pair whose latency varies nearly monotonically with elevation from about 400μs at $\phi = -\pi/4$ to 100μs at $\phi = \pi/2$. The sharply positive sloping diagonal events are due to a torso reflection and its replication by pinna effects. The delay between the direct and the reflected sound from the torso is maximum above the head and decreases with elevation, as one would expect from geometrical considerations.

Note that the scheme depicted in Fig 4.33 corresponds to HRIRs measured in the median plane ($\theta = 0$). Torso echoes vary significantly with azimuth also. On the contrary, pinna events exhibit very limited azimuth dependence.

The main torso reflection is relatively straightforward to estimate, either from measured HRIRs, or from numerical simulations of simplified models where both head and torso are approximated as

[16] We are using the interaural polar coordinate system here.

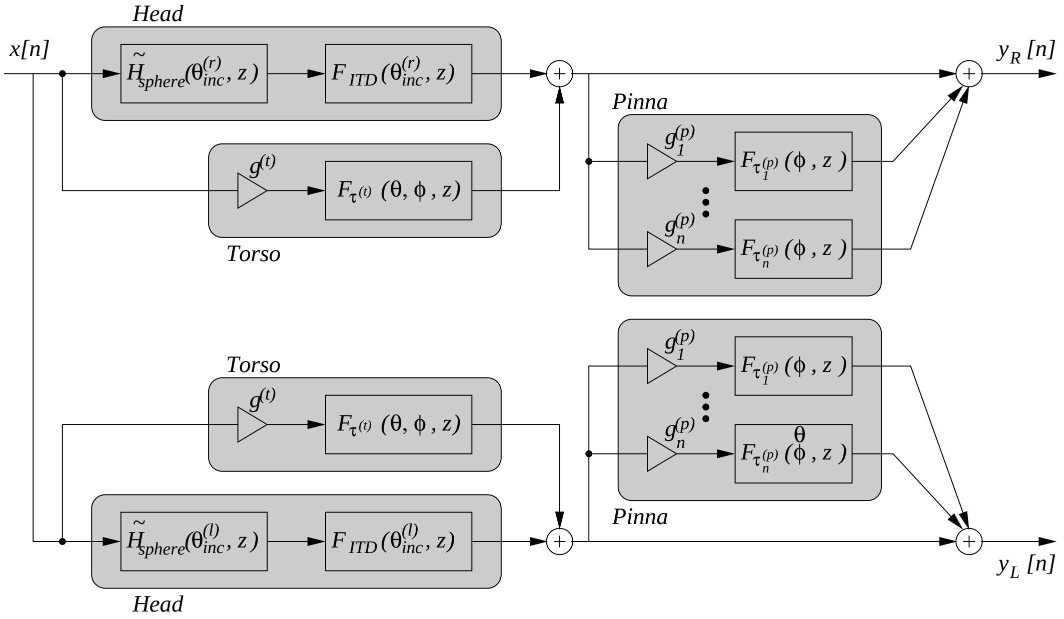

Head
x[n]
$\tilde{H}_{sphere}(\theta^{(r)}_{inc}, z)$
$F_{ITD}(\theta^{(r)}_{inc}, z)$
$y_R\,[n]$
Pinna
$g^{(p)}_1$
$F_{\tau^{(p)}_1}(\phi, z)$
$g^{(t)}$
$F_{\tau^{(t)}}(\theta, \phi, z)$
$g^{(p)}_n$
$F_{\tau^{(p)}_n}(\phi, z)$
Torso
Torso
$g^{(t)}$
$F_{\tau^{(t)}}(\theta, \phi, z)$
$g^{(p)}_1$
$F_{\tau^{(p)}_1}(\phi, z)$
$g^{(p)}_n$
θ
$F_{\tau^{(p)}_n}(\phi, z)$
Pinna
$\tilde{H}_{sphere}(\theta^{(l)}_{inc}, z)$
$F_{ITD}(\theta^{(l)}_{inc}, z)$
$y_L\,[n]$
Head

Figure 4.34: *A simple yet complete structural model.*

ellipsoids (so-called “snowman” models). With these methods one can estimate the θ- and ϕ-dependent delay $\tau^{(t)}(\theta, \phi)$ of the main torso reflection. In conclusion, torso effects can be modeled with a single fractional delay filter $g^{(t)}F_{\tau^{(t)}}(\theta, \phi, z)$, where $g^{(t)}$ is the torso reflection coefficient. Note that this model is only valid for positive ϕ values: as the source descends in elevation, a point of grazing incidence is reached, below which torso reflections disappear and torso shadowing emerges.

Pinna effects are harder to model, since it is more difficult to automatically extract filter parameters from measured data. Time-domain analysis (i.e., identification of reflections in the HRIR) is in this case not reliable. Frequency-domain analysis is preferable, and consists in identifying notch series in the HRTF. If such series can be identified, they can then be related to ear anatomy. More precisely, the delay $\tau_i^{(p)}(\phi)$ of the ith pinna reflection causes periodic notches in the spectrum,[17] with frequencies $\omega_{i,n}^{(p)}(\phi) = 2\pi(2n+1)/\tau_i^{(p)}(\phi)$ (with $n \geq 0$). Moreover, $\tau_i^{(p)}(\phi) = 2d_i^{(p)}(\phi)/c$, where $d(\phi)$ is the distance between the ith reflecting surface of the pinna (e.g. the cavum conchae) and the ear canal. The frequency $\omega_{i,0}^{(p)}(\phi)$ of the first notch and $d_i^{(p)}(\phi)$ are then related through the equation

$$\omega_{i,0}^{(p)}(\phi) = \pi c / 2 d_i^{(p)}(\phi). \tag{4.77}$$

Therefore, given an estimate of the function $\omega_{i,0}^{(p)}(\phi)$ obtained from analysis of HRTFs, the function $d_i^{(p)}(\phi)$ can be estimated through this equation and consequently the measured notches can be be related to anatomical details of the pinna. In conclusion, pinna effects can be modeled with a set of n fractional delay filters $g_i^{(p)}F_{\tau_i^{(p)}}(\phi, z)$.

4.6.2.3 A complete structural model

The components that we have analyzed in the previous sections can be combined to form the simple yet complete structural model depicted in Fig. 4.34, which explodes the general block scheme presented

[17] See comb filters in *Sound modeling: source based approaches*.

in Fig. 4.31. The rationale for this structure is that sound can reach the ear pinna via two major paths: diffraction around the head, and reflection from the torso. In both cases, the sound waves that reach the pinna are altered by pinna reflections before entering the ear canal.

M-4.17

Realize the structural model of Fig. 4.34.

This model can be refined in many respects. The first-order head-shadow filter $\tilde{H}_{\text{sphere}}$ can be replaced by more accurate filters. In particular, $\tilde{H}_{\text{sphere}}$ is derived in the far-field limit: in order to model near-field effects we should substitute it with a filter that approximates Eq. (4.57) directly, and takes into account dependence on range.

Some parameters can be made direction-dependent: in particular, careful examination of torso echo patterns reveals that torso reflection coefficients $g^{(t)}$ vary with elevation. Finally, note that in this model sound diffracted from the head and sound reflected from the torso are processed through the same pinna models: this is not entirely correct since the torso echoes arrive at the ear from a different direction than the direct sound, and therefore they should really pass through a different pinna model. On the other hand the actual perceptual relevance of torso reflections is not clear, as already mentioned in Sec. 4.5.2, therefore this approximate description can be considered to be acceptable.

4.7 Commented bibliography

Wallace C. Sabine has in a way invented the science of concert hall acoustics in the early '900s. For a review of his work and early literature on concert hall acoustics see [Sabine, 1939] (note that the Paul E. Sabine author of this paper is the cousin of Wallace). A very complete discussion of physical aspects of room acoustics is provided by Kuttruff [1991]: Section 4.2.1 is almost entirely based on this book. We have not discussed techniques for impulse response (and particularly RIR) measurement, for a review see e.g. [Stan et al., 2002]. Farina and coworkers have worked extensively on RIR measurements but also on the simulation of the acoustics of closed spaces; the RIR plotted in Fig. 4.3 is one of the publicly available RIRs on the group webpage.[18]

Concerning the research on perceptual attributes of reverberation, the tutorial paper by Beranek [1992] summarizes the main results obtained up to 1992. Research at IRCAM tried to provide a minimal set of independent parameters that give an exhaustive characterization of room acoustic quality [Jot, 1999]. These parameters are divided into three categories, that relate to room perception, source/room interaction, and source perception, respectively.

The first artificial reverberators were electromechanical devices such as *plate reverberators* and *spring reverberators*, in which mechanical elements like plates and springs were fed with a dry sound signal, and an output signal was read at a different point of the element. Despite their limited ability in simulating real environments, plate and spring reverbs have become through the years some of the most sought after effects in digital audio [Bilbao and Parker, 2010].

The first artificial reverberator based on filters was proposed by Manfred Schroeder in the early '60's. The reverberator realized in our example M-4.3 is in fact the Schroeder [1962] reverberator. Schroeder also provided a method for measuring the reverberation time [Schroeder, 1965], which can be used to realize the code in example M-4.1. Moreover, Schroeder [1970] proposed the combination of early reflections and late reverberation depicted in our Fig. 4.12(a).

An extensive experimentation on structures for artificial reverberation was conducted by Andy Moorer in the late '70's. He extended the Schroeder's work in relating some basic computational structures (e.g.,

[18] See http://pcfarina.eng.unipr.it/.

tapped delay lines, comb and allpass filters) with the physical behavior of actual rooms. The reverberator realized in our example M-4.4 is in fact the Moorer [1979] reverberator. He also proposed the combination of early reflections and late reverberation depicted in our Fig. 4.12(b).

Gardner [1998] has explored the use of structures based on all-pass and nested all-pass filters (see in particular Figs. 4.10 and 4.11). This reference, together with [Rocchesso, 2002], also provides an general extensive overview of reverberation algorithms, including binaural reverberation. Research on binaural reverberation techniques includes work by [Begault, 1994, Chapter 4] and by Griesinger [1997]. Our Fig. 4.13(b) is based on this latter reference.

Feedback Delay Networks were first suggested for artificial reverberation by Gerzon [1971, 1972], who noted that several comb filters could "sound good" when cross-coupled. He proposed an orthogonal matrix feedback around a parallel bank of delay lines, as a means of maximizing cross-coupling. Some years later Stautner and Puckette [1982] independently suggested similar ideas and proposed a four-channel FDN reverberator based on the feedback matrix given in our Eq. (4.38). Jot [Jot and Chaigne, 1991, Jot, 1991, 1997] developed a systematic FDN design methodology allowing largely independent setting of reverberation time in different frequency bands. Rocchesso and Smith [1997] have provided further insights about the structures of feedback matrices in FDNs, and discussed analogies between FDNs and DWNs. General discussions of the use of FDNs for artificial reverberation are provided by Gardner [1998], Rocchesso [2002], Smith [2008]

Waveguide meshes were first studied by Van Duyne and Smith [1993, 1995]. Since then many studies have focused on techniques for reducing dispersion errors. Savioja and Välimäki [2000, 2003] have proposed interpolation and frequency-warping techniques to reduce dispersion as function of both frequency and propagation direction. Fontana and Rocchesso [1998, 2001] have focused on 2-D meshes, and provided results both about applications to membrane modeling and about general numerical aspects: they compared square, triangular, and hexagonal meshes in terms of sampling efficiency and dispersion error. Bilbao [2004] has also investigated in details many numerical and computational properties of the waveguide mesh, in particular he analyzed dispersion properties of various mesh topologies using von Neumann analysis and he provided a unified view of the digital waveguide mesh and wave digital filters as particular classes of energy invariant finite difference schemes. Finally, another topic addressed in the literature is the design of mesh boundaries, with a special focus on modeling diffusion. This problem was addressed by Laird et al. [1999], and later by Lee and Smith [2004], who used quadratic residue sequences to design maximally diffusing boundaries.

Three general and valuable books on spatial hearing are [Blauert, 1996], which is the traditional reference on the psychophysics of three-dimensional hearing, [Carlile, 1996], which not surveys the physics and psychophysics of 3-D auditory perception and also addresses the synthesis of spatial sound, and [Begault, 1994], which is focused on 3-D sound rendering techniques and applications to virtual reality and multimedia.

One of the pioneers in spatial hearing research was John Strutt, better known as Lord Rayleigh. He first described quantitatively the shadow effects of a sphere in [Strutt, 1904], and subsequently presented in [Strutt, 1907] the Duplex Theory that we have described in Sec. 4.5.2. The acoustic effects of the pinna have been studied in later years. Edgar A. G. Shaw and coworkers developed mechanical models of the external ear and measured their acoustic properties in several works (see e.g. [Teranishi and Shaw, 1968]). In the same years Dwight W. Batteau studied the role of the pinna in sound localization [Batteau, 1967]. More recently pinna effects have been studied through computational models, e.g. by Katz [2001].

As already mentioned, auditory cues for distance perception are still not completely understood. A recent review on the subject is provided by Zahorik et al. [2005]. The perceptual relevance of intensity scaling with distance han been known for a long time (see [Coleman, 1963]). Begault [1991] has shown that the preferred scaling of intensity with distance depends on the stimulus type. The R/D ratio have been cited in many studies as a relevant cue to distance since Rabinovich [1936]. Other relevant studies about the role of reverberation, familiarity, and expectation in distance perception include those by Mershon and Bowers [1979] and by Gardner [1969]. Butler et al. [1980] have studied distance-dependent spectral effects due to air absorption. Sound source localization in the near-field is another open research topic. Recent studies include the work by Shinn-Cunningham et al. [2000] and by Brungart [2002].

Studies on the importance of dynamic cues for sound localization date back to Wallach [1940]. Since then many studies have shown that active motion helps especially in azimuth estimation and to a lesser extent in elevation estimation [Thurlow and Runge, 1967, Perrett and Noble, 1997]. Wightman and Kistler [1999] have provided evidence of the disappearing of front-back reversal when listeners are allowed to turn their heads during the localization task. Loomis et al. [1998] have studied the role of dynamic cues, specifically motion parallax and acoustic tau, on the perception of distance.

A tutorial of HRTF-based rendering techniques is [Cheng and Wakefield, 2001], while a review paper more focused on the evaluation of 3-D sound systems is [Martens, 2003]. Huopaniemi [1999] also provides an extensive overview, especially on synthetic HRTFs and pole-zero models. The first attempt to develop a pole-zero HRTF model is reportedly Asano et al. [1990]. Other relevant contributions include work by Wakefield and coworkers (see e.g. [Durant and Wakefield, 2002]), and by Kulkarni and Colburn [2004]. The Interface Lab. at UC Davis has created a public-domain database of high-spatial-resolution HRTF measurements for 45 different subjects, including the KEMAR mannequin with both small and large pinnae. The database is described in Algazi et al. [2001]. The HRTFs plotted in our Fig. 4.25 have been taken from this database.

The first attempt to apply PCA techniques to series expansions of HRTFs appears to be [Martens, 1987]. Other relevant contributions include in particular work by Kistler and Wightman [1992]. Middlebrooks and Green [1992] have studied the relation between basis vectors obtained from PCA and anthropometric data. PCA is the oldest technique in multivariate analysis. It was originally developed by Pearson [1901] and further generalized by other authors. A general introduction to PCA can be found e.g. in [Duda et al., 2000].

Concerning HRTF interpolation: direct bilinear interpolation on FIR coefficients is described by Huopaniemi [1999]. Other recent contribution include [Zotkin et al., 2004], where an interpolation method that uses only three grid points is proposed, and [Freeland et al., 2004], where an interpolation procedure similar to the bilinear method, but based on auxiliary "interpositional transfer functions" (IPTFs), is proposed. Interpolation of pole-zero HRTF models is addressed e.g. by Hacıhabiboğlu et al. [2005] and by Larcher [2001] Interpolation of HRTF models based on PCA expansions has been investigated by Chen et al. [1995].

The origin of research on structural HRTF models is probably to be found in the work of Genuit [1984]. Even though it was based on very crude approximations of human geometries, the model incorporated static features of the HRTF (ear-canal resonance and eardrum impedance), as well as azimuth-dependent (ITD, IID) and elevation-dependent (pinna and torso reflections) features. The Interface Lab. at UC Davis has been working on the topic since the early '90's and has produced a number of relevant research papers. Much of our Sec. 4.6.2 is based on their work. For a start, see [Brown and Duda, 1998]. An interesting work that relates resonant properties of the pinna to anthropometry is [Raykar et al., 2005].

References

V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendano. The CIPIC HRTF database. In *Proc. IEEE Workshop Appl. Signal Process. to Audio and Acoust. (WASPAA01)*, pages 99–102, Mohonk Mountain House, New Paltz, NY, Oct. 2001.

Futoshi Asano, Yoiti Suzuki, and Toshio Sone. Role of spectral cues in median plane localization. *J. Acoust. Soc. Am.*, 88(1): 159–168, July 1990.

Dwight W. Batteau. The role of the pinna in human localization. *Proc. R. Soc. London. Series B, Biological Sciences*, 168, (1011):158–180, Aug. 1967.

Durand R. Begault. Preferred sound intensity increase for sensation of half distance. *Perceptual and Motor Skills*, 72:1019–1029, June 1991.

Durand R. Begault. *3-D Sound for Virtual Reality and Multimedia*. Academic Press Inc., 1994.

Leo L. Beranek. Concert hall acoustics. *J. Acoust. Soc. Am.*, 92(1):1–39, July 1992.

Stefan Bilbao. *Wave and Scattering Methods for Numerical Simulation*. ohn Wiley and Sons, Inc., New York, 2004.

Stefan Bilbao and Julian Parker. A virtual model of spring reverberation. *IEEE Trans. Speech Audio Process.*, 2010. In press.

Jens Blauert. *Spatial Hearing: Psychophysics of Human Sound Localization*. MIT Press, Cambridge, Mass., 2nd edition, 1996.

C. Phillip Brown and Richard O. Duda. A structural model for binaural sound synthesis. *IEEE Trans. Speech Audio Process.*, 6(5):476–488, Sep. 1998.

Douglas S. Brungart. Near-field virtual audio displays. *Presence: Teleoperators and Virtual Environment*, 11(1):93–106, Feb. 2002.

Robert A. Butler, Elena T. Levy, and William D. Neff. Apparent distance of sounds recorded in echoic and anechoic chambers. *J. Experimental Psychology*, 6(4):745–750, Nov. 1980.

Simon Carlile. *Virtual Auditory Space: Generation and Applications*. Chapman and Hall, New York, 1996.

Jiashu Chen, Barry D. Van Veen, and Kurt E. Hecox. A spatial feature extraction and regularization model for the head-related transfer function. *J. Acoust. Soc. Am.*, 97(1):439–452, Jan. 1995.

Corey I. Cheng and Gregory H. Wakefield. Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in time, frequency, and space. *J. Audio Eng. Soc.*, 49(4):231–249, Apr. 2001.

Paul D. Coleman. An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60(3):302–315, May 1963.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, Nov. 2000.

Eric A. Durant and Gregory H. Wakefield. Efficient model fitting using a genetic algorithm: pole-zero approximations of HRTFs. *IEEE Trans. Speech Audio Process.*, 10(1):18–27, Jan. 2002.

Federico Fontana and Davide Rocchesso. Physical modeling of membranes for percussion instruments. *Acta Acustica united with Acustica*, 84(3):529–542, May 1998.

Federico Fontana and Davide Rocchesso. Signal-Theoretic Characterization of Waveguide Mesh Geometries for Models of Two-Dimensional Wave Propagation in Elastic Media. *IEEE Trans. Speech Audio Process.*, 9(2):152–161, Feb. 2001.

Fabio P. Freeland, Luiz W. P. Biscainho, and Paulo S. R. Diniz. Interpositional transfer function for 3d-sound generation. *J. Audio Eng. Soc.*, 52(9):915–930, Sep. 2004.

Mark B. Gardner. Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. *J. Acoust. Soc. Am.*, 45 (1):47–53, Jan. 1969.

William G. Gardner. Reverberation algorithms. In Mark Kahrs and Karl-Heinz Brandenburg, editors, *Applications of Digital Signal Processing to Audio and Acoustics*, pages 85–131. Kluwer Academic Publishers, New York, Mar. 1998.

Klaus Genuit. *A model for the description of outer-ear transmission characteristics*. PhD thesis, Rheinisch-Westfalischen Technichen Hochschule Aachen, Aachen, Germany, Dec. 1984.

Michael A. Gerzon. Synthetic stereo reverberation, Part I. *Studio Sound*, 13:632–635, Dec. 1971.

Michael A. Gerzon. Synthetic stereo reverberation, Part II. *Studio Sound*, 14:24–28, Jan. 1972.

David Griesinger. The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. *Acta Acustica united with Acustica*, 83(4):721–731, 1997.

Hüseyin Hacıhabiboğlu, Banu Günel, and Ahmet M. Kondoz. Head-related transfer function filter interpolation by root displacement. In *Proc. IEEE Workshop Appl. Signal Process. to Audio and Acoust. (WASPAA05)*, pages 134–137, Mohonk Mountain House, New Paltz, NY, Oct. 2005.

Jyri Huopaniemi. *Virtual acoustics and 3-D sound in multimedia signal processing*. PhD thesis, Helsinki University of Technology, Faculty of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing, Espoo, 1999.

Jean-Marc Jot. An analysis/synthesis approach to real-time artificial reverberation. In *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, volume 2, pages 221–224, S. Francisco, Feb. 1991.

Jean-Marc Jot. Efficient models for reverberation and distance rendering in computer music and virtual audio reality. In *Proc. Int. Computer Music Conf.*, pages 236–243, Thessaloniki, 1997.

Jean-Marc Jot. Real-time spatial processing of sounds for music, multimedia, and interactive human-computer interfaces. *Multimedia Systems*, 7(1):55–69, Jan. 1999.

Jean-Marc Jot and Antoine Chaigne. Digital delay networks for designing artificial reverberators. In *Proc. Audio Engineering Society Convention*, Paris, Feb. 1991. Preprint 3030.

Brian F. G. Katz. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *J. Acoust. Soc. Am.*, 110(5):2440–2448, Nov. 2001.

Doris J. Kistler and Frederic L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91(3):1637–1647, Mar. 1992.

Abhijit Kulkarni and H. Steven Colburn. Infinite-impulse-response models of the head-related transfer function. *J. Acoust. Soc. Am.*, 115(4):1714–1728, Apr. 2004.

Heinrich Kuttruff. *Room Acoustics*. Elsiever Applied Science, London and New York, 3rd edition, 1991.

Joel Laird, Paul Masri, and Nishan Canagarajah. Modelling diffusion at the boundary of a digital waveguide mesh. In *Proc. Int. Computer Music Conf.*, pages 492–495, Beijing, Oct. 1999.

Véronique Larcher. *Techniques de spatialisation des sons pour la réalité virtuelle*. PhD thesis, Universite de Paris VI, Paris, May 2001.

Kyogu Lee and Julius O. Smith. Implementation of a highly diffusing 2-D digital waveguide mesh with a quadratic residue diffuser. In *Proc. Int. Computer Music Conf.*, Miami, Nov. 2004.

Jack M. Loomis, Roberta L. Klatzky, John W. Philbeck, and Reginald G. Golledge. Assessing auditory distance perception using perceptually directed action. *Perception and Psychophysics*, 60(6):966–980, 1998.

William L. Martens. Principal components analysis and resynthesis of spectral cues to perceived direction. In *Proc. Int. Computer Music Conf. (ICMC87)*, pages 274–281, Champaine-Urbana, IL, Sep. 1987.

William L. Martens. Perceptual evaluation of filters controlling source direction: Customized and generalized HRTFs for binaural synthesis. *Acoust. Sci. and Tech.*, 24(5):220–232, 2003. Special Issue on Spatial Hearing.

Donald H. Mershon and John N. Bowers. Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, 8(3):311–322, Mar. 1979.

John C. Middlebrooks and David M. Green. Observations on a principal components analysis of head-related transfer functions. *J. Acoust. Soc. Am.*, 92(1):597–599, July 1992.

Jame A. Moorer. About this reverberation businness. *Computer Music J.*, 3(2):13–18, Summer 1979.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2(6):559–572, 1901.

Stephen Perrett and William Noble. The effect of head rotations on vertical plane sound localization. *J. Acoust. Soc. Am.*, 102 (4):2325–2332, Oct. 1997.

A. V. Rabinovich. The effect of distance in the broadcasting studio. *J. Acoust. Soc. Am.*, 7(3):199–203, Jan. 1936.

Vikas C. Raykar, Ramani Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *J. Acoust. Soc. Am.*, 118(1):364–374, July 2005.

Davide Rocchesso. Spatial effects. In Udo Zölzer, editor, *Digital Audio Effects*, pages 137–200. John Wiley & Sons, Chirchester Sussex, UK, 2002.

Davide Rocchesso and Julius O. Smith. Circulant and elliptic feedback delay networks for artificial reverberation. *IEEE Trans. Speech Audio Process.*, 5(1):51–63, Jan. 1997.

Paul E. Sabine. Architectural acoustics: Its past and its possibilities. *J. Acoust. Soc. Am.*, 11(1):21–28, July 1939.

Lauri Savioja and Vesa Välimäki. Reducing the dispersion error in the digital waveguide mesh using interpolation and frequency-warping techniques. *IEEE Trans. Speech Audio Process.*, 8(2):184–194, Mar. 2000.

Lauri Savioja and Vesa Välimäki. Interpolated rectangular 3-d digital waveguide mesh algorithms with frequency warping. *IEEE Trans. Speech Audio Process.*, 11(6):783–790, Nov. 2003.

Manfred R. Schroeder. Natural-sounding artificial reverberation. *J. Audio Eng. Soc.*, 10(3):219–233, July 1962.

Manfred R. Schroeder. New method of measuring reverberation time. *J. Acoust. Soc. Am.*, 37(6):1187–1188, June 1965.

Manfred R. Schroeder. Digital simulation of sound transmission in reverberant spaces. *J. Acoust. Soc. Am.*, 47(2):424–431, Feb. 1970.

Barbara G. Shinn-Cunningham, Scott Santarelli, and Norbert Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *J. Acoust. Soc. Am.*, 107(3):1627–1636, Mar. 2000.

Julius O. Smith. *Physical Audio Signal Processing: for Virtual Musical Instruments and Digital Audio Effects, December 2008 Edition*. `http://ccrma.stanford.edu/˜jos/pasp/`, 2008. Accessed 15/12/2008.

Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *J. Audio Eng. Soc.*, 50(4):249–262, Apr. 2002.

John Stautner and Miller Puckette. Designing multichannel reverberators. *Computer Music J.*, 3(2):52–65, 1982. Reprinted in The Music Machine, Curtis Roads (Ed.). Cambridge, The MIT Press, 1989. (pp. 569-582).

(Lord Rayleigh) John W. Strutt. On the acoustic shadow of a sphere. *Philos. Trans. R. Soc. London*, A-203:87–89, 1904.

(Lord Rayleigh) John W. Strutt. On our perception of sound direction. *Philos. Mag.*, 13:214–232, 1907.

R. Teranishi and Edgar A. G. Shaw. External-ear acoustic models with simple geometry. *J. Acoust. Soc. Am.*, 44(1):357–263, July 1968.

Willard R. Thurlow and Philip S. Runge. Effect of induced head movements on localization of direction of sounds. *J. Acoust. Soc. Am.*, 42(2):480–488, Aug. 1967.

Scott A. Van Duyne and Julius O. Smith. Physical modeling with the 2-d digital waveguide mesh. In *Proc. Int. Computer Music Conf.*, pages 40–47, Tokio, 1993.

Scott A. Van Duyne and Julius O. Smith. The tetrahedral digital waveguide mesh. In *Proc. IEEE Workshop Appl. Signal Process. to Audio and Acoust.*, pages 234–237, Mohonk, Oct. 1995.

Hans Wallach. The role of head movement and vestibular and visual cues in sound localization. *J. Experimental Psychology*, 27:339–368, 1940.

Frederic L. Wightman and Doris J. Kistler. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *J. Acoust. Soc. Am.*, 105(5):2841–2853, May 1999.

Pavel Zahorik, Douglas S. Brungart, and Adelbert W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420, May 2005.

Dmitry N. Zotkin, Ramani Duraiswami, and Larry S. Davis. Rendering localized spatial audio in a virtual auditory space. *IEEE Trans. Multimedia*, 6(4):553–562, Aug. 2004.

Contents