5

Spatial effects

V. Pulkki, T. Lokki and D. Rocchesso

5.1 Introduction

A listener is capable of sensing his surroundings in some degree using only hearing, for example directions, distances, and spatial extents of sound sources, and also some characteristics of the rooms. This information is obtained by comparing the sound signals in ear canals to a set of spatial cues, used by the brain. Understanding the cues used by the hearing system helps the audio engineer to introduce some artificial features in the sound material in order to project the sound events in space. In the first half of this chapter, the most important techniques for sound projection are described, both for individual listeners using headphones and for an audience listening through a set of loudspeakers.

In natural listening conditions, sounds propagate from a source to the listener and during this trip they are widely modified by the environment. Therefore, there are some spatial effects imposed by the physical and geometric characteristics of the environment on the sound signals arriving to the listener's ears. Generally speaking, we call reverberation the kind of processing operated by the environment. The second half of this chapter illustrates this kind of effect and describes audio-processing techniques that have been devised to imitate and extend the reverberation that occurs in nature.

The importance of space has been largely emphasized in electroacoustic compositions, with the result that sophisticated spatial orchestrations often result in poor musical messages to the listener. Indeed, space cannot be treated as a composition parameter in the same way as pitch or timbre are orchestrated, just because space for sounds is not an "indispensable attribute" [KV01] as it is for images. This relative weakness is well explained if we think of two loudspeakers playing the same identical sound track: the listener will perceive one apparent source. The phenomenon is analogous to two colored spotlights that fuse to give one new, apparent, colored spot. In fact, color is considered a non-indispensable attribute for visual perception. However, just as color is a very important component in visual arts, the correct use of space can play a fundamental role

DAFX: Digital Audio Effects, Second Edition. Edited by Udo Zölzer.
© 2011 John Wiley & Sons, Ltd. Published 2011 by John Wiley & Sons, Ltd. ISBN: 978-0-470-66599-2

in music composition, especially for improving the effectiveness of other musical parameters of sound, such as pitch, timbre, and intensity.

5.2 Concepts of spatial hearing

As already mentioned, using only hearing, humans can localize sound sources, and they are also able to perceive some properties of the space they are in. This section considers both physical and perceptual issues which are related to such perception of spatial sound.

5.2.1 Head-related transfer functions

As a sound signal travels from a sound source to the ear canals of the listener, the signals in both ear canals will be different from the original sound signal and from each other. The transfer functions from a sound source to the ear canals are called the head-related transfer functions (HRTF) [Bla97]. They are dependent on the direction of a sound source related to the listener, and they yield temporal and spectral differences between left and right ear canals. Due to the fact that the ears are located on different sides of the skull, the arrival times of a sound signal vary with direction. Also, the skull casts an acoustic shadow that causes the contralateral ear signal to be attenuated. The shadowing is most prominent at frequencies above about 2 kHz, and does not exist when the frequency is below about 800 Hz. The pinna and other parts of the body may also change the sound signal. In some cases, it is advantageous to think about these filtering effects in the time domain, thus considering them *head-related impulse responses* (HRIR). Several authors have measured HRTFs by means of manikins or human subjects. A popular collection of measurements was taken by Gardner and Martin using a KEMAR dummy head, and made freely available [GM94, Gar97a]. A large set of HRTFs measured from humans have also been made available [ADDA01]. The HRTFs are also dependent on distance [BR99] with sources close to the listener. If the distance is more than about 1 m the dependence can be omitted. It will always be assumed in this chapter that the sources are in far field.

5.2.2 Perception of direction

Humans decode the differences of sound between the ear channels and use them to localize sound sources. These differences are called binaural directional cues. Temporal difference is called the interaural time difference (ITD) and spectral difference is called the interaural level difference (ILD) [Bla97]. Humans are sensitive to ILD at all frequencies, and to ITD mainly at frequencies lower than about 1.5 kHz. At higher frequencies, humans are also slightly sensitive to ITDs between signal envelopes, and not at all to ITD between the carriers of the signals. In typical HRTFs there exists a region near 2 kHz, where ILD is not monotonic with azimuth angle, and listeners easily localize the sources erroneously if the ITD between signal envelopes does not provide information of sound source direction [MHR10].

ITD and ILD provide information on where a sound source is in the left–right dimension. The angle between the sound source direction and the median plane can thus be decoded by the listener. The median plane is the vertical plane which divides the space related to a listener into left and right parts. The angle between the median plane and the sound source defines the cone of confusion, which is a set of points that all satisfy the following condition: the difference in distance from both ears to any point on the cone is constant, as shown in Figure 5.1. The angular coordinate system used in this chapter is also shown in the figure, which utilizes clockwise azimuth angle θ, being zero in front of the listener, and elevation angle ϕ, which defines the angle between the horizontal plane and the sound source direction, where positive is above the horizontal plane.

The information of the cone of confusion provided by the ITD and ILD is only an intermediate phase in the localization process. It is known that there are two mechanisms, which refine the

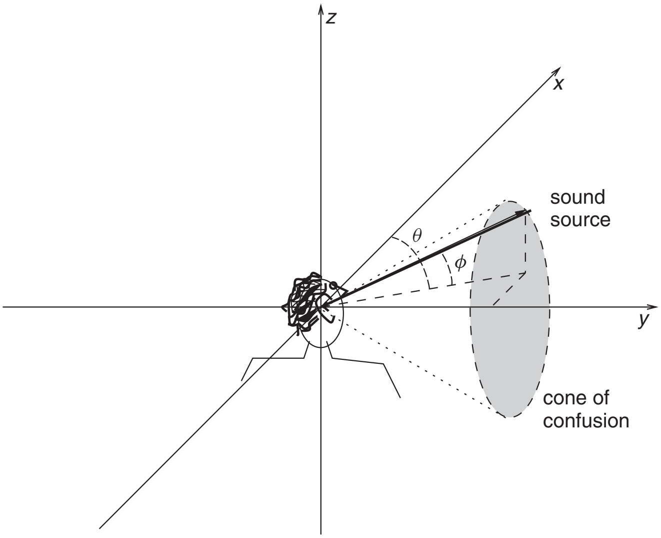


Figure 5.1 The azimuth-elevation coordinate system and cone of confusion.

perceived direction. One is related to monaural spectral cues, and the other is the monitoring of the effect of head rotation to binaural cues.

The monaural spectral cues are caused by the pinna of the listener, which filters the sound depending on the direction of arrival. For example, the concha, which is the cavity just around the ear canal opening, is known to have a direction-dependent resonance around 5–6 kHz [Bla97]. This effect with other direction-dependent filtering by the pinna and torso of the listener introduces spectral changes into the sound signal entering the ear canal at frequencies above 1–2 kHz. This provides information on the direction within the cone of confusion obtained from the ITD and ILD cues. Note that this mechanism is thus dependent on the spectrum of the signal, and a sufficiently broad and locally smooth spectrum is needed to decode the direction from the monaural spectrum. If the signal has too narrow a bandwidth, monaural spectral cues cannot be decoded. For example, some birds have a narrow bandwidth in their calls, and their localization using only hearing is relatively hard.

The effect of head movements on binaural cues, and how humans use this information in sound source localization [Bla97, GA97] are now discussed. For example, when a source is in front of the listener, and the listener rotates his head to the right, the left ear becomes closer the source, and the ITD and ILD cues change favoring the left ear. If the source is in the back of the listener, the cues would change favoring the right ear. This dynamic cue gives information on the source direction. Humans seem to use the information in a relatively coarse manner, such as if the source is in front, back or above of the listener. However, it is a very strong cue. A simple and very effective spatial effect can be composed by switching the ear canal signals of the listener in dynamic conditions either with tubes or microphones and loudspeakers. In this device, the sound signal captured on one side of the head of the listener is delivered to the ear on the other side. When wearing such a device, a striking directional effect is obtained, where the perceived direction of the voice of the visible speaker in front is perceived at the back of the listener.

5.2.3 Perception of the spatial extent of the sound source

It is also possible to perceive the extent of a sound source in some cases. For example, the sea shore and grand piano can be perceived to have a substantial width only using audition. Unfortunately, the knowledge of the corresponding perceptual phenomena and mechanisms is relatively sparse.

A basic result is, that point-like broadband sound sources are perceived to be point-like, and when incoherent broadband sound arrives from multiple directions evenly it is perceived to surround the listener [Bla97]. In these cases, the perception corresponds well to the physical situation. When the frequency content is narrower, or the duration of the stimulus is short, the perceived widths of the sources are perceived to be narrower than in reality [PB82, CT03, Hir07, HP08]. When the frequency bands of a broad sound signal are presented using loudspeakers in different directions, the listener perceives the source to be wide, though not as wide as the loudspeaker ensemble is [Hir07].

5.2.4 Room effect

So far we have discussed only the direct sound coming from the source to the listener. In real rooms and in many outdoor spaces there exist reflections and reverberation, which do not carry information on the direction of the sound. A mechanism has evolved which helps to localize sources in such environments. The precedence effect [Bla97, Zur87, LCYG99] is a suppression of early delayed versions of the direct sound in source direction perception. This has been researched a lot in classical studies, where a direct sound and a delayed sound are presented to a listener in anechoic conditions with two loudspeakers. When the delay is about 0–3 ms, no echo is perceived, and the perceived direction depends on the amplitude relationship and on the delay between the loudspeakers. The perceived direction may also be dependent on the frequency content of the sound. When the delay is about 5–30 ms, the presence of the lagging sound may be perceived, but it is not localized correctly. With larger delays, the delayed loudspeaker starts to be localizable. The effect is dependent on the signal, in principle: the more transient-like the nature of the signal, the more the precedence effect is salient. The precedence effect manifests itself in the Franssen effect, where the rapid onset of a sinusoid with a slow fadeout in one loudspeaker is interleaved with a slow fade in of the same sinusoid in another loudspeaker. The listener does not perceive that the second loudspeaker is emitting sound, but he erroneously perceives that the first loudspeaker is still active [Bla97].

Humans can also perceive the effect of the room in some manner. Indeed, in real life, a free-field condition very seldom occurs and sound always contains some reverberation, composed of reflections from surfaces. Humans can estimate the size of a room and even surface materials by listening to sounds. The perception relies on the density of the reflections and the length of the reverberation. Consider a concert hall and a bathroom, which both can have a reverberation time of 2 to 3 seconds, i.e., the sound is audible 2 to 3 seconds after the source has stopped emitting sound. The density of reflections, as well as their frequency characteristics modify the sound color, based on which humans can tell the size of space, even though the reverberation time in both cases is the same. The shape of the space can be also perceived to some extent, at least if it is a long narrow corridor, or a big concert hall.

5.2.5 Perception of distance

Humans also perceive the distance of sound sources to some extent [Bla97]. There are some main cues used for this. The perceived loudness created by a sound source has been proven to affect the perceived distance: the softer the auditory event, that farther away it is perceived. However, the signal has to be somewhat known by the listener, and is effective only with sources in about 1 m–10 m distances.

Listeners use the acoustical room effect caused by the source in perception of distance: the more the room effect is present in the ear canals of the listener, the farther away is the source perceived. This is quantified with the direct-to-reverberant ratio (DRR) of sound energies expressed in decibels. Besides the DRR, the room also has another well-known effect on perceived distance. If the impulse response of the room contains no strong early reflections, the source is perceived to be relatively near. This is utilized in studio reverberators with a predelay parameter, which controls the delay between the direct sound and the reverb tail. If the value of the predelay is long enough, the source is perceived at the distance of the loudspeakers, and if it is very short, then the source is perceived to be farther away.

When the source is very near to the listener, there are also some binaural effects which are used in distance perception [DM98]. With close sources the magnitude of ILD is higher and appears at lower frequencies than a far source with the same direction, which creates the perception of a nearby source.

5.3 Basic spatial effects for stereophonic loudspeaker and headphone playback

The most common loudspeaker layout is the two-channel setup, called the standard stereophonic setup. It was widely taken into use after the development of single-groove 45°/45° two-channel records in the late 1950s. Two loudspeakers are positioned in front of the listener, separated by 60° from the listener's viewpoint, as presented in Figure 5.2. The setup of two loudspeakers is very common, though quite often the setup is not as shown in the figure. In domestic use, or in car audio typically the listener is not situated in the centre, but the loudspeakers are located in different directions and distances from him than in Figure 5.2. However, even then two-channel reproduction is preferred from monophonic presentation in most cases. On the other hand, in some cases the listener can be assumed to be in the best listening position, as in computer audio.

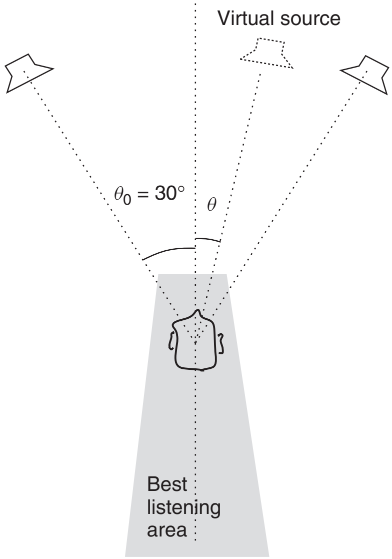


Figure 5.2 Standard stereophonic listening configuration.

This section deals with the spatial effects obtainable with such two-channel reproduction and simple processing. Two types of effects are presented, the creation of point-like sources, and the creation of spatially spread sources. More advanced methods for loudspeaker reproduction with HRTF processing are discussed in Section 5.4.5, which provide some more degrees of freedom, but unfortunately also introduce some limitations in listening position and listening room acoustics.

5.3.1 Amplitude panning in loudspeakers

Amplitude panning is the most frequently used virtual-source-positioning technique. In it a sound signal is applied to loudspeakers with different amplitudes, which can be formulated as

$$x_i(t) = g_i x(t), \quad i = 1, \ldots, N, \tag{5.1}$$

where $x_i(t)$ is the signal to be applied to loudspeaker i, g_i is the gain factor of the corresponding channel, N is the number of loudspeakers, and t is the time. The listener perceives a virtual source, the direction of which is dependent on the gain factors.

If the listener is located equally distant from the loudspeakers, the panning law estimates the perceived direction θ from the gain factors of loudspeakers. The estimated direction is called the panning direction or panning angle. In [Pul01] it has been found that amplitude panning provides consistent ITD cues up to 1.1 kHz, and roughly consistent ILD cues above 2 kHz for a listener in the best listening position. The *level* differences between the loudspeakers are changed a bit surprisingly to *phase* differences between the ears, which is due to the fact that the sound arrives from both loudspeakers to both ears, which is called cross-talk. This effect is valid at low frequencies. At high frequencies, the level differences of the loudspeakers turn into level differences due to lack of the cross-talk caused by the shadowing of the head.

There exist many published methods to estimate the perceived direction. In practice, all the proposed methods are equally good for audio effects, and the tangent law by Bennett *et al.* [BBE85] is formulated as

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}, \tag{5.2}$$

which has been found to estimate perceived direction best in listening tests in anechoic listening [Pul01]. There are also other panning laws, reviewed in [Pul01].

The panning laws set only the ratio between the gain factors. To prevent undesired changes in loudness of the virtual source depending on panning direction, the sum-of-squares of the gain factors should be normalized:

$$\sqrt{\sum_{n=1}^{N} g_n^2} = 1. \tag{5.3}$$

This normalization equation is used in real rooms with some reverberation. Depending on listening room acoustics, different normalization rules may be used [Moo90].

The presented analysis is valid only if the loudspeakers are equidistant from the listener, and if the base angle is not larger than about 60°. This defines the best listening area where the virtual sources are localized between the loudspeakers. The area is located around the axis of symmetry of the setup, as shown in Figure 5.2. When the listener moves away from the area, the virtual source is localized towards the nearest loudspeaker which emanates a considerable amount of sound, due to the precedence effect.

In principle, the amplitude-panning method creates a comb-filter effect in the sound spectrum, as the same sound arrives from both loudspeakers to each ear. However, this effect is relatively mild, and when heard in a normal room, the room reverberation smooths the coloring effect prominently. The sound color is also very similar when heard in different positions in the room. The lack of prominent coloring and the relatively robust directional effect provided by it are very probably the reasons why amplitude panning is included in all mixing consoles as "panpot" control, which makes it the most widely used technique to position virtual sources.

M-file 5.1 `(stereopan.m)`

```
% stereopan.m
% Author: V. Pulkki

% Stereophonic panning example with tangent law
Fs=44100;
theta=-20; % Panning direction
% Half of opening angle of loudspeaker pair
lsbase=30;
```

```
% Moving to radians
theta=theta/180*pi;
lsbase=lsbase/180*pi;
% Computing gain factors with tangent law
g(2)=1; % initial value has to be one
g(1)=- (tan(theta)-tan(lsbase)) / (tan(theta)+tan(lsbase)+eps);
% Normalizing the sum-of-squares
g=g/sqrt(sum(g.^2));
% Signal to be panned
signal=mod([1:20000]',200)/200;
% Actual panning
loudsp_sig=[signal*g(1) signal*g(2)];
% Play audio out with two loudspeakers
soundsc(loudsp_sig,Fs);
```

5.3.2 Time and phase delays in loudspeaker playback

When a constant delay is applied to one loudspeaker in stereophonic listening, virtual sources with transient signals are perceived to migrate towards the loudspeaker that radiates the earlier sound signal [Bla97]. Maximal effect is achieved asymptotically when the delay is approximately 1.0 ms or more. However, the effect depends on the signal used. With continuous signals containing low frequencies, the effect is much less prominent than with modulated signals containing high frequencies.

In such processing the *phase* or *time* delays between the loudspeakers are turned at low frequencies into *level* differences between the ears, and at high frequencies to *time* differences between the ears. This all makes the virtual source direction depend on frequency [Coo87, Lip86]. The produced binaural cues vary with frequency, and different cues suggest different directions for virtual sources [PKH99]. It may thus generate a "spread" perception of direction of sound, which is desirable in some cases. The effect is dependent on listening position. For example, if the sound signal is delayed by 1 ms in one loudspeaker, the listener can compensate the delay by moving 30 cm towards the delayed loudspeaker.

M-file 5.2 (delaypan.m)

```
% delaypan.m
% Author: V. Pulkki
% Creating spatially spread virtual source by delaying one channel
Fs=44100;
% Delay parameter for channel 1 in seconds
delay=0.005;
% Corresponding number of delayed samples
delaysamp=round(delay*Fs)
% Signal to be used
signal=mod([1:20000]',400)/400;
signal(1:2000)=signal(1:2000).*[1:2000]'/2000; % Fade in
% Delaying first channel
loudsp_sig=[[zeros(delaysamp,1); signal(1:end-delaysamp)] signal];
% Play audio with loudspeakers
soundsc(loudsp_sig,Fs);
```

A special case of a phase difference in stereophonic reproduction is the use of antiphasic signals in the loudspeakers. In such a technique, the same signal is applied to both loudspeakers, however, the polarity of the other loudspeaker signal is inverted, which produces a constant 180°

phase difference between the signals at all frequencies. This changes the perceived sound color, and also spreads the virtual sources. Depending on the listening position, the low frequencies may be cancelled out. At higher frequencies this effect is milder. This effect is also milder in rooms with longer reverberation. The directional perception of the antiphasic virtual source depends on the listening position. In the sweet spot, the high frequencies are perceived to be at the center, and low frequencies in random directions. Outside the sweet spot, the direction is either random, or towards the closest loudspeaker. In the language of professional audio engineers this effect is called "phasy", or "there is phase error in here".

M-file 5.3 (phaseinvert.m)

```
% phaseinvert.m
% Author: V. Pulkki
% Create a spread virtual source by inverting phase in one loudspeaker
Fs=44100;
signal=mod([1:20000]',400)/400; %signal to be used
% Inverting one loudspeaker signal
loudsp_sig=[-signal signal];
% Play audio out with two loudspeakers
soundsc(loudsp_sig,Fs);
```

A further method to spread the virtual source between the loudspeakers is to change the phase spectrum of the sound differently at different frequencies. A basic method is to convolve the signal for the loudspeakers with two different short bursts of white noise. Another method is to apply a different delay to different frequencies. This effectively spreads out the virtual source between the loudspeakers, and the effect is audible over a large listening area. Unfortunately, the processing changes the temporal response slightly, which may be audible as temporal smearing of transients of the signal.

Below is a example creating spread virtual sources for stereophonic listening by convolving the sound with short noise bursts:

M-file 5.4 (spreadnoise.m)

```
% spreadnoise.m
% Author: V. Pulkki
% Example how to spread a virtual source over N loudspeakers
Fs=44100;
signal=mod([1:20000]',400)/400; % Signal to be used
NChan=2; % Number of channels
% Generate noise bursts for all channels
nois=rand(round(0.05*Fs),NChan)-0.5;
% Convolve signal with bursts
loudsp_sig=conv(signal,nois(:,1));
for i=2:NChan
    loudsp_sig=[loudsp_sig conv(signal,nois(:,i))];
end
if NChan == 2
    % Play audio out with  loudspeakers
    soundsc(loudsp_sig,Fs);
else
    % Write file to disk
    loudsp_sig=loudsp_sig/max(max(loudsp_sig))*0.9;
    wavwrite([loudsp_sig],Fs,16,'burstex.wav');
end
```

5.3.3 Listening to two-channel stereophonic material with headphones

The headphone listening is significantly different to loudspeaker listening. In headphones the cross-talk present in loudspeaker listening is missing, meaning that the sound from the left headphone enters only to the left ear canal, and similarly with the right side. Typically, the audio engineers create the stereophonic audio content in studios with two-channel loudspeaker listening. It is then relevant to ask how the spatial perception of the content changes, when listened to over headphones.

With amplitude-panned virtual sources the level difference between headphone channels is turned directly into ILD, and ITD remains zero. This is very different from loudspeaker listening, where the direction of amplitude-panned sources relies on ITD cues, and ILD remains zero at low frequencies. Although this seems a potential source for large differences in spatial perception of resulting virtual sources, the resulting spatial image is similar. The virtual sources are ordered from the left to right in about the same order as in loudspeaker listening, however in headphone listening the sources are perceived inside the listener's head. This internalization is due to two facts: the dynamic cues propose internalized sources since the ITD and ILD do not change with listener movements, and also the monaural spectral cues do not suggest external sources, since the spectral cues are very different from the cues produced with distant sources.

If the stereophonic material includes virtual sources which have been spatialized by applying time delays, as in Section 5.3.2, this may result in a vastly different spatial perception in headphone listening, e.g., a 5 ms delay in the left loudspeaker may produce a spread perception of the sound in loudspeaker listening, but in headphone listening the sound can be perceived to originate only from the right headphone.

In Section 5.3.2 the technique to spread out virtual sources by convolution with noise burst was also described. This effect provides a similar effect in both headphone and loudspeaker listening. The frequency-dependent alteration of signal phase and magnitude creates ITD and ILD cues which change as a function of frequency in both loudspeaker and headphone listening. Of course, the effect is not the same, as in headphone listening the sound is perceived inside the head, and in loudspeaker listening it is perceived between the loudspeakers.

5.4 Binaural techniques in spatial audio

Binaural techniques are loosely defined to be methods which aim to control directly the sound in the ear canals to match a recorded real case or with a simulated virtual case. This is done by careful binaural recordings, or by utilizing measured or modeled head-related transfer functions (HRTFs) and acoustical modeling of the listening space.

5.4.1 Listening to binaural recordings with headphones

The basic technique is to reproduce a recorded binaural sound-track with headphones. The recording is made by inserting miniature microphones in to the ear canals of a real human listener, or by using a manikin with microphones in the ears [Bla97]. This recording is reproduced by playing the recorded signals in the ears of the listener. This is a very simple technique in principle, and can provide effective results. A simple implementation is to replace the transducers of in-ear headphones with insert miniature microphones, to use a portable audio recorder to record the sounds of the surroundings, and to play back the sound with headphones. Already without any equalization, a nice spatial effect is achieved, as the left–right directions of the sound sources and reverberant sound field are reproduced naturally. Especially, if the person who did the recording is listening, the effect can be striking.

Unfortunately, there are also problems with the technique. The sound may appear colored, the perceived directions move from front to back, and everything may be localized inside head. To partially avoid these problems, the recording and the reproduction should be equalized carefully to get a flat frequency response from the ear drum of the person in the recording position to the ear drum of the listener. Such equalization requires very careful measurements, and is not discussed further here.

A further problem in listening to binaural recordings is the fact that listeners use also dynamic cues to localize sound. When a binaural recording is listened with headphones, the movements of the listener do not naturally change the binaural recording at all. This is also a reason why binaural recordings easily tend to be localized inside head of the listener.

Another issue is the problem of individuality. Each listener has different pinna, and head size, and the sound in similar conditions appears different in different individuals' ears. When a binaural recording made by another individual is listened to, similar problems occur as with non-optimal equalization.

5.4.2 Modeling HRTF filters

Modeling the structural properties of the system pinna–head–torso gives us the possibility to research spatial hearing. Much of the physical/geometric properties can be understood by careful analysis of the HRIRs, plotted as surfaces, functions of the variables time and azimuth, or time and elevation. This is the approach taken by Brown and Duda [BD98] who came up with a model which can be structurally divided into three parts: (1) head shadow and ITD, (2) shoulder echo, and (3) pinna reflections.

Starting from the approximation of the head as a rigid sphere that diffracts a plane wave, the shadowing effect can be effectively approximated by a first-order continuous-time system, i.e., a pole-zero couple in the Laplace complex plane:

$$s_z = \frac{-2\omega_0}{\alpha(\theta)} \tag{5.4}$$

$$s_p = -2\omega_0 \ , \tag{5.5}$$

where ω_0 is related to the effective radius a of the head and the speed of sound c by

$$\omega_0 = \frac{c}{a} \ . \tag{5.6}$$

The position of the zero varies with the azimuth θ according to the function

$$\alpha(\theta) = 1.05 + 0.95 \cos\left(\frac{\theta + \frac{\pi}{2}}{150^\circ} 180^\circ\right) . \tag{5.7}$$

The pole-zero couple can be directly translated into a stable IIR digital filter by bilinear transformation [Mit98], and the resulting filter (with proper scaling) is

$$H_{\mathrm{hs}} = \frac{(\omega_0 + \alpha F_s) + (\omega_0 - \alpha F_s) z^{-1}}{(\omega_0 + F_s) + (\omega_0 - F_s) z^{-1}} \ . \tag{5.8}$$

The ITD can be obtained by simple delay in seconds as is the following function of the azimuth angle θ:

$$\tau_{\mathrm{h}}(\theta) = \begin{cases} -\frac{a}{c}\cos\left(\theta + \frac{\pi}{2}\right) & \text{if } 0 \le |\theta + \frac{\pi}{2}| < \frac{\pi}{2} \\ \frac{a}{c}\left(|\theta + \frac{\pi}{2}| - \frac{\pi}{2}\right) & \text{if } \frac{\pi}{2} \le |\theta + \frac{\pi}{2}| < \pi \end{cases} \ . \tag{5.9}$$

The overall magnitude responses of the block responsible for head shadowing is reported in Figure 5.3.

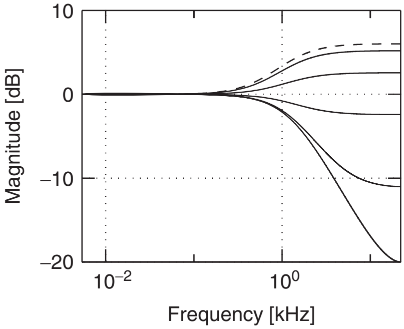


Figure 5.3 Magnitude responses of the simplified HRTFs for the ear in negative azimuth side. Azimuth ranging from $-\frac{\pi}{2}$ (dashed line) to $\frac{\pi}{2}$ at steps of $\pi/6$.

The M-file implementing the head-shadowing filter as a time-domain HRIR is:

M-file 5.5 (simpleHRIR.m)

```
function [output] = simpleHRIR(theta, Fs)
% [output] = simpleHRIR(theta, Fs)
% Author: F. Fontana and D. Rocchesso, V.Pulkki
%
% computes simplified HRTFs with only simple ITD-ILD approximations
% theta is the azimuth angle in degrees
% Fs is the sample rate
theta = theta + 90;
theta0 = 150 ;
alfa_min = 0.05 ;
c = 334; % speed of sound
a = 0.08; % radius of head
w0 = c/a;
input=zeros(round(0.003*Fs),1); input(1)=1;
alfa = 1+ alfa_min/2 + (1- alfa_min/2)* cos(theta/ theta0* pi) ;
B = [(alfa+w0/Fs)/(1+w0/Fs), (-alfa+w0/Fs)/(1+w0/Fs)] ;
    % numerator of Transfer Function
A = [1, -(1-w0/Fs)/(1+w0/Fs)] ;
    % denominator of Transfer Function
if (abs(theta) < 90)
 gdelay = round(- Fs/w0*(cos(theta*pi/180) - 1))  ;
else
 gdelay = round(Fs/w0*((abs(theta) - 90)*pi/180 + 1) );
end;
out_magn = filter(B, A, input);
output = [zeros(gdelay,1); out_magn(1:end-gdelay);  ];
```

The function `simpleHRIR` gives a rough approximation of HRIR of one ear with one direction. To obtain a HRIR for the left and right ears the same function has to be used with opposite values of argument `theta`. An example is presented in Section 5.4.3.

5.4.3 HRTF processing for headphone listening

A monophonic sound signal can be positioned virtually in any direction in headphone listening, if HRTFs for both ears are available for the desired virtual source direction [MSHJ95, Beg94]. The result of the measurements is a set of HRIRs that can be directly used as coefficients of a pair of FIR filters. Since the decay time of the HRIR is always less than a few milliseconds, 256 to 512 taps are sufficient at a sampling rate of 44.1 kHz. A sound signal is filtered with a digital filter modeling the measured HRTFs. The method simulates the ear-canal signals that would have been produced if a sound source existed in a desired direction.

A point-like virtual source is created with this example:

M-file 5.6 (simplehrtfconv.m)

```
function [binauralsig] = simplehrtfconv(theta)
% [binauralsig] = simplehrirconv(theta)
% Author: V. Pulkki
% Convolve a signal with HRIR pair corresponding to direction theta
% Theta is azimuth angle of virtual source
Fs =44100; % Sample rate
HRTFpair=[simpleHRIR(theta,Fs) simpleHRIR(-theta,Fs)];
signal=rand(Fs*5,1);
% Convolution
binauralsig=[conv(HRTFpair(:,1),signal) conv(HRTFpair(:,2),signal)];
%soundsc(binauralsig,Fs);% Uncomment to play sound for headphones
```

The demonstration above produces very probably the perception of inside-head virtual source, which may also sound colored. If a head tracker is available, much more realistic perception of external sound sources can be obtained with headphones. In head tracking, the direction of the listener's head is monitored about 10–100 times a second, and the HRTF filter is changed dynamically to keep the perceived direction of sound constant with the space where the listener is. In practice, the updating of the HRTF filter has to be done carefully in order not to produce audible artifacts.

The technique discussed above simulates anechoic listening of a distant sound source. It is also possible to simulate with the same technique the binaural listening of a sound source in a real room. In this approach, the binaural room impulse responses (BRIRs) are measured from the ear canals of a subject in a room with a relatively distant loudspeaker. The main difference to HRTFs is that typically the lengths of HRTFs are of the order of a few milliseconds and include only the acoustical response of the subject, whereas the BRIRs include the room responses, and they can be even few seconds long. The same MATLAB® example can also be used to process sound with BRIRs, although the required processing is much heavier in that case, since the convolution with such long responses is computationally a complex process.

HRTFs can also be used for cross-talk-canceled loudspeaker listening. In that case, the binaural signals are computed as shown in this section, and then played back with a stereo dipole, as shown in Section 5.4.5.

5.4.4 Virtual surround listening with headphones

An interesting application for HRTF technologies with headphones is listening to existing multichannel audio material. In such cases, each loudspeaker in the multichannel loudspeaker layout is simulated using an HRTF pair. For example, a signal meant to be applied to the loudspeaker in a 30° direction is convolved with the HRTF pair measured from the same direction, and the convolved signals are applied to the headphones. The usage of HRTFs measured in anechoic conditions is often suboptimal in this case, and the use of BRIRs is beneficial, which have similar responses to the room in which the subject is located. This can be done using measured room impulse responses, or by simulating the effect or room with a reverberator, see Section 5.6.

An example of virtual loudspeaker listening with headphones:

M-file 5.7 (virtualloudspeaker.m)

```
% virtualloudspeaker.m
% Author: V. Pulkki
% Virtual playback of 5.0 surround signal over headphones using HRIRs
Fs=44100;
% generate example 5.0 surround signal
cnt=[0:20000]’;
signal=[(mod(cnt,200)/200) (mod(cnt,150)/150) (mod(cnt,120)/120)...
    (mod(cnt,90)/90) (mod(cnt,77)/77)];
i=1;
% go through the input channels
outsigL=0; outsigR=0;
for theta=[30 -30 -110 110 0]
  HRIRl=simpleHRIR(theta,Fs);
  HRIRr=simpleHRIR(-theta,Fs);
  outsigL=outsigL+conv(HRIRl,signal(:,i));
  outsigR=outsigR+conv(HRIRr,signal(:,i));
  i=i+1;
end
% sound output to headphones
soundsc([outsigL outsigR],Fs)
```

5.4.5 Binaural techniques with cross-talk canceled loudspeakers

Binaural recordings are meant to be played back in such a way that the sound which originates from the left ear is played back only to the left ear, and correspondingly with the right ear. If such a recording is played back with stereophonic setup of loudspeakers, the sound from the left loudspeaker also travels to the right ear, and vice versa, called *cross-talk*, which ruins the spatial audio quality.

In order to be able to listen to binaural recordings over two loudspeakers, some methods have been proposed [CB89, KNH98]. In these methods, the loudspeakers are driven in such a way that in practice the cross-talk is canceled as much as possible.

A system can be formed as presented in Figure 5.4 to deliver binaurally recorded signals to the listener’s ears using two closely spaced loudspeakers with cross-talk cancellation. The binaural signals are represented as a 2x1 vector in $\mathbf{x}(n)$, and the produced ear canal signals also as 2x1 vector $\mathbf{d}(n)$. The system can be formulated in the z-domain

$$\mathbf{d}(z) = \mathbf{C}(z)\mathbf{H}(z)\mathbf{x}(z), \tag{5.10}$$

where $\mathbf{C}(z) = \begin{bmatrix} C_{11}(z) & C_{12}(z) \\ C_{21}(z) & C_{22}(z) \end{bmatrix}$ contains the electro-acoustical responses of the loudspeakers measured in the ear canals, as shown in the figure, and $\mathbf{H}(z) = \begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{bmatrix}$ contains the responses for performing inverse filtering to minimize the cross-talk.

Ideally, $\mathbf{x}(z) = \mathbf{d}(z)$, which can be obtained if $\mathbf{H}(z) = \mathbf{C}(z)^{-1}$. Unfortunately, the direct inversion is not feasible due to unidealities of the loudspeakers and the listening conditions. A regularized method to find an optimal $\mathbf{H}_{opt}(z)$ has been proposed in [KNH98],

$$\mathbf{H}_{opt}(z) = \left[\mathbf{C}^T(z^{-1})\mathbf{C}(z) + \beta\mathbf{I}\right]^{-1} \mathbf{C}^T(z^{-}1)z^{-m}, \tag{5.11}$$

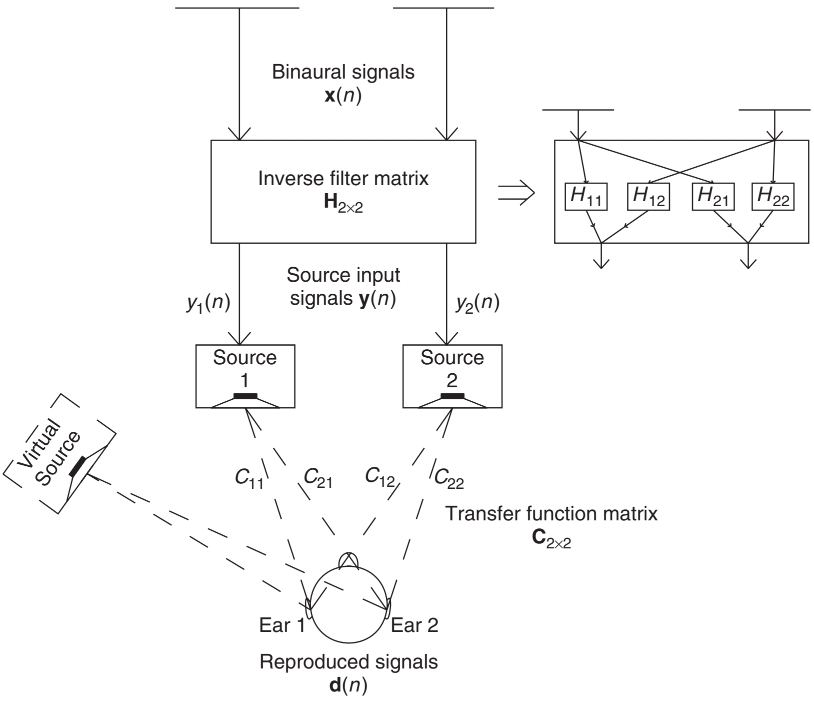


Figure 5.4 Presentation of binaurally recorded signals with loudspeakers with cross-talk canceling [Pol07].

where β is a positive scalar regularization factor, and z^{-m} models the time delay due to the sound reproduction system. If β is selected very low, there will be sharp peaks in the resulting time-domain inverse filters, which may exceed the dynamic range of the loudspeakers. If β is selected to be higher, the inverse filter will have longer duration in time, which is less demanding on the loudspeakers, but unfortunately the inversion is also less accurate [KNH98].

A **MATLAB** example is provided in the following to compute inverse filters for a cross-talk canceling system:

- The responses in **C** are moved into the frequency domain with discrete Fourier transform (DFT) with the desired length of time window.
- The filter responses are computed by $\mathbf{H}_{opt}(k) = \left[\mathbf{C}^H(k)\mathbf{C}(k) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^H(k)$, where k presents the frequency bin indexes and H Hermitian transposition.
- The inverse DFT is taken of **H**, resulting in the inverse filters for cross-talk cancellation.
- A circular shift of half of the applied time-window length is implemented on the inverse filters.

M-file 5.8 (crosstalkcanceler.m)

```
% crosstalkcanceler.m
% Author: A. Politis, V. Pulkki
% Simplified cross-talk canceler
theta=10;  % spacing of stereo loudspeakers in azimuth
Fs=44100; % sample rate
```

```
b=10^-5;   % regularization factor
% loudspeaker HRIRs for both ears (ear_num,loudspeaker_num)
% If more realistic HRIRs are available, pls use them
HRIRs(1,1,:)=simpleHRIR(theta/2,Fs);
HRIRs(1,2,:)=simpleHRIR(-theta/2,Fs);
HRIRs(2,1,:)=HRIRs(1,2,:);
HRIRs(2,2,:)=HRIRs(1,1,:);
Nh=length(HRIRs(1,1,:));
%transfer to frequency domain
for i=1:2;for j=1:2
        C_f(i,j,:)=fft(HRIRs(i,j,:),Nh)
    end;end
% Regularized inversion of matrix C
H_f=zeros(2,2,Nh);
for k=1:Nh
    H_f(:,:,k)=inv((C_f(:,:,k)'*C_f(:,:,k)+eye(2)*b))*C_f(:,:,k)';
end
% Moving back to time domain
for k=1:2; for m=1:2
        H_n(k,m,:)=real(ifft(H_f(k,m,:)));
        H_n(k,m,:)=fftshift(H_n(k,m,:));
    end; end
% Generate binaural signals.  Any binaural recording shoud also be ok
binauralsignal=simplehrtfconv(70);
%binauralsignal=wavread('road_binaural.wav');
% Convolve the loudspeaker signals
loudspsig=[conv(reshape(H_n(1,1,:),Nh,1),binauralsignal(:,1)) + ...
    conv(reshape(H_n(1,2,:),Nh,1),binauralsignal(:,2)) ...
    conv(reshape(H_n(2,1,:),Nh,1),binauralsignal(:,1)) + ...
    conv(reshape(H_n(2,2,:),Nh,1),binauralsignal(:,2))];
soundsc(loudspsig,Fs)       % play sound for loudspeakers
```

In practice, this method works best with loudspeakers close to each other, as a larger loudspeaker base angle would lead to coloration at lower frequencies. The listening area in which the effect is audible is very small, as if the listener departs from the mid line between the loudspeakers by about 1–2 cm, the effect is lost.

A nice feature of this technique is that the sound is typically externalized. This may be due to the fact that head movements of the listener produce somewhat relevant cues, and since the sound is reproduced using far-field loudspeakers generating plausible monaural spectral cues. However, although the sound is externalized, a surrounding spatial effect is hard to obtain with this technique. With a stereo dipole in the front, the reproduced sound scene is typically perceived only at the front.

The technique also is affected by the reflections and reverberation of the listening room. It works best only in spaces without prominent reflections. To get the best results, the HRTFs of the listener should be known, however already very plausible results can be obtained with generic responses.

5.5 Spatial audio effects for multichannel loudspeaker layouts

5.5.1 Loudspeaker layouts

In the history of multichannel audio [Ste96, Dav03, Tor98] multiple different loudspeaker layouts with more than two loudspeakers have been specified. The most frequently used layouts are presented in this chapter. In the 1970s, the quadraphonic setup was proposed, in which four loudspeakers are placed evenly around the listener at azimuth angles ±45° and ±135°. This layout was never successful because of problems related to reproduction techniques of that time, and because the layout itself had too few loudspeakers to provide good spatial quality in all directions around the listener [Rum01].

For cinema sound, a system was evolved in which the frontal image stability of the standard stereophonic setup was enhanced by one extra center channel, and two surround channels were used to create atmospheric effects and room perception. This kind of setup was first used in Dolby's surround sound system for cinemas from 1976 [Dav03]. Later, the layout was investigated [The91], and ITU gave a recommendation about the layout in 1992 [BS.94]. In the late 1990s, this layout also became common in domestic use. It is widely referred to as the 5.1 system, where 5 stands for the number of loudspeakers, and .1 stands for the low-frequency channel. In the recommendation, three frontal loudspeakers are at directions 0° and ±30°, and two surround channels at $\pm(110 \pm 10)^\circ$, as shown in Figure 5.5. The system has been criticized for not delivering good directional quality anywhere but in front [Rum01]. To achieve better quality, it can be extended by adding loudspeakers. Layouts with 6–12 loudspeakers have been proposed, and are presented in [Rum01].

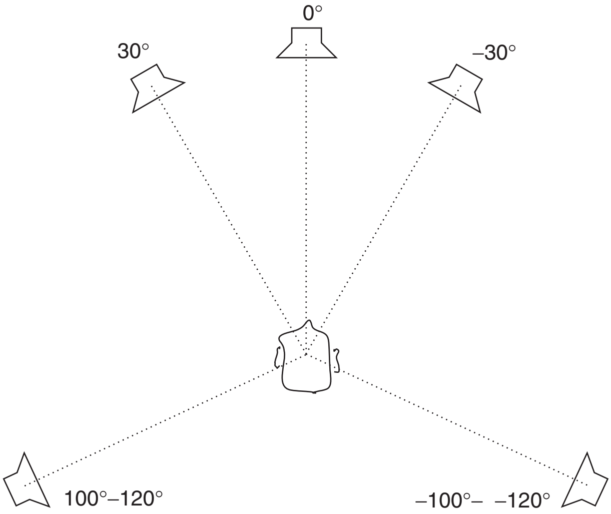


Figure 5.5 Five channel surround loudspeaker layout based on ITU recommendation BS775.1.

In computer music, media installations and in academic projects, loudspeaker setups, in which the loudspeakers have equal spacing, have been used. In horizontal arrays the number of loudspeakers can be, for example, six (hexagonal array) or eight (octagonal array). In wave-field synthesis, see Section 5.5.6, the number of loudspeakers is typically between 20 and 200. In theaters and in virtual environment systems there exist systems in which loudspeakers are also placed above and/or below the listener.

5.5.2 2-D loudspeaker setups

In 2-D loudspeaker setups all loudspeakers are on the horizontal plane. Pair-wise amplitude panning [Cho71] is the best method to position virtual sources with such setups, when the number of loudspeakers is less than about 20. In pair-wise panning the sound signal is applied only to two adjacent loudspeakers of the loudspeaker setup at one time. The pair between which the panning direction lies is selected. Different formulations for pair-wise panning are Chowning's law [Cho71], which is not based on any psychoacoustic criteria, or 2-D vector-base amplitude panning (VBAP) [Pul97], which is a generalization of the tangent law (Equation (5.2)) for stereophonic panning.

In VBAP a loudspeaker pair is specified with two vectors. The unit-length vectors $\mathbf{l}_m$ and $\mathbf{l}_n$ point from the listening position to the loudspeakers. The intended direction of the virtual source (panning direction) is presented with a unit-length vector $\mathbf{p}$. Vector $\mathbf{p}$ is expressed as a linear weighted sum of the loudspeaker vectors

$$\mathbf{p} = g_m\mathbf{l}_m + g_n\mathbf{l}_n. \tag{5.12}$$

Here g_m and g_n are the gain factors of the respective loudspeakers. The gain factors can be solved as

$$\mathbf{g} = \mathbf{p}^T\mathbf{L}_{mn}^{-1}, \tag{5.13}$$

where $\mathbf{g} = [g_m\ g_n]^T$ and $\mathbf{L}_{mn} = [\mathbf{l}_m\ \mathbf{l}_n]$. The calculated factors are used in amplitude panning as gain factors of the signals applied to respective loudspeakers after suitable normalization, e.g., $||\mathbf{g}|| = 1$.

The directional quality achieved with pair-wise panning was studied in [Pul01]. When the loudspeakers are symmetrically placed on the left and right of the listener, VBAP estimates the perceived angle accurately. When the loudspeaker pair is not symmetrical with the median plane, the perceived direction is biased towards the median plane [Pul01], which can be more or less compensated [Pul02].

When there is a loudspeaker in the panning direction, the virtual source is sharp, but when panned between loudspeakers, the binaural cues are unnatural to some degree. This means that the directional perception of the virtual source varies with panning direction, which can be compensated by always applying sound to more than one loudspeaker [Pul99, SK04]. As in pair-wise panning, outside the best listening position the perceived direction collapses to the nearest loudspeaker producing a specific sound. This implies that the maximal directional error is of the same order of magnitude with the angular separation of loudspeakers from the listener's viewpoint in pair-wise panning. In practice, when the number of loudspeakers exceeds about eight, the virtual sources are perceived to be in the similar directions in a large listening area.

A basic implementation of 2-D VBAP is included here as a **MATLAB** function.

M-file 5.9 (VBAP2.m)

```
function [gains] = VBAP2(pan_dir)
% function [gains] = VBAP2(pan_dir)
% Author: V. Pulkki
% Computes 2D VBAP gains for horizontal loudspeaker setup.
% Loudspeaker directions in clockwise or counterclockwise order.
% Change these numbers to match with your system.
ls_dirs=[30 -30 -90 -150 150 90];
ls_num=length(ls_dirs);
ls_dirs=[ls_dirs ls_dirs(1)]/180*pi;
% Panning direction in cartesian coordinates.
panvec=[cos(pan_dir/180*pi) sin(pan_dir/180*pi)];
for i=1:ls_num
    % Compute inverse of loudspeaker base matrix.
    lsmat=inv([[cos(ls_dirs(i)) sin(ls_dirs(i))];...
        [cos(ls_dirs(i+1)) sin(ls_dirs(i+1))]]);
    % Compute unnormalized gains
    tempg=panvec*lsmat;
    % If gains nonnegative, normalize the gains and stop
    if min(tempg) > -0.001
        g=zeros(1,ls_num);
        g([i mod(i,ls_num)+1])=tempg;
        gains=g/sqrt(sum(g.^2));
        return
    end
end
```

5.5.3 3-D loudspeaker setups

A 3-D loudspeaker setup denotes here a setup in which all loudspeakers are not in the same plane as the listener. Typically this means that there are some elevated and/or lowered loudspeakers added to a horizontal loudspeaker setup. Triplet-wise panning can be used in such setups [Pul97]. In it, a sound signal is applied to a maximum of three loudspeakers at one time that form a triangle from the listener's viewpoint. If more than three loudspeakers are available, the setup is divided into triangles, one of which is used in the panning of a single virtual source at one time, as shown in Figure 5.6. 3-D vector-base amplitude panning (3-D VBAP) is a method to formulate such setups [Pul97]. It is formulated in an equivalent way to pair-wise panning in the previous section. However, the number of gain factors and loudspeakers is naturally three in the equations. The angle between the median plane and virtual source is estimated correctly with VBAP in most cases, as in pair-wise panning. However, the perceived elevation of a sound source is individual to each subject [Pul01].

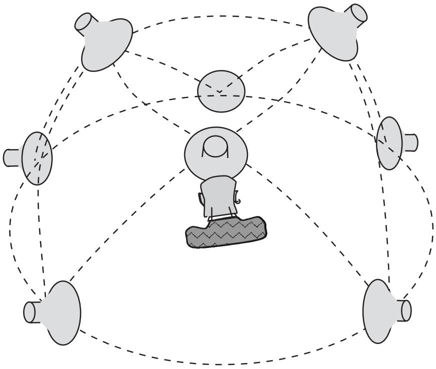


Figure 5.6 A 3-D triangulated loudspeaker system for triplet-wise panning.

A basic implementation of 3-D VBAP for the loudspeaker setup in Figure 5.6 is shown here as a **MATLAB** function.

M-file 5.10 (VBAP3.m)

```
function [gains] = VBAP3(pan_dir)
% function [gains] = VBAP3(pan_dir)
% Author: V. Pulkki
% Computes 3D VBAP gains for loudspeaker setup shown in Fig.6.4
% Change the lousdpeaker directions to match with your system,
% the directions are defined as azimuth elevation; pairs
loudspeakers=[0 0; 50 0; 130 0; -130 0; -50 0;  40 45; 180 45;-40 45];
ls_num=size(loudspeakers,1);
```

```
% Define the triangles from the loudspeakers here
ls_triangles=[1 2 6; 2 3 6; 3 4 7; 4 5 8; 5 1 8; 1 6 8;
    3 6 7; 4 7 8; 6 7 8];
% Go through all triangles
for tripl=1:size(ls_triangles,1)
    ls_tripl=loudspeakers(ls_triangles(tripl,:),:);
    % Panning direction in cartesian coordinates
    cosE=cos(pan_dir(2)/180*pi);
    panvec(1:2)=[cos(pan_dir(1)/180*pi)*cosE sin(pan_dir(1)/180*pi)*cosE];
    panvec(3)=sin(pan_dir(2)/180*pi);
    % Loudspeaker base matrix for current triangle.
    for i=1:3
        cosE=cos(ls_tripl(i,2)/180*pi);
        lsmat(i,1:2)=[cos(ls_tripl(i,1)/180*pi)*cosE...
            sin(ls_tripl(i,1)/180*pi)*cosE];
        lsmat(i,3)=sin(ls_tripl(i,2)/180*pi);
    end
    tempg=panvec*inv(lsmat); % Gain factors for current triangle.
    % If gains nonnegative, normalize g and stop computation
    if min(tempg) > -0.01
        tempg=tempg/sqrt(sum(tempg.^2));
        gains=zeros(1,ls_num);
        gains(1,ls_triangles(tripl,:))=tempg;
        return
    end
end
```

5.5.4 Coincident microphone techniques and Ambisonics

In coincident microphone technologies first- or higher-order microphones positioned ideally in the same position are used to capture sound for multichannel playback. The ambisonics technique [Ger73, ED08] is a form of this. The most common microphone for these applications is the first-order four-capsule B-format microphone, producing a signal $w(t)$ with omnidirectional characteristics, which has been scaled down by $\sqrt{2}$. The B-format microphone also outputs three signals $x(t)$, $y(t)$, and $z(t)$ with figure-of-eight characteristics pointing to corresponding Cartesian directions. The microphones are ideally in the same position. Higher-order microphones have also been proposed and are commercially available, which have much more capsules than the first-order microphones.

In most cases, the microphones are of first order. When the loudspeaker signals are created from the recorded signal, the channels are simply added together with different gains. Thus each loudspeaker signal can be considered as a virtual microphone signal having first-order directional characteristics. This is expressed as

$$s(t) = \frac{2-\kappa}{2}w(t) + \frac{\kappa}{2\sqrt{2}}[\cos(\theta)\cos(\phi)x(t) + \sin(\theta)\cos(\phi)y(t) + \sin(\phi)z(t)], \qquad (5.14)$$

where $s(t)$ is the produced virtual microphone signal having an orientation of azimuth θ and elevation ϕ. The parameter $\kappa \in [0, 2]$ defines the directional characteristics of the virtual microphone from omnidirectional to cardioid and dipole, as shown in Figure 5.7.

In multichannel reproduction of such first-order B-format recordings, a virtual microphone signal is computed for each loudspeaker. In practice such methods provide good quality only in certain loudspeaker configurations [Sol08] and at frequencies well below 1 kHz. At higher frequencies the high coherence between the loudspeaker signals, which is caused by the broad

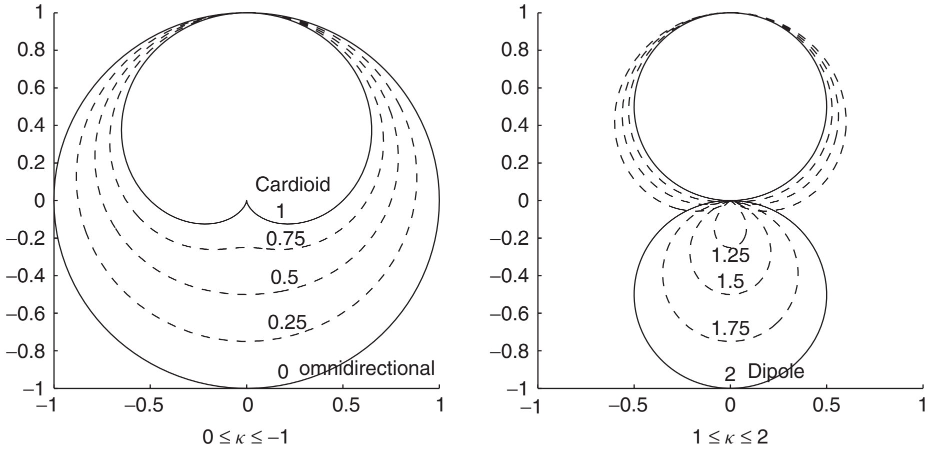


Figure 5.7 The directional pattern of virtual microphones which can be generated from first-order B-format recordings. Figure reprinted with permission from [Vil08].

directional patterns of the microphones, leads to undesired effects such as coloration and loss of spaciousness. Using higher-order microphones these problems are less severe, though some problems with microphone self-noise may appear.

The decoding of first-order B-format microphone signals to a horizontal loudspeaker layout with directional characteristics controllable by κ is shown in the following example:

M-file 5.11 (kappa.m)

```
% kappa.m
% Author: V. Pulkki
% Simple example of cardioid decoding of B-format signals
Fs=44100;
% mono signal
signal=(mod([0:Fs*2],220)/220);
% Simulated horizontal-only B-format recording of single
% sound source in direction of theta azimuth.
% This can be replaced with a real B-format recording.
theta=0;
w=signal/sqrt(2);
x=signal*cos(theta/180*pi);
y=signal*sin(theta/180*pi);
% Virtual microphone directions
% (In many cases the values equal to the directions of loudspeakers)
ls_dir=[30 90 150 -150 -90 -30]/180*pi;
ls_num=length(ls_dir);
% Compute virtual cardioids (kappa = 1) out of the B-format signal
kappa=1;
for i=1:ls_num
    LSsignal(:,i)=(2-kappa)/2*w...
        +kappa/(2*sqrt(2))*(cos(ls_dir(i))*x+sin(ls_dir(i))*y);
end
% File output
wavwrite(LSsignal,Fs,16,'firstorderB-formatexample.wav')
```

The previous example was of reproducing a recorded sound scenario. Such coincident recording can also be simulated to perform synthesis of spatial audio [MM95] for 2-D or 3-D loudspeaker setups. In this case it is an amplitude-panning method in which a sound signal is applied to all loudspeakers placed evenly around the listener with gain factors

$$g_i = \frac{1}{N}\sum_{m=1}^{M}\{1 + 2p_m \cos(m\alpha_i)\}, \tag{5.15}$$

where g_i is the gain of ith speaker, N is the number of loudspeakers, α is the angle between loudspeaker and panning direction, $\cos(m\alpha_i)$ represents a single spherical harmonic with order m, M is the order of Ambisonics, and p_m are the gains for each spherical harmonic [DNM03, Mon00]. When the order is low, the sound signal emanates from all the loudspeakers, which causes some spatial artifacts due to unnatural behavior of binaural cues [PH05]. In such cases, when listening outside the best listening position, the sound is also perceived at the nearest loudspeaker which produces the sound. This effect is more prominent with first-order ambisonics than with pair- or triplet-wise panning, since in ambisonics virtually all loudspeakers produce the same sound signal. The sound is also colored, for the same reason, i.e., multiple propagation paths of the same signal to ears produce comb-filter effects. Conventional microphones can be used to realize first-order ambisonics.

When the order is increased, the cross-talk between loudspeakers can be minimized by optimizing gains of spherical harmonics for each loudspeaker in a listening setup [Cra03]. Using higher-order spatial harmonics increases both directional and timbral virtual source quality, since the loudspeaker signals are less coherent. The physical wave field reconstruction is then more accurate, and different curvatures of wavefronts, as well as planar wavefronts can be produced [DNM03], if a large enough number of loudspeakers is in use. The selection of the coefficients for different spherical harmonics has to be done carefully for each loudspeaker layout.

A simple implementation of computing second-order ambisonic gains for a hexagonal loudspeaker layout is shown here:

M-file 5.12 (ambisonics.m)

```
% ambisonics.m
% Author: V. Pulkki
% Second-order harmonics to compute gains for loudspeakers
% to position virtual source to a loudspeaker setup
theta=30;% Direction of virtual source
loudsp_dir=[30 -30 -90 -150 150 90]/180*pi; % loudspeaker setup
ls_num=length(loudsp_dir);
harmC=[1 2/3 1/6]; % Coeffs for harmonics "smooth solution", [Mon00]
theta=theta/180*pi;
for i=1:ls_num
    g(i)= (harmC(1) + 2*cos(theta-loudsp_dir(i))*harmC(2) +...
        2*cos(2*(theta-loudsp_dir(i)))*harmC(3));
end
% use gains in g for amplitude panning
```

5.5.5 Synthesizing the width of virtual sources

A loudspeaker layout having loudspeakers around the listener can also be used to control the width of the sound source, or even to produce an enveloping perception of the sound source. A simple demonstration can be made by playing back pink noise with all loudspeakers, where the noises are independent of each other [Bla97]. The sound source is then perceived to surround the listener totally.

The precise control of the width of the virtual source is a complicated topic. However, simple and effective spatial effects can be performed by using some of the loudspeakers to generate a wide virtual source. The example of decorrelating a monophonic input signal for stereophonic listening shown in Section 5.3.2 can be also used with multichannel loudspeaker setups. In such cases, the number of columns in the noise matrix corresponds to the number of loudspeakers to which the decorrelated sound is applied, and the convolution has to be computed for each loudspeaker channel. The virtual source will be more or less perceived to originate evenly from the loudspeakers and from the space between them. Unfortunately, the drawback of time-smearing of transients is also present in this case.

Note that this is a different effect than the effect generated with reverberators, which are discussed in Section 5.6, although the processing is very similar. The reverberators simulate the room effect, and try to create the perception of a surrounding reverberant tail, and in contrast the effect discussed in this section generates the perception of a surrounding non-reverberant sound. However, such effects cannot be achieved with all types of signals, due to human perceptual capabilities, as discussed briefly in Section 5.2.3.

5.5.6 Time delay-based systems

There are a number of methods proposed, which instead of using amplitude differences use time-delays in positioning the virtual sources. The most complete one, the wave field synthesis is presented first, after which Moore's room-inside-room approach is discussed.

Wave-field synthesis is a technique that requires a large number of carefully equalized loudspeakers [BVV93, VB99]. It aims to reconstruct the whole sound field in a listening room. When a virtual source is reproduced, the sound for each loudspeaker is delayed and amplified in a manner that a desired circular or planar sound wave occurs as a superposition of sounds from each loudspeaker. The virtual source can be positioned far behind the loudspeakers, or in some cases even in the space inside the loudspeaker array, as shown in Figure 5.8. The loudspeaker signals have to be equalized depending on virtual source position [ARB04].

Theoretically the wave-field synthesis is superior as a technique, as the perceived position of the sound source is correct within a very large listening area. Unfortunately, to create a desired wave field in the total area inside the array, it demands that the loudspeakers are at a distance of maximally a half wavelength from each other. The area in which a perfect sound field synthesis is achieved

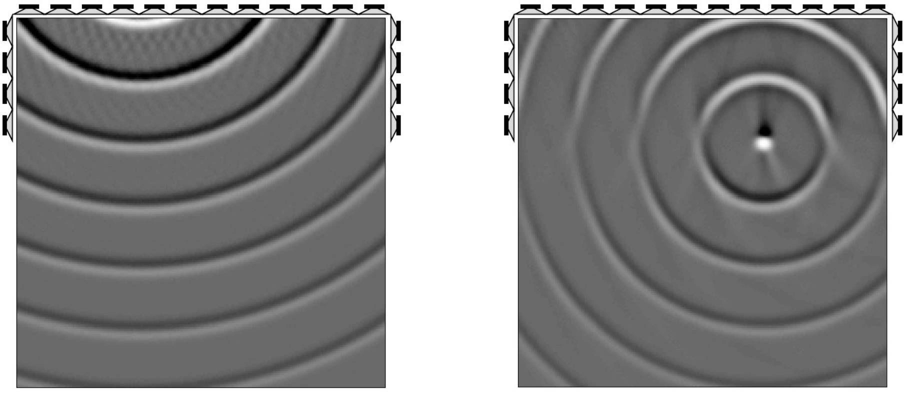


Figure 5.8 Wave-field synthesis concept. A desired 2-D sound field can be constructed using a large loudspeaker array. Figure reprinted with IEEE permission from [VB99].

shrinks with increasing frequency [DNM03]. In practice, due to the perception mechanisms of humans, more error can be allowed above approximately 1.5 kHz. Arrays for wave-field synthesis have been built for room acoustics control and enhancement to be used in theaters and multipurpose auditoria [VB99].

Time delays also can be used when creating spatial effects with loudspeaker setups with fewer loudspeakers than used in wave-field synthesis. The use of time delays for positioning virtual sources was considered for a two-channel stereophonic layout in Section 5.3.2. It was noted that short delays of a few milliseconds make the virtual sources perceived as spatially spread. This effect of course also applies with multi-loudspeaker setups. With longer delays, the precedence effect makes the sound direction perceived to be form the loudspeaker where the sound arrived first. This technique has also been used for multichannel loudspeaker layouts, as in Moore's approach, where the relative time delay between the loudspeaker feeds is controlled. A model supporting this approach was introduced by Moore [Moo83], and can be described as a physical and geometric model. The metaphor underlying the Moore model is that of a room-within-a-room, where the inner room has holes in the walls, corresponding to the positions of the loudspeakers, and the outer room is the virtual room where the sound events take place. When a single virtual source is applied in the virtual world, all the loudspeakers emanate the same sound with different amplitudes and delays. This is a similar idea to wave-field synthesis, though the mathematical basis is less profound, and the number of loudspeakers utilized is typically smaller. The recreated sound field will be limited to low frequencies, and at higher frequencies comb-filter effects and unstable directions for virtual sources will appear. However, if the delays are large enough, the virtual source will be perceived at all listening positions to the loudspeaker which emanates the sound first. The effects introduced by the Moore model are directly related with the artifacts one would get when listening to the outside world through windows.

5.5.7 Time-frequency processing of spatial audio

The directional resolution of spatial hearing is limited within auditory frequency bands [Bla97]. In principle, all sound within one critical band can be only perceived as a single source with broader or narrower extent. In some special cases a binaural narrow-band sound stimulus can be perceived as two distinct auditory objects, but the perception of three or more concurrent sources is generally not possible. This is different from visual perception, where already one eye can detect the directions of a large number of visual objects sharing the same color.

The limitations of spatial auditory perception imply that the spatial realism needed in visual reproduction is not needed in audio. In other words, the spatial accuracy in reproduction of acoustical wave field can be compromised without decrease in perceptual quality. There are some recent technologies which exploit this assumption. Methods to compress multichannel audio files to 1–2 audio signals with metadata have been proposed [BF08, GJ06], where the level and time differences between the channels are analyzed in the time-frequency domain, and are coded as metadata to the signal. A related technology for spatial audio, directional audio coding (DirAC) [Pul07], is a signal-processing method for spatial sound, which can be applied to spatial sound reproduction for any multichannel loudspeaker layout, or for headphones. The other applications suggested for it include teleconferencing and perceptual audio coding.

In DirAC, it is assumed that at one time instant and at one critical band the spatial resolution of auditory system is limited to decoding one cue for direction and another for inter-aural coherence. It is further assumed that if the direction and diffuseness of sound field is measured and reproduced correctly, a human listener will perceive the directional and coherence cues correctly.

The concept of DirAC is illustrated in Figure 5.9. In the analysis phase, the direction and diffuseness of the sound field is estimated in auditory frequency bands depending on time, forming the metadata transmitted together with a few audio channels. In the "low-bitrate" approach shown in the figure, only one channel of audio is transmitted. The audio channel may also be further compressed to obtain a lower transmission data rate. The version with more channels is shown

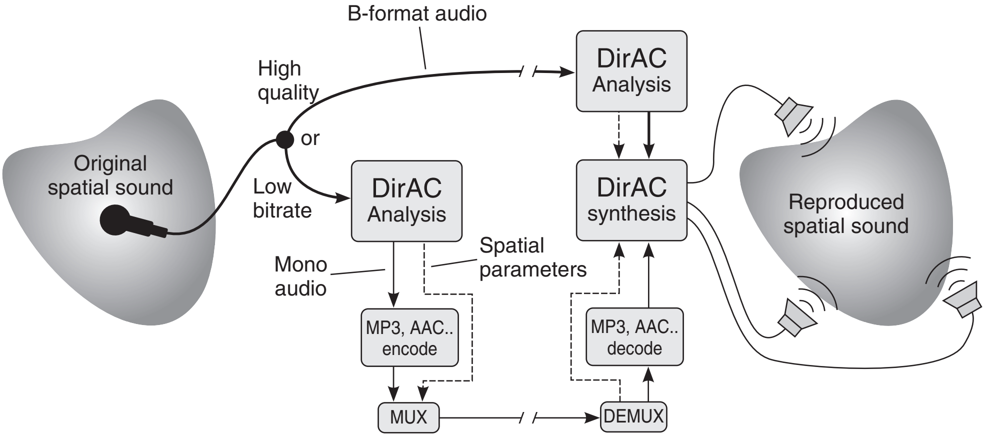


Figure 5.9 Two typical approaches of DirAC: High quality (top) and low bitrate (bottom). Figure reprinted with AES permission from [Vil08].

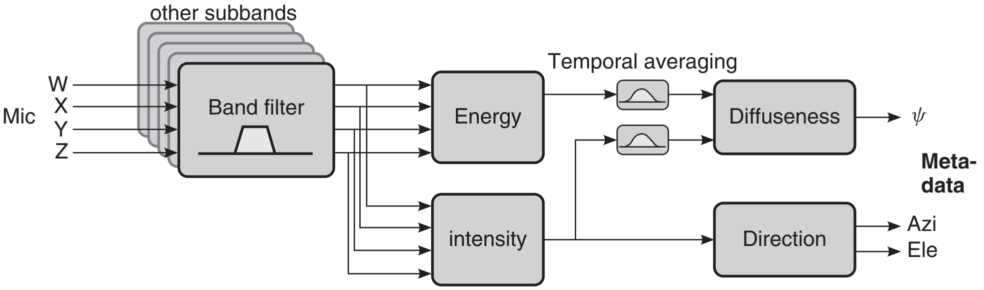


Figure 5.10 DirAC analysis. Figure reprinted with AES permission from [Vil08].

as a "high-quality version," where the number of transmitted channels is three for horizontal reproduction, and four for 3-D reproduction. In the high-quality version the analysis may be conducted at the receiving end. It has been proven that the high-quality version of DirAC produces better perceptual quality in loudspeaker listening than other available techniques using the same microphone input [VLP09].

The processing in DirAC and in other time-frequency-domain processing for spatial audio is typically organized as an analysis-transmission-synthesis chain. To give an idea of the analysis process, the analysis part of DirAC is now discussed in detail, followed by a code example.

The aim of directional analysis, which is shown in Figure 5.10, is to estimate at each frequency band the direction of arrival of sound, together with an estimate of if the sound is arriving from one or multiple directions at the same time. In principle this can be performed with a number of techniques, however, an energetic analysis of the sound field has been found to be suitable, which is shown in Figure 5.10. The energetic analysis can be performed when the pressure signal and velocity signals in one, two or three dimensions are captured from a single position.

The first-order B-format signals described in Section 5.5.4 can be used for directional analysis. The sound pressure can be estimated using the omnidirectional signal $w(t)$ as $P = \sqrt{2}W$, expressed in the STFT domain. The figure-of-eight signals $x(t)$, $y(t)$, and $z(t)$ are grouped in the STFT domain into a vector $\mathbf{U} = [X, Y, Z]$, which estimates the 3-D sound field velocity vector. The energy E of sound field can be computed as

$$E = \frac{\rho_0}{4}||\mathbf{U}||^2 + \frac{1}{4\rho_0 c^2}|P|^2, \tag{5.16}$$

where ρ_0 is the mean density of air, and c is the speed of sound. The capturing of B-format signals can be obtained with either coincident positioning of directional microphones, or with closely spaced set of omnidirectional microphones. In some applications, the microphone signals may be formed in the computational domain, i.e., simulated. The analysis is repeated as frequently as is needed for the application, typically with an update frequency of 100–1000 Hz.

The intensity vector **I** expresses the net flow of sound energy as a 3-D vector, and can be computed as

$$\mathbf{I} = \overline{P}\mathbf{U}, \tag{5.17}$$

where $\overline{(\cdot)}$ denotes complex conjugation. The direction of sound is defined as the opposite direction to the intensity vector at each frequency band. The direction is denoted as the corresponding angular azimuth and elevation values in the transmitted metadata. The diffuseness of the sound field is computed as

$$\psi = 1 - \frac{\|\mathrm{E}\{\mathbf{I}\}\|}{c\mathrm{E}\{E\}}, \tag{5.18}$$

where E is the expectation operator. Typically the expectation operator is implemented with temporal integration, as in the example below. This process is also called “smoothing.” The outcome of this equation is a real-valued number between zero and one, characterizing whether the sound energy is arriving from a single direction, or from all directions.

An example of directional analysis of B-format signals is presented in the following. The synthesis of sound is not shown here, however the reader is encouraged to use the analysis results to build different spatial effects of his own.

M-file 5.13 (diranalysis.m)

```
% diranalysis.m
% Author: V. Pulkki
% Example of directional analysis of simulated B-format recording
Fs=44100; % Generate signals
sig1=2*(mod([1:Fs]’,40)/80-0.5) .* min(1,max(0,(mod([1:Fs]’,Fs/5)-Fs/10)));
sig2=2*(mod([1:Fs]’,32)/72-0.5) .* min(1,max(0,(mod([[1:Fs]+Fs/6]’,...
Fs/3)-Fs/6)));
% Simulate two sources in directions of -45 and 30 degrees
w=(sig1+sig2)/sqrt(2);
x=sig1*cos(50/180*pi)+sig2*cos(-170/180*pi);
y=sig1*sin(50/180*pi)+sig2*sin(-170/180*pi);
% Add fading in diffuse noise with  36 sources evenly in the horizontal plane
for dir=0:10:350
    noise=(rand(Fs,1)-0.5).*(10.^((([1:Fs]’/Fs)-1)*2));
    w=w+noise/sqrt(2);
    x=x+noise*cos(dir/180*pi);
    y=y+noise*sin(dir/180*pi);
end
hopsize=256; % Do directional analysis with STFT
winsize=512; i=2; alpha=1./(0.02*Fs/winsize);
Intens=zeros(hopsize,2)+eps; Energy=zeros(hopsize,2)+eps;
for time=1:hopsize:(length(x)-winsize)
```

```
    % moving to frequency domain
    W=fft(w(time:(time+winsize-1)).*hanning(winsize));
    X=fft(x(time:(time+winsize-1)).*hanning(winsize));
    Y=fft(y(time:(time+winsize-1)).*hanning(winsize));
    W=W(1:hopsize);X=X(1:hopsize);Y=Y(1:hopsize);
    %Intensity computation
    tempInt = real(conj(W) * [1 1 ] .* [X Y])/sqrt(2);%Instantaneous
    Intens = tempInt * alpha + Intens * (1 - alpha); %Smoothed
    % Compute direction from intensity vector
    Azimuth(:,i) = round(atan2(Intens(:,2), Intens(:,1))*(180/pi));
    %Energy computation
    tempEn=0.5 * (sum(abs([X Y]).^2, 2) * 0.5 + abs(W).^2 + eps);%Inst
    Energy(:,i) = tempEn*alpha + Energy(:,(i-1)) * (1-alpha); %Smoothed
    %Diffuseness computation
    Diffuseness(:,i) = 1 - sqrt(sum(Intens.^2,2)) ./ (Energy(:,i));
    i=i+1;
end
% Plot variables
figure(1); imagesc(log(Energy)); title('Energy');set(gca,'YDir','normal')
xlabel('Time frame'); ylabel('Freq bin');
figure(2); imagesc(Azimuth);colorbar; set(gca,'YDir','normal')
title('Azimuth'); xlabel('Time frame'); ylabel('Freq bin');
figure(3); imagesc(Diffuseness);colorbar; set(gca,'YDir','normal')
title('Diffuseness'); xlabel('Time frame'); ylabel('Freq bin');
```

5.6 Reverberation

5.6.1 Basics of room acoustics

Most of the techniques presented in previous sections have concentrated on reproducing one sound source in a free field from a certain direction. However, in a real space reverberation is always present. Reverberation is composed of reflections which are delayed and attenuated copies of the direct sound. The frequency content of each reflection is also modified due to the directivity of the sound source and due to the material absorption of reflecting surfaces.

The most important concept in room acoustics is an impulse response which describes the acoustics of a room from a static sound source to a listening position. Engineers often divide the impulse response into three parts, the direct sound, early reflections and late reverberation. This division is illustrated with a simulated impulse response in Figure 5.11. The direct sound is the sound reaching the listener first. The early reflections are the first reflections that are not perceived separately as human hearing integrates them with the direct sound. Although their individual directions are not perceived due to the precedence effect, they contribute to the perception of the sound color and the size of the sound source, as well as the size of the room. Late reverberation is considered after a time moment, which is sometimes called the mixing time, when the reflection density is so high that individual reflections cannot be seen in the response. Late reverberation gives cues of the size of the room as well as the distance of the sound source.

The room impulse response can be measured [MM01] or modeled with room acoustics modeling techniques [SK02, Sil10]. Section 5.7 presents various ways to create the impulse response, but before that the basic method to add the room effect to a sound signal is presented.

5.6.2 Convolution with room impulse responses

If the impulse response of a target room is readily available, the most faithful reverberation method would be to convolve the input signal with such a response. Direct convolution can be done by

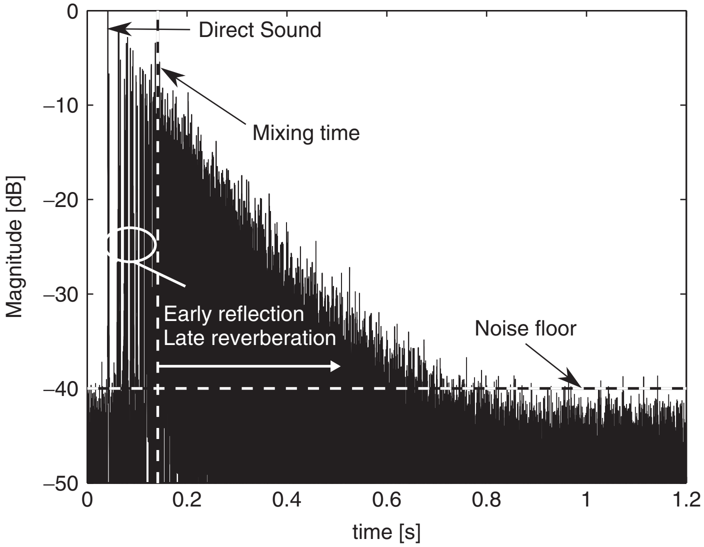


Figure 5.11 Simulated room energy response illustrating the direct sound, early reflections and late reverberation.

storing each sample of the impulse response as a coefficient of an FIR filter whose input is the dry signal. Direct convolution becomes easily impractical if the length of the target response exceeds small fractions of a second, as it would translate into several hundreds of taps in the filter structure. A solution is to perform the convolution block by block in the frequency domaing given the Fourier transform of the impulse response, and the Fourier transform of a block of input signal, the two can be multiplied point by point and the result transformed back to the time domain. As this kind of processing is performed on successive blocks of the input signal, the output signal is obtained by overlapping and adding the partial results [OS89]. Thanks to the FFT computation of the discrete Fourier transform, such a technique can be significantly faster. A drawback is that, in order to be operated in real time, a block of N samples must be read and then processed while a second block is being read. Therefore, the input–output latency in samples is twice the size of a block, and this is not tolerable in practical real-time environments.

The complexity–latency tradeoff is illustrated in Figure 5.12, where the direct-form and the block-processing solutions can be located, together with a third efficient yet low-latency solution [Gar95, MT99]. This third realization of convolution is based on a decomposition of the

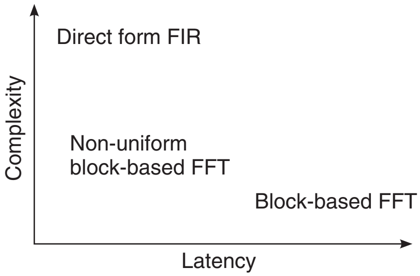


Figure 5.12 Complexity vs. latency tradeoff in convolution.

impulse response into increasingly large chunks. The size of each chunk is twice the size of its predecessor, so that the latency of prior computation can be occupied by the computations related to the following impulse response chunk. Details and discussion on convolution were presented in Section 2.2.7.

When the reproduction is monaural the convolution is a trivial task, but when spatial sound reproduction is applied the process is more complicated. In real life the reflections reach the listening position (or a microphone) from all directions and the reflections can be considered to be virtual sources (discussed in previous sections). Thus, each reflection should be reproduced from the correct direction, which means that the impulse response has to contain such information. Currently, the most common technique is to measure or model the first or higher-order B-format response that is convolved with source signals. Then the reproduction with multi loudspeaker setups is performed with spatial-sound rendering techniques (see Sections 5.5.4 and 5.5.7).

5.7 Modeling of room acoustics

The room effect, i.e., the room impulse response, can be created by modeling how sound propagates and reflects from surfaces if the geometry of a room is available. Such a process is called room acoustics modeling and several techniques are discussed in Section 5.7.4. In many cases the geometry of a room is not needed since an artificial room impulse response can be created from a perceptual point of view. In fact, the human hearing is not very sensitive to details in the reverberant tail and any decaying response can be used as an effect.

Even if we have enough computer power to compute convolutions by long impulse responses in real time, there are still reasons to prefer reverberation algorithms based on feedback delay networks in many practical contexts. The reasons are similar to those that make a CAD description of a scene preferable to a still picture whenever several views have to be extracted or the environment has to be modified interactively. In fact, it is not easy to modify a room impulse response to reflect some of the room attributes, e.g., its high-frequency absorption. If the impulse response is coming from a room acoustics modeling algorithm, these manipulations can be operated at the level of room description, and the coefficients of the room impulse response are transmitted to the real-time convolver. In low-latency block-based implementations, we can even have faster update rates for the smaller early chunks of the impulse response, and slower update rates for the reverberant tail. Still, continuous variations of the room impulse response are easier to render using a model of reverberation operating on a sample-by-sample basis. For this purpose dozens of reverberation algorithms have been developed and in the following some of them are introduced in more detail.

5.7.1 Classic reverb tools

In the second half of the twentieth century, several engineers and acousticians tried to invent electronic devices capable of simulating the long-term effects of sound propagation in enclosures. The most important pioneering work in the field of *artificial reverberation* has been that of Manfred Schroeder at the Bell Laboratories in the early sixties [Sch61, Sch62, Sch70, Sch73, SL61]. Schroeder introduced the recursive *comb filters* and the delay-based *allpass filters* as computational structures suitable for the inexpensive simulation of complex patterns of echoes. In particular, the allpass filter based on the recursive delay line has the form

$$y(n) = -g \cdot x(n) + x(n-m) + g \cdot y(n-m) \;, \tag{5.19}$$

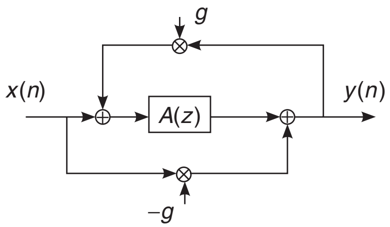


Figure 5.13 The allpass filter structure.

where m is the length of the delay in samples. The filter structure is depicted in Figure 5.13, where $A(z)$ is usually replaced by a delay line. This filter allows one to obtain a dense impulse response and a flat frequency response. Such a structure became rapidly a standard component used in almost all the artificial reverberators designed up to now [Moo79]. It is usually assumed that the allpass filters do not introduce coloration in the input sound. However, this assumption is valid from a perceptual viewpoint only if the delay line is much shorter than the integration time of the ear, i.e., about 50 ms [ZF90]. If this is not the case, the time-domain effects become much more relevant and the timbre of the incoming signal is significantly affected.

In the seventies, Michael Gerzon generalized the single-input single-output allpass filter to a multi-input multi-output structure, where the delay line of m samples has been replaced by an order-N unitary network [Ger76]. Examples of trivial unitary networks are orthogonal matrices, and parallel connections of delay lines or allpass filters. The idea behind this generalization is that of increasing the complexity of the impulse response without introducing appreciable coloration in frequency. According to Gerzon's generalization, allpass filters can be nested within allpass structures, in a telescopic fashion. Such embedding is shown to be equivalent to lattice allpass structures [Gar97b], and it is realizable as long as there is at least one delay element in the block $A(z)$ of Figure 5.13. An example **MATLAB** code with a delay of 40 samples is:

M-file 5.14 `(comballpass.m)`

```
% Author: T. Lokki
% Create an impulse
x = zeros(1,2500); x(1) = 1;
% Delay line and read position
A = zeros(1,100);
Adelay=40;
% Output vector
ir = zeros(1,2500);
% Feedback gain
g=0.7;
% Comb-allpass filtering
for n = 1:length(ir)
    tmp = A(Adelay) + x(n)*(-g);
    A = [(tmp*g + x(n))’ A(1:length(A)-1)];
    ir(n) = tmp;
end
% Plot the filtering result
plot(ir)
```

Extensive experimentation on structures for artificial reverberation was conducted by Moorer in the late seventies [Moo79]. He extended the work done by Schroeder [Sch70] in relating some basic computational structures (e.g., tapped delay lines, comb and allpass filters) with the physical behavior of actual rooms. In particular, it was noticed that the early reflections have great importance in the perception of the acoustic space, and that a direct-form FIR filter can reproduce these early reflections explicitly and accurately. Usually this FIR filter is implemented as a tapped delay line, i.e., a delay line with multiple reading points that are weighted and summed together to provide a single output. This output signal feeds, in Moorer's architecture, a series of allpass filters and parallel comb filters. Another improvement introduced by Moorer was the replacement of the simple gain of feedback delay lines in comb filters with lowpass filters resembling the effects of air absorption and lossy reflections.

An original approach to reverberation was taken by Julius Smith in 1985, when he proposed *digital waveguide networks* (*DWNs*) as a viable starting point for the design of numerical reverberators [Smi85]. The idea of *waveguide reverberators* is that of building a network of waveguide branches (i.e., bidirectional delay lines simulating wave propagation in a duct or a string) capable of producing the desired early reflections and a diffuse, sufficiently dense reverb. If the network is augmented with lowpass filters it is possible to shape the decay time with frequency. In other words, waveguide reverberators are built in two steps: the first step is the construction of a prototype lossless network, the second step is the introduction of the desired amount of losses. This procedure ensures good numerical properties and good control over stability [Smi86, Vai93]. In ideal terms, the quality of a prototype lossless reverberator is evaluated with respect to the whiteness and smoothness of the noise that is generated in response to an impulse. The fine control of decay time at different frequencies is decoupled from the structural aspects of the reverberator.

Among the classic reverberation tools we should also mention the structures proposed by Stautner and Puckette [SP82], and by Jot [Jot92]. These structures form the basis of feedback delay networks, which are discussed in detail in Section 5.7.2.

Clusters of comb/allpass filters

The construction of high-quality reverberators is half an art and half a science. Several structures and many parameterizations have been proposed in the past, especially in non-disclosed form within commercial reverb units [Dat97]. In most cases, the various structures are combinations of comb and allpass elementary blocks, as suggested by Schroeder in the early work. As an example, we briefly describe Moorer's preferred structure [Moo79], depicted in Figure 5.14. The block (a) of Moorer's reverb takes care of the early reflections by means of a tapped delay line. The resulting signal is forwarded to the block (b), which is the parallel of a direct path on one branch, and a delayed, attenuated diffuse reverberator on the other branch. The output of the reverberator is delayed in such a way that the last of the early echoes coming out of block (a) reaches the output before the first of the non-null samples coming out of the diffuse reverberator. In Moorer's preferred implementation, the reverberator of block (b) is best implemented as a parallel of six comb filters, each with a first-order lowpass filter in the loop, and a single allpass filter. In [Moo79], it is suggested to set the allpass delay length to 6 ms and the allpass coefficient to 0.7. Despite the fact that any allpass filter does not add coloration in the magnitude frequency response, its time response can give a metallic character to the sound, or add some unwanted roughness and granularity. The feedback attenuation coefficients and the lowpass filters of the comb filters can be tuned to resemble a realistic and smooth decay. In particular, the attenuation coefficients g_i determine the overall decay time of the series of echoes generated by each comb filter. If the desired decay time (usually defined for an attenuation level of 60 dB) is T_d, the gain of each comb filter has to be set to

$$g_i = 10^{-3\frac{T_d F_s}{m_i}} , \tag{5.20}$$

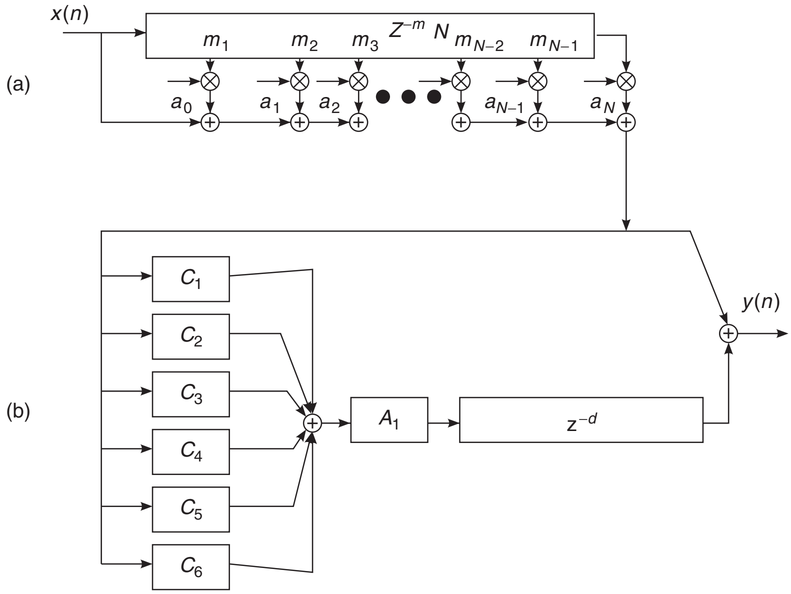


Figure 5.14 Moorer's reverberator.

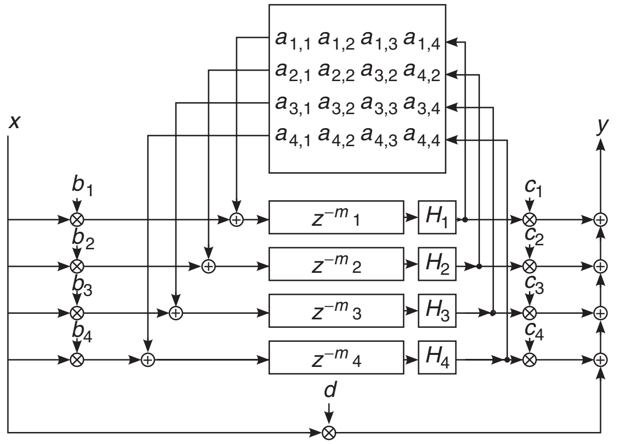
where F_s is the sample rate and m_i is the delay length in the samples. Further attenuation at high frequencies is provided by the feedback lowpass filters, whose coefficient can also be related to decay time at a specific frequency or fine tuned by direct experimentation. In [Moo79], an example set of feedback attenuation and allpass coefficients is provided, together with some suggested values of the delay lengths of the comb filters. As a general rule, they should be distributed over a ratio 1:1.5 between 50 and 80 ms. Schroeder suggested a number-theoretic criterion for a more precise choice of the delay lengths [Sch73]: the lengths in samples should be mutually coprime (or incommensurate) to reduce the superimposition of echoes in the impulse response, thus reducing the so-called flutter echoes. This same criterion might be applied to the distances between each echo and the direct sound in early reflections. However, as was noticed by Moorer [Moo79], the results are usually better if the taps are positioned according to the reflections computed by means of some geometric modeling technique, such as the image method. As is explained next, even the lengths of the recirculating delays can be computed from the geometric analysis of the normal modes of actual room shapes.

5.7.2 Feedback delay networks

In 1982, J. Stautner and M. Puckette [SP82] introduced a structure for artificial reverberation based on delay lines interconnected in a feedback loop by means of a matrix (see Figure 5.15). Later, structures such as this have been called *feedback delay networks* (*FDN*s). The Stautner–Puckette FDN was obtained as a vector generalization of the recursive comb filter

$$y(n) = x(n-m) + g \cdot y(n-m) \ , \tag{5.21}$$

where the m-sample delay line was replaced by a bunch of delay lines of different lengths, and the feedback gain g was replaced by a feedback matrix $\mathbf{G}$. Stautner and Puckette proposed the



$a_{1,1}$ $a_{1,2}$ $a_{1,3}$ $a_{1,4}$
$a_{2,1}$ $a_{2,2}$ $a_{3,2}$ $a_{4,2}$
$a_{3,1}$ $a_{3,2}$ $a_{3,3}$ $a_{3,4}$
$a_{4,1}$ $a_{4,2}$ $a_{4,3}$ $a_{4,4}$
x
y
b_1
b_2
b_3
b_4
z^{-m} 1
z^{-m} 2
z^{-m} 3
z^{-m} 4
H_1
H_2
H_3
H_4
c_1
c_2
c_3
c_4
d

Figure 5.15 Fourth-order feedback delay network.

following feedback matrix:

$$\mathbf{G} = g \begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix} /\sqrt{2}\,. \tag{5.22}$$

Due to its sparse special structure, **G** requires only one multiple per output channel.

An example of FDN without lowpass filters H_n is:

M-file 5.15 (delaynetwork.m)

```
% delaynetwork.m
% Author: T. Lokki
fs=44100;
gain=0.97;
% Create an impulse
x = zeros(1,1*fs); x(1) = 1;
y = zeros(1,fs);
b = [1 1 1 1];
c = [0.8 0.8 0.8 0.8];
% Feedback matrix
a(1,:)=[0 1 1 0];
a(2,:)=[-1 0 0 -1];
a(3,:)=[1 0 0 -1];
a(4,:)=[0 1 -1 0];
a2=a*(1/sqrt(2)) * gain;
% Delay lines, use prime numbers
m=[149 211 263 293]’;
z1=zeros(1,max(max(m)));
z2=zeros(1,max(max(m)));
z3=zeros(1,max(max(m)));
z4=zeros(1,max(max(m)));
```

```
for n = 1:length(y)
    tmp = [z1(m(1)) z2(m(2)) z3(m(3)) z4(m(4))];
    y(n) = x(n) + c(1)*z1(m(1)) + c(2)*z2(m(2)) ...
           + c(3)*z3(m(3)) + c(4)*z4(m(4));
    z1 = [(x(n)*b(1) + tmp*a2(1,:)') z1(1:length(z1)-1)];
    z2 = [(x(n)*b(2) + tmp*a2(2,:)') z2(1:length(z2)-1)];
    z3 = [(x(n)*b(3) + tmp*a2(3,:)') z3(1:length(z3)-1)];
    z4 = [(x(n)*b(4) + tmp*a2(4,:)') z4(1:length(z4)-1)];
end
plot(y)
```

More recently, Jean-Marc Jot has investigated the possibilities of FDNs very thoroughly. He proposed to use some classes of unitary matrices allowing efficient implementation. Moreover, he showed how to control the positions of the poles of the structure in order to impose a desired decay time at various frequencies [Jot92]. His considerations were driven by perceptual criteria and the general goal was to obtain an ideal diffuse reverb. In this context, Jot introduced the important design criterion that all the modes of a frequency neighborhood should decay at the same rate, in order to avoid the persistence of isolated, ringing resonances in the tail of the reverb [JC91]. This is not what happens in real rooms though, where different modes of close resonance frequencies can be differently affected by wall absorption [Mor91]. However, it is generally believed that the slow variation of decay rates with frequency produces smooth and pleasant impulse responses.

General structure

Referring to Figure 5.15, an *FDN* is built starting from N delay lines, each being $\tau_i = m_i T_s$ seconds long, where $T_s = 1/F_s$ is the sampling interval. The FDN is completely described by the following equations:

$$\begin{aligned} y(n) &= \sum_{i=1}^{N} c_i s_i(n) + d x(n) \\ s_i(n+m_i) &= \sum_{j=1}^{N} a_{i,j} s_j(n) + b_i x(n), \end{aligned} \tag{5.23}$$

where $s_i(n)$, $1 \le i \le N$, are the delay outputs at the nth time sample. If $m_i = 1$ for every i, we obtain the well-known *state space description* of a discrete-time linear system [Kai80]. In the case of FDNs, m_i are typically numbers on the orders of hundreds or thousands, and the variables $s_i(n)$ are only a small subset of the system state at time n, being the whole state represented by the content of all the delay lines.

From the state-variable description of the FDN it is possible to find the system transfer function [Roc96, RS97] as

$$H(z) = \frac{Y(z)}{X(z)} = \mathbf{c}^T [\mathbf{D} z^{-1} - \mathbf{A}]^{-1} \mathbf{b} + d. \tag{5.24}$$

The diagonal matrix $\mathbf{D}(z) = \mathrm{diag}\left(z^{-m_1}, z^{-m_2}, \ldots z^{-m_N}\right)$ is called the *delay matrix*, and $\mathbf{A} = [a_{i,j}]_{N\times N}$ is called the *feedback matrix*.

The stability properties of a FDN are all ascribed to the feedback matrix. The fact that $\|A\|^n$ decays exponentially with n ensures that the whole structure is stable [Roc96, RS97].

The poles of the FDN are found as the solutions of

$$\det[\mathbf{A} - \mathbf{D}z^{-1}] = 0\ . \tag{5.25}$$

In order to have all the poles on the unit circle it is sufficient to choose a unitary matrix. This choice leads to the construction of a *lossless prototype*, but this is not the only choice allowed.

The zeros of the transfer function can also be found [Roc96, RS97] as the solutions of

$$\det\left[\mathbf{A} - \mathbf{b}\frac{1}{d}\mathbf{c}^T - \mathbf{D}z^{-1}\right] = 0\ . \tag{5.26}$$

In practice, once we have constructed a lossless FDN prototype, we must insert attenuation coefficients and filters in the feedback loop. For instance, following the indications of Jot [JC91], we can cascade every delay line with a gain

$$g_i = \alpha^{m_i}\ . \tag{5.27}$$

This corresponds to replacing $D(z)$ with $D(z/\alpha)$ in (5.24). With this choice of the attenuation coefficients, all the poles are contracted by the same factor α. As a consequence, all the modes decay with the same rate, and the reverberation time (defined for a level attenuation of 60 dB) is given by

$$T_d = \frac{-3T_s}{\log\alpha}\ . \tag{5.28}$$

In order to have a faster decay at higher frequencies, as happens in real enclosures, we must cascade the delay lines with lowpass filters. If the attenuation coefficients g_i are replaced by lowpass filters, we can still get a local smoothness of decay times at various frequencies by satisfying the condition (5.27), where g_i and α have been made frequency dependent:

$$G_i(z) = A^{m_i}(z), \tag{5.29}$$

where $A(z)$ can be interpreted as per-sample filtering [JSer, JC91, Smi92].

It is important to notice that a uniform decay of neighbouring modes, even though commonly desired in artificial reverberation, is not found in real enclosures. The *normal modes* of a room are associated with stationary waves, whose absorption depends on the spatial directions taken by these waves. For instance, in a rectangular enclosure, axial waves are absorbed less than oblique waves [Mor91]. Therefore, neighboring modes associated with different directions can have different reverberation times. Actually, for commonly found rooms having irregularities in the geometry and in the materials, the response is close to that of a room having diffusive walls, where the energy rapidly spreads among the different modes. In these cases, we can find that the decay time is quite uniform among the modes [Kut95].

Parameterization

The main questions arising after we established a computational structure called FDN are: What are the numbers that can be put in place of the many coefficients of the structure? How should these numbers be chosen?

The most delicate part of the structure is the feedback matrix. In fact, it governs the stability of the whole structure. In particular, it is desirable to start with a lossless prototype, i.e., a reference structure providing an endless, flat decay. The reader interested in general matrix classes that might work as prototypes is referred to the literature [Jot92, RS97, Roc97, Gar97b]. Here we only mention the class of *circulant matrices*, having the general form

$$\mathbf{A} = \begin{bmatrix} a(0) & a(1) & \dots & a(N-1) \\ a(N-1) & a(0) & \dots & a(N-2) \\ \dots & & & \\ a(1) & \dots & a(N-1) & a(0) \end{bmatrix} . \tag{5.30}$$

The stability of an FDN is related to the magnitude of its eigenvalues, which can be computed by the discrete Fourier transform of the first row, in the case of a circulant matrix. By keeping these eigenvalues on the unit circle (i.e., magnitude one) we ensure that the whole structure is stable and lossless. The control over the angle of the eigenvalues can be translated into a direct control over the degree of diffusion of the enclosure that is being simulated by the FDN. The limiting cases are the diagonal matrix, corresponding to perfectly reflecting walls, and the matrix whose rows are sequences of equal-magnitude numbers and (pseudo-)randomly distributed signs [Roc97].

Another critical set of parameters is given by the lengths of the delay lines. Several authors suggested to use lengths in samples that are mutually coprime numbers in order to minimize the collision of echoes in the impulse response. However, if the FDN is linked to a physical and geometrical interpretation, as it is done in the ball-within-a-box model [Roc95], the delay lengths are derived from the geometry of the room being simulated and the resulting digital reverb quality is related to the quality of the actual room. A delay line is associated with a harmonic series of normal modes, all obtainable from a plane-wave loop that bounces back and forth within the enclosure. The delay length for the particular series of normal modes is given by the time interval between two consecutive collisions of the plane wavefront along the main diagonal, i.e., twice the time taken to travel the distance

$$l = \frac{c}{2f_0} , \tag{5.31}$$

being f_0 the fundamental frequency of the harmonic modal series.

5.7.3 Time-variant reverberation

Reverberation algorithms are usually time invariant, meaning that the response does not change as a function of time. This is reasonable, since reverberation algorithms model an LTI system, an impulse response. However, in live performances and installations, it is sometimes beneficial to have a time-variant reverberation to prevent and reduce the coloration and instability due to the feedback caused by the proximity of microphones and loudspeakers. The frequency response of such a system is not ideally flat, which easily leads to acoustical feedback at the frequency with the highest loop gain. Several algorithms exist [NS99] to modify the frequency response of the system so that resonance frequencies vary over time.

One efficient implementation of time-variance to an FDN type reverberator has been proposed [LH01]. The FDN is modified to contain a comb-allpass filter at each delay line. The time variance is implemented by modulating the feedback coefficient of this comb-allpass filter with a few Hertz modulation frequency. Such modulations change the group delay of each delay line, resulting in the frequency shift of resonant frequencies. However, this shift is not constant at all frequencies and if all delay lines in the FDN have different modulation frequencies no audible pitch shift is perceived. Such an algorithm has been successfully applied in the creation of a physically small, but sonically large rehearsal space for a symphony orchestra [LPPS09].

5.7.4 Modeling reverberation with a room geometry

In some applications, it is beneficial to have a room effect based on the defined room geometry. Then, the impulse response is created with computational room acoustics modeling methods.

The methods can be divided into ray-based and wave-based methods, based on the underlying assumptions of the sound propagation [Sil10].

Wave-based methods

Wave-based acoustic modeling aims to numerically solve the wave equation. Traditional techniques are the finite element (FEM) and the boundary element (BEM) methods [SK02]. However, these techniques are computationally too heavy for the whole audible frequency range, although at low frequencies they could be applied in combination with other techniques more suitable at higher frequencies. The digital waveguide mesh method is a newer wave-based technique, being computationally less expensive and thus more suitable for room impulse response creation or even for real-time auralization [MKMS07, Sav10]. A novel, very interesting wave-based method is the adaptive rectangular decomposition method [RNL09].

Ray-based methods

In ray-based acoustic modeling sound is assumed to behave similarly to light. This approximation is valid at high frequencies, and makes it possible to utilize plenty of algorithms developed in computer graphics in the field of global illumination. All the ray-based methods are covered by the room acoustic rendering equation [SLKS07], and all methods can be seen as a special solution for this equation. The detailed presentation of the room acoustic rendering equation is outside the scope of this book, but the most common ray-based modeling methods are briefly introduced here.

The *image-source method* and *beam-tracing methods* are the most common techniques to find specular reflection paths. The image-source method [AB79, Bor84] is based on recursive reflection of sound sources against all the surfaces in the room. This results in a combinatory explosion, and in practice only low-order, i.e., early reflections can be found. Figure 5.16 illustrates the process by showing image sources up to fourth order in a very simple 2-D geometry. Beam-tracing methods, such as [FCE$^+$98, LSLS09], are optimized versions for the same purpose capable of dealing with more complicated geometries and higher reflection orders. A related approach to beam tracing is frustum tracing [LCM07], which scales even better to very large models. For image-source computation with **MATLAB** see [CPB05][1] and [LJ08].[2]

Ray tracing [KSS68] is the most popular offline algorithm for modeling sound propagation since it enables more advanced reflection modeling than the image-source method. A common approach is to shoot rays from the sound source in every direction, according to the directivity pattern of the source, and trace each ray until its energy level has decayed below a given threshold and at the same time keep track of instants when the ray hits a receiver, see Figures 5.17a and 5.17b.

The *acoustic radiance transfer* is a recently presented acoustic modeling technique based on progressive radiosity and arbitrary reflection models [SLKS07]. The acoustic energy is shot from the sound source to the surfaces of the model, which have been divided into patches, as illustrated in Figure 5.17c. Then, the propagation of the energy is followed from patch to patch and the intermediate results are stored on the patches. Finally, when the desired accuracy is achieved, the energy is collected from the patches to the listener.

These three methods have different properties for room-effect simulation. The image source method can only model specular reflections, but it is very efficient in finding perceptually relevant early reflections. As a reverberation effect the image source method is suitable for real-time processing, since the image sources, i.e., early reflections, can be spatially rendered and the late reverberation can be added with, for example, an FDN reverberator [SHLV99, LSH$^+$02]. The ray-tracing method is not suitable for real-time reverberation processing. However, it is good at off line creation of the whole impulse response, which can be applied later with a real-time convolver.

[1] http://media.paisley.ac.uk/∼campbell/Roomsim/

[2] http://www.eric-lehmann.com/ism_code.html

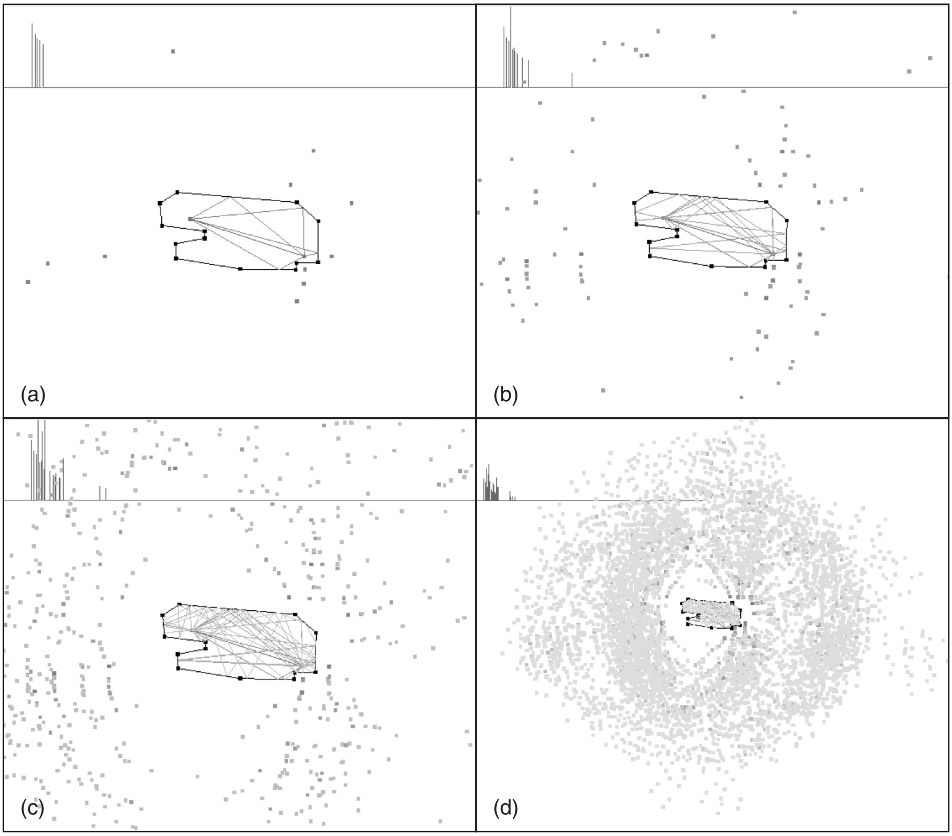


Figure 5.16 Visualization of the image source method. The dots outside of the concert hall silhouette are the image sources. The response above is the energy response at the receiver position, which is on the balcony. Figures (a)–(d) contain first–fourth-order image sources, respectively.

The acoustic radiance transfer method is the most advanced ray-based room acoustics modeling method, since it can handle both specular and diffuse reflections. Although the method is computationally extensive, the usage of the GPUs makes it possible to run the final gathering and sound rendering in real time, thus enabling interactive reverberation effects of environments with arbitrary reflection properties [SLS09].

5.8 Other spatial effects

5.8.1 Digital versions of classic reverbs

In the past, before the era of digital signal processing, many systems were used to create a reverberation effect. In studios the common way was to replay the recorded signals in a reverberant room or a corridor and record it again in that particular space. In addition, plates and springs were applied to create a decaying tail to the sound. Recently, researchers have modeled the physical principles of old-school reverberators and they have proposed digital implementations of them. For example, plate reverbs have been implemented with finite difference schemes [BAC06, Bil07]. An extension to digital plate reverbs to handle objects of any shape has been made with modal

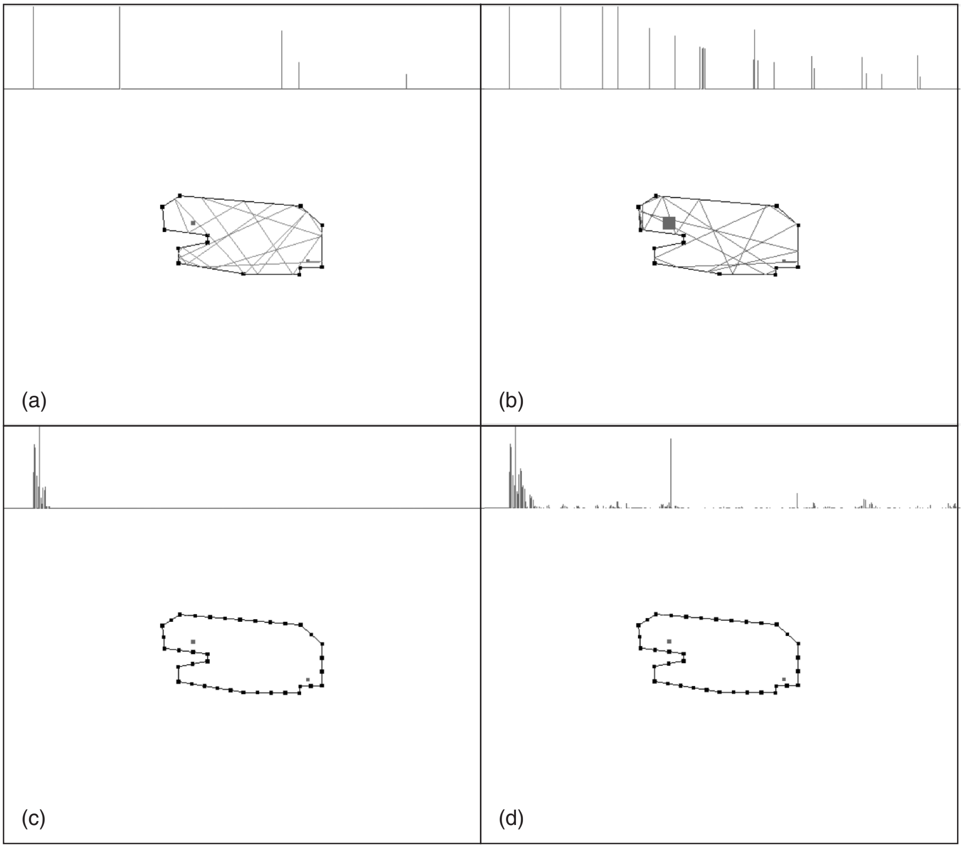


Figure 5.17 The ray-tracing method produces (a) a sparse response. (b) If the receiver area is larger more reflections are modeled. (c) The acoustics radiance transfer method when the initial energy is distributed to the surface patches. (d) The whole response after 100 energy distributions.

synthesis methods. Thus, simulations of the vibration of many different shapes and materials can be performed in real time [Max07]. Spring reverbs have also been modeled and it seems that efficient implementation can be achieved with parallel waveguides, which include dispersive all pass filters [ABCS06].

5.8.2 Distance effects

In digital audio effects, the control of apparent distance can be effectively introduced even in monophonic audio systems. In fact, the impression of distance of a sound source is largely controllable by insertion of artificial wall reflections or reverberant room responses.

There are not reliable cues for distance in anechoic or open spaces. Familiarity with the sound source can provide distance cues related with air absorption of high frequency. For instance, familiarity with a musical instrument tells us what is the average intensity of its sound when coming from a certain distance. The fact that timbral qualities of the instrument will change when playing loud or soft is also a cue that does help the identification of distance. These cues seem to vanish when using unfamiliar sources or synthetic stimuli that do not resemble any physical-sounding

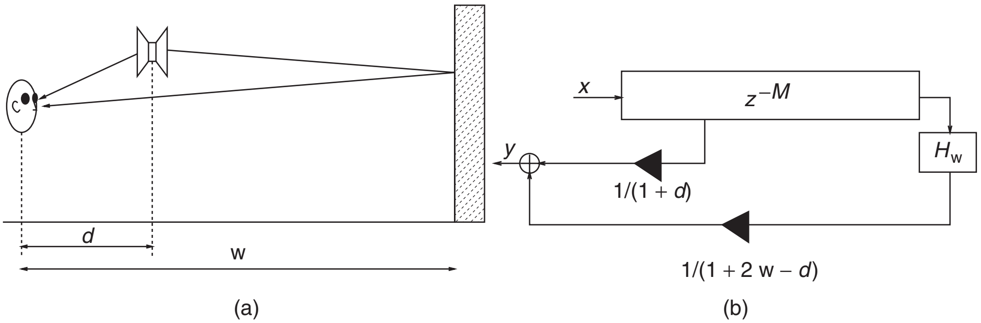


Figure 5.18 Distance rendering via single wall reflection: (a) physical situation, (b) signal-processing scheme

object. Conversely, in an enclosure the ratio of reverberant to direct acoustic energy has proven to be a robust distance cue [Bla97]. It is often assumed that in a small space the amplitude of the reverberant signal changes little with distance, and that in a large space it is roughly proportional to $1/\sqrt{\text{distance}}$ [Cho71]. The direct sound attenuates as 1/distance if spherical waves are propagated.

A single reflection from a wall can be enough to provide some distance cues in many cases. The physical situation is illustrated in Figure 5.18a, together with the signal-processing circuit that reproduces it. A single delay line with two taps is enough to reproduce this basic effect. Moreover, if the virtual source is close enough to the listening point, the first tap can be taken directly from the source, thus reducing the signal-processing circuitry to a simple non-recursive comb filter. To be physically consistent, the direct sound and its reflection should be attenuated as much as the distance they travel, and the wall reflection should also introduce some additional attenuation and filtering in the reflected sound, represented by the filter H_w in Figure 5.18b. The distance attenuation coefficients of Figure 5.18b have been set in such a way that they become one when the distance goes to zero, just to avoid the divergence to infinity that would come from the physical laws of a point source.

From this simple situation it is easy to see how the direct sound attenuates faster than the reflected sound, as long as the source approaches the wall.[3] This idea can be generalized to closed environments adding a full reverberant tail to the direct sound. An artificial yet realistic reverberant tail can be obtained just by taking an exponentially decayed gaussian noise and convolving it with the direct sound. The reverberant tail should be added to the direct sound after some delay (proportional to the size of the room) and should be attenuated with distance to a lesser extent than the direct sound. Figure 5.19 shows the signal-processing scheme for distance rendering via room reverberation.

The following M-file allows one to experiment with the situations depicted in Figures 5.18 and 5.19, with different listener positions, provided that x is initialized with the input sound, and y, z, and w are long-enough vectors initialized to zero.

M-file 5.16 (distfx.m)

```
% distfx.m
% Author: T. Lokki
h = filter([0.5,0.5],1, ...
    random('norm',0,1,1,lenh).*exp(-[1:lenh]*0.01/distwall)/100);
    % reverb impulse response
```

[3] Indeed, in this single-reflection situation, the intensity of the reflected sound increases as the source approaches the wall.

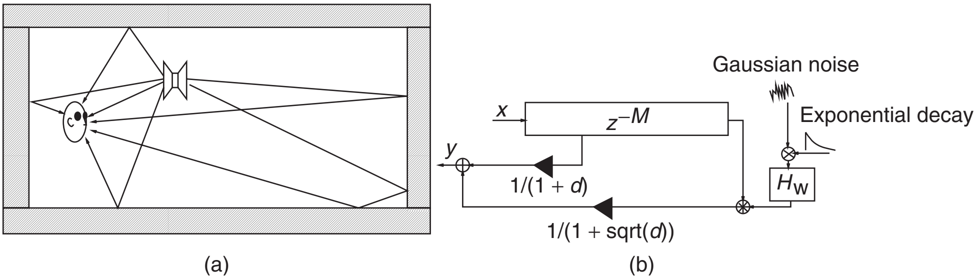


Figure 5.19 Distance rendering via room reverberation: (a) physical situation, (b) signal-processing scheme

```
offset = 100;
st = Fs/2;

for i = 1:1:distwall-1 % several distances listener-source
  del1 = floor(i/c*Fs);
  del2 = floor((distwall*2 - i)/c*Fs);
  y(i*st+1:i*st+del1) = zeros(1,del1);
  y(i*st+del1+1:i*st+del1+length(x)) = x./(1+i); % direct signal
  w(i*st+del2+1:i*st+del2+length(x)) = ...
         y(i*st+del2+1:i*st+del2+length(x)) + ...
         x./(1+(2*distwall-i));    % direct signal + echo
  z(i*st+del2+1:i*st+del2+length(x)+lenh-1+offset) = ...
         y(i*st+del2+1:i*st+del2+length(x)+lenh-1+offset) + ...
         [zeros(1,offset),conv(x,h)]./sqrt(1+i);
                                   % direct signal + delayed reverb
end
```

5.8.3 Doppler effect

Movements of the sound sources are detected as changes in direction and distance cues. The Doppler effect is a further (strong) cue that intervenes whenever there is a radial component of motion between the sound source and the listener. In a closed environment, radial components of motion are likely to show up via reflections from the walls. Namely, even if a sound source is moving at constant distance from the listener, the paths taken by the sound waves via wall reflections are likely to change in length. If the source motion is sufficiently fast, in all of these cases we will have transpositions in frequency of the source sound.

The principle of the Doppler effect is illustrated in Figure 5.20, where the listener is moving toward the sound source with speed c_s. If the listener meets f_s wave crests per second at rest, it ends up meeting crests at the higher rate

$$f_d = f_s \left(1 + \frac{c_s}{c}\right) \tag{5.32}$$

when the source is moving. Here c is the speed of sound in air. We usually appreciate the pitch shift due to Doppler effect in non-musical situations, such as when an ambulance or a train is passing by. The perceived cue is so strong that it can evocate the relative motion between source and listener even when other cues indicate a constant relative distance between the two. In fact, ambulance or insect sounds having a strong Doppler effect are often used to demonstrate how good a spatialization system is, thus deceiving the listener who doesn't think that much of the spatial

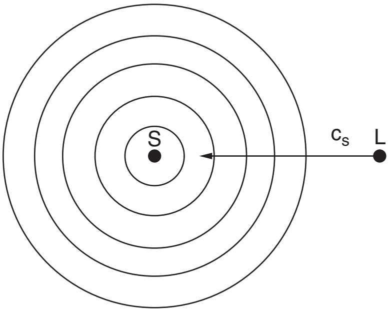


Figure 5.20 Illustration of the Doppler effect.

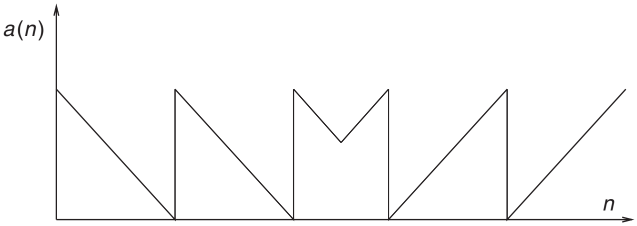


Figure 5.21 Control signal for simulating the Doppler effect with a delay-based pitch shifter

effect is already in the monophonic recording. Research into psychoacoustics has also shown how the perception of pitch can be strongly affected by dynamic changes in intensity, as they are found in situations where the Doppler effect occurs [Neu98]. Namely, a sound source approaching the listener at constant speed produces a rapid increase in intensity when it traverses the neighborhood of the listener. On the other hand, while the frequency shift is constant and positive before passing the listener, and constant and negative after it has passed, most listeners perceive an increase in pitch shift as the source is approaching. Such apparent pitch increase is due to the simultaneous increase in loudness.

The Doppler effect can be faithfully reproduced by a pitch shifter (see Chapter 6) controlled by the relative velocity between source and listener. In particular, the circuit in Figure 6.13 can be used with sawtooth control signals whose slope increases with the relative speed. Figure 5.21 shows the signal used to control one of the delays in Figure 6.13 for a sound source that approaches the listening point and passes it. Before the source reaches the listener, the sound is raised in pitch, and it is lowered right after.

Any sound-processing model based on the simulation of wave propagation, implements an implicit simulation of the Doppler effect. In fact, these models are based on delay lines that change their length according to the relative position of source and listener, thus providing positive or negative pitch transpositions.

In general, the accuracy and naturalness of a Doppler shift reproduced by digital means depends on the accuracy of interpolation in variable-length delays. If this is not good enough, modulation products affect the transposed signal, producing remarkable artifacts.

5.9 Conclusion

Playing with the spatial attributes of sound has been an intriguing and challenging task for many musicians and sound designers. The multiplicity of techniques developed so far has been roughly overviewed in the previous pages. Despite the length of this chapter, we have certainly missed many important contributions to the field. However, we have tried to communicate which are the main structural, perceptual, or technological limitations and possibilities of spatial audio. We hope that the sound designer, after reading this chapter, will be able to model some spatial features of sound or, at least, to be conscious of those features that will be part of the aesthetics of the design process rather than part of the sonic outcome.

Acknowledgements

The Academy of Finland, project [218238], have supported this work. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreements no. [240453] and no. [203636].

References

[AB79] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am*., 65(4): 943–950, 1979.

[ABCS06] J. S. Abel, D. P. Berners, S. Costello, and J. O. Smith. Spring reverb emulation using dispersive allpass filters in a waveguide structure. In *the 121st Audio Eng. Soc. (AES) Conv.,* 2006. Paper # 6954.

[ADDA01] V. R. Algazi, R. O Duda, Thompson D. M., and C. Avendano. Parameters for auditory display of height and size. In *Proc. 2001 IEEE Workshop Appl. Signal Proces. Audio Electroacoust.,* Mohonk Mountain House, New Paltz, NY, 2001.

[ARB04] A. Apel, T. Röder, and S. Brix. Equalization of wave field synthesis systems. In *Proc. 116th AES Convention, 2004 Paper # 6121.*

[BAC06] S. Bilbao, K. Arcas, and A. Chaigne. A physical model for plate reverberation. In *Proc. Int. Conf. Acoust., Speech, Signal Proces. (ICASSP 2006)*, volume V, pp. 165–168, 2006.

[BBE85] J. C. Bennett, K. Barker, and F. O. Edeko. A new approach to the assessment of stereophonic sound system performance. *J. Audio Eng. Soc*., 33(5): 314–321, 1985.

[BD98] C. P. Brown and R. O. Duda. A structural model for binaural sound synthesis. *IEEE Trans. Speech and Audio Process.,* 6(5): 476–488, 1998.

[Beg94] D. R. Begault. *3-D Sound For Virtual Reality and Multimedia*. AP Professional, 1994.

[BF08] J. Breebaart and C. Faller. *Spatial Audio Processing: MPEG Surround and Other Applications*. Wiley-Interscience, 2008.

[Bil07] S. Bilbao. A digital plate reverberation algorithm. *J. Audio Eng. Soc.,* 55(3): 135–144, 2007.

[Bla97] J. Blauert. *Spatial Hearing. The Psychophysics of Human Sound Localization*, 2nd edition MIT Press, 1997.

[Bor84] J. Borish. Extension of the image model to arbitrary polyhedra. *J. Acoust. Soc. Am*., 75(6): 1827–1836, 1984.

[BR99] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. head-related transfer functions. *J. Acoust. Soc. Am*., 106(3): 1465–1479, 1999.

[BS.94] ITU-R Recommendation BS.775-1. Multichannel stereophonic sound system with and without accompanying picture. Technical report, International Telecommunication Union, Geneva, Switzerland, 1992-1994.

[BVV93] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustics control by wave field synthesis. *J. Acoust. Soc. Am*., 93(5): 2764–2778, May 1993.

[CB89] D. H. Cooper and J. L. Bauck. Prospects for transaural recording. *J. Audio Eng. Soc*., 37(1/2): 3–39, 1989.

[Cho71] J. Chowning. The simulation of moving sound sources. *J. Audio Eng. Soc*., 19(1): 2–6, 1971.

[Coo87] D. H. Cooper. Problems with shadowless stereo theory: Asymptotic spectral status. *J. Audio Eng. Soc*., 35(9): 629–642, 1987.

[CPB05] D. R. Campbell, K. J. Palomäki, and G. Brown. A matlab simulation of "shoebox" room acoustics for use in research and teaching. *Comp. Inform. Syst. J*., 9(3), 2005.

[Cra03] P. G. Craven. Continuous surround panning for 5-speaker reproduction. In *AES 24th Int. Conf. Multichannel Audio*, 2003.

[CT03] D. Cabrera and S. Tilley. Parameters for auditory display of height and size. In *Proc. ICAD*, 2003.

[Dat97] J. Dattorro. Effects design, part 1: Reverberator and other filters. *J. Audio Eng. Soc.*, 45(9): 660–683, 1997.

[Dav03] M. F. Davis. History of spatial coding. *J. Audio Eng. Soc.*, 51(6): 554–569, 2003.

[DM98] R. Duda and W. Martens. Range-dependence of the HRTF of a spherical head. *Appl. Signal Process. Audio Acoust.* 104(5): 3048–3058, November 1998.

[DNM03] J. Daniel, R. Nicol, and S. Moreau. Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging. In *Proc. 114th AES Conv.*, 2003. Paper # 5788.

[ED08] N. Epain and J. Daniel. Improving Spherical Microphone Arrays. In *Proc. 124th AES Convention*, 2008. Paper #7479.

[FCE+98] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West. A beam tracing approach to acoustics modeling for interactive virtual environments. In *Proc. 25th Ann. Conf. Comp. Graphics Interactive techniques (SIGGRAPH'98)*, pp. 21–32, 1998.

[GA97] R. H. Gilkey and T. R. Anderson (eds). *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Assoc., 1997.

[Gar95] W. G. Gardner. Efficient convolution without input-output delay. *J. Audio Eng. Soc.,* 43(3): 127–136, 1995.

[Gar97a] W. G. Gardner. *3-D Audio Using Loudspeakers*. PhD thesis, MIT Media Lab, 1997.

[Gar97b] W. G. Gardner. Reverberation algorithms. In M. Kahrs and K. Brandenburg (eds), *Applications of Digital Signal Processing to Audio and Acoustics*, pp. 85–131. Kluwer Academic Publishers, 1997.

[Ger73] M. J. Gerzon. Periphony: With height sound reproduction. *J. Audio Eng. Soc.*, 21(1): 2–10, 1973.

[Ger76] M. A. Gerzon. Unitary (energy preserving) multichannel networks with feedback. *Electron. Lett. V*, 12(11): 278–279, 1976.

[GJ06] M. M. Goodwin and J.-M. Jot. A frequency-domain framework for spatial audio coding based on universal spatial cues. In *Proc. 120th AES Convention*, 2006. Paper # 6751.

[GM94] W. G. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. Technical Report 280, MIT Media Lab Perceptual Computing, 1994.

[Hir07] T. Hirvonen. *Perceptual and Modeling Studies on Spatial Sound*. PhD thesis, Helsinki University of Technology, 2007. http://lib.tkk.fi/Diss/2007/isbn9789512290512/.

[HP08] T. Hirvonen and V. Pulkki. Perceived Spatial Distribution and Width of Horizontal Ensemble of Independent Noise Signals as Function of Waveform and Sample Length signals as Function of Waveform and Sample length. In *Proc. 124th AES Convention,* 2008. Paper # 7408.

[JC91] J.-M. Jot and A. Chaigne. Digital delay networks for designing artificial reverberators. In *Proc. AES Convention*, 1991. Preprint no. 3030.

[Jot92] J.-M. Jot. *Etude et réalisation d'un spatialisateur de sons par modèles physique et perceptifs*. PhD thesis, l'Ecole Nationale Superieure des Telecommunications, Télécom Paris 92 E 019, 1992.

[JSer] D. Jaffe and J. O. Smith. Extensions of the Karplus-Strong plucked string algorithm. *Comp. Music J.*, 7(2): 56–69, 1983 Summer. Reprinted in C. Roads (ed.), *The Music Machine*. (MIT Press, 1989, pp. 481–49.

[Kai80] T. Kailath. *Linear Systems*. Prentice-Hall, 1980.

[KNH98] O. Kirkeby, P. A. Nelson, and H. Hamada. Local sound field reproduction using two closely spaced loudspeakers. *J. Acoust. Soc. Am.*, 104: 1973–1981, 1998.

[KSS68] A. Krokstad, S. Strom, and S. Sorsdal. Calculating the acoustical room response by the use of a ray tracing technique. *J. Sound Vibr.*, 8: 118–125, 1968.

[Kut95] H. Kuttruff. A simple iteration scheme for the computation of decay constants in enclosures with diffusely reflecting boundaries. *J. Acous. Soc. Am.*, 98(1): 288–293, 1995.

[KV01] M. Kubovy and D. Van Valkenburg. Auditory and visual objects. *Cognition*, 80(1–2): 97–126, 2001.

[LCM07] C. Lauterbach, A. Chandak, and D. Manocha. Interactive sound rendering in complex and dynamic scenes using frustum tracing. *IEEE Trans. Visualization Comp. Graphics*, 13(6): 1672–1679, 2007.

[LCYG99] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *J. Acoust. Soc. Am.*, 106: 1633, 1999.

[LH01] T. Lokki and J. Hiipakka. A time-variant reverberation algorithm for reverberation enhancement systems. In *Proc. Digital Audio Effects Conf. (DAFx-01)*, 2001, pp. 28–32.

[Lip86] S. P. Lipshitz. Stereophonic microphone techniques... are the purists wrong? *J. Audio Eng. Soc.*, 34(9): 716–744, 1986.

[LJ08] E. A. Lehmann and A. M. Johansson. Prediction of energy decay in room impulse responses simulated with an image-source model. *J. Acoust. Soc. Am.*, 124(1): 269–277, 2008.

[LPPS09] T. Lokki, J. Pätynen, T. Peltonen, and O. Salmensaari. A rehearsal hall with virtual acoustics for symphony orchestras. In *Proc. 126th AES Conv.*, 2009. paper no. 7695.

[LSH$^+$02] T. Lokki, L. Savioja, J. Huopaniemi, R. Väänänen, and T. Takala. Creating interactive virtual auditory environments. *IEEE Comp. Graphics Appl.*, 22(4): 49–57, 2002.

[LSLS09] S. Laine, S. Siltanen, T. Lokki, and L. Savioja. Accelerated beam tracing algorithm. *Appl. Acoust.*, 70(1): 172–181, 2009.

[Max07] C. B. Maxwell. Real-time reverb simulation for arbitrary object shapes. In *Proc. 10th Int. Conf. Digital Audio Effects (DAFX-07)*, 15–20, 2007.

[MHR10] E. J. Macaulay, W. M. Hartmann, and B. Rakerd. The acoustical bright spot and mislocalization of tones by human listeners. *J. Acoust. Soc. Am.*, 127: 1440, 2010.

[Mit98] S. K. Mitra. *Digital Signal Processing: A Computer-Based Approach*. McGraw-Hill, 1998.

[MKMS07] D. Murphy, A. Kelloniemi, J. Mullen, and S. Shelley. Acoustic modeling using the digital waveguide mesh. *IEEE Signal Proces. Mag.*, 24(2): 55–66, 2007.

[MM95] D. G. Malham and A. Myatt. 3-D sound spatialization using ambisonic techniques. *Comp. Music J.*, 19(4): 58–70, 1995.

[MM01] S. Müller and P. Massarani. Transfer function measurement with sweeps. *J. Audio Eng. Soc.*, 49(6): 443–471, 2001.

[Mon00] G. Monro. In-phase corrections for ambisonics. In *Proc. Int. Comp. Music Conf.,* 2000, pp. 292–295.

[Moo79] J. A. Moorer. About this reverberation business. *Comp. Music J.*, 3(2): 13–28, 1979.

[Moo83] F. R. Moore. A general model for spatial processing of sounds. *Comp. Music J.*, 7(3): 6–15, 1983.

[Moo90] F. R. Moore. *Elements of Computer Music*. Prentice Hall, 1990.

[Mor91] P. M. Morse. *Vibration and Sound*. American Institute of Physics for the Acoustical Society of America, 1991.

[MSHJ95] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen. Head-related transfer functions of human subjects. *J. of Audio Eng. Soc.*, 43(5): 300–321, May 1995.

[MT99] C. Müller-Tomfelde. Low-latency convolution for real-time applications. In *Proc. 16th AES Int. Conf*, 1999 pp. 454–459.

[Neu98] J. G. Neuhoff. A perceptual bias for rising tones. *Nature*, 395(6698): 123–124, 1998.

[NS99] J. L. Nielsen and U. P. Svensson. Performance of some linear time-varying systems in control of acoustic feedback. *J. Acoust. Soc. Am.*, 106(1): 240–254, 1999.

[OS89] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, Inc., 1989.

[PB82] D. R. Perrott and T. N. Buell. Judgments of sound volume: Effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise. *J. Acoust. Soc. Am.*, 72: 1413, 1982.

[PH05] V. Pulkki and T. Hirvonen. Localization of virtual sources in multi-channel audio reproduction. *IEEE Trans. Speech Audio Proc.*, 2005.

[PKH99] V. Pulkki, M. Karjalainen, and J. Huopaniemi. Analyzing virtual sound source attributes using a binaural auditory model. *J. Audio Eng. Soc.*, 47(4): 203–217, April 1999.

[Pol07] A. Politis. *Subjective evaluation of the performance of virtual acoustic imaging systems under suboptimal conditions of implementation*, MSc thesis. University of Southampton, 2007.

[Pul97] V. Pulkki. Virtual source positioning using vector base amplitude panning. *J. Audio Eng. Soc.*, 45(6): 456–466, 1997.

[Pul99] V. Pulkki. Uniform spreading of amplitude panned virtual sources. In *1999 IEEE Workshop Appl. Signal Proces. Acoust.*, 1999.

[Pul01] V. Pulkki. *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2001.

[Pul02] V. Pulkki. Compensating displacement of amplitude-panned virtual sources. In *22nd AES Int. Conf. on Virtual, Synth. Enter. Audio*, 2002 pp. 186–195.

[Pul07] V. Pulkki. Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc.*, 55(6): 503–516, June 2007.

[RNL09] N. Raghuvanshi, R. Narain, and M. Lin. Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE Trans. Visual. Comp. Graphics*, 15(5): 789–801, 2009.

[Roc95] D. Rocchesso. The ball within the box: a sound-processing metaphor. *Comp Music J.*, 19(4): 47–57, 1995.

[Roc96] D. Rocchesso. *Strutture ed Algoritmi per l'Elaborazione del Suono basati su Reti di Linee di Ritardo Interconnesse. Tesi sottoposta per il conseguimento del titolo di dottore di ricerca in ingegneria informatica ed elettronica industriali*. PhD thesis, Universita di Padova, Dipartimento di Elettronica e Informatica, 1996.

[Roc97] D. Rocchesso. Maximally diffusive yet efficient feedback delay networks for artificial reverberation. *IEEE Signal Process. Lett*., 4(9): 252–255, Sep. 1997.

[RS97] D. Rocchesso and J. O. Smith. Circulant and elliptic feedback delay networks for artificial reverberation. *IEEE Trans. Speech Audio Process*., 5(1): 51–63, 1997.

[Rum01] F. Rumsey. *Spatial Audio*. Focal Press, 2001.

[Sav10] L. Savioja. Real-time 3D finite-difference time-domain simulation of low- and mid-frequency room acoustics. In *13th Int. Conf. Digital Audio Effects*, 2010.

[Sch61] M. R. Schroeder. Improved quasi-stereophony and "colorless" artificial reverberation. *J. Acoust. Soc. Am*., 33(8): 1061–1064, 1961.

[Sch62] M. R. Schroeder. Natural-sounding artificial reverberation. *J. Audio Eng. Soc*., 10(3): 219–223, 1962.

[Sch70] M. R. Schroeder. Digital simulation of sound transmission in reverberant spaces. *J. Acoust. Soc. Am*., 47(2 (Part 1)): 424–431, 1970.

[Sch73] M. R. Schroeder. Computer models for concert hall acoustics. *Am. J. Physics*, 41: 461–471, 1973.

[SHLV99] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environments. *J. Audio Eng. Soc*., 47(9): 675–705, 1999.

[Sil10] S. Siltanen. *Efficient physics-based room-acoustics modeling and auralization*. PhD thesis, Aalto University School of Science and Technology, 2010. Available at http://lib.tkk.fi/Diss/2010/isbn9789522482645/.

[SK02] U. P. Svensson and U. R. Kristiansen. Computational modeling and simulation of acoustic spaces. In *Proc. 22nd AES Int. Conf. on Virtual, Synth. Entert. Audio*, 2002 pp. 11–30.

[SK04] R. Sadek and C. Kyriakakis. A novel multichannel panning method for standard and arbitrary loudspeaker configurations. In Proc. *117th AES Conv.,* 2004 Paper #6263.

[SL61] M. R. Schroeder and B. Logan. "Colorless" artificial reverberation. *J. Audio Eng. Soc*., 9: 192–197, 1961.

[SLKS07] S. Siltanen, T. Lokki, S. Kiminki, and L. Savioja. The room acoustic rendering equation. *J. Acoust. Soc. Am*., 122(3): 1624–1635, 2007.

[SLS09] S. Siltanen, T. Lokki, and L. Savioja. Frequency domain acoustic radiance transfer for real-time auralization. *Acta Acust. United AC*, 95(1): 106–117, 2009.

[Smi85] J. O. Smith. A new approach to digital reverberation using closed waveguide networks. In *Proc. Int. Comp. Music Conf. (ICMC'85)*, 1985 pp. 47–53.

[Smi86] J. O. Smith. Elimination of limit cycles and overflow oscillations in time-varying lattice and ladder digital filters. In *Proc. IEEE Conf. Circuits Syst*., 197–299, 1986.

[Smi92] J. O. Smith. Physical modeling using digital waveguides. *Comp. Music J*., 16(4): 74–87, 1992.

[Sol08] A. Solvang. Spectral impairment of two-dimensional higher order ambisonics. *J. Audio Eng. Soc*., 56(4): 267–279, 2008.

[SP82] J. Stautner and M. Puckette. Designing multi-channel reverberators. *Comp. Music J*., 6(1): 569–579, 1982.

[Ste96] G. Steinke. Surround sound–the new phase. an overview. In *Proc. 100th AES Conv*, 1996. Preprint #4286.

[The91] G. Theile. HDTV sound systems: How many channnels ? In *Proc. 9th AES Int. Conf. "Television Sound Today and Tomorrow"*, 1991. pp. 217–232.

[Tor98] E. Torick. Highlights in the history of multichannel sound. *J. Audio Eng. Soc*., 46(1/2): 27–31, 1998.

[Vai93] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.

[VB99] D. Vries and M. Boone. Wave field synthesis and analysis using array technology. In *Proc. 1999 IEEE Workshop Appl. Signal Proces. Audio Acoust*., 1999. pp. 15–18.

[Vil08] J. Vilkamo. *Spatial sound reproduction with frequency band processing of b-format audio signals*, MSc thesis. Helsinki University Technology, 2008.

[VLP09] J. Vilkamo, T. Lokki, and V. Pulkki. Directional audio coding: virtual microphone-based synthesis and subjective evaluation. *J. Audio Eng. Society*, 57(9): 709–724, 2009.

[ZF90] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Springer-Verlag, 1990.

[Zur87] P. M. Zurek. The precedence effect. In W. A. Yost and G. Gourewitch (eds), *Directional Hearing*, pp. 3–25. Springer-Verlag, 1987.