NORTHWESTERN UNIVERSITY

A Musically Motivated Approach to Spatial Audio for Large Venues

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Music Technology

By

David Etlinger

EVANSTON, ILLINOIS

December 2009

Abstract

A Musically Motivated Approach to Spatial Audio for Large Venues

David Etlinger

Spatial audio refers to those aspects of sound reproduction associated with the perceived direction and expanse of sound images. The vast majority of research on spatial audio considers a single listener, ideally positioned in the listening area among the speakers. From that position, spatial effects can be quite convincing. However, this approach cannot be extended to large or public spaces, or to non-centered listeners. Many existing solutions are limited to cinematic special effects or have extensive hardware requirements.

This dissertation presents a new approach to large-venue spatialization, particularly appropriate for musical and creative artistic works. Digital signal processing (DSP) is used to explicitly control the region of focus in the listening area, and hence which audience members are receiving spatial cues. By moving the target area over time, the entire audience can receive maximally effective spatial audio at some point. Target location processing is guided by consideration of the entire soundfield. The emphasis is not on absolute localization accuracy but on potential aesthetic uses. The system is not limited to localization cues; decorrelation effects are treated with equal attention.

A spherical head model is used, both for synthetic directional cues and for predicting the physical acoustic response. Targeting of listener locations is performed using several methods, including crosstalk cancellation. This work also makes the novel assertion that overhead loudspeakers create an optimal soundfield, both by maximizing the target area and by minimizing undesirable properties elsewhere. Evidence is provided to quantify and support this claim.

A historical review of prior approaches to spatial audio in large venues is given first. The relevant literature on spatial hearing perception is then summarized, with emphasis on large-space loudspeaker reproduction. Next, the signal processing techniques used for implementation are discussed. Evaluation is done primarily through simulations of soundfields using recorded Head-Related Transfer Functions (HRTFs). In addition, listening demonstrations were conducted to confirm the correct operation of the system.

Acknowledgements

I was fortunate indeed to work with my advisor and committee chair, Gary Kendall. Gary was instrumental in completing this dissertation. Besides his important research in the field, he has a deeply musical relationship with spatial sound, from which I have benefited greatly. I would also like to thank the two other members of my committee, Chris Mercer and Don Ellis. Chris's approach served as a reminder to focus on practice as well as theory. He is an inventive composer whose pieces have informed my understanding of sound in space. Don's perspective was a valuable counterpoint. A physicist and chemist as well as a musician, his incisive comments often revealed for me the aesthetic implications of technical choices. I cannot imagine a better committee to have worked with—I am truly honored to call these three colleagues and friends.

The people or institutions shown in table 0.1 were kind enough to provide permission to use the indicated images, for which I am grateful.

Source	*Image*	*Figure*	*Page*
Ina/Maurice Lecardent	Pierre Henry	2.6	39
Philips/Iannis Xenakis	Philips Pavilion	2.9	45
Ina/Ruszka Laslo	Acousmonium	2.10	46
Kevin Busby	BEAST	2.13	51

TABLE 0.1. Image credits.

Author	*Title*	*Purpose*
James McCartney	SuperCollider	Sound Synthesis
H. James Harkins	dewdrop_lib	Realtime Mixing and Routing
Wouter Snoei	wslib	Many Utilities
Scott Wilson	PV_Decorrelate	Decorrelation UGen
Donald E. Knuth	TEX	Document Typesetting
Leslie Lamport	LATEX	Document Typesetting
Hàn Thế Thành	pdfTEX	PDF Generation and Microtypography
Sven Wiegand	TEXnicCenter	LATEX IDE
Miguel Lerma	nuthesis.cls	Northwestern Dissertation Formatting
Philipp Lehman	`biblatex`	Bibliography Generation
Robert Schlicht	`microtype`	Microtypography
Sebastian Rahtz & Heiko Oberdiek	`hyperref`	PDF Hyperlinks
Simon Fear	`booktabs`	Table Typesetting
Axel Sommerfeldt	`caption`	Caption Formatting
Steven Douglas Cochran	`subfig`	Multipart Figures
Till Tantau	PGF/TikZ	Vector Graphics
Christian Feuersäger	PGFPLOTS	Plot Graphics
Gene Ressler	Sketch	3D Graphics

TABLE 0.2. Authors of some free software packages used for this dissertation.

I made extensive use of free software, both for my research and during the preparation of this document. I would like to take the lamentably rare step of thanking a few of the authors of this code (table 0.2).

Thanks to the Northwestern School of Music and the Graduate School for their financial support. The opportunity to conduct research at SARC was partially supported by a Graduate Research Grant from the University Research Grants Committee at Northwestern University. I would also like to thank Sergio Gregorio for his extensive assistance during my time there, particularly his insights during our listening sessions.

In no particular order, my thanks and gratitude to the following people. First, to Virgil Moorefield and Scott Lipscomb—I hope that one day we will have the chance to work together again. Second, to my fellow Music Technology students, particularly Casey Farina, Brett Masteller, Theron Humiston and Scott Jaeger. I will always remember time spent in the basement Grad Lab ("Rat Lab") fondly. I owe Dan DiPaolo and Dan Harrison a debt for getting me into music in the first place, and Sergio Monteiro for making me a halfway decent player. To Caroline and James Davis and to Anna Yankee, thank you for your friendship and moral support, especially when preparing for my defense. For the many others who have helped me in one way or another, thanks. I cannot name everyone, but you know who you are.

To my parents, my brother Daniel, my sister Rebecca, and my entire family, no thanks will ever be enough, though I say it anyway—thank you. Kara, you have listened to countless hours of me swearing at code and never once complained. You are very kind and I am profoundly grateful for you (and for Ganieda too!). Persephone, you have listened to even more swearing and never complained either, because you are a cat and that would not be dignified. Thank you for keeping my lap warm while I wrote.

Finally, thank you to all those who have ever shared their music with me. You have changed my life, whether you know it or not.

To my parents,
Who have always loved me.

To Kara,
Whom I will always love.

To Persephone,
Who's a good kitty. Yes you are!

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction and Overview

The relationships between sound and space hold enormous potential for musical expression and meaning. The general term *spatial audio* refers to any electronic sound that is processed to convey spatial information. The vast majority of spatial audio research has focused on the ideal case of a single, properly centered listener in a small room such as a typical living room [17]. The larger size and listening area of public spaces present numerous technical hurdles for spatial sound. Yet public performance is a critical part of music's social and artistic function. The attempted solutions have suffered numerous drawbacks; primarily, they often focus on a limited set of percepts designed for cinematic special effects [32]. Even these basic effects only work well for a small group of listeners centered in the room. Other approaches rely on multiple loudspeakers, often eight or more for electroacoustic music. This dissertation describes a strategy which addresses these issues in a novel way.

1.1. Spatial Hearing

Prior work on spatial hearing is extensive [17]. The domain is usually divided into two loose categories: localization and environmental information. *Localization* refers to the physical location of a sound source: azimuth angle (front, back or side angle); elevation angle, and distance. A related percept, *apparent source width* (ASW), describes the perceived size of the source. Though the details are complex, there are three major localization cues. *Interaural Time Difference* (ITD) refers to the delay between a sound's arrival at the left and right ears

(usually around 0.8 ms or less). Similarly, *Interaural Intensity Difference* (IID) describes the difference in sound pressure level at the left and right ears. Finally, the head, torso and outer ear act as filters, altering the spectrum of sound and leading to frequency-dependent IIDs and ITDs. As the orientation of source to listener changes, so does this filtering effect. This orientation-dependent filtering is called a *Head-Related Transfer Function* (HRTF). Almost all directional hearing relies on these three cues. They are well understood, and can be used quite convincingly in a small space.

In addition to localization, all sounds provide information about the listening environment through reflections and reverberation. Of primary importance is *envelopment*, the sense of being immersed in sound, provided by superior concert halls. Envelopment is often aesthetically crucial to music performance [56]. Envelopment and apparent source width are not rigidly distinct. Other environmental cues include room size, shape and material type. These in turn contribute to intelligibility, timbral warmth, and streaming (the ability to separate simultaneous sounds from different sources).

A more complete review of spatial hearing perception is found in chapter 3.

1.2. Design Principles

This section outlines the aesthetic goals and basic structural components of the system. Specific implementation details are given in chapter 4. Evidence to support system effectiveness is found in chapter 5.

1.2.1. Musically Motivated Spatial Sound

Often, the success of spatial audio reproduction is equated with objective localization performance. But there are many more kinds of spatial hearing which contribute to the quality of subjective experience [116]. In music, and for electroacoustic music in particular, the manipulation of spatial attributes can be of essential aesthetic significance [75]. While direct exploration of the aesthetics of spatial audio is outside the scope of this document, the relationships of music to space served as context and motivation for the directions of this research.

1.2.2. Dynamic Target Location

The overwhelming majority of the literature on signal processing for spatial audio deals with a single listener. The notion of a single "sweet spot" for stereo is well entrenched. Crosstalk cancellation techniques are particularly known for a narrow listening area. From a quantitative perspective, two louspeakers permit one to specify precise signals for two points in space (the left and right ears of a single listener). The rest of the soundfield is usually considered only in terms of reducing errors due to relatively small head motions.

Here we are concerned with the experience of listeners over a wide area. A central principle of the described system is the *targeting of spatial cues*: the idea that, for any given time, a certain region of the audience can receive convincing spatial audio. Contrast this to many conventional methods, where a small region of the audience is always receiving optimal spatial cues, while most audience members are always receiving cues ranging from weaker to completely invalid. In fact, a static target is a special case of the system we describe.

The approach we have taken is to sweep the target location across the audience along a path from left to right (or right to left), in a line parallel to the front and rear walls. Generally, a target location moving left to right produces a sound image that moves right to left. This is because, for a target to the far left, the right loudspeaker must emit a signal earlier and louder, to compensate for the greater distance (see section 5.2.3). Although the target location is technically a point, in many cases the region of the audience receiving the desired cues has a significant extent front-to-back. Consequently, most members of the audience will receive near optimal spatial cues at some point during the path's duration, though some later than others. This would be problematic if, for example, we wanted to present a sound effect synchronously with a visual. In many nonrepresentative musical contexts, however, this is far less of a concern.

1.2.3. Overhead Loudspeakers

A novel assertion of this work is that elevated loudspeakers offer numerous advantages for large-venue spatial audio, when compared to ear-level speakers. Most importantly, they can be placed symmetrically with respect to the audience area, which distributes the quality of spatial cues most evenly across the audience. Because they can be located more freely while retaining symmetry, they provide more flexibility in controlling the soundfield properties that influence reproduction quality. The area of the target zone can be maximized according to a chosen measure, or a balance found between several competing factors. Additionally, HRTFs for higher elevations are generally flatter, and therefore easier to model and compensate for. Overhead loudspeakers also afford a path to each listener that is largely unobstructed by other audience members. This is not usually addressed theoretically but is a very real

practical issue. The advantages of overhead loudspeakers are supported throughout chapter 5, particularly in section 5.2.2. It should be noted however that the signal processing described in chapter 4 is equally applicable to non-elevated speakers.

1.2.4. Spherical Head Model

With multiple listeners of varying anatomy, using exact HRTFs is not even possible, let alone practical. Therefore the system must rely on some form of synthetic HRTF, usually termed a Directional Transfer Function (DTF). Though seemingly inferior to a true HRTF, DTFs offer a number of advantages [29, 89]. The transfer functions for the system described in this document are derived from a simple spherical approximation, detailed in section 4.1. The most attractive feature of this model is that its response varies very smoothly, across both frequency and direction. Approximations with sharper features that are numerically closer to real HRTFs potentially perform better when the listener is in the target location, and when the model is a good match for a given listener. When either of these is not the case however, these same features create spectral artifacts and phase distortions. The spherical head model tends to degrade smoothly without "drawing attention to itself." Other advantages include computational efficiency and the possibility of deriving analytical results. The model can also be used to synthesize spatial cues, and we extend this to include cues for sound sources close to the head.

1.3. Organization of this Document

This dissertation consists of six chapters and two appendices. Chapter 1, "Introduction," provides an overview of spatial audio, as well as aesthetic and technical motivation. It then

briefly outlines the proposed solution. Chapter 2, "Historical Review of Prior Approaches," documents the various methods of spatialization in large spaces that others have developed. Chapter 3, "Perceptual and Technical Background," summarizes the relevant existing literature on spatial hearing. Chapter 4, "System Implementation," describes the technical details of the DSP algorithms used. Chapter 5, "System Evaluation," provides evidence supporting the success of the system in meeting the stated goals. Finally, chapter 6, "Conclusions," summarizes the results and suggests areas of possible future research. Appendix A provides some standard audio signal processing definitions. Appendix B describes methods of obtaining decorrelated audio.

CHAPTER 2

Historical Review of Prior Approaches

The rapid progress of technology in the 20th century allowed sound to be amplified, controlled and projected into space. This new power naturally lead many to consider the interrelationship of space and sound. Even as early as the 1880s, experiments at Bell Labs investigated binaural sound techniques [32]. But it would take another 70 years before infrastructure and technical advancements permitted any serious widespread use of spatial audio. In this chapter we investigate both the quantitative (technical) and qualitative (aesthetic) aspects of numerous spatialization efforts since the mid-1950s.

2.1. Spatial Sound for Cinema

While acknowledging the live music of "silent" pictures, film sound truly began with the 1927 release of Warner Bros.' *The Jazz Singer* [32]. The soundtrack was monophonic. In 1935, Alan Blumlein created the first stereo short films. His work was not immediately noticed, and stereo films were not widely used. However, development continued and film sound was soon to experience a drastic leap.

One consideration specific to cinema sound is the need for a center channel. This is motivated by the overriding need for reliable speech intelligibility. In early trials, a typical two-channel format was tried. Experimenters quickly confronted "the law of the first wavefront" (see section 3.2). This psychoacoustic phenomenon describes what happens when the same signal arrives from two separate sources. In this case, the two-channel setup works well for

centered listeners. For those on either side, however, the far speaker's signal is substantially delayed because of the width of a typical movie theater. This delay is large enough that the far signal is heard as an echo, causing interference with the near signal and hindering sound quality [17]. Because speech recognition is so critical for movie viewing, this interference was deemed unacceptable. The solution was to retain a mono center channel for dialogue, reserving the left and right channels (and later surround channels) mostly for music and sound effects. Although center-channel dialogue cannot move with the onscreen action, the psychoacoustic "fusion" effect suppresses this discrepancy [25].

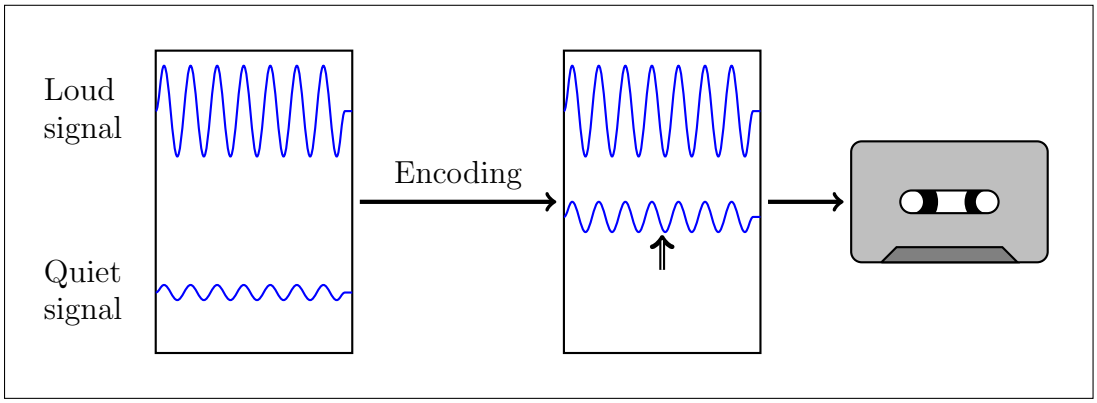
2.1.1. Early Surround Technologies

Cinematic surround sound actually dates back to the 1940 release of Disney's *Fantasia*. The composer, Leopold Stokowski, was inspired by experiments at Bell Labs. He and Walt Disney spearheaded the effort to include surround technology in the movie. Their solution, dubbed Fantasound, was actually quite sophisticated. It employed three front channels (left, right and center), three surround channels, and a number of auxiliary control channels, to control relative gains automatically. The project resulted in the development of the pan-pot. Unfortunately, the technology was proprietary and wildly expensive, with the result that only two complete systems were ever sold [78].

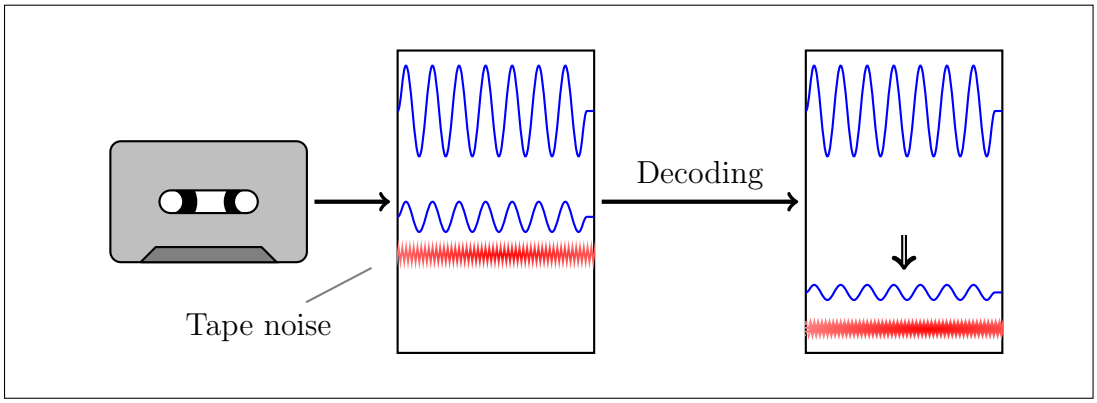
In the early- to mid-1950s, several competing surround formats sprung up. This was primarily driven by a desire to move to widescreen formats; sound was a secondary consideration. Cinerama debuted in 1952 with the film *This is Cinerama*. It featured five front channels (left, mid-left, center, mid-right and right) as well as two surround channels, switchable between left- and right-surround or mid- and side-surround. The MS configuration could be used to

project sounds "above" the audience. Cinerama was a three-projector widescreen format, and because of the expense and maintainance involved, it was essentially abandoned in 1963. A competing format, Cinemascope, appeared in 1953. Like the later Dolby Stereo, it used four channels: left, center, right and mono surround. Advances in lens technology quickly dated Cinemascope, and by 1967 it was completely out of use. Finally, a third format called Todd-AO appeared in 1955, debuting with the release of *Oklahoma!* The system combined the five front channels of Cinerama with the mono surround of Cinemascope. Studios stopped releasing Todd-AO films in 1971, but the format influenced later work at Dolby Labs. All three of these formats used magnetic soundtracks, which at the time offered fidelity superior to optical (though more expensive to produce, and more prone to damage) [72].

The 1960s saw few advances for the end consumer, but did produce one major technical breakthrough: Dolby-A noise reduction, released in 1965. The general principle is as follows: filmstrip audio noise is more or less white noise, with equal energy across equal linear frequency bins. But hearing is logarithmic, and each octave doubles the linear frequency width, so each higher octave has double the energy—in other words, white noise is most noticeable in higher frequencies. Loud sounds obscure the noise because of the psychoacoustic *masking* effect, but quiet sounds are close to the noise floor and therefore lose definition. We can overcome this with an encode/decode scheme that raises the level of quiet sounds when recording, then lowers their level appropriately on playback (see figure 2.1). This can be accomplished by a compressor (with makeup gain) on encoding, with the inverse expander as a decoder (the pairing is called a *compander*). Indeed, this simple scheme does work, but it introduces a new problem. The noise floor now rises with loud passages and lowers during quiet passages—it is correlated with the signal, making it far more noticeable (this is called *noise modulation*).



(A) Encoding



(B) Decoding

FIGURE 2.1. Conceptual view of a basic noise-reduction scheme. After [37].

The solution is to perform noise reduction independently across several frequency bands. Default behavior (i.e., behavior with no input signal) is to perform maximum noise reduction in each band. When the amplitude in a specific band becomes greater, noise reduction in that band is lowered, but masking obscures the correlation. Other bands are unaffected, and no (audible) modulation occurs. Dolby A noise reduction uses four frequency bands: 9 kHz highpass; 3 kHz highpass; 80 Hz–3 kHz bandpass; and 80 Hz lowpass. The scheme actually uses "bilinear" compression, yielding unity gain for low- and high-amplitude signals, and $2:1$

compression for middle levels. This is done to avoid audible artifacts during abrupt volume changes. The system offers a fixed 10 dB noise reduction, small by today's standards but a major breakthrough at the time [37, 41, 60].

2.1.2. Dolby Stereo

Dolby noise reduction allowed stereo optical tracks with an acceptable noise level. Naturally studios tried to add even more optical tracks, but the decreased track width and resulting alignment issues, together with increased noise, made this impractical. However, new technologies known generally as *matrix encoding* were being developed to store four logical channels using only two physical tracks. In 1967 Peter Scheiber, a Masters student at Indiana University, worked out the basic equations [120]. On the consumer side, this technology became the infamous "quadraphonic sound," actually a number of incompatible formats. Dolby decided to adopt the process to the needs of cinema by incorporating the center channel. In 1976 Dolby released the technology, then called Dolby Stereo (later renamed Dolby Surround). Originally for 35mm optical soundtracks, a similar system was soon released for 70mm magnetic tracks. It consisted of left, right and center stereo channels and a mono surround channel. The left and right channels are recorded normally. The center output is recorded on both physical channels, down 3dB. Using a dedicated decoder, the common information is extracted and sent to the center channel. Conveniently, since the two copies are in phase, standard stereo playback simply results in the familiar phantom center channel. The surround channel is encoded similarly, but the two copies are 180° out of phase [40]. On mono playback, the out-of-phase signals cancel and the surround channel is not heard (one good reason not to put crucial information in the surround channel). In a typical theater

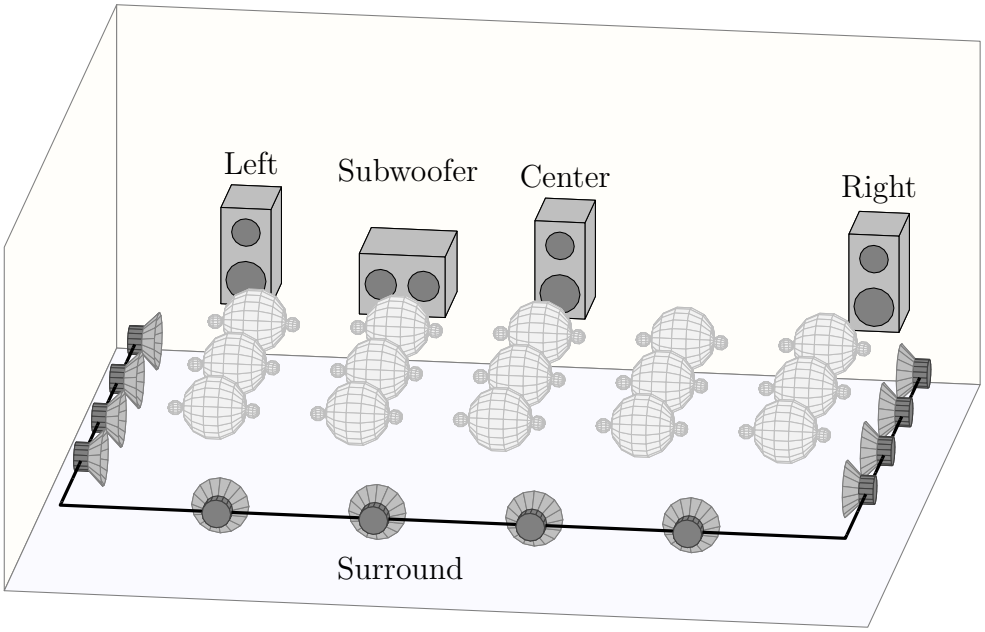


FIGURE 2.2. Typical Dolby Stereo environment. After [38].

environment, the front speakers are positioned normally. The surround channel is heard from an array of speakers across the side and rear of the space (see figure 2.2).

The physical compatibility of Dolby Stereo with existing equipment (requiring only new speakers and decoders/amplifiers, rather than new projection equipment) resulted in the rapid widespread adoption of the new system. In 1977 *Star Wars* became the first major release with full Dolby Stereo sound, obviously another economic influence. In 1986, Dolby Stereo was replaced with Dolby SR, essentially the same system with superior noise reduction, incorporating elements of Dolby A, B and C (B and C noise reduction are consumer formats). The system uses both the fixed-band, variable-gain systems of Dolby A and the sliding-band, fixed-gain approach of Dolby B/C. The resulting combination offers noise reduction approaching 25dB [41, 42].

However, this technical improvement could not overcome the scheme's major drawback: a mono surround channel. The goal of the surround channel was to provide an "enveloping" listening experience, but Dolby Stereo failed to accomplish this because of a psychoacoustic principle called the *precedence effect* (see section 3.2). When a listener hears a sound followed by a slightly delayed copy, the later copy is suppressed and not heard consciously. Without the precedence effect, a sound and its reflection would become two distinct events. Under Dolby Stereo, a listener first hears the surround signal from the closest speaker, and sound from further speakers is then suppressed—not "enveloping" sound but a single point source! Overcoming this problem requires *decorrelated* signals: two signals that sound the same as macro-events but possess independent micro-structure, such as happens with natural reverb or mono-to-stereo reverb units (see section 3.5). Decorrelation in turn requires stereo surround channels, and so Dolby Stereo ultimately fails as a surround scheme (but certainly not as a commercial product). A secondary weakness of Dolby Stereo and SR is that the surround channel is bandpass filtered between 100 Hz and 7 kHz [40]. The highpass filtering is not much of a problem, given the current near-universal use of subwoofers, but the loss of high frequencies does limit fidelity. Finally, all matrix encoding schemes suffer from problems of channel separation and possible phase issues. Good decoders can overcome much of this, but suboptimal encoding or any equipment misalignment can create audible problems [32].

2.1.3. THX

Strictly speaking, THX is not a format but rather a quality certification program. It was developed in 1982–83 by Tom Holman at Lucasfilm to ensure acceptable technical infrastructure for *Return of the Jedi*. THX does not specify a particular surround encoding

scheme, but it does mandate certain acoustic features: isolation from outside noise, non-parallel walls with sound-dampening material, low HVAC and background noise (below NC-30), extended frequency bandwidth and a perforated screen to allow center-channel fidelity. The system also specifies a sophisticated fourth-order crossover [32]. THX does not specifically guarantee a sophisticated surround sound system, but it does at least ensure a moderately high-quality screening environment. Though its impact is difficult to measure precisely, it certainly was a major factor in the widespread deployment of quality cinema surround systems (if for no other reason than the industry clout behind it).

2.1.4. Digital Surround Sound

Matrix encoding schemes were tremendously successful at their job: moving surround sound into the cinematic mainstream, cheaply and effectively. Nonetheless, from a technical standpoint there was much to be desired, and as the 1990s approached Dolby Stereo was showing its age. There were a number of attempts to create movies with a 5.1-channel soundtrack. The first was actually *Superman* in 1978, using six magnetic tracks on a 70 mm print. In 1979, *Apocalypse Now* had a limited surround-sound release. But the track count was simply too high for analog techniques, even with advanced Dolby SR noise-reduction. The first digital soundtrack was called Cinema Digital Soundtrack (CDS). Backed by Kodak and Optical Radiation Corporation, it premiered with *Dick Tracy* in 1990 but quickly failed.

In 1992 Dolby released the Dolby Digital format (also called Dolby SR-D), first used on *Batman Returns* and *Star Trek VI*. The scheme allows five full-bandwidth (3 Hz–20 kHz) channels, plus a low-frequency channel limited to sounds below 120 Hz [39]. Dolby Digital is encoded with a lossy encoding scheme known generally as AC-3. This is one of a large number

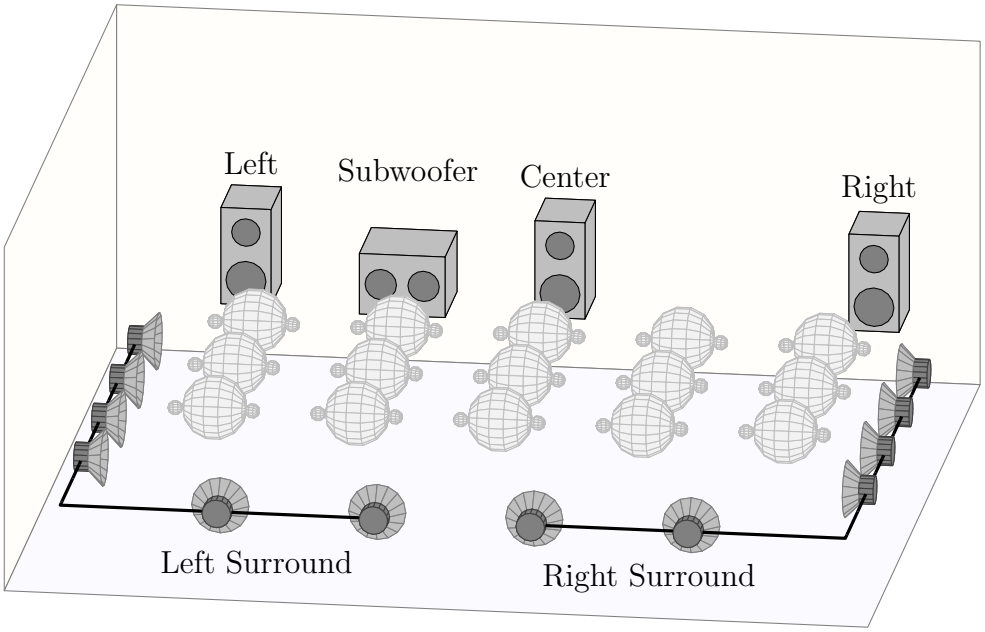


FIGURE 2.3. Typical Dolby Digital speaker placement. After [38].

of *perceptual codecs* that attempts to reduce data rate by discarding information that the human hearing mechanism will not perceive. Higher or lower amounts of compression can be achieved by keeping more or less data. To maintain universal compatibility, Dolby ingeniously left the analog track untouched, placing the digital data between the sprocket holes on the outer edge of the filmstrip. A CCD reads the optical data, converts it to electrical impulses, and sends it to the decoder. However, the drawback to interleaving data with the sprocket holes is that physical track space is halved, requiring aggressive compression. Cinematic AC-3 runs at a fixed rate of 320 kb/s [132]. Figure 2.3 shows a typical speaker setup for a Dolby Digital theater.

Competing formats quickly emerged. In 1993, Digital Theater Sound (DTS) debuted with *Jurassic Park*. Rather than storing audio data directly on the filmstrip, it was placed on separate CD-ROMs. An optical sync code is placed between the analog soundtrack and the

visual data. Like Dolby Digital, DTS uses the 5.1-channel format. The theatrical standard (technically known as apt-X100) does not use perceptual coding (later consumer variants do, at least for low bitrates). Although references are sparse due to its proprietary nature, a few papers are available [79, 124]. Basically, DTS divides the audio into four subbands and uses Linear Predictive Coding to compress each band. More accurate predictions allow greater bit-reduction; below a certain threshold the prediction is just as costly as the original data and is hence disabled. The net result is an 882 kb/s data stream with approximately $4:1$ data compression. DTS is still a lossy format, but some subjective accounts suggest that it offers superior quality to Dolby Digital. One study finds DTS superior to AC-3, but in a consumer context and with consumer codec parameters, limiting its applicability [8].

A third format, Sony Dynamic Digital Sound (SDDS) was first used for *The Last Action Hero*. It places the audio data outside the sprocket holes, printed during the cyan ink process. Hearkening back to the glory days of Cinerama, SDDS uses five front channels and discrete stereo surrounds, as well as a LFE channel. For smaller venues the front channels are downmixed to the usual left, center, right. Historically the full eight channels are rarely used [123]. SDDS uses Sony's ATRAC1 perceptual codec, achieving about $5:1$ compression. Although some believe SDDS offers the highest technical quality of the three original formats, cost of entry and industry inertia have kept its market share lowest. However all three formats (Dolby Digital, DTS and SDDS) are in active use today. Figure 2.4 shows a filmstrip featuring all major analog and digital soundtrack formats.

Further refinements have led to modest improvements in some codecs. Released in 1999, Dolby Digital EX uses matrix technology to add a rear surround, embedded with the left and right surrounds [36]. In response, DTS released a matrixed 6.1 format, dubbed DTS-ES, as

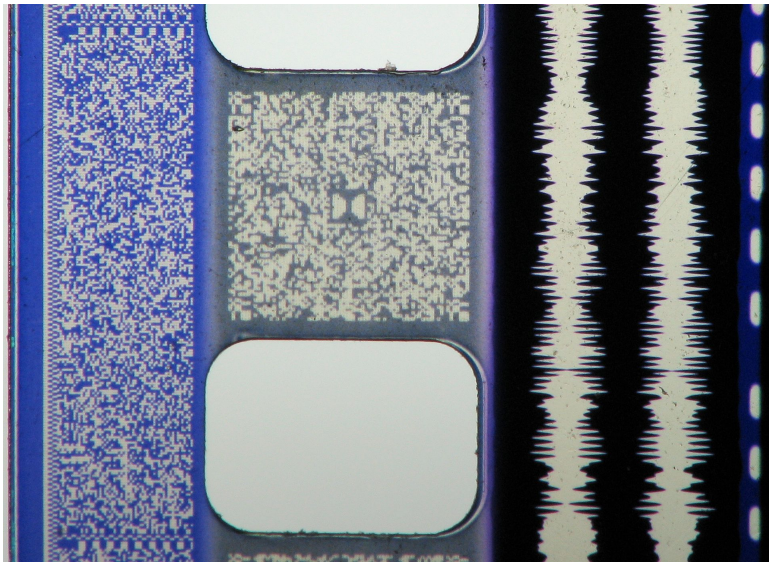


FIGURE 2.4. Filmstrip featuring multiple soundtrack formats. Outside-to-inside (left-to-right): SDDS, Dolby Digital (note the embedded Dolby logo), analog optical, DTS timecode [126].

well as a version using a true discrete sixth channel [35]. Numerous variants of both formats have been released for home use, featuring increased bit depth, sampling rate, and discrete channels, but their implications for cinema sound are not yet clear. In addition, many new formats will require major equipment upgrades for many movie theaters. The most obvious chance to do this in a widespread manner seems to be with the adoption of digital cinema, but that technology is still not widespread. One other workaround could be DTS, which—because it stores the sound on separate discs, using the filmstrip only for sync—might offer an easier upgrade path. Some recent proposals suggest a radical expansion in the number of channels. The best known is the "10.2" format [136, 140], created by Tom Holman (inventor of THX). The system features stereo subwoofers for more diffuse bass, and elevated speakers to add height (see figure 2.5). The commercial viability of 10.2 is questionable however.

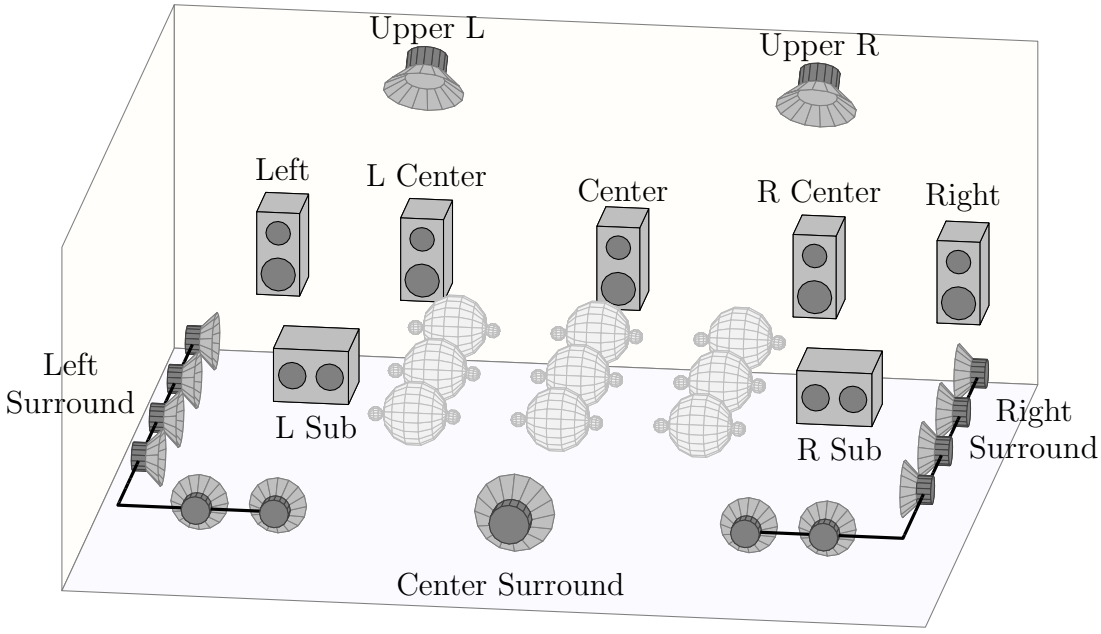


FIGURE 2.5. Speaker layout for the 10.2 format. After [136].

2.1.5. Low-Frequency Effects

In film sound, very low frequencies are often used for effects—earthquake rumblings, explosions, etc. Portraying these with sufficient sensory impact is therefore important. Because regular speakers are not always large enough to handle these frequencies, dedicated subwoofers are often used. Since directionality does not depend strongly on extreme bass frequencies, surround systems often use only one subwoofer [17]. There are two ways to generate subwoofer signals. One is to use a crossover after the soundtrack decoder. These crossovers simply split the frequency range into two bands, sending the low band to the subwoofer and the other band to the regular speakers. This is used in Dolby Stereo, for example. Digital formats tend to incorporate a dedicated, bandlimited channel sent directly to the subwoofer. Research by William Martens shows that two subwoofers provide a more diffuse, enveloping image [88],

but very few real-world systems incorporate this. Two exceptions are 10.2, above, and IMAX, below.

2.1.6. IMAX

It is worth mentioning, if only briefly, the sound systems for IMAX, the most well-known high-resolution large-venue film format. IMAX was developed in the late 1960s and is often used at parks, museums, and city centers. IMAX screens occupy a large portion of the viewer's field of vision, and the films are often designed to impress the senses. Originally IMAX used a six-channel magnetic tape. This has largely been replaced with the DTAC six-channel digital audio system, notable for using no compression and full audio bandwidth. The speakers follow a 5.1 format with the interesting addition of a sixth, elevated center channel. IMAX specifies an NC-25 noise rating and high absorptivity, and uses speakers with a dispersion pattern designed to match the shape of the theater [33, 68]. Recently, IMAX introduced the Personal Sound Environment (PSE) consisting of two speakers on their 3D glasses, providing customized sound:

> Binaural sound emanates from the headsets' two small speakers, just above and slightly in front of your ears; they cover all but the frequencies below 100 Hz. Low bass is handled by a pair of subwoofers behind the giant screen. Four full-range speakers, also behind the screen, keep sounds tied solidly to the film's images even if you turn your head; if you have trouble imaging binaurally (as some people do), these speakers will prevent front sounds from seeming to come from the sides or rear. Two more speakers, in the rear of the theater, carry only surround ambience; the headset's binaural speakers

> carry sounds that are supposed to originate behind you. Eight channels of an 18,000-watt, 10-channel amplification system feed the speakers; the other two channels feed the binaural signals to the headsets. These amps are fed from four audio CDs, computer-synchronized with one another and with the projectors. The headsets can receive four separate soundtracks, so a movie could be presented in different languages simultaneously if the theater provides enough channels [15].

2.2. Spatial Sound for Art Music

Spatialization for cinema has been driven by the relatively straightforward and homogeneous goals (and large budgets) of the major motion picture industry: namely, consistent deployment, enveloping soundfields and high-impact special effects, while remaining largely subservient to the visuals. The story of spatialization for art music (and the most relevant subset, electroacoustic music) stands in stark contrast. For one, there is far more work by individuals and small groups. On the downside, this makes funding a real issue; even major collaborations like IRCAM have miniscule budgets compared to the film industry. On the positive, it allows for highly customized solutions and real innovation. Since the audience is smaller, performances are fewer and the composer is often involved with staging the work, standardization is much less critical. The goals of art music also lead to different solutions: localization becomes a less overriding goal, replaced by a desire to create or augment an aesthetic experience, or even make philosophical statements about sound and music. The non-issue of speech intelligibility and the lack of a concrete visual anchor allow for much more radical diffusion, but certainly composers care as much or more about sonic micro-detail as filmmakers. Rather than a

spatialization that is fixed, sound diffusion is often conceived of as an active part of the live performance. Finally, there is not even uniform agreement as to whose needs should be served—composer, performer or listener. As with all music, weighting these needs differently will lead to different solutions.

2.2.1. Early Work

One of the first efforts at electroacoustic spatialization was a sophisticated system used by Pierre Schaeffer and designed primarily by his technician, Jacques Poullin, in the early 1950s. Motivated by a new five-track tape recorder, the system used four output channels: left, right, (center) rear surround, and a fourth directly overhead. Special wide-focus loudspeaker cones were used to compensate for suboptimal listening positions. Four of the tape channels had fixed spatialization. The fifth could be steered in realtime to any speaker, using an absolutely ingenious device dubbed a *potentiomètre d'espace* (or *pupitre d'espace*):

> "[It] consisted of a small hand-held transmitting coil, and four wire receiving loops arranged around the performer in a tetrahedron, representing in miniature the location of the loudspeakers in the auditorium. Moving the coil about within this receiving area induced signals of varying strengths in the loops, this information being applied to attenuators regulating the distribution of the fifth track between the four channels [87]."

Figure 2.6 shows another electroacoustic pioneer, Pierre Henry, operating the device in concert.

Karlheinz Stockhausen experimented with a number of spatialization techniques. Beginning with *Gesang der Jünglinge* in 1956, all of his studio electronic pieces used multi-channel

FIGURE 2.6. Pierre Henry at the potentiomètre d'espace, 1955. © Ina/ Maurice Lecardent.

formats [61]. That work was originally conceived for five speakers in a circle about the audience and one overhead for the boy's voice, but the overhead was ultimately eliminated [85]. Many of his later pieces, such as *Kontakte*, use a four-channel format. *Kontakte* was also his first piece to treat the channels as adjacent and move sounds continuously between them. This was accomplished with the aid of his invention, the *rotation table*: a mono speaker was mounted on a rotating turntable, surrounded by microphones fixed at the edge of the table. As the center speaker rotated, its sound was picked up by the outside mics and hence could be rotated across their individual outputs. Although the piece uses only four channels, Stockhausen required two speakers per channel, facing in different directions, to achieve what he considered a satisfactory effect (see figure 2.7).

In the mid-1960s, Stockhausen turned his attention to live electronic performance. One of his first attempts, *Mikrophonie I*, clearly dealt strongly with space, requiring performers to move microphones across a large tam-tam to amplify different resonances. However, the

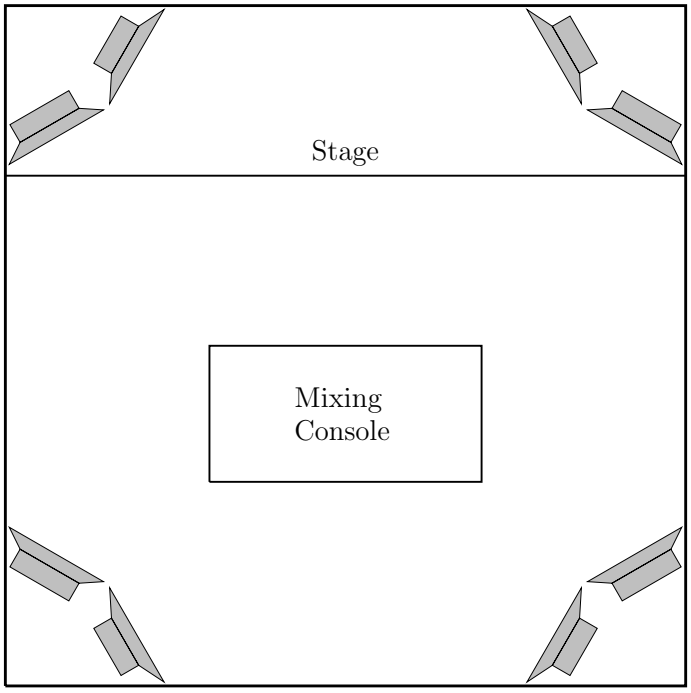


FIGURE 2.7. Speaker setup for *Kontakte*. After [129].

actual projection was in stereo. For *Mikrophonie II*, four independent channels were used, but spatialization relied in large part on the physical location of the performers (12 singers sit in a semicircle, their backs to the audience, facing an elevated organist). An interesting piece is 1968's *Musik für ein Haus*: musicians were situated throughout four different rooms on two floors. The sound from each room was mic'ed, processed and fed into the other room. In the basement, sound from all four rooms was played over four channels. Listeners wandered freely throughout the house [141]. The idea reached a new high point in 1969 at the Beethovenhalle in Bonne. Four orchestral groups throughout three concert halls and the intervening corridors and lobbies performed for four hours. In addition to amplified live orchestral music, the evening included numerous rare multi-channel play-backs of studio and

live electronic work. A newly composed work, *Fresco*, filled the public spaces with sounds that "colour an environment acoustically but are themselves relatively featureless [85]."

In 1968 Stockhausen was invited to consult on the construction of the West German Pavilion at the 1970 World's Fair Expo in Osaka, Japan. It was designed to feature live performances of many of his pieces. The end result was a spherical performance space. In the score to *Spiral*, Stockhausen describes the details of the building and sound reinforcement:

> "In the auditorium the public sat somewhat below the level of the sphere's equator on a circular, sound transparent platform .... The diameter of the auditorium was c. 28 metres....
>
> The soloists in SPIRAL played or sang either on a podium at one side ... or on one of six balconies ....
>
> All instrumental or vocal sounds and all short-wave sounds were fed into a control desk ... and from there they were projected into the auditorium over 50 loudspeakers.
>
> The loudspeakers were arranged in 7 circles one above another giving 10 vertical rows of speakers. At the control desk 14 microphone inputs could be switched in any combination to 7 master channels by means of 14×7 push-buttons—and they could be switched around in any way even during a performance.
>
> ... The voice/instrument and/or short-wave receiver were connected to a so-called *rotation mill*. This rotation mill had 1 input and 10 outputs, and the sound fed into it could be connected with the 10 contacts for the outputs

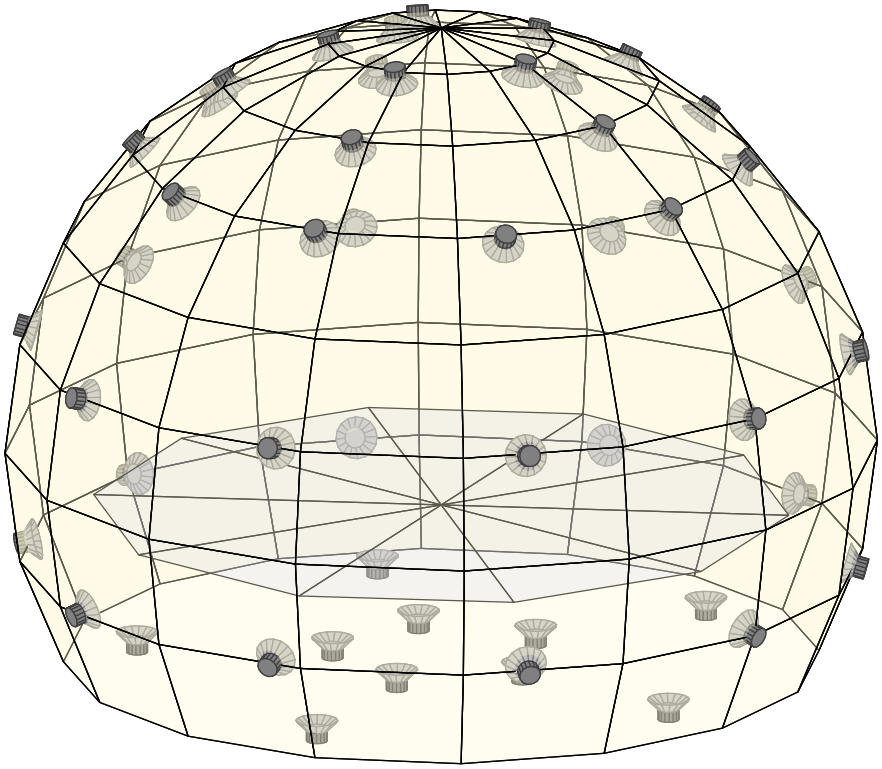


FIGURE 2.8. Speaker distribution used by Stockhausen at the 1970 World's Fair West German Pavilion. The lowest levels are below the acoustically transparent audience platform. After [128].

consecutively by means of a handle turned manually in a circular fashion. The maximum speed of rotation was c. 4 revolutions per second.

The outputs of each of the 7 master channels could be plugged into any combination of loudspeakers, whereby the 10 output plugs of the rotation mill were connected to 10 loudspeakers: either in a circular pattern ... or in a spiral pattern ... or in a diagonal loop pattern [128]."

Figure 2.8 shows the distribution of loudspeakers.

Also in 1969, Stockhausen was commissioned to design and compose for a "sound tunnel" in Los Angeles. The final design included 200 small speakers and several large speakers

at either end. The tunnel was 40 feet long by 6 feet high by 3.5 feet wide. Stockhausen's sound design caused the signal to spiral around and down the tunnel, taking 16 seconds to complete the trip. Most of Stockhausen's subsequent work is concerned with space in one way or another, from the cosmic orchestra of *Ylem* (1972) to the sprawling epic opera *LICHT* (composed between 1977 and 2002, comprising 29 total hours of music [127]). Stockhausen's use of space is clearly centered on the interrelationships between sound, music and space, often blurring the boundaries between purely musical composition and sound installations.

Other composers pursued various spatialization schemes. Iannis Xenakis composed for four tracks from his earliest electronic work, *Diamorphoses* (1957). Later works such as *Hibiki Hana Ma* (1969–70) used as many as 12 tracks [142]. Morton Subotnick created a number of pieces for the quad format; later at IRCAM he explored real-time spatialization control [63]. John Cage, in his typically prescient way, quickly made spatialization an integral part of the performance. An early example is *0′00″(4′33″ No. 2)* (1962). The piece originally consisted of a single instruction: "In a situation provided with maximum amplification (no feedback), perform a disciplined action [113]." At a May 1965 concert, this translated to Cage on stage, typing letters with microphones magnifying his actions and sending them about the hall. In *Variations VII* (1966), contact mics attached to four assistants as well as radios, TVs and telephones were routed to 17 different speakers. In *Variations VI* (1966), loudspeakers become represented by tokens on a page, connected by chance operations to different inputs [47, 113]. Cage's work leads inevitably to the world of Happenings, performance art and installation pieces, and the valuable contributions they make to the dialog between sound and space.

2.2.2. *Poème Électronique*

Indisputably, one of the most ambitious and influential spatialization efforts was the Philips Pavilion at the 1958 Worlds Fair in Brussels. The Pavilion was a collaboration between the composer Edgard Varèse (with additional material from Iannis Xenakis), the architect le Corbusier, and the filmmaker Philippe Agostini. It was conceived by executives at the Philips company as a venue to showcase their advances in lighting and sound systems, but the artists involved had grander visions: the first "electronic poem." Le Corbusier convinced Philips to commision Varèse and give him virtually free reign with the sound. His final result was 480 seconds long (the only external requirement given to him), and was strictly a tape piece. Varése had thought deeply about sound and space since his earliest compositions, and he exploited the possibilities of the Pavilion to the fullest. In this he was assisted by Willem Tak, the project's acoustical advisor, who was largely in charge of the actual technical details of the spatialization. In the final design, tweeters were arranged in clusters along the walls of the interior; subwoofers were concealed at floor level. The tweeters were arranged along certain "sound routes," some following the curving walls at audience level, and others ascending linearly to the room's apex (see figure 2.9). The exact speaker count is not known—estimates range from 150 to 450, with 300 a plausible guess. The interior of the space was sprayed with absorbent asbestos, creating a very dry room; hence most ambience was electronic. The actual sound was stored mainly on a single track, with two additional tracks for "reverberation and stereophonic effects." A 15-track control tape was used to automate spatial distribution via twenty 120 W amplifiers [133]. Reaction to the final experience was decidedly mixed, but the Pavilion's innovation (it is a strong contender for the first modern multimedia work) has become clear with time. Though *Poème Électronique* is considered a classic work of

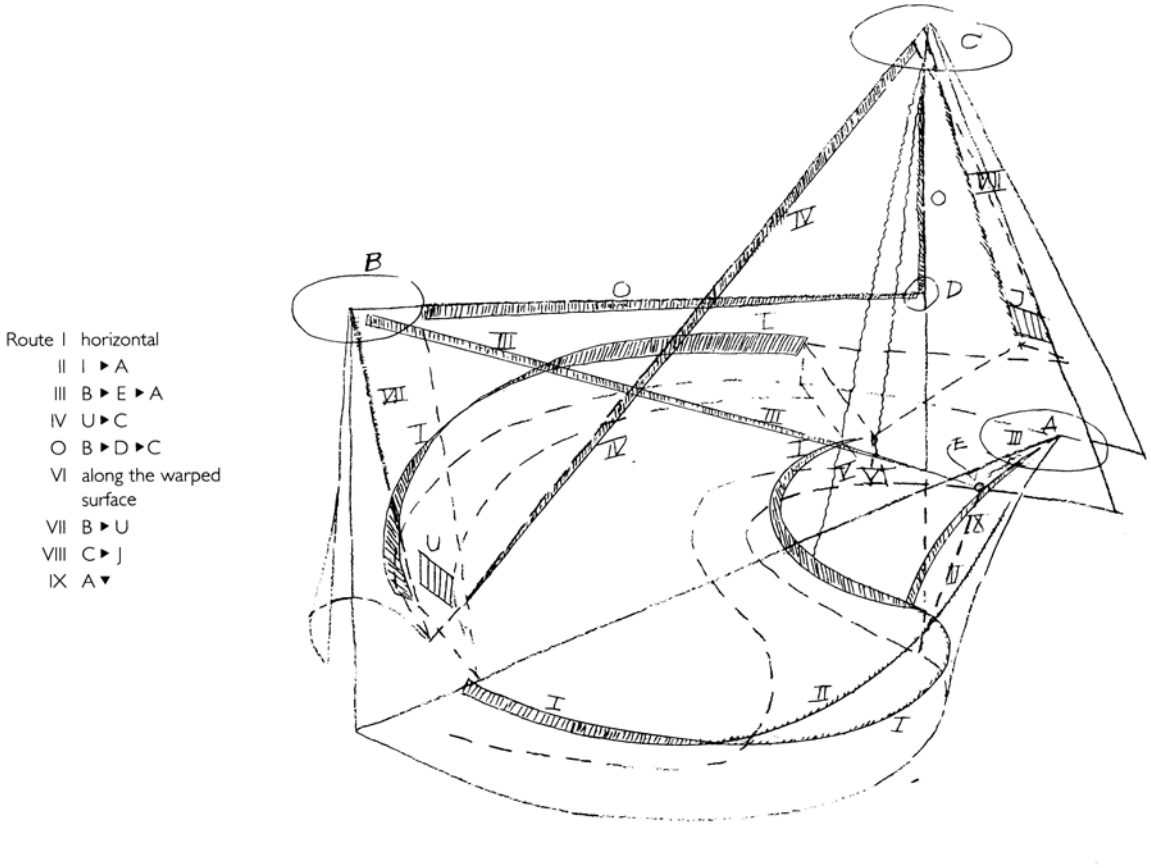


FIGURE 2.9. Diagram of the "Sound Routes" within the Pavilion. © Philips/ Iannis Xenakis.

electroacoustic music, it was composed expressly for a specific space, and only a few have ever experienced it as such. Several projects have attempted to recreate the Pavilion; a recent effort using Ambisonics has the potential to be the most successful [144].

2.2.3. Acousmonium

One of the first designs for a general-purpose, reusable electroacoustic diffusion array was the Acousmonium, shown in figure 2.10. It was premiered in 1973 by its creator, François Bayle, an early member of Pierre Schaeffer's "Groupe de Recherches Musicales," or GRM [64].

FIGURE 2.10. François Bayle and the Acousmonium. Taken during "Les Concerts du Patrimoines," Salle Olivier Messiaen, Maison de Radio France, Paris, December 1980. © Ina/Ruszka Laslo.

Prior to that, Bayle had been using a six-speaker setup (front, side and rear pairs) [34]. The Acousmonium is usually described as a "loudspeaker orchestra" involving over 80 individual speakers of different types, placed at varying heights and distributed about the stage. Describing its effect, Bayle stated: "It puts you inside the sound .... It's like the interior of a sound universe" [125]. Unfortunately, the Acousmonium is most well-known in France, making English-language references difficult to find. A few additional details are given in [99].

2.2.4. Gmebaphone and Cybernéphone

Beginning in 1973, researchers at the Groupe de Musique Expérimentale de Bourges (GMEB) began developing an elaborate system for live diffusion, dubbed the Gmebaphone (renamed the Cybernéphone in 1997). Figure 2.11 shows the earliest version of the system. An article by Christian Clozier, the main force behind the project, describes the history, technology and aesthetics of the Gmebaphone [28]. The layout primarily consists of four speaker arrays, called "V" arrays (see figure 2.12). Each array has six left and six right speakers (twelve total). The speakers are not identical but are specialized for one of six frequency registers. V1 is the main array and serves to fill the entire hall with sound. V2 serves to augment V1; "Its relationship to V1 ... is that of establishing variations in responses, in expanding and reducing, and creating 'zooms' and macros." V3 encircles the audience "to resynthesize acoustic space in the listener's head rather than that of the hall." Finally, W4 (a "double" array with 24 speakers) creates a vertical plane of motion.

In addition to the V arrays, there are two auxiliary groups called "reference networks." Their purpose is to "configure and reconfigure spaces that are conventional, arbitrary, or paradoxical. These spaces illuminate and enhance the V systems." One reference network has four stereo pairs and corrects for room idiosyncrasies; the second network has three pairs and defines distance. The Gmebaphone is now operated via a custom digital console (obviously the early iterations were analog) with 36 faders and two screens for additional operations. The system is explicitly intended for live use, but automation is also available (ostensibly to allow use by composers who cannot actually be present).

Clozier makes numerous references to the aesthetic intentions of the system:

FIGURE 2.11. Gmebaphone 1, 1973 [28]

"Diffusion and interpretation involves transmitting the work to the audience, enhanced by an instrument such as the Gmebaphone, that allows the performer a personal interpretation while taking into account the expectations of the audience.... The Cybernéphone may be defined as a huge acoustic synthesizer, an interpretation instrument that the composer plays in concert.... It is not a question of putting the music into motion, but of allowing the spaces contained within the music to unfold and be revealed.... When the Cybernéphone is properly played, the ear cannot pinpoint any single sound-source. Instead, spaces and the relationships between them are heard. The loudspeakers on stage make up an ensemble of abstract volumes in which the music is generated, that movement of colored time developing its own space."

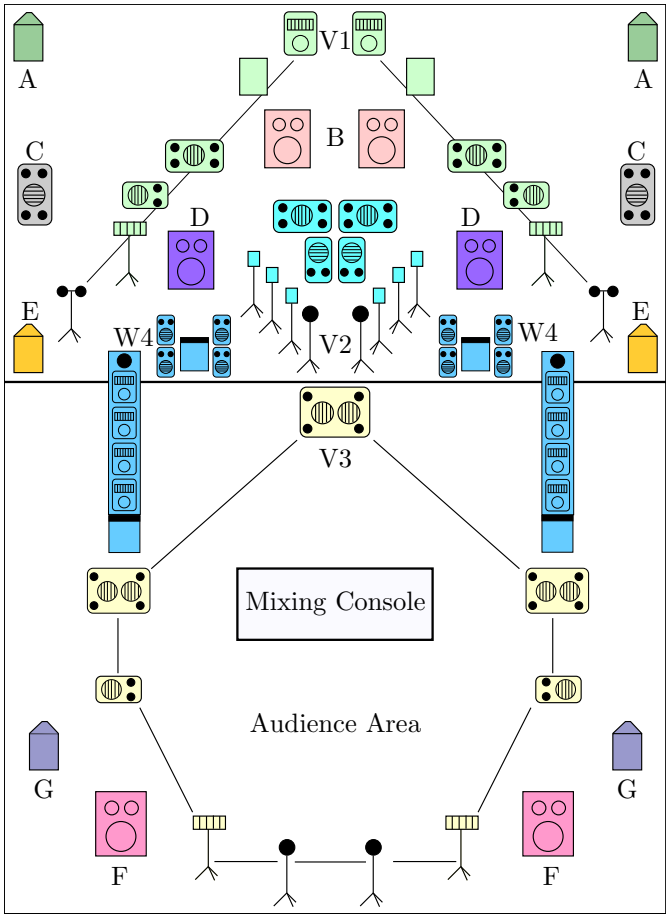


FIGURE 2.12. Overall layout of the Cybernéphone. After [28].

Although Clozier does an admirable job of describing the system and its goals, he does not fully justify its effectiveness or the necessity for such a complex design. For example, simply stating that V2 "creat[es] 'zooms' and macros" is not very helpful: what are these effects? Why does this relationship between V1 and V2 help project them? Why would simpler setups be ineffective? Clearly the goal is subjective experience and not quantitative placement of sounds in objective space. In that sense, the system is successful if it creates an enjoyable listening experience. But the perceptual impact could use much more concrete treatment.

2.2.5. BEAST

BEAST (Birmingham ElectroAcoustic Sound Theatre) is another adaptable multi-speaker system designed especially for live electroacoustic sound diffusion in public spaces (see figure 2.13). It was developed at the University of Birmingham around 1982 and is described in detail in an article by Jonty Harrison [61]. The most basic setup consists of what he calls the *main eight*: two stereo pairs, one close to eliminate a center hole and one wide to fill the space; a rear surround pair, raised 2–3 m above ear height; and a forward distance pair, also raised above ear height, placed behind the stereo pairs and angled in toward each other to maintain a distant image. More elaborate installations include subwoofers for 80 Hz and below; raised clusters of tweeters; side fills for long halls; a third stereo pair for very wide halls; additional raised speakers to add a controllable height dimension; and punch speakers for dynamic emphasis.

The entire setup is controlled in real time by a mixer operator; center faders are used for the main eight speakers, while outlying faders control auxiliary channels. Tellingly, there is no automation available. BEAST is not an "aesthetically neutral" system; the diffuser is intended to act as another performer and interpreter, not a transparent mechanical operator. This stems from the old dichotomy between *musique concrète* and *elektronische Musik*. Harrison writes:

> "Sound material approached as organic matter to be sculpted, shaped, coaxed, caressed into participating in a piece of 'sonic art' generally ... behaves well in diffusion ... because diffusion is an extension of the compositional approach.... Works exhibiting architectonic structure and space

FIGURE 2.13. BEAST being set up for its 20th anniversary concert. © Kevin Busby.

> are not well suited to diffusion, whilst those displaying organic structure and space certainly are."

Although the physical layouts are quite different, the philosophical similarities between BEAST and the Gmebaphone are clear. BEAST has been used by such artists as Christian Clozier, Larry Austin, Denis Smalley, and Francis Dhomont [24].

2.2.6. Ambisonics

Ambisonics is an unusual technology with a storied history, almost dying off before a recent rally. It was made practical primarily by Michael Gerzon in the 1970s, to overcome limitations of quadraphonic sound [52, 53]. It is an encode/decode system designed to be "reproduction-neutral:" authors create one source encoding that can be played back over a wide variety of systems, degrading gracefully on simpler setups. In practice, there is one codec family for two-channel transmission, and a preferred codec (B-format) for true spatial playback. Surround effects require at least four speakers, and eight if height is desired.

The four-channel encoding scheme is based on a spherical coordinate system with the following components: W (0-order, omnidirectional); X (front-back); Y (left-right); Z (up-down). The Z channel allows inclusion of height information. Sounds can either be recorded with spatial information using special microphone techniques, or processed for placement within the sphere. Note that when recording a signal, all spatial information is mixed, and individual elements cannot be manipulated inpedendently. Because the encoding is based on a simple transformation matrix, complex motions such as rotating about any axis are computationally easy [86]. Ambisonic decoding combines aspects of wavefront synthesis (which ideally recreates the original 3D sound field of the recording) and the psychoacoustics of localization.

Ambisonics has been used in live performance by a number of composers, including Ambrose Field [6]. There are approximations in first-order (B-format) ambisonics that are potential problems for large spaces, but solutions are being explored [7]. Further references on Ambisonics are found in Michael Gerzon's extensive bibliography [54].

2.2.7. Alternative Systems

Barry Truax has developed an 8×8 matrix router called the DM-8, compatible with both stereo and discrete-multichannel sources. The mixer is software-controlled and fully automatable [135]. The system is mostly used by Truax but a number of other composers have also explored it, including Peter Manning [134]. The Center for Research in Electronic Art Technology (CREATE), a group at UCSB under Dr. JoAnn Kuchera-Morin and Dr. Curtis Roads, is developing a system called the Creatophone. It consists of between 12 and 32 loudspeakers, and strongly features height by requiring elevated speakers. It is designed to be adaptable to different environments, and to engage the spatial mixer as an active performer [112]. A few other systems are available; several are described in [99].

2.3. Spatial Sound for Popular Music and Entertainment

While cinematic sound is characterized by standardization and the clear purpose of supporting visuals, and electroacoustic diffusion by diverse experiments, commercial live surround sound is designed mostly for entertainment and popular appeal. Unfortunately, it is not characterized by extensive documentation! Although many rock sound engineers are quite scientific, there is also a large collection of ad hoc techniques that are not catalogued. Similarly, though many working engineers have written journal articles, their primary focus is fieldwork and not publishing. Much of the available literature focuses on recording concerts for later surround-encoded release, rather than actually presenting concerts in surround. Nonetheless, some information is available.[1]

[1] Recent AES conventions (115, New York; 117, San Francisco; 119, New York; 121, San Francisco; 123, New York; 125, San Francisco) have featured a series of symposia on live surround sound [5]. However their content is not available to the author at this time.

2.3.1. Overview

There are Internet references to uses of quadraphonic sound by U2 [107] and ELP [139]. An online article mentions early quad tours by Yes, and describes a recent tour using DTS 5.1 surround sound with the standard speaker layout:

> "The mixing engineers use a specially designed surround matrix control grid, consisting of 30 joystick-like pan-pots and custom electronics to localize individual instruments and vocals anywhere in the concert venue. Creativity with taste and descrete [*sic*] sound are the biggest challenges here, as the 360-degree mix is produced in real-time during every YES performance. Due to the variables in arrival time in a concert situation, the surround sound portion on this night consisted primarily of secondary effects and other minor spatial cues worked out with the band in advance [139]."

Many other major acts are also using 5.1 surround sound, including Celine Dion [27]. A recent Neil Diamond tour used an eight-channel setup (left, center, right, left/right side, left/right rear, center rear) [91].

Two articles by Michael Miles offers a number of practical suggestions (most of which are described in various other sources) for implementing live concert surround sound [96, 97]. For example, he recommends using two arrays for the front left and right channels: a wide-angle array for near listeners, and an elevated directional array with a gain boost aimed at the far audience. The opposition of a gain boost and a longer delay time can help restore the stereo balance. Alternatively, the audience can be divided into smaller sections, each covered by dedicated directional speakers (delayed to match the main signal). He also recommends adding a center channel, and confining panning motions to adjacent channels—left-and-center

or right-and-center, but not left-and-right. Many of these techniques are based on rapid recent progress in the design of PA speaker arrays.

2.3.2. Pink Floyd

Pink Floyd was one of rock's live sound innovators. Early on they became interested in multichannel sound—meaning quadraphonic. On May 12, 1967 they became the first band in Britain to use live quadraphony, thanks to a now-legendary device called the "Azimuth Coordinator." Made by Bernard Speight of Abbey Road Studios, it was essentially nothing more than four rheostats (variable resistors) driving six speakers. It was used to project sound effects in a circle around the audience [31]. Their live sound rig advanced rapidly, and by the *Dark Side of the Moon* tour they had a highly usable setup. The album, though originally released in stereo, had been recorded with an ear toward quad. Studio engineer Alan Parsons was also given the task of adapting the mix to live performance. Rather than use the canonical quad layout, Parsons rotated the arrangement to give front, back, left and right channels (see figure 2.14). This was done for several reasons: the stereo stage PA was quite loud, and two rear quad channels could not hope to compete. Instead, the quad system was designed to augment the stereo setup. The front quad channel was raised, supporting the sense of motion. The setup was controlled by a custom Allen and Heath console with a pair of Penny and Giles quad-panning joysticks (adapted from an earlier console used by The Who). Live sounds were processed in quad:

> "Any individual instrument carrying a microphone to the mixing console can be individually switched into the quad system. This is used to the most

> advantage on quiet passages where the main stereo [PA] can be virtually shut down and the maximum use made of quad pan pots for movement [109]."

In addition, recorded material (such as the original "clock" tape loop from the studio recording of "Time") was fed through a Teac four-track (earlier a Sony model was used, and later an eight-track Brennell). Describing the aesthetic, Parsons writes:

> "The quad system for a live hall has to be used in a rather subtle manner, the aim being in this case to add impact at relevant points in the piece being performed. In the long introduction to *Dark Side*, the heart beat fades on slowly as the house lights dim and synthesizers and voices swirl around the hall. In one of the Floyd's old favorites the effect is slightly less subtle: in 'Careful With That Axe Eugene' the quad carries no sound at all until the famous horrific scream at which every amplifier and speaker is driven to its absolute maximum, accompanied by an explosion of flash powder behind the stage [109]."

For their "In the Flesh" tour, the band eliminated the rear quad channel, resulting in the curiously-named "three-channel quad" system; it was then reinstated for "The Wall" tour. Of course, the entire sound system was always growing in complexity, reaching well over 130 input channels, making quad mixing more challenging. Liberal use of subgroups and sub-consoles was necessary; eventually a custom Midas XL3 board was used for all quad spatialization [31]. Through continual pursuit of new speaker technology, and extensive trial-and-error, the Pink Floyd tours were consistently praised for sound quality far exceeding the usual rock show.

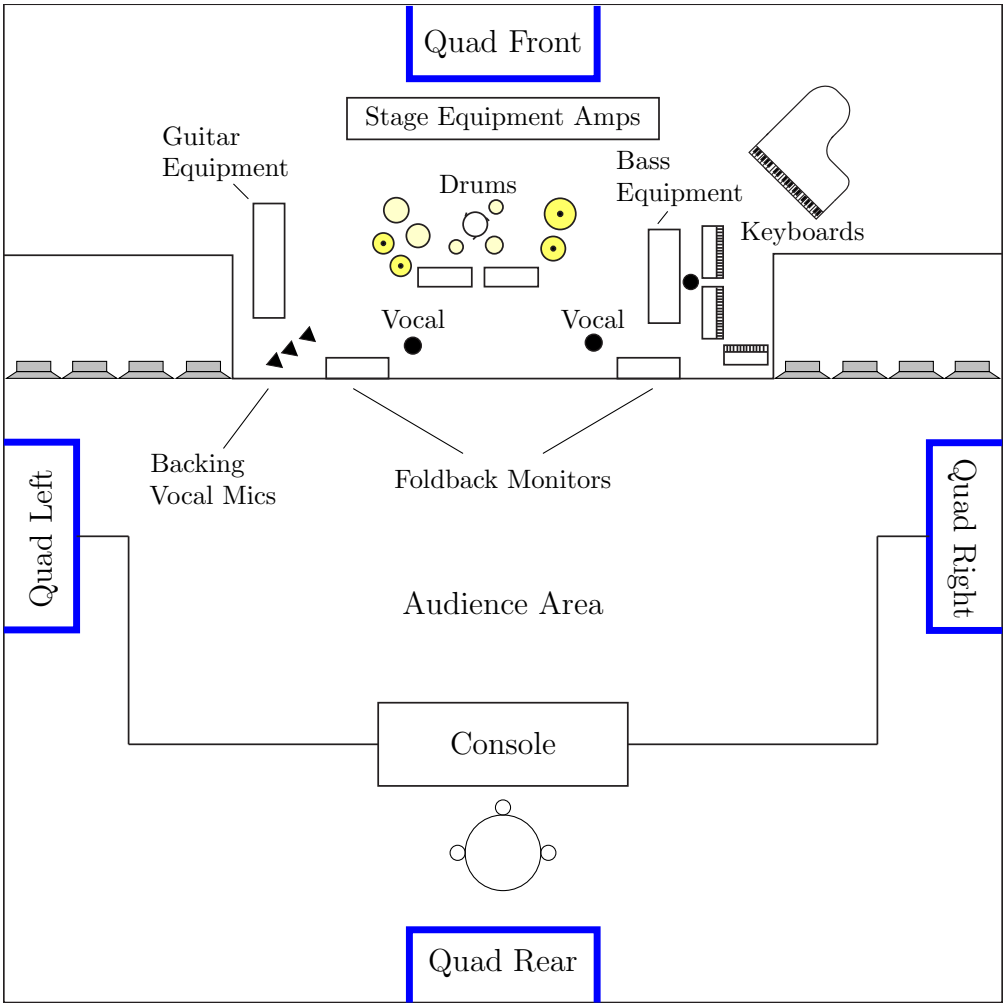


FIGURE 2.14. Quad system used for live *Dark Side of the Moon* shows. After [109].

2.3.3. Blue Man Group

The Blue Man Group uses bizarre handmade instruments and a surreal stageshow to portray an alternate reality. Their show *Live at Luxor* integrated rock concert elements along with their usual theatrics. The sound reinforcement system used both left/right stereo and surround. The stereo system used arrays on stage and at the room's midpoint, along with (downstream)

subwoofers. The surround system utilized four channels for the room's four corners, as well as a fifth channel sent to a ceiling array at the room's midpoint.

> " 'Everything that happens on the stage, such as when Blue Man is playing, comes through the mains,' [sound designer Todd] Pearlmutter explains. 'But when there are things happening in the audience, especially at the end of the show, the surround system is used, because that's where the action is. There are other times when the two work together, such as when the voiceover comes through the mains, and the incidental music comes through the surround.' "

A Yamaha 03D digital mixer handles the surrounds, while 4(!) consoles drive the main system [71].

2.3.4. Cirque du Soleil

Cirque du Soleil's *Love* is an extravagant tribute to the Beatles' music. It uses an extremely involved custom surround system. The show takes place in a dedicated in-the-round theater. The audio started from a 5.1 mix created from the original multitrack recordings by Giles Martin (son of Sir George Martin, the original Beatles producer). The show uses 30 Meyer speakers on the stage, along with additional Meyer surround speakers and JBL ceiling speakers. Most unusual though, is that each audience seat has its own set of left, right and center Innovox Audio speakers—6039 in all. They are built out of "cardboard and magnets" to better emulate Beatles-era speakers (though the degree of authenticity is unknown). The designers claim around 25 total sound channels:

> "Starting, typically, with the 5.1 mix, the team would slowly identify elements of the recording to extract out and place in the sound picture—either forward in the [Meyer] M1-D front arrays, in speaker seats, surrounds, or elsewhere. 'If there was a sound or a voice that corresponded to a character onstage, we might place it lower, higher, in the air—we'd focus the voice in the area where the audience would expect to hear it,' [sound designer Jonathan] Deans explains."

Because of the in-the-round format, the team divided the theater into eight separate zones, replicating the same mix in each of the zones, with tweaks for any asymmetries [67].

CHAPTER 3

Perceptual and Technical Background

This chapter provides an overview of the cues used by the auditory system to form spatial percepts, with an emphasis on the issues that arise specifically in the context of loudspeaker reproduction in large spaces. Prior work on these cues is extensive; this is necessarily only a brief summary. More details and extensive references are available in several sources [17, 73]. Here and throughout this document we occasionally make use of the *median*, *frontal* and *horizontal* planes, shown in figure 3.1.

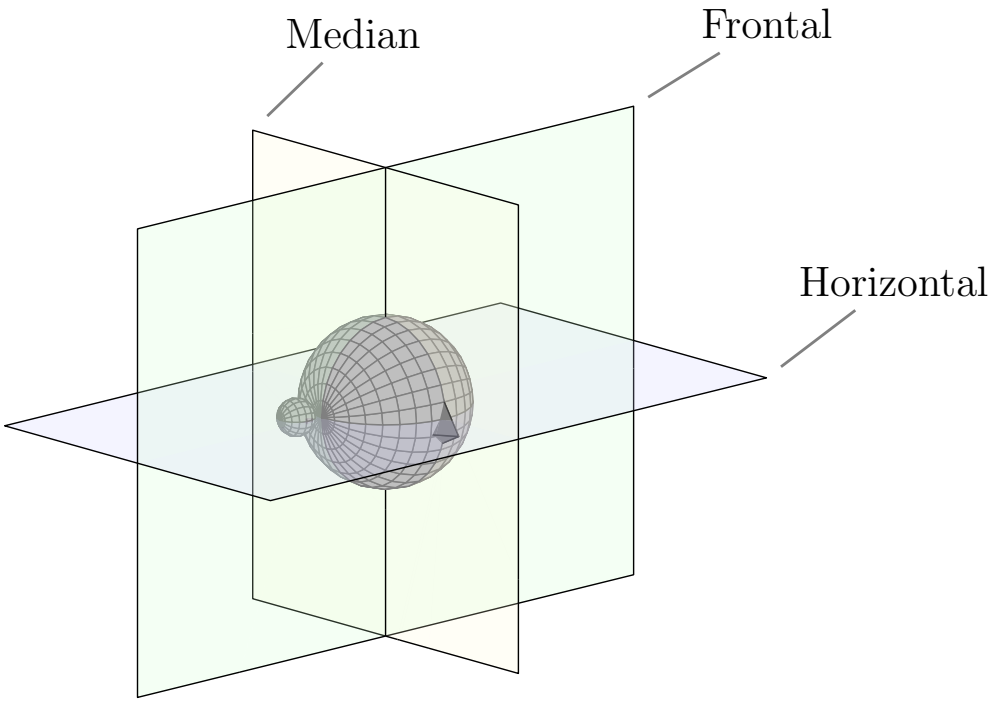


FIGURE 3.1. Median, horizontal and frontal planes.

3.1. Spatial Hearing Cues

Spatial hearing is conventionally divided into two aspects: localization and environmental cues. Localization refers to the location of the auditory image in space: azimuth angle, elevation angle, and distance. Additionally, the Apparent Source Width (ASW), or size of the auditory image, is often included here. Environmental cues refers to information about the size, shape, and reverberant properties of the listening space. This distinction is only a generalization; the categories are not sharply defined perceptually.

Localization relies on three primary cues (see figure 3.2):

- Interaural Time Difference (ITD)
- Interaural Intensity Difference (IID)
- Head-Related Transfer Functions (HRTFs)

A given sound will be perceived by both the *ipsilateral* (near) and *contralateral* (far) ears. ITD and IID arise because the signal at the contralateral ear is typically delayed and attenuated relative to the signal at the ipsilateral ear. The delay is caused by the longer path length when travelling around the head. The ITD is usually less than 0.8 ms; figure 5.4 on page 122 shows typical values as a function of source angle. Attenuation occurs primarily because the head blocks the sound energy ("head-shadowing"). Both IID and ITD vary as the location of the sound source varies. In addition, they are frequency-dependent, though a single aggregate value can also be obtained. These interaural differences are the primary cues for azimuth; they play only a minor role for elevation perception. In nearfield monitoring situations, convincing perceptual results can be obtained using only frequency-independent delay and attenuation. This is known as "summing localization" (see section 3.2). The

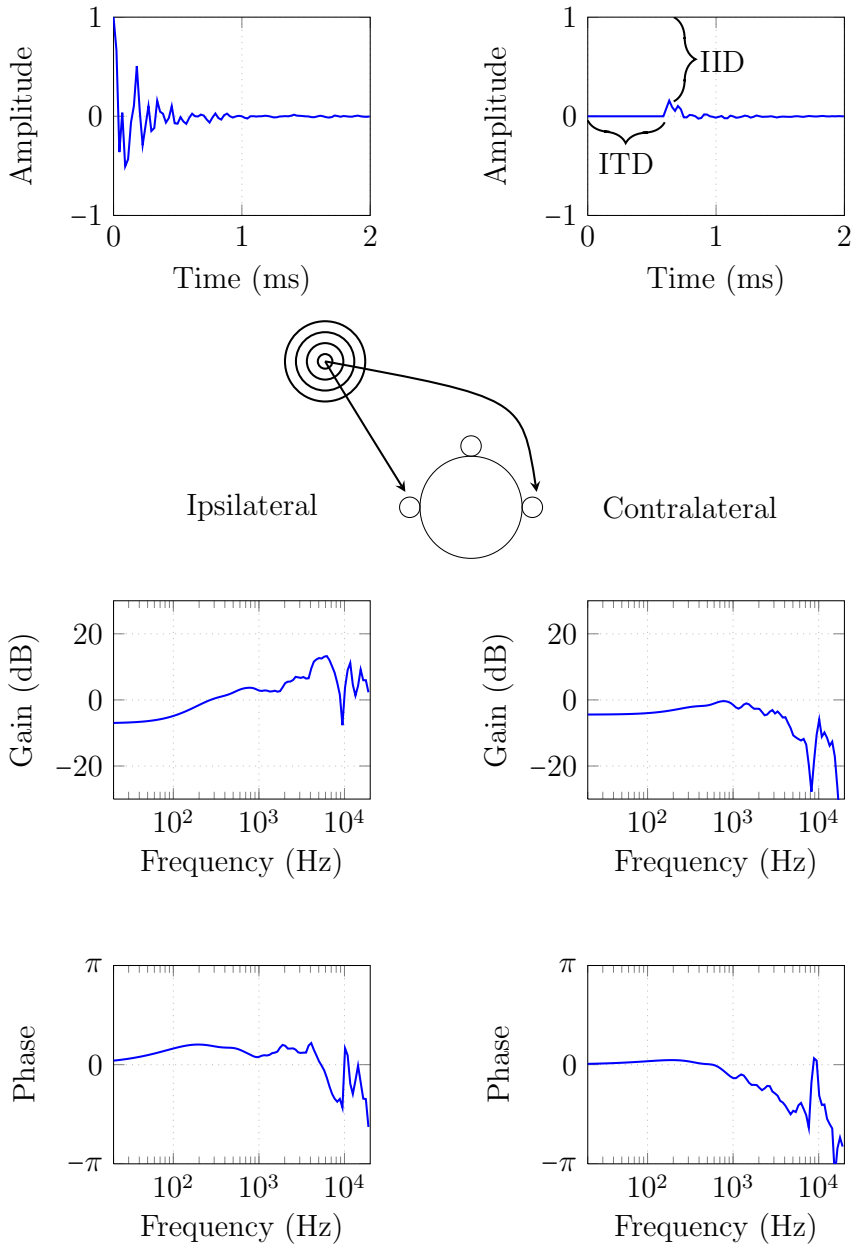


FIGURE 3.2. IID, ITD, and Time- and frequency-domain views of HRTFs.

great majority of left-right panning in recorded music makes use of intensity-based summing localization.

The well-known "duplex theory" of localization, first proposed by Lord Rayleigh, describes the relative importance of ITD and IID across frequency. While it is now considered incomplete, it is still a useful source of intuition. At low frequencies, wavelengths are very long compared to the head, so there is little head-shadowing and IID is near 0 dB. Therefore, ITD serves as the dominant azimuth cue. At higher frequencies, above about 1500 Hz, ITD becomes ambiguous because the delay can be longer than a single period of the wave. However, since the wavelength now becomes small relative to the head, sound at the contralateral ear is significantly attenuated, by as much as 20 dB at maximum lateral angles. In summary, ITD tends to dominate at low frequencies, and IID at high frequencies. Again, this is only a first approximation.

While IID and ITD work quite well for azimuth, they are less successful for elevation. This is often explained as the "cone of confusion," the locus of points which produce the same interaural differences for a spherical head model (see figure 3.3). While real heads are not this symmetric of course, interaural differences are still quite similar for points on the cone.[1] To resolve this, the auditory system turns to the spectral properties of head-related transfer functions. By contrast to ITD and IID, HRTFs are monaural, affecting the sound arriving at a single ear. The head, torso, and especially pinnae (outer ears) all act as complex filters (with impact in both the time and frequency domains) on the incoming sound. This filtering action changes as the source location changes. Elevation cues in particular are primarily based on notches and peaks in the frequency spectrum caused by constructive/destructive

[1] We use a range-dependent head model with ears set at $\theta = \pm 100°$, which creates small but meaningful variations in interaural differences on the cone. See section 4.1.

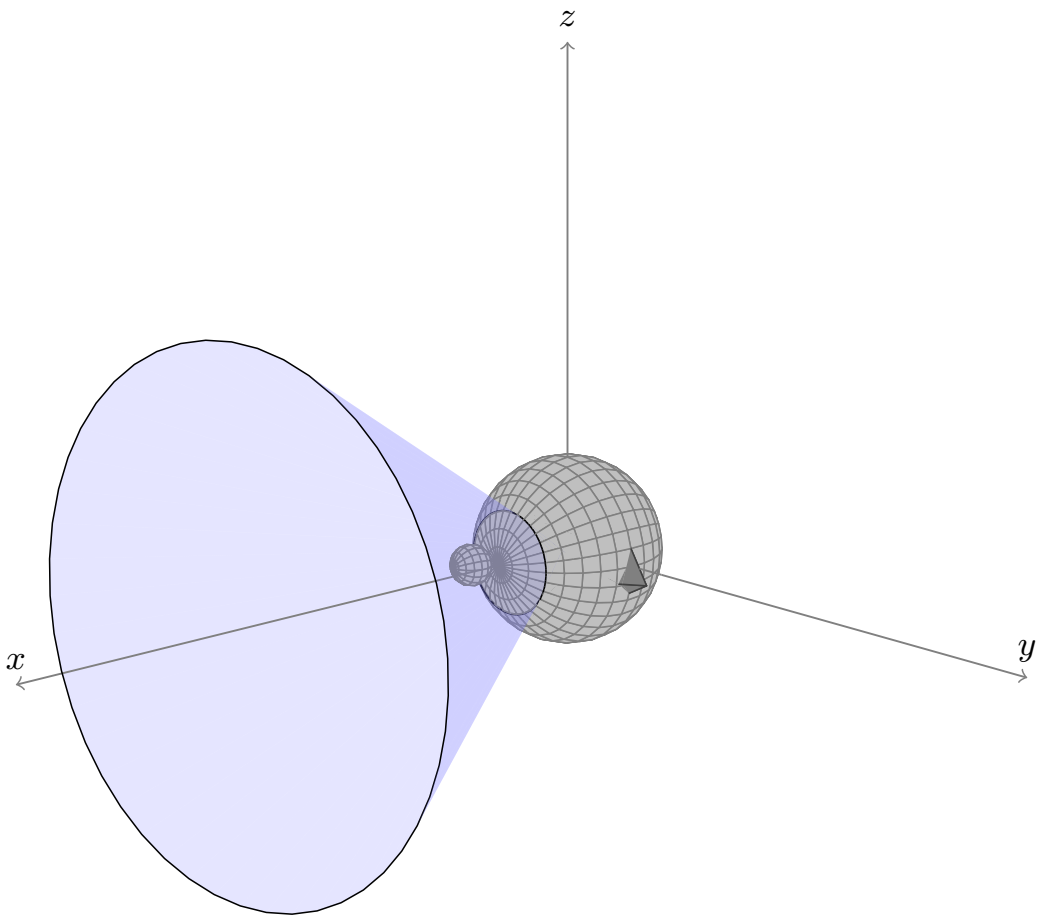


FIGURE 3.3. The "cone of confusion." Points on this cone all result in the same IID and ITD for a spherical head model.

interference from pinna reflections (figure 3.4). HRTFs also influence azimuth detection somewhat, though they are dominated by ITD/IID except for sound sources to the sides.

Distance perception is a combination of many factors. For close sources (less than about 1 m), HRTFs exhibit a range dependence. At most distances the effect is slight but it becomes quite pronounced when very close to the head (see section 4.1.4). At longer ranges, the ratio of direct to reverberant sound is central: typically, a near sound source will be much louder than the reverberant field, while a far source can easily create a direct sound that has a lower

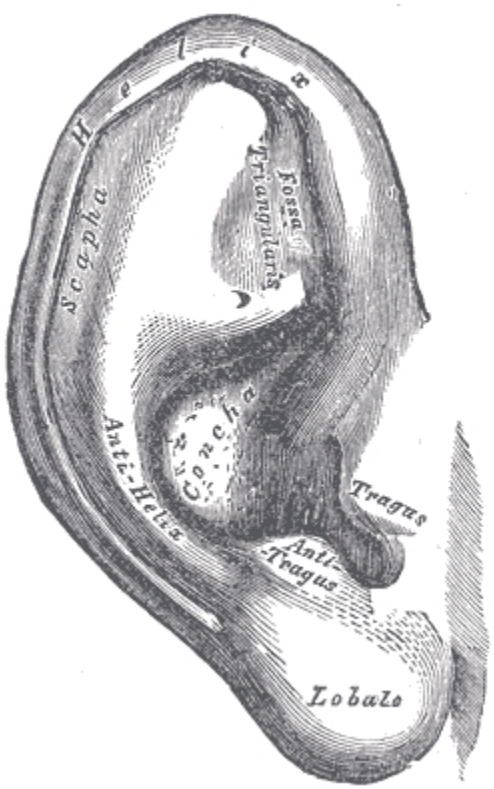


FIGURE 3.4. Anatomy of the pinna. Sound is reflected in a way that depends strongly on the source elevation, creating a localization cue. Image source: [55].

amplitude than the indirect sound. The relative timing of direct and reflected sound is a further cue. Finally, prior knowledge of the source and its average volume plays a large role.

In an enclosed space, sound reaches the listener both through the direct air path, and through reflections off the enclosing surfaces. The first sparse echoes are called "early reflections." Once the reflections reach a certain density (typically 1000/s), they are indistinguishable and form a uniform reverberation field (see figure 3.5). The reverberation field conveys significant spatial information. Relative timing of early reflections conveys information about source distance and room size. The ratio of direct to reverberant sound also influences distance perception, as mentioned. The intensity of reflected sound can suggest

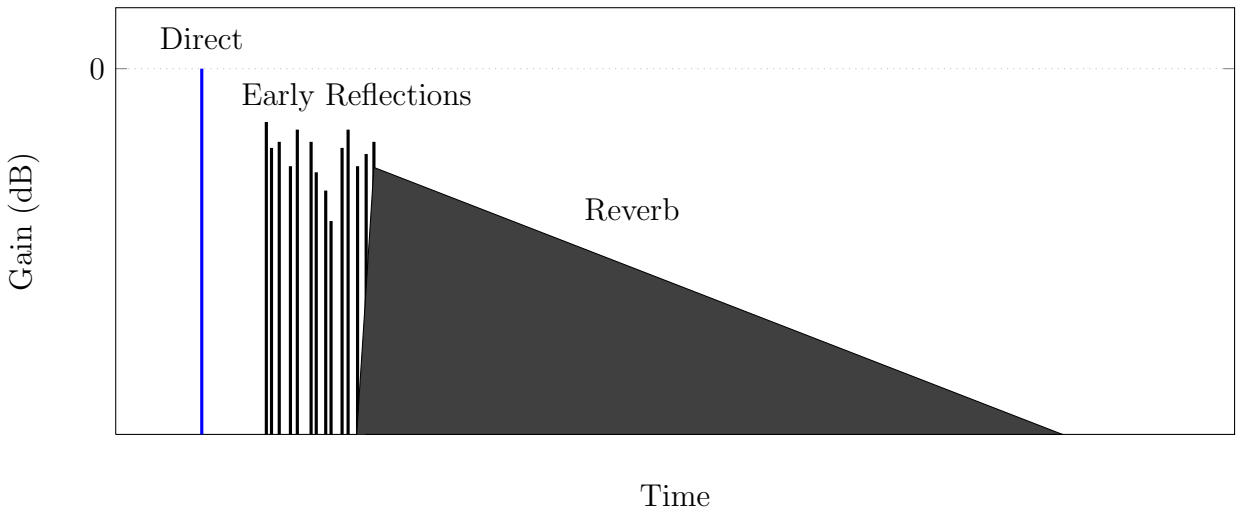


FIGURE 3.5. Typical time evolution of reverberation.

material properties of the surrounding surfaces. Particularly important for spatialization, lateral (side) reflections influence perception of source width and envelopment [57]. They are also yet another distance cue [74].

3.2. The Precedence Effect

Understanding the precedence effect is an important prerequisite for understanding large-venue spatial audio. The term describes a group of phenomena related to "lead-lag" scenarios: an initial sound is followed by one or more delayed, gain-shifted copies [17, 84]. In acoustics, this describes the reflections that comprise reverberation. In sound reproduction, this is primarily related to summing localization of multiple loudspeaker signals.

An important article by Barron describes the phenomena that arise when a direct, musical sound is followed by a single reflection 40° from the median plane [9]. Figure 3.6 shows the perceptual regions that arise. Barron defines the regions as follows:

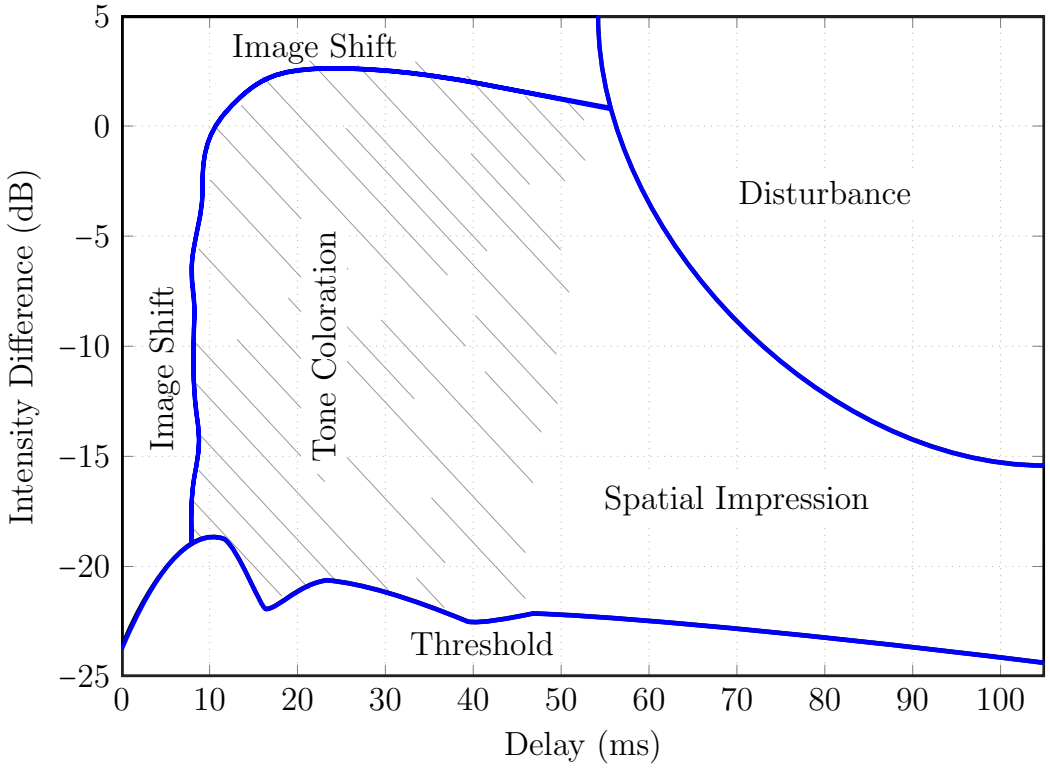


FIGURE 3.6. Perceptual regions for a direct sound and single side reflection. After [9, 73].

Threshold: Below this line, the reflection has no audible effect.

Image Shift: The image moves away from the direct sound, toward the reflected sound (summing localization).

Tone Coloration: The timbre becomes "sharper" due to comb filtering.

Spatial Impression: The sound becomes fuller, enveloping, and three-dimensional. This is often aesthetically highly desirable [56].

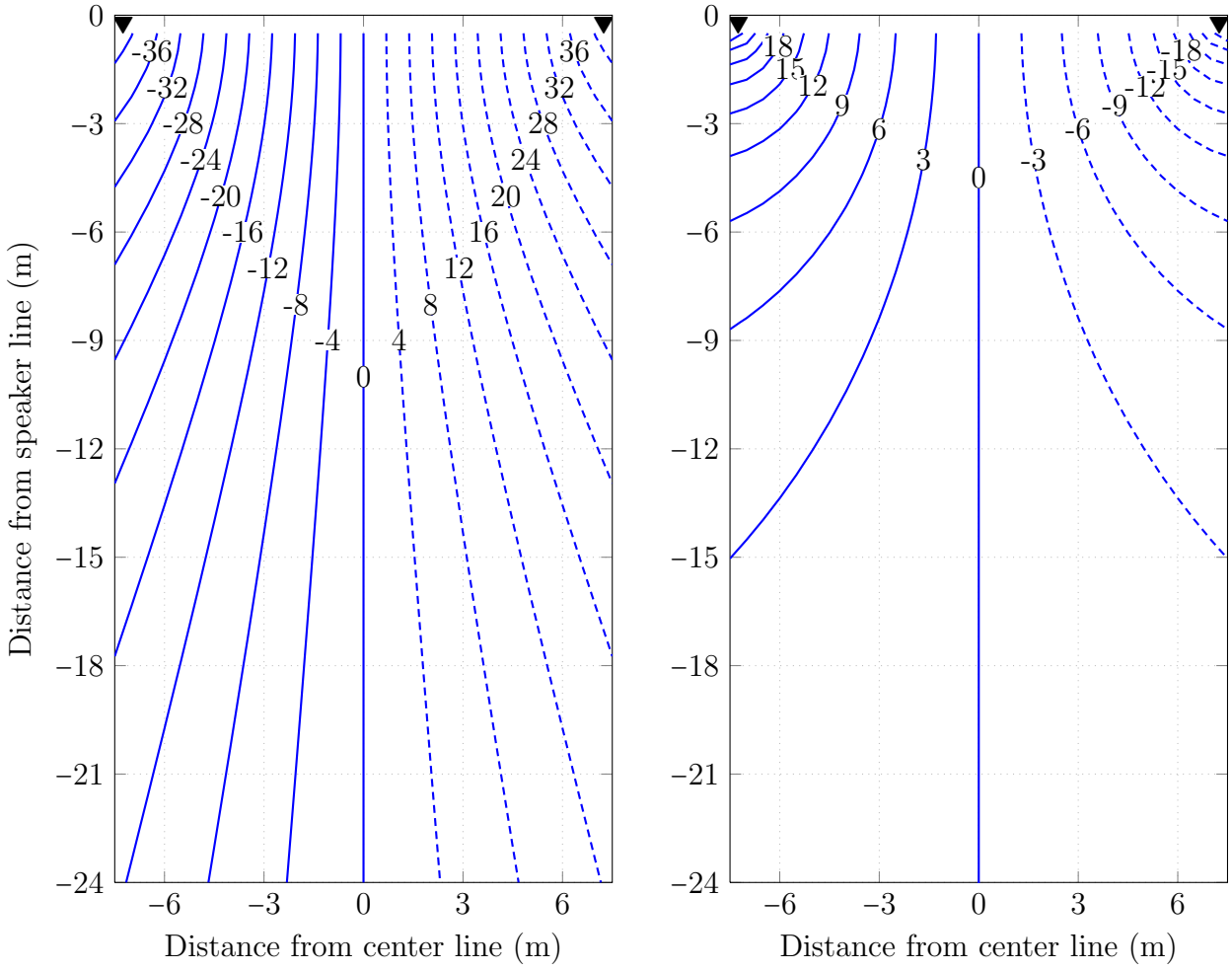
Echo Disturbance: The reflection becomes audible as a distinct echo, interfering negatively with intelligibility.

These effects explain why off-centered listening to stereo recordings causes the image to move toward the closer loudspeaker, and eventually to collapse into it entirely. They also offer one explanation for the use of intensity-based rather than delay-based panning. The perceptual zone for image shift is much narrower in the time dimension. In fact, in some circumstances a delay of only 1 ms is sufficient to move the image entirely into one speaker [84]. Intensity panning is much more robust, requiring up to a 30 dB imbalance for the same effect in typical situations. Of course, off-centered listening will create both a delay and a level imbalance.

The precedence effect is not as simple as this section suggests. Perceptual thresholds can change based on prior sonic history, transient characteristics, and number of reflections, among other factors. However, these are higher-order considerations that are not pursued here.

3.3. Interaural Cues in Large Venues

The overall soundfield for two-loudspeaker reproduction in large venues is dominated by the distances involved. For off-centered listeners, large differences in path length to the two speakers are common. As illustration, figure 3.7 shows the arrival time and intensity differences over the audience area for two speakers 15 m apart, radiating identical signals. The intensity difference for practical locations is at most 20 dB or so. The delay reaches a maximum of about 40 ms. Figure 3.8 shows similar information for a number of speaker separation distances. Notice that the range of intensity values does not change, but the range of delay values increases with wider speakers. This is because intensity is a relative measure, whereas the delay depends on the absolute difference in timing.



(A) Time of arrival difference (ms)

(B) Arrival intensity difference (dB)

FIGURE 3.7. Timing and intensity differences for speakers 15 m apart.

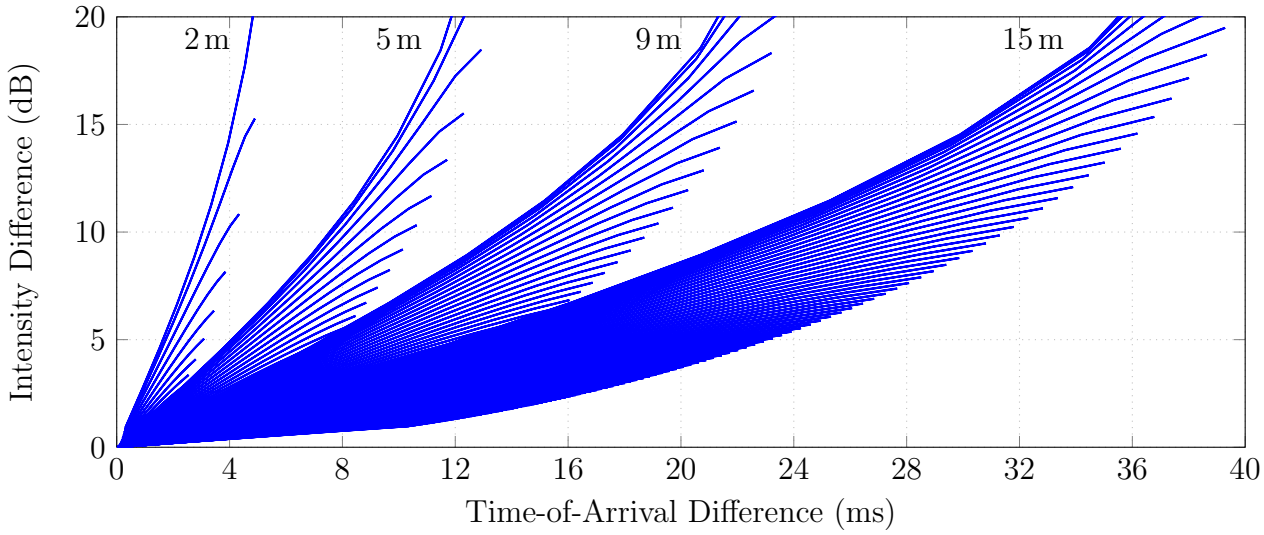


FIGURE 3.8. Timing and intensity differences for speaker separations ranging from 2 m to 15 m. After [73].

To understand the perceptual consequences of this, we must again refer to the precedence effect. Figure 3.9 superimposes the range of arrival differences for both 2 m and 15 m speaker separation onto Barron's diagram of perceptual regions. Recall that summing localization is frequently used for azimuth control of stereo sound. For centered listeners, identical loudspeaker signals will result in a centered auditory image. In the small venue (typical of a home listening environment), off-center listeners will hear the image shift toward the near speaker. Even when directly in front of one speaker, the far speaker is delayed only about 8 ms. By contrast, this delay reaches 40 ms in the larger venue. This is more than sufficient to cause timbral distortions, rather than a simple image shift. Only a small percentage of the audience will not experience timbral colorations; an even narrower band will perceive the full range of left-right panning (for listeners to either side, most images will collapse to the near loudspeaker). In other words, a majority of the audience will experience timbral distortion,

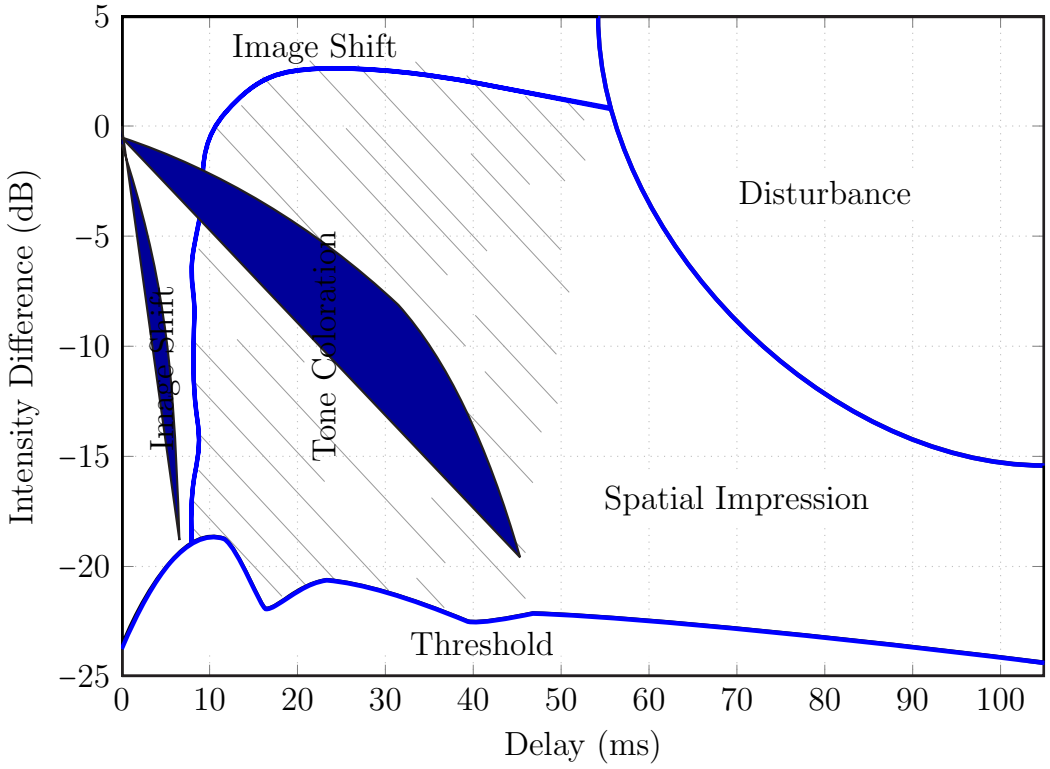


FIGURE 3.9. Perceptual phenomena activated by loudspeaker playback. The narrower darkened region corresponds to a 2 m speaker separation; the broader region to a 15 m separation. After [73].

and the great majority will receive poor spatial cues. Therefore, traditional stereo panning fails in large venues.

3.4. Spectral Cues in Large Venues

Reliable spectral cues are particularly difficult to project in large venues. The explanation is a phenomenon known as *crosstalk*: signals intended for one ear are inevitably received by the opposite ear (see figure 3.10). The signal actually received will be the sum of the intended signal and the crosstalk signal (slightly delayed because of the longer travel path). Normally, listeners simply ignore crosstalk, but the impact is sufficient to disrupt sensitive

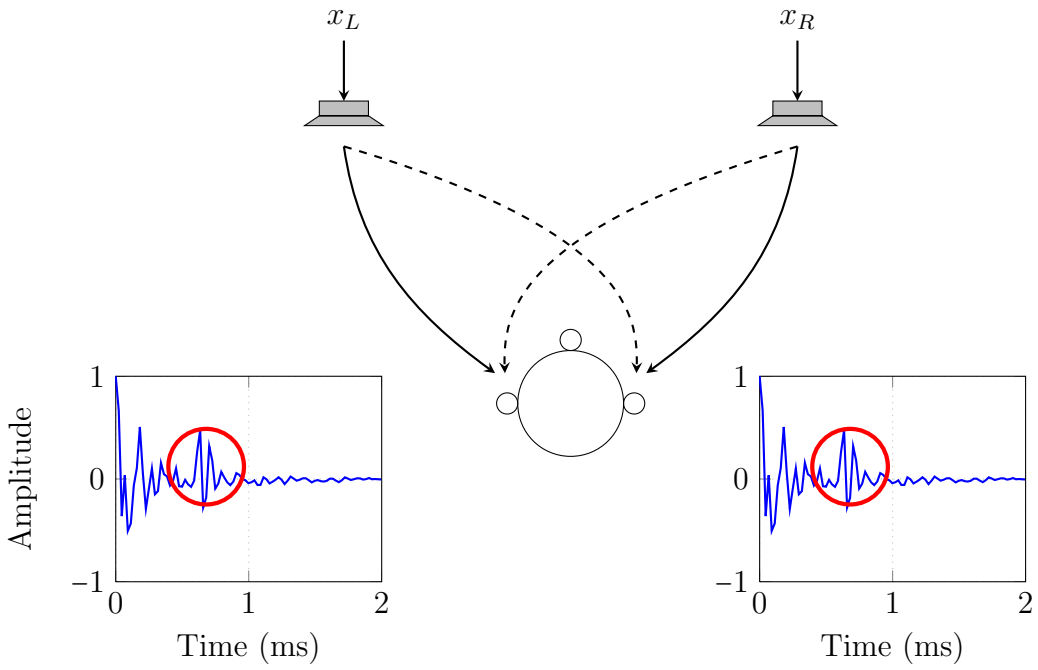


FIGURE 3.10. Illustration of the crosstalk phenomenon.

spatial cues. The first successful solution was presented by Schroeder and Atal in 1963 [121]. They reasoned that the crosstalk signal could be cancelled by its inverse signal, appropriately timed. However this inverse signal would also need crosstalk cancellation, and so on. By performing the infinite sum, they calculated the exact set of filters needed for complete crosstalk cancellation (see figure 3.11). While this scheme is technically correct, it presents numerous real-world difficulties, most noticeably a very narrow listening area. Numerous improvements on and alternatives to the original Schroeder-Atal formulation exist. Cooper and Bauck found that reducing crosstalk for high frequencies allows a wider target area; however the scheme still suffers from imaging problems [30]. A version of crosstalk cancellation is used as one component of this research and is described in section 4.2.3.

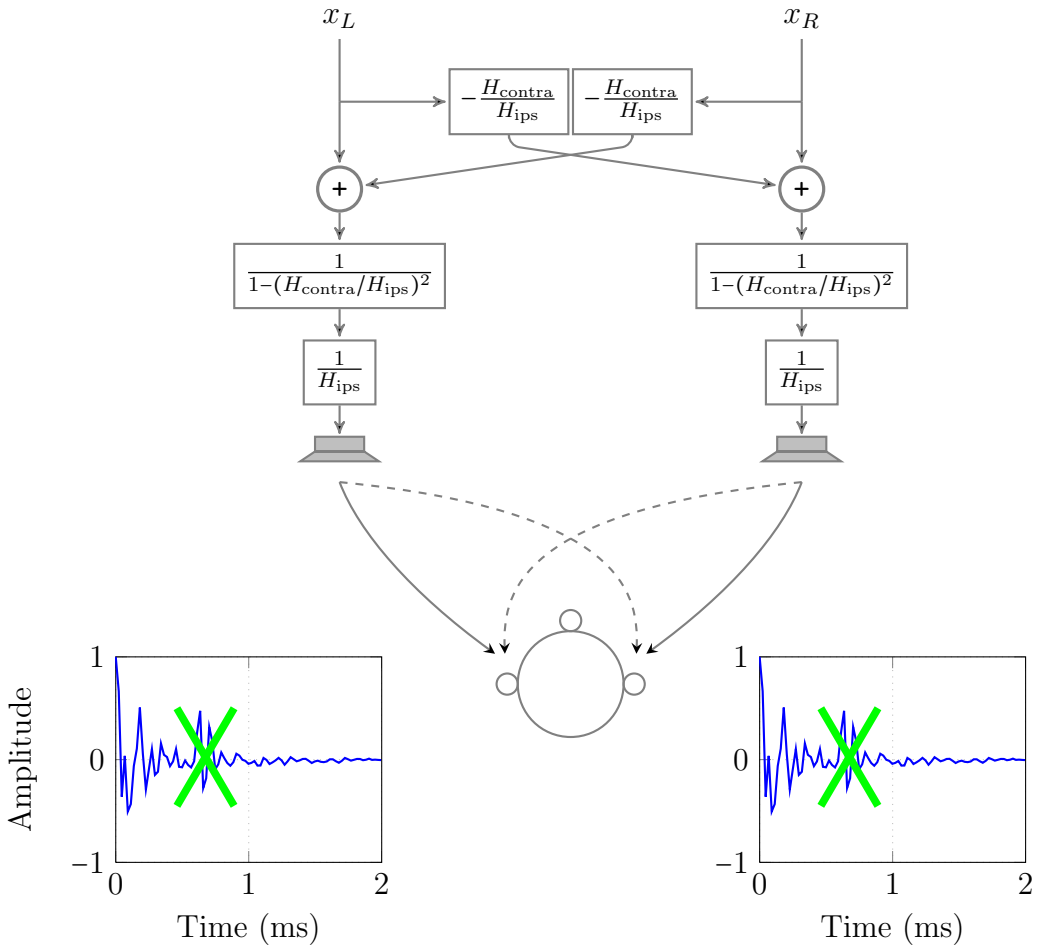


FIGURE 3.11. Simplified diagram of Schroeder-Atal crosstalk cancellation.

We have omitted discussion of a related major area of current research: wavefront (or wave field) synthesis (Ambisonics can be considered a type of wavefront synthesis, but it has other aspects as well). Wavefront synthesis strives to use a finite number of loudspeakers to recreate the entire 3D sound field of the recording (or create a new imagined sound field) [16]. While this technology is seeing rapid advances and certainly holds potential, generalizing it to large spaces is non-trivial. It requires many transducers to cover a given audience area, and suffers from limited bandwidth and artifacts under most practical conditions.

3.5. Decorrelation

Decorrelation is another important concept for spatial audio, and large venue reproduction in particular. Conceptually, it is a measure of similarity between two signals (section A.4 provides a quantitative definition). For our purposes, decorrelated signals are those which sound the same, but whose waveforms have different micro-level features. In concert halls, for example, decorrelation arises from lateral (side-wall) reflections [59]. A reflection from the left wall will arrive primarily at the left ear, and create a signal which is very different from that at the right ear. Perceptually however, the ear recognizes both ear signals as arising from the same source. Appendix B provides methods for creating decorrelated sources through signal processing.

Kendall describes several consequences of using decorrelation [74; see also 81]. Two are particularly important in this context. First, decorrelation can defeat the precedence effect and stabilize the sound image with respect to listener motion. This happens at least in part because localization cues are supressed when correlation is low [46, 93]. This phenomenon effectively broadens the width of the audience region receiving spatial audio cues. Depending on the precise circumstances, delay tolerance can improve four times or more using decorrelated signals. This potentially translates to an equal improvement in terms of area (see for example, figure 5.6 on page 127).

Secondly, decorrelation creates a diffuse sound image; that is, an image with a broad spatial extent, one that is immersive and potentially envelops the listener.[2] These properties are usually considered very pleasing attributes; for example decorrelation is related to

[2] While there is a distinction between apparent source width, immersion and envelopment [57], this precision is not needed here.

favorable subjective ratings of concert halls [122]. Normally, decorrelation is a consequence of reverberation. From a musical perspective, the composer usually has little control over acoustic reverberation. Artificial reverberation can limit intelligibility, or simply not be what is desired. Decorrelation creates diffusion without requiring reverberation.

CHAPTER 4

System Implementation

Implementing the design principles given in section 1.2 requires extensive signal processing. This chapter first gives an extensive analysis of the spherical head model. Next, the acoustics of two-speaker reproduction are outlined. Based on this, we develop several techniques for targeting spatial cues to specific listener locations. Finally, we discuss some aspects of the software that was written to realize this system. The definitions found in appendix A are used throughout this chapter, as are standard signal processing concepts such as the z-transform [105].

4.1. Head Model

Brown and Duda give a simple yet effective model for the response of a rigid spherical head to an infinitely distant point source [21, 22]. Table 4.1 summarizes the various parameters used. The ears are set back slightly (but centered vertically) to better match human anthropometry [17]. The model consists of a one-pole, one-zero minimum-phase filter cascaded with a simple delay. The minimum-phase portion simulates head-shadowing, and provides for the increased ITD found at low frequencies [80]. We give three forms of the range-indepedent model: continuous-time, discrete-time, and inverse discrete-time. In section 4.1.4 we extend the model to include range dependency.

TABLE 4.1. Parameters of the range-independent spherical head model.

Parameter	*Description*	*Value*
c	Speed of sound	$343.23\,\mathrm{m/s}^*$
a	Head radius	$0.087\,\mathrm{m}^\dagger$
ω_0^{cont}	Characteristic frequency, c/a	$3945.17\,\mathrm{Hz}$
β	$2\omega_0^{\mathrm{cont}} = 2c/a$	$7890.34\,\mathrm{Hz}$
θ	Angle from the median plane	
θ_{ear}	Azimuth angle of the ears	$\pm 100^\circ = \pm\frac{5}{9}\pi$
$T_d(\theta)$	Time delay	
$\alpha(\theta)$	Filter zero location	
α_{min}	Minimum value of α	0.1
θ_{min}	Angle of greatest head-shadowing	$150^\circ = \frac{5}{6}\pi$
F_s	Sampling frequency	$44\,100\,\mathrm{samples/s}$
ω_0	Discrete-time characteristic frequency, $\omega_0^{\mathrm{cont}}/F_s$	$0.09\,\mathrm{cycles/sample}$

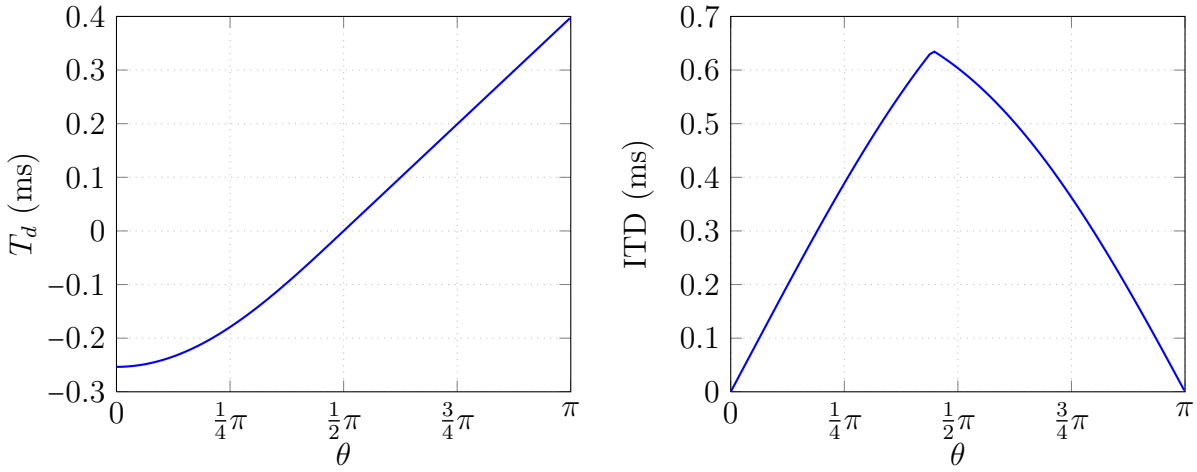
* Speed of sound at 20 °C, sea level, in the U.S. Standard Atmosphere [137].
† Optimal radius for a spherical head model, as determined by [1].

4.1.1. Continuous-Time Head Filter

In the continuous-time domain, the model is given by:

$$H^{\mathrm{cont}}(s,\theta) = \frac{\alpha(\theta)\, s + \beta}{s + \beta} e^{-T_d(\theta)s}. \tag{4.1}$$

Here $\beta \equiv \frac{2c}{a}$, with c the speed of sound and a the radius of the sphere. The angle θ is the incident angle, i.e., the angle between two rays from the center of the sphere to the sound source and the ear location. The magnitude response of the model gives the ratio between the observed response and the free-field response, which is the response that would have occurred at the location of the center of the head, were there no head present. T_d is the time delay, again relative to the time the sound would have arrived at the location of the sphere's center, had the sphere been absent. Note that this can lead to negative delays. The true



(A) T_d as a function of incident angle (B) ITD as a function of source angle

FIGURE 4.1. Frequency-independent time delay used in the model.

time delay is approximated well by the following formula, due to Woodworth and Schlosberg:

$$T_d(\theta) = \begin{cases} -\frac{a}{c}\cos\theta & \text{if } 0 \le |\theta| < \frac{\pi}{2} \\ \frac{a}{c}\left(|\theta| - \frac{\pi}{2}\right) & \text{if } \frac{\pi}{2} \le |\theta| \le \pi. \end{cases} \tag{4.2}$$

Figure 4.1 shows this frequency-independent delay and the resulting ITD. The head-shadowing filter adds an additional group delay at low frequencies (see figure 4.4).

The function $\alpha(\theta)$ controls the location of the filter zero, which must vary with incident angle. It is given by:

$$\alpha(\theta) = \left(1 + \frac{\alpha_{\min}}{2}\right) + \left(1 - \frac{\alpha_{\min}}{2}\right)\cos\left(\frac{\theta}{\theta_{\min}}\pi\right). \tag{4.3}$$

This function has a maximum of $\alpha(0) = 2$ and a minimum of $\alpha(\theta_{\min}) = \alpha_{\min}$. We use the values suggested by Brown and Duda. Figure 4.2 shows the function.

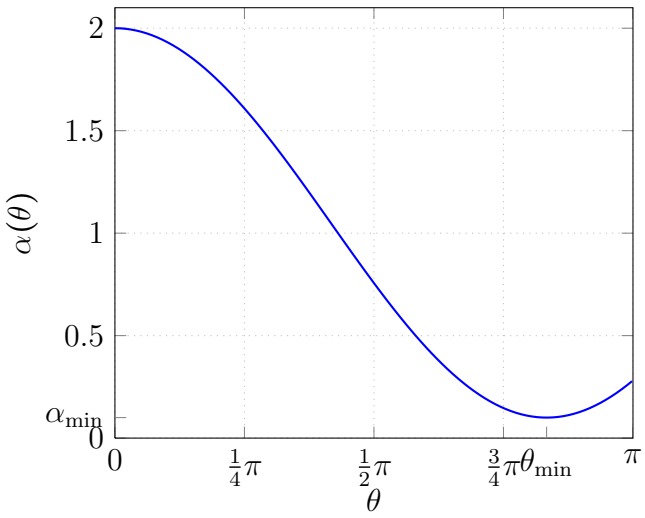


FIGURE 4.2. Variation of the head-model parameter α with respect to the incident angle.

Basic transfer function analysis will demonstrate the frequency response of the model (see e.g. [106]). We can write the minimum-phase portion as

$$H_{hs}^{\text{cont}}(s) = \alpha \frac{s + \beta/\alpha}{s + \beta} \tag{4.4}$$

(omitting the dependence on θ for simplicity). There is a fixed pole at $s = -\beta$ and an angle-dependent zero at $s = -\beta/\alpha(\theta)$. When $\theta = 0$, $\alpha(\theta) = 2$ and there is a high-frequency boost. When $\theta = \theta_{\min}$, $\alpha(\theta) = \alpha_{\min}$ and there is significant high-frequency attenuation. Figure 4.3 shows the pole-zero plot for the model.

The magnitude of the frequency response is given by:

$$|H_{hs}^{\text{cont}}(j\Omega)| = \alpha \cdot \sqrt{\frac{(\beta/\alpha)^2 + \Omega^2}{\beta^2 + \Omega^2}}. \tag{4.5}$$

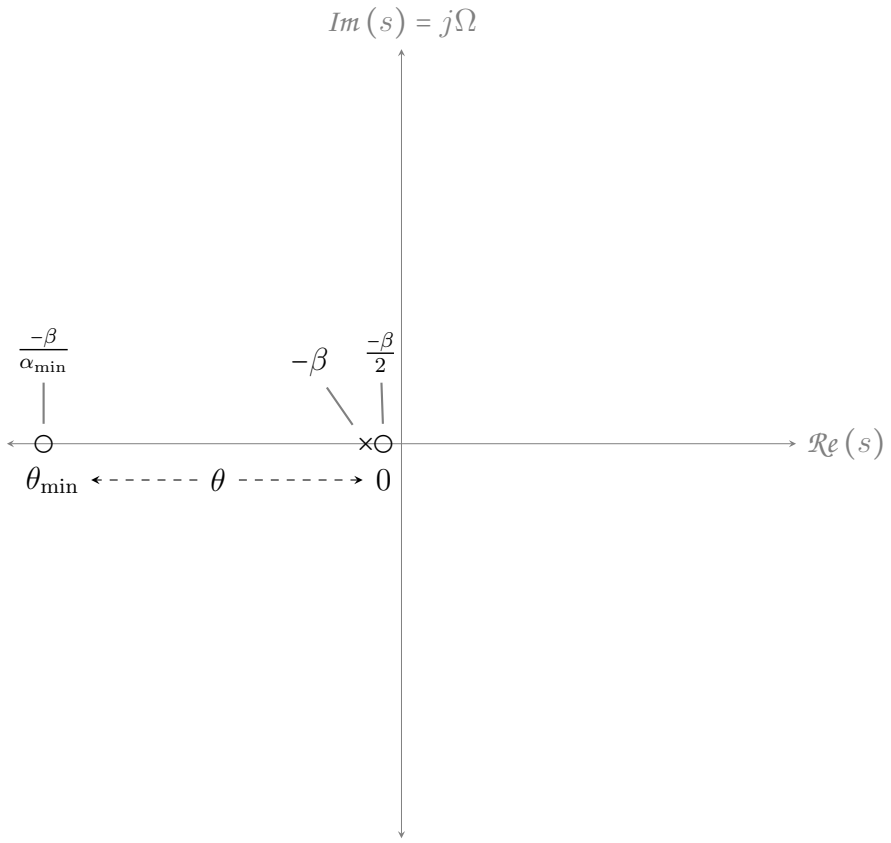


FIGURE 4.3. Pole-zero plot for the continuous-time head model.

The phase of the frequency response is given by:

$$\measuredangle H_{hs}^{\text{cont}}(j\Omega) = \tan^{-1}\left(\frac{\Omega}{\beta/\alpha}\right) - \tan^{-1}\left(\frac{\Omega}{\beta}\right). \tag{4.6}$$

To find the full frequency response of the model, we must include the phase of the delay:

$$\measuredangle H^{\text{cont}}(j\Omega) = \measuredangle H_{hs}^{\text{cont}}(j\Omega) - T_d\Omega. \tag{4.7}$$

We can then compute the phase delay:

$$\begin{aligned}\tau_\phi &= -\frac{\measuredangle H^{\text{cont}}(j\Omega)}{\Omega} \\ &= -\frac{\measuredangle H_{hs}^{\text{cont}}(j\Omega)}{\Omega} + T_d.\end{aligned} \tag{4.8}$$

With a bit of computation, the group delay also follows:

$$\begin{aligned}\tau_g &= -\frac{\mathrm{d}}{\mathrm{d}\Omega}\measuredangle H^{\text{cont}}(j\Omega) \\ &= -\frac{\beta/\alpha}{(\beta/\alpha)^2 + \Omega^2} + \frac{\beta}{\beta^2 + \Omega^2} + T_d.\end{aligned} \tag{4.9}$$

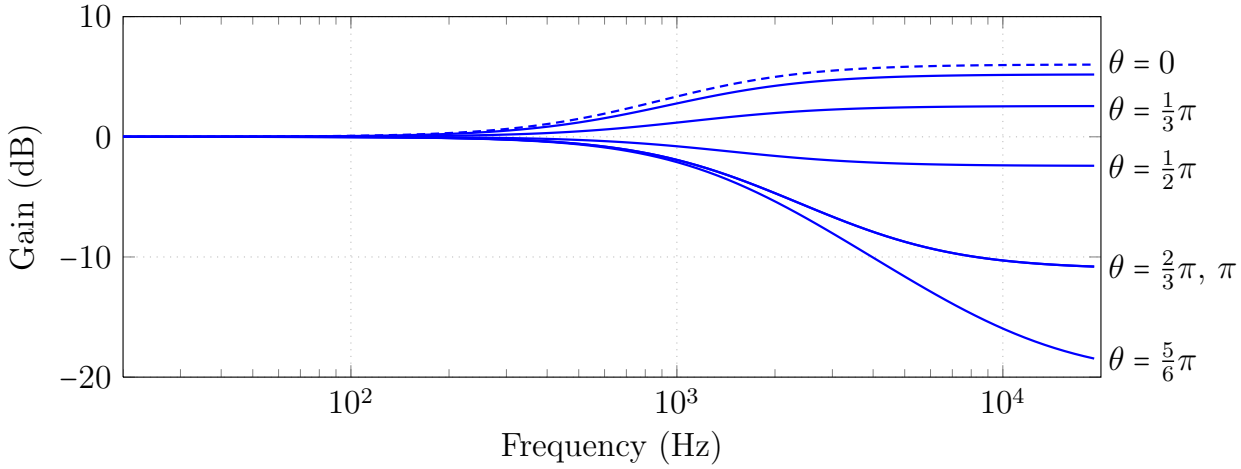
Figure 4.4 shows the magnitude response and group delay of the continuous-time head filter.

4.1.2. Discrete-Time Head Filter

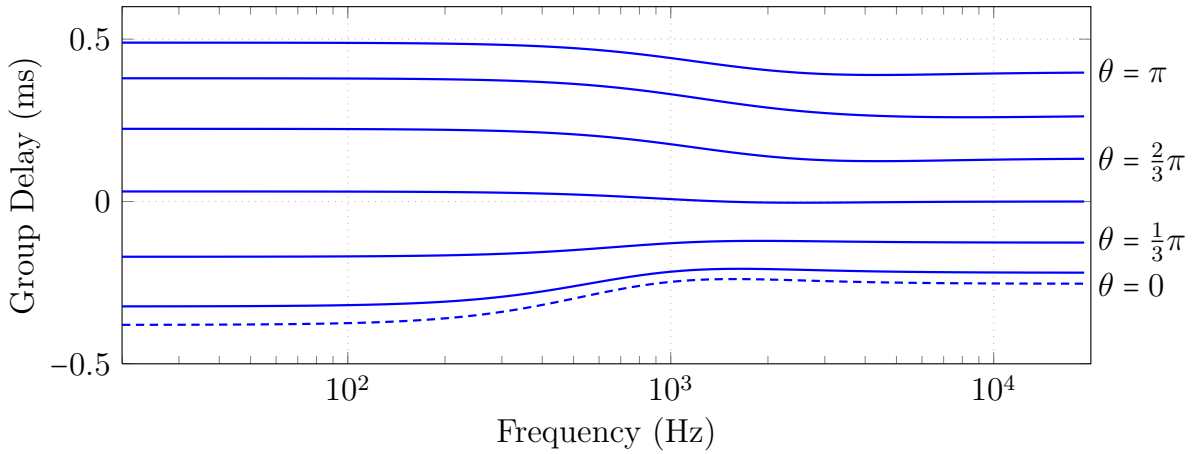
To implement the head-shadowing filter, we must first convert it to the discrete-time z-domain. To do this, we use the standard bilinear transform and make the following substitution [105]:

$$s \Rightarrow 2F_s\frac{1 - z^{-1}}{1 + z^{-1}}, \tag{4.10}$$

where F_s is the sampling frequency. This transform maps the continuous-time frequency axis (the $j\Omega$-axis) onto the frequency curve in the z-domain, namely the unit circle. Since we are mapping an infinite frequency range onto a finite one, there is unavoidable distortion. Fortunately, this distortion is primarily confined to the high-frequency range, where the model's response is relatively flat, and where we do not expect great accuracy compared to



(A) Magnitude Response



(B) Group Delay

FIGURE 4.4. Continuous-time head filter: magnitude response and group delay. Shown for incident angles from $\theta = 0$ (dashed line) to $\theta = \pi$, in steps of $\frac{\pi}{6}$.

actual HRTFs. Performing the bilinear substitution leads to the following discrete-time filter:

$$H_{hs}(z) = \frac{(\omega_0 + \alpha) + (\omega_0 - \alpha)\, z^{-1}}{(\omega_0 + 1) + (\omega_0 - 1)\, z^{-1}}, \tag{4.11}$$

where

$$\omega_0 = \frac{\omega_0^{\text{cont}}}{F_s} = \frac{c}{aF_s} \tag{4.12}$$

is the characteristic frequency in samples per second. Using partial fraction expansion [105], we can write the filter as:

$$H_{hs}(z) = \frac{\omega_0 - \alpha}{\omega_0 - 1} + \frac{\left(\frac{\omega_0+\alpha}{\omega_0+1}\right) - \left(\frac{\omega_0-\alpha}{\omega_0-1}\right)}{1 + \left(\frac{\omega_0-1}{\omega_0+1}\right) z^{-1}}. \tag{4.13}$$

The corresponding impulse response is given by:

$$h_{hs}[n] = \frac{\omega_0 - \alpha}{\omega_0 - 1}\delta[n] + \left(\frac{\omega_0 + \alpha}{\omega_0 + 1} - \frac{\omega_0 - \alpha}{\omega_0 - 1}\right)\left(-\frac{\omega_0 - 1}{\omega_0 + 1}\right)^n u[n]. \tag{4.14}$$

This must still be cascaded with the pure delay to give the complete head model:

$$H(z) = H_{hs}(z) z^{-T_D(\theta)F_s}. \tag{4.15}$$

In general the delay is not an integer number of samples; we are making the natural notational extension to noninteger sample delays. In the SuperCollider language used for implementation, fractional delays are handled automatically. In other environments, there are a number of techniques for implementing fractional delay lines [83].

The head-shadowing filter can also be written in a "zero-pole-gain" form as

$$H_{hs}(z) = \left(\frac{\omega_0 + \alpha}{\omega_0 + 1}\right)\left(\frac{1 - \circledcirc z^{-1}}{1 + \otimes z^{-1}}\right). \tag{4.16}$$

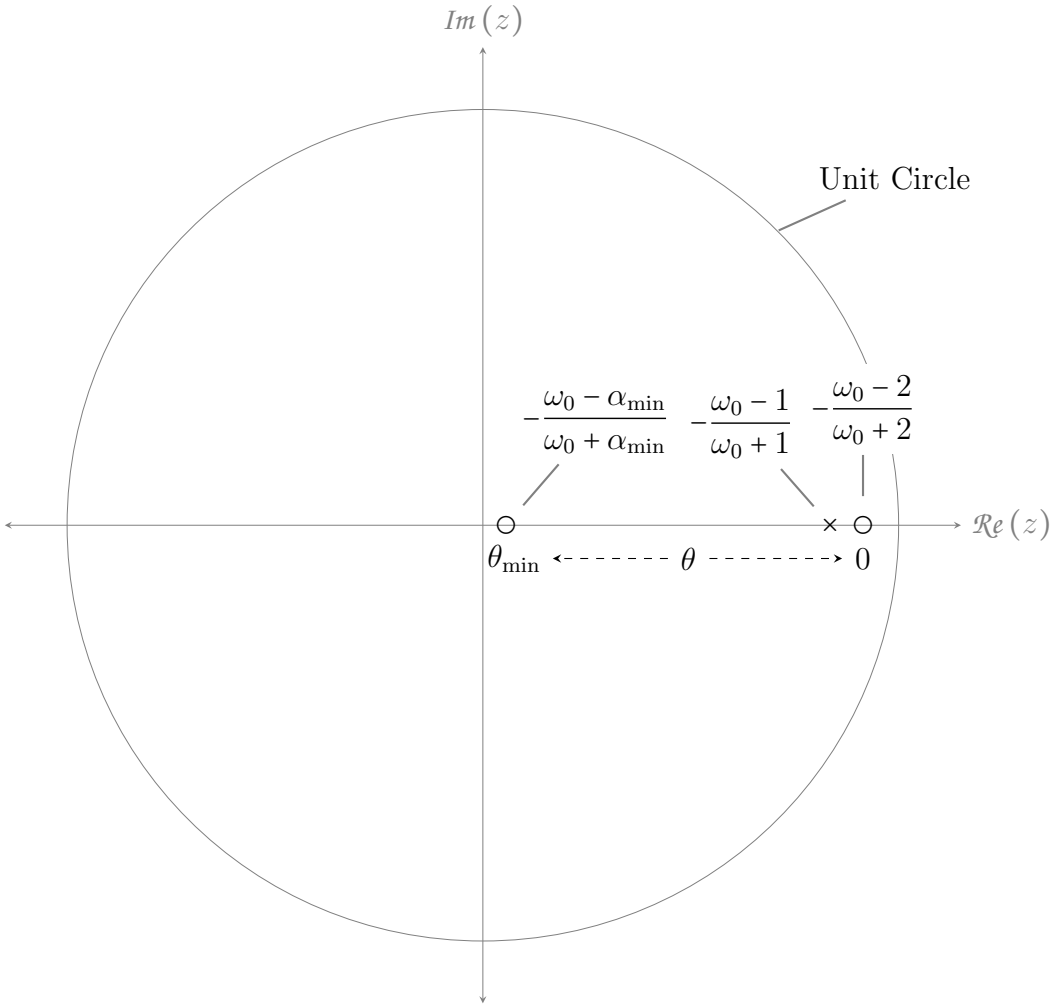


FIGURE 4.5. Pole-zero plot for the discrete-time head model.

There is a stationary pole at $z = -\frac{\omega_0-1}{\omega_0+1}$ (whose value we denote by $\otimes$) and an angle-dependent zero at $z = -\frac{\omega_0-\alpha}{\omega_0+\alpha}$ (denoted $\circledcirc$). Figure 4.5 shows the pole-zero plot.

Using a geometric pole-zero analysis we can find the magnitude of the frequency response [105]:

$$\left|H_{hs}\left(e^{j\omega}\right)\right| = \frac{\omega_0+\alpha}{\omega_0+1} \cdot \sqrt{\frac{1+\circledcirc^2-2\circledcirc\cos(\omega)}{1+\otimes^2-2\otimes\cos(\omega)}}. \tag{4.17}$$

The phase of the frequency response is given by:

$$\measuredangle H_{hs}\left(e^{j\omega}\right) = \tan^{-1}\left(\frac{\sin(\omega)}{\cos(\omega) - \circledcirc}\right) - \tan^{-1}\left(\frac{\sin(\omega)}{\cos(\omega) - \otimes}\right). \tag{4.18}$$

The group delay is:

$$\tau_g = \frac{\circledcirc^2 - 1}{2\left(1 + \circledcirc^2 - 2\circledcirc\cos(\omega)\right)} - \frac{\otimes^2 - 1}{2\left(1 + \otimes^2 - 2\otimes\cos(\omega)\right)}. \tag{4.19}$$

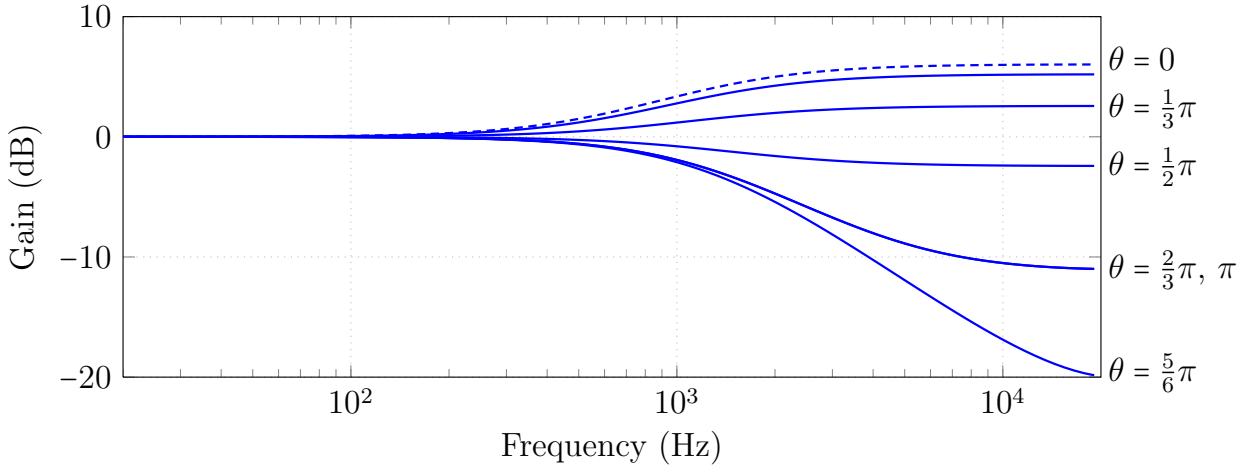
Figure 4.6 shows the magnitude response and group delay of the discrete-time head filter.

4.1.3. Discrete-Time Inverse Head Filter

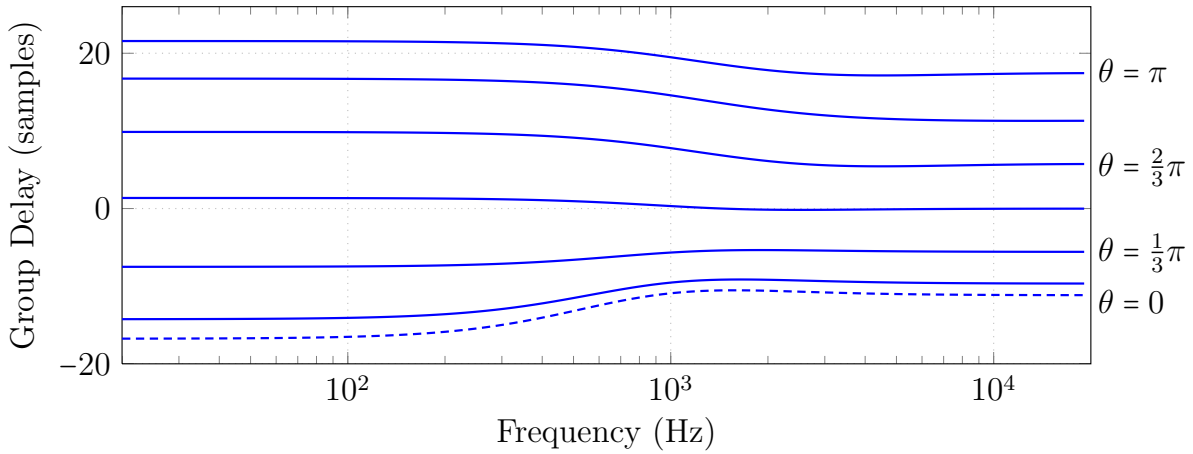
Compensating for the acoustic filtering of the listener's head requires discrete-time inverse head filters (see section 4.2). This is readily obtained by taking the reciprocal of the discrete-time filter:

$$H^{-1}(z) = \frac{1}{H_{hs}(z)} z^{T_d(\theta)F_s}. \tag{4.20}$$

This has the effect of interchanging the locations of the pole and zero. The magnitude $|H^{-1}(z)|$ is the reciprocal of $|H(z)|$; the phase, phase delay and group delay are negated. Figure 4.7 shows the magnitude response and group delay of the discrete-time inverse head filter.

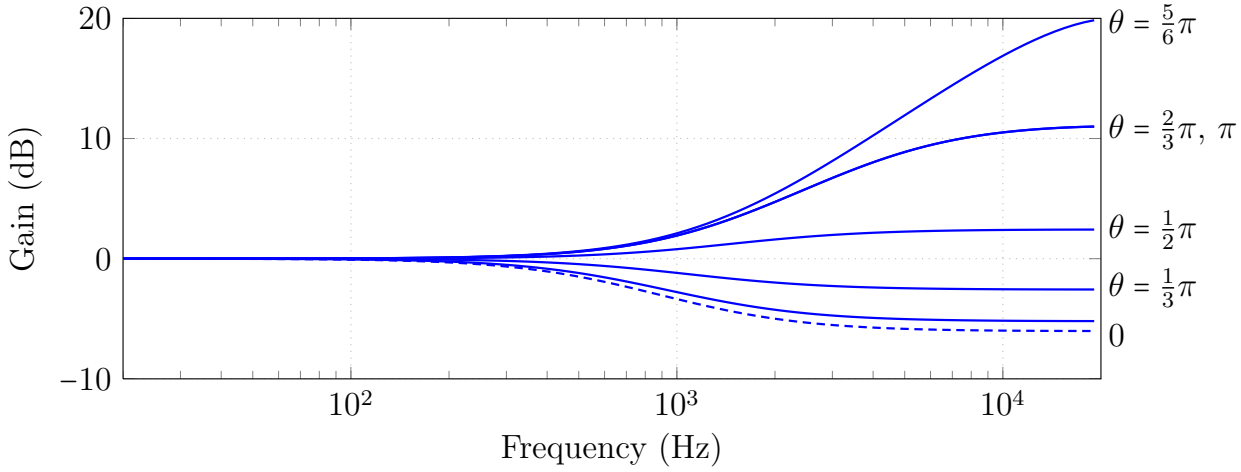


(A) Magnitude Response

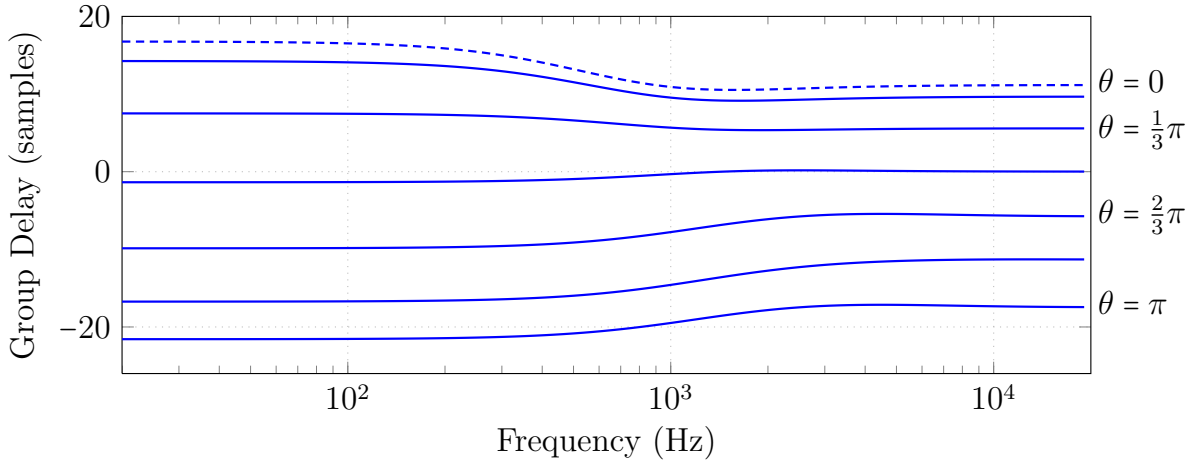


(B) Group Delay

FIGURE 4.6. Discrete-time head filter: magnitude response and group delay. Shown for incident angles from $\theta = 0$ (dashed line) to $\theta = \pi$, in steps of $\frac{\pi}{6}$.



(A) Magnitude Response



(B) Group Delay

FIGURE 4.7. Discrete-time inverse head filter: magnitude response and group delay. Shown for incident angles from $\theta = 0$ (dashed line) to $\theta = \pi$, in steps of $\frac{\pi}{6}$.

Parameter	*Description*	*Value*
r	Distance to source	
ρ	Normalized distance	r/a
$\gamma(\theta, \rho)$	Frequency-independent gain	
$\alpha(\theta, \rho)$	Filter zero location	
$\alpha_{\max}$	Maximum value of α	$2 - 0.9/\rho$
$\alpha_{\min}$	Minimum value of α	$0.1 - 0.08/\rho$
$T_d(\theta, \rho)$	Time delay	

TABLE 4.2. Range-dependent parameters of the spherical head model.

4.1.4. Range-Dependent Discrete-Time Head Model

The model developed above is valid for an infinitely distant source. For binaural synthesis involving distance cues, or for improved accuracy of compensation, a range-dependent discrete-time head model is needed. Theoretical results for the ideal rigid sphere are given in the literature [44]. There are three key findings: first, at close distances, the overall gain is increased for small incident angles, and decreased for angles near $\theta_{\min}$. Second, the high-frequency response is somewhat attenuated; slightly for small incident angles, and significantly for angles near $\theta_{\min}$. Third, the maximum ITD is slightly increased at close distances. These results can be incorporated into the simplified model by modifying a few parameters, summarized in table 4.2. For a distance r from the center of the sphere to the point source, we define the normalized distance $\rho = r/a$. The frequency-independent gain factor can be obtained through interpolation: given the theoretical gain for very close and very far distances, and for minimum and maximum incident angles, we can construct a function to estimate the intermediate values. The following empirical formula is found:

$$\gamma(\theta, \rho) = \left(\frac{10^{1.125/\rho} + 10^{-0.5/\rho}}{2} \right) + \left(\frac{10^{1.125/\rho} + 10^{-0.5/\rho}}{2} \right) \cos\left(\frac{\theta}{\theta_{\min}} \pi \right). \tag{4.21}$$

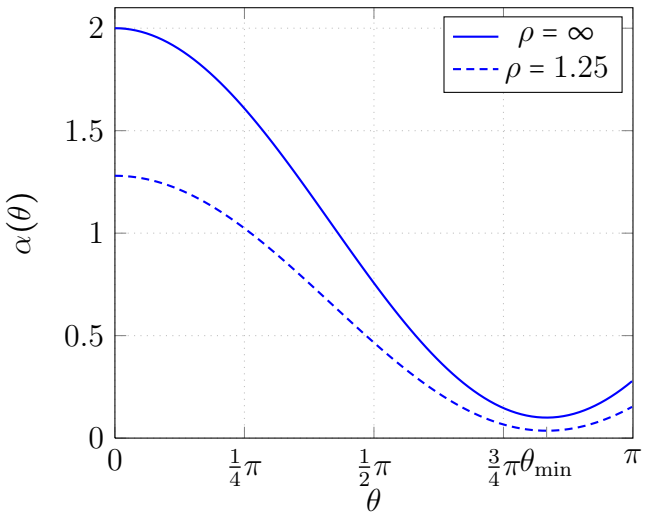


FIGURE 4.8. The head-model parameter α at $\rho = \infty$ (solid line) and $\rho = 1.25$ (dashed line).

The MATLAB code in listing 4.1 is somewhat more transparent. Figure 4.9 below demonstrates the effect.

The high-frequency attenuation can be incorporated by decreasing the maximum and minimum values of $\alpha(\theta, \rho)$ (see figure 4.8). This will in turn move the location of the filter zero to the left, toward the high-frequency region of the unit circle. The following empirical formulas match the exact solution well:

$$\alpha_{\max} = 2 - 0.9/\rho$$

$$\alpha_{\min} = 0.1 - 0.08/\rho$$

$$\alpha(\theta, \rho) = \left(\frac{\alpha_{\max} + \alpha_{\min}}{2}\right) + \left(\frac{\alpha_{\max} - \alpha_{\min}}{2}\right) \cos\left(\frac{\theta}{\theta_{\min}}\pi\right). \tag{4.22}$$

Figure 4.9 shows the effect both of γ and of the modified α function on the magnitude response of the head model.

LISTING 4.1. MATLAB code to calculate the gain factor γ.

```
% We map [150deg 0deg] --> [-1 1]. At rho=1.25, we have
% gain(0deg=1) = 18dB, gain(150deg=-1) = -8dB.
% We need amplitude gain, not dB so,
% g0   = 10^(18dB/20) = 7.9433
% g150 = 10^(-8dB/20) = 0.3981
% Slope m = (7.9433-0.3981)/(1- -1) = 7.5452/2 = 3.7726
% y-intercept b = y0 - m*x0 = 7.9433 - 3.7726*1 = 4.1707
%
% To generalize to other dists, we first scale endpoints:
% we scale the dBs linearly, not the gain.
% Want 1/rho=[0.8 0] --> g0=[18dB 0dB]
%  or g0_dB = rho * (18/0.8) = rho * 22.5
%       g0 = 10^(g0_dB/20)
% want 1/rho=[0.8 0] --> g150=[-8dB 0dB]
%  or g150_dB = rho * (-8/0.8) = rho * -10
%
% Then, m = (g0-g150)/2 and b=g0-m

function g = gamma(theta, rho)

% maps [150 0] --> [-1 1]
th = cos( theta*pi / deg2rad(150) );

rh = 1 / rho;                    % maps [1.25 inf] --> [0.8 0]

g0 = 10^( 22.5 *rh / 20 );       %22.5 = 18/0.8
g150 = 10^( -10 *rh / 20 );      %-10 = -8/0.8

m = (g0-g150)/2;
b = g0-m;

g = m *th + b;
```

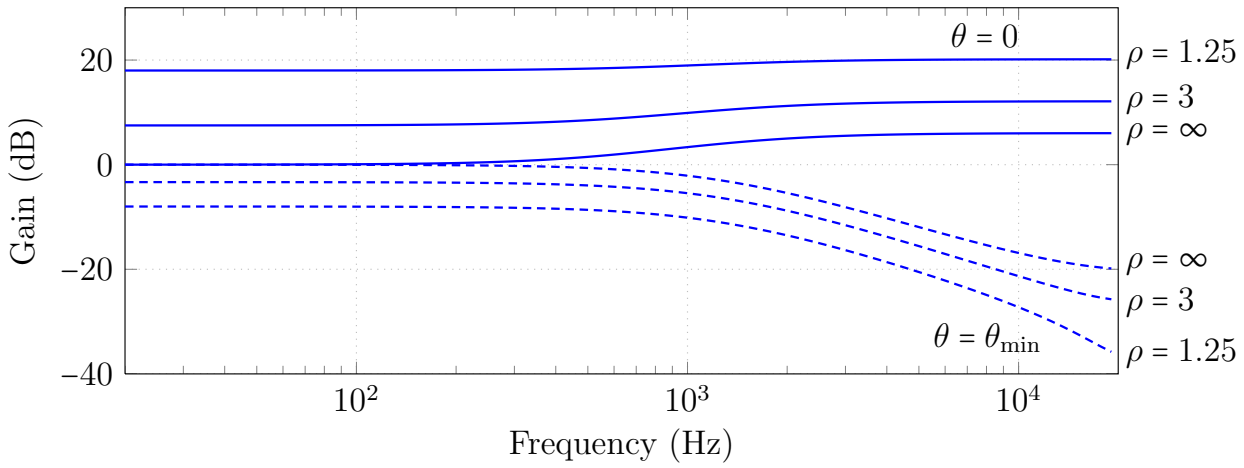


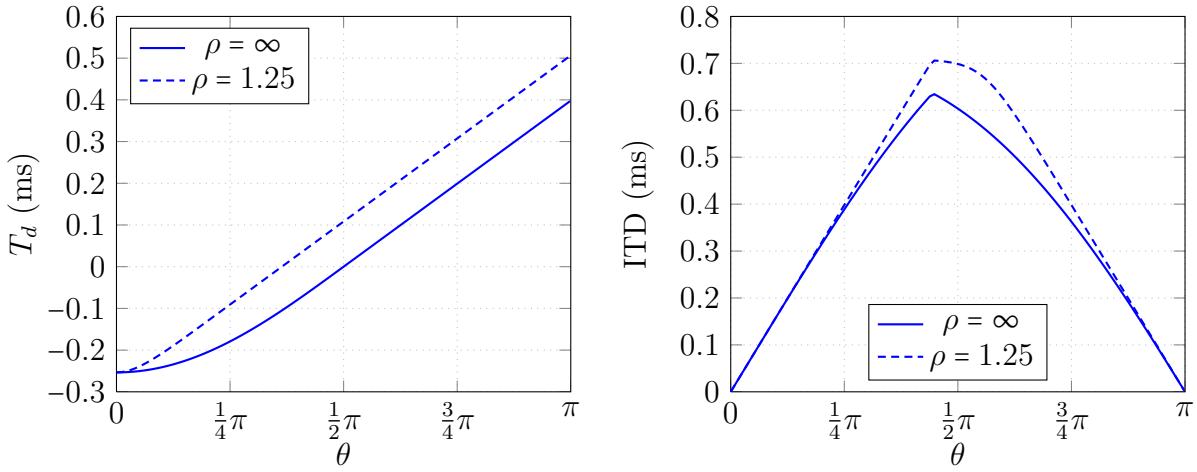
FIGURE 4.9. Magnitude response of the head model for $\theta = 0$ (solid lines) and $\theta = \theta_{\min}$ (dashed lines), for three ranges. Each doubling of ρ roughly halves the value of γ. The effect of the range-dependent α function is visible by comparing the high-frequency response of large and small values of ρ.

Except for the frequency-indepedent delay, the exact response of an ideal rigid sphere is minimum-phase at all ranges. Using the range-dependent α function, values of ρ less than about 1.15 lead to a head-shadowing filter with a pole outside the unit circle, and therefore no longer minimum-phase. However it seems unlikely that this simple model could be accurate at such a close distance. Filters with a value of ρ below about 1.25 have not been tested and are probably not valid models. For all values of $\rho \geq 1.25$, the minimum-phase property means that the phase is uniquely determined by the magnitude response. Therefore, by empirically modelling the magnitude response we have also correctly modelled the group delay.

Finally, the frequency-independent time delay becomes:

$$T_d(\theta, \rho) = \begin{cases} \frac{a}{c}\left(\sqrt{\rho^2 - 2\rho\cos\theta + 1} - \rho\right) & \text{if } 0 \leq |\theta| < \vartheta \\ \frac{a}{c}\left(\theta - \vartheta + \sqrt{\rho^2 - 1} - \rho\right) & \text{if } \vartheta \leq |\theta| \leq \pi, \end{cases} \tag{4.23}$$

where $\vartheta = \cos^{-1}(1/\rho)$. Figure 4.10 shows the delay and resulting ITD, for $\rho = \infty$ and $\rho = 1.25$.



(A) T_d as a function of incident angle

(B) ITD as a function of source angle

FIGURE 4.10. Frequency-independent time delay, for $\rho = \infty$ (solid line) and $\rho = 1.25$ (dashed line).

The full discrete-time head model is obtained by altering α and T_d and adding the gain-scaling factor. Explicitly,

$$H(z,\theta,\rho) = \gamma(\theta,\rho)\frac{(\omega_0 + \alpha(\theta,\rho)) + (\omega_0 - \alpha(\theta,\rho))\, z^{-1}}{(\omega_0 + 1) + (\omega_0 - 1)\, z^{-1}} z^{-T_d(\theta,\rho)F_s}. \tag{4.24}$$

This is the form of the model that is used for system implementation.

4.2. Target Location Processing

This section first discusses the physical acoustics of two-speaker sound reproduction. The various signal processing methods which compensate for these acoustics are then described and analyzed. We suggest new extensions for crosstalk cancellation equalization and the high-frequency energy-based model.

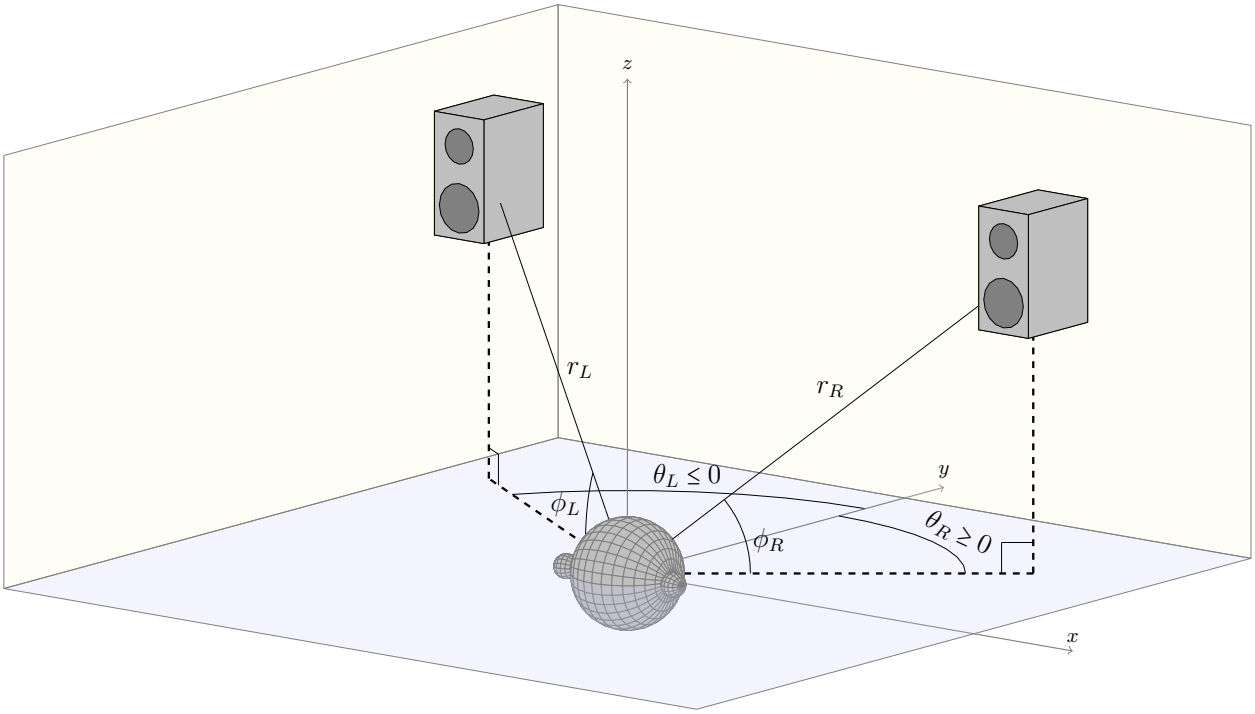


FIGURE 4.11. Basic geometry of the listening environment.

4.2.1. Geometry and Acoustics of Loudspeaker Reproduction

The basic configuration for two-channel loudspeaker playback is shown in figure 4.11. The head-centered coordinate system places positive x to the right, positive y forward, and positive z up. Azimuth angles are measured starting at the y-axis and ending on the ray from the center of the head to the projection of the loudspeaker into the xy-plane. Values range from $-180° \leq \theta \leq 180°$, with angles to the right of the median plane positive and angles to the left negative. Elevation angles are measured starting at the ray from the center of the head to the projection of the loudspeaker into the xy-plane, and ending at the ray from the center of the head to the loudspeaker. Values are in the range $0° \leq \phi \leq 90°$. As mentioned, the ears are at $\theta = \pm 100°$, $\phi = 0°$.

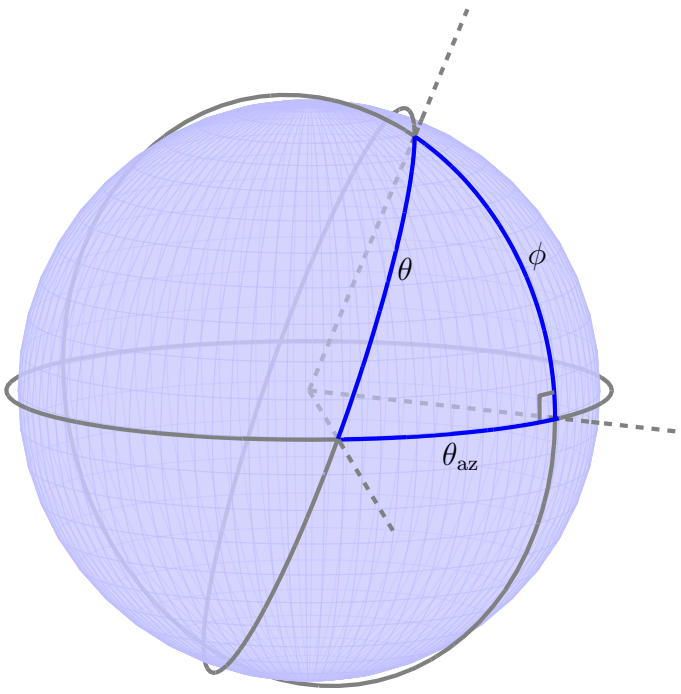


FIGURE 4.12. The Pythagorean theorem on a sphere.

The head model makes no distinction between azimuth and elevation angles, but depends only on the net (smallest) incident angle between source and ear location. Given azimuth angle θ_{az} and elevation angle ϕ, we must compute the incident angle θ. This is readily done using the "Pythagorean theorem on a sphere" [49]:

$$\cos(\theta) = \cos(\theta_{\text{az}})\cos(\phi). \tag{4.25}$$

Figure 4.12 illustrates. In a slight abuse of notation, we use θ for the net incident angle when referring to the head model, but for the azimuth angle only when discussing the geometrical relationship between listener and loudspeaker. Context will always make this distinction clear.

Suppose we have two input signals which we wish to convey to the left and right ears of a listener. We cannot send these signals directly to the loudspeakers, because of acoustic crosstalk: the left loudspeaker signal y_L is transmitted to the left ear, but is (undesirably) sent to the right ear as well; similarly with the right loudspeaker signal y_R. Hence the signal at each ear is the sum of the two received signals (see figure 4.13). Another reason we cannot send the desired signals directly to the loudspeaker is that the paths from the loudspeakers to the ears are not acoustically transparent. They include filtering both from air propagation and from HRTFs. We must introduce a processing matrix $\mathbf{C}$ which compensates for the physical acoustics of loudspeaker playback. We can denote the situation using matrix notation as follows:

$$\begin{bmatrix} e_L \\ e_R \end{bmatrix} = \begin{bmatrix} A_{LL} & A_{RL} \\ A_{LR} & A_{RR} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} x_L \\ x_R \end{bmatrix}, \tag{4.26}$$

where A_{xy} denotes the transmission path from speaker x to ear y. More compactly:

$$\mathbf{e} = \mathbf{ACx}. \tag{4.27}$$

Hence the loudspeaker signals $\mathbf{y} = \mathbf{Cx}$, and the ear signals $\mathbf{e} = \mathbf{Ay}$. Though obvious, it is worth pointing out that the speaker signals are the same for all listeners, while the transfer matrix varies.

The acoustic-transfer matrix $\mathbf{A}$ can be factored as:

$$\begin{aligned} \mathbf{A} &= \mathbf{HS} \\ &= \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix} \begin{bmatrix} S_L A_L & 0 \\ 0 & S_R A_R \end{bmatrix}. \end{aligned} \tag{4.28}$$

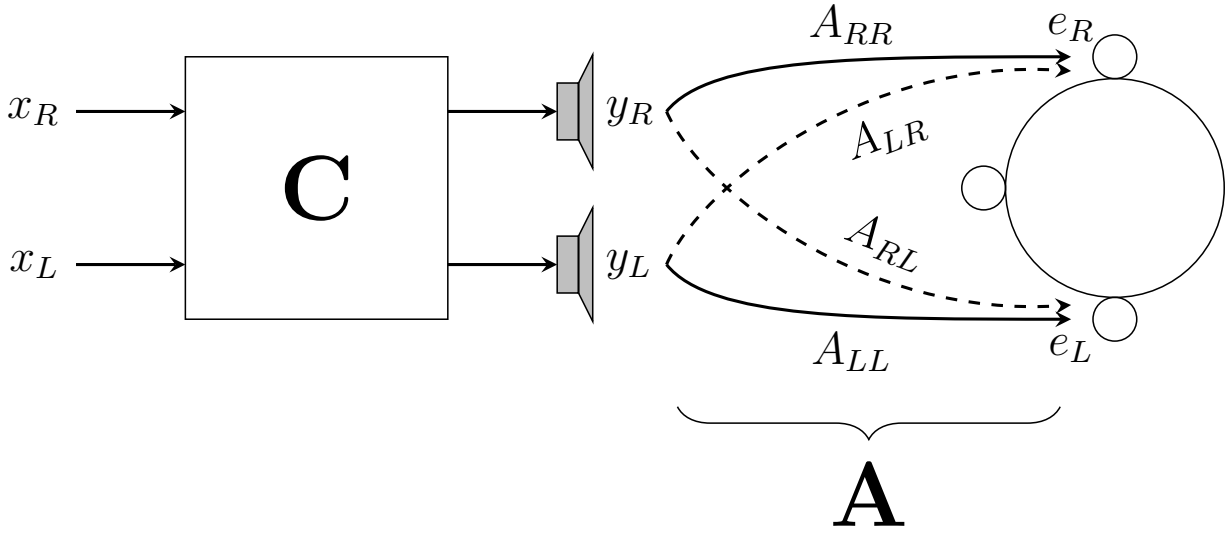


FIGURE 4.13. Transmission paths from system input to the ears.

H represents head-related transfer effects (normalized by the free-field response), and **S** represents the speaker- and air-transfers. S_x is the transfer function of speaker x (in practice, this is often set to unity, though it can have a significant impact); A_x is the transfer function through the air, from speaker x to the center point of the head, as if were there no head present. For most purposes we can assume free-field propagation, modelled as a frequency-independent delay and attenuation, ignoring high-frequency rolloff and other air effects [45]:

$$A_x = \frac{1}{r_x} e^{-jkr_x}, \tag{4.29}$$

where r_x is the distance from speaker x to the center of the head and k is the wavenumber, $k = 2\pi f/c = \omega/c$.

In the discussion that follows, we assume that the head model used for processing is identical to the head of the intended listener. Chapter 5 addresses the more practical case when this is not true.[1]

4.2.2. Direct Path Compensation

Consider choosing a specific target location. We adopt the convention that functions related to this spot are denoted with a subscript or superscript (0). The simplest approach is to compensate for the uneven path lengths to the speakers and for the speaker transfer functions:

$$\begin{aligned}\mathbf{C} &= \mathbf{S}_{(0)}^{-1} \\ &= \begin{bmatrix} \frac{1}{S_L^{(0)} A_L^{(0)}} & 0 \\ 0 & \frac{1}{S_R^{(0)} A_R^{(0)}} \end{bmatrix}. \end{aligned} \tag{4.30}$$

This leads to target ear signals

$$\mathbf{e}_{(0)} = \mathbf{H}_{(0)}\mathbf{x}. \tag{4.31}$$

We refer to this as *path length compensation*. In the context of target location processing, path length compensation is always implied.

The direct paths are still not acoustically transparent, because of head effects. We can include their inverses in the system matrix, which we term *direct path compensation*:

$$\mathbf{C} = \mathbf{S}_{(0)}^{-1} \begin{bmatrix} \frac{1}{H_{LL}^{(0)}} & 0 \\ 0 & \frac{1}{H_{RR}^{(0)}} \end{bmatrix}. \tag{4.32}$$

[1] Unfortunately, due to limited resources, we were unable to obtain a listener with a perfectly spherical head.

The target ear signals become:

$$\mathbf{e}_{(0)} = \begin{bmatrix} 1 & H_{RL}^{(0)}/H_{RR}^{(0)} \\ H_{LR}^{(0)}/H_{LL}^{(0)} & 1 \end{bmatrix} \mathbf{x}. \tag{4.33}$$

To better understand this result, we define the *interaural transfer functions* as:

$$\mathrm{ITF}_L = \frac{H_{LR}}{H_{LL}}, \qquad \mathrm{ITF}_R = \frac{H_{RL}}{H_{RR}}. \tag{4.34}$$

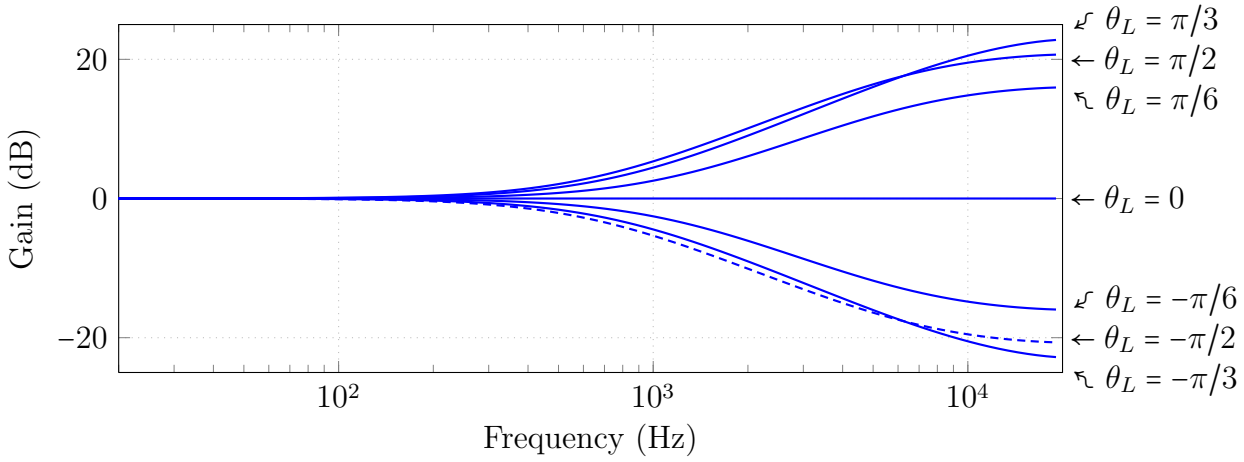
The ITF ratios compare, for a given loudspeaker, the sound that reaches the unintended ear to the sound reaching the intended ear. Figure 4.14 shows the magnitude response and group delay of ITF_L. Notice that, for non-centered listeners, it is quite possible for the unintended ear to be closer to the source than the intended ear. Using this definition, we can write the target ear signals as the sum of the unmodified intended channel plus the unintended channel, filtered by the appropriate ITF:

$$\begin{bmatrix} e_L^{(0)} \\ e_R^{(0)} \end{bmatrix} = \begin{bmatrix} x_L + \mathrm{ITF}_R^{(0)} x_R \\ \mathrm{ITF}_L^{(0)} x_L + x_R \end{bmatrix}. \tag{4.35}$$

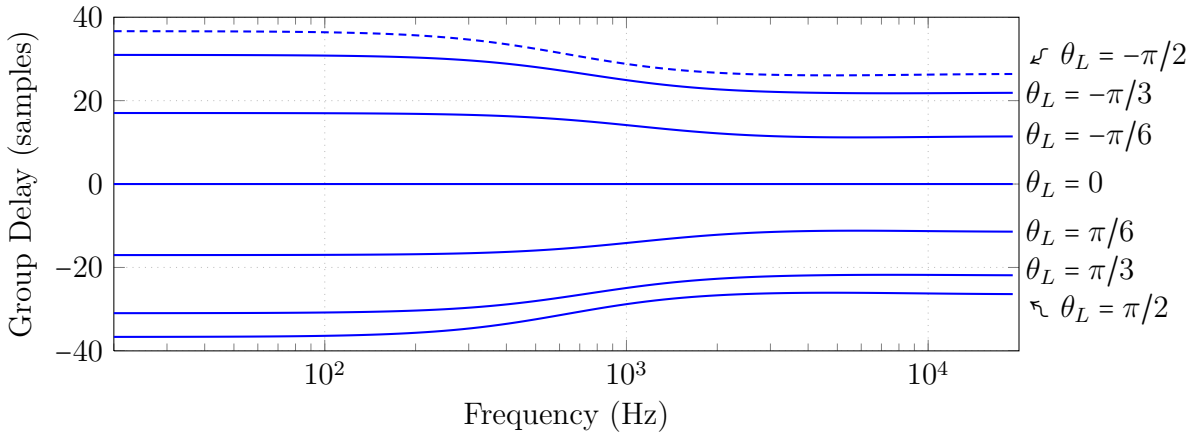
From figure 4.14 it is clear that, for low frequencies, the crosstalk signal is equally as significant as the direct signal.

4.2.3. Crosstalk Cancellation

The mathematics of ideal crosstalk cancellation are well-known [12, 48, 98]. Suppose we wish to convey the input signals to the ears exactly. It follows immediately from equation (4.27) that, in order to achieve our goal of $\mathbf{e}_{(0)} = \mathbf{x}$, we must have $\mathbf{C} = \mathbf{A}_{(0)}^{-1}$. Based on equation (4.28),



(A) Magnitude Response



(B) Group Delay

FIGURE 4.14. ITF_L: magnitude response and group delay. Shown for loudspeaker angles from $\theta_L = -\pi/2$ (dashed line) to $\theta_L = \pi/2$ in steps of $\pi/6$, with $\phi_L = 0$. (Notice that the independent variable is the loudspeaker angle, and not the incident angle.)

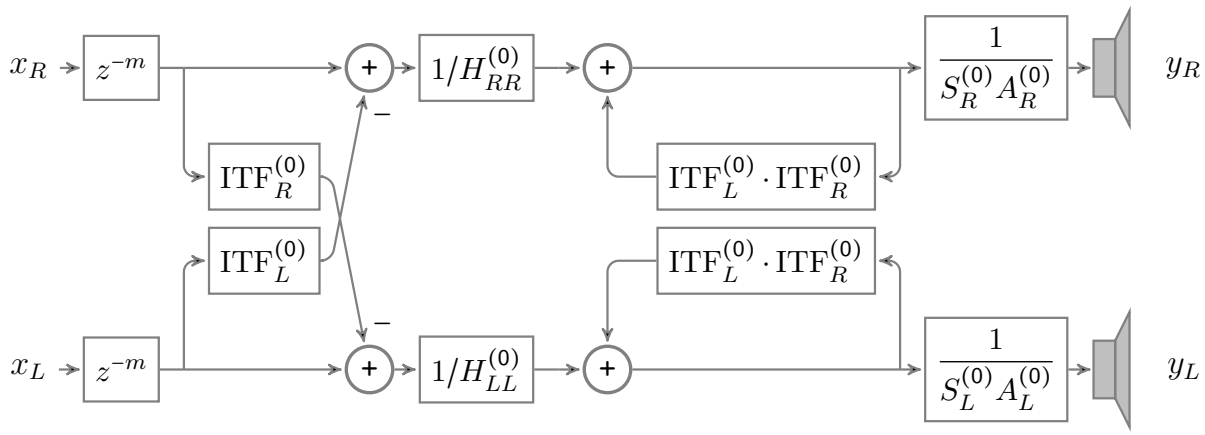


FIGURE 4.15. Crosstalk cancellation circuit.

we can write

$$\begin{aligned}\mathbf{C} &= \mathbf{S}_{(0)}^{-1}\mathbf{H}_{(0)}^{-1} \\ &= \begin{bmatrix} \frac{1}{S_L^{(0)} A_L^{(0)}} & 0 \\ 0 & \frac{1}{S_R^{(0)} A_R^{(0)}} \end{bmatrix} \begin{bmatrix} H_{RR}^{(0)} & -H_{RL}^{(0)} \\ -H_{LR}^{(0)} & H_{LL}^{(0)} \end{bmatrix} \frac{1}{D^{(0)}},\end{aligned} \tag{4.36}$$

where $D^{(0)} = H_{LL}^{(0)} H_{RR}^{(0)} - H_{LR}^{(0)} H_{RL}^{(0)}$ is the determinant of $\mathbf{H}_{(0)}$. For realtime implementations, we must allow a certain amount of modelling delay to ensure causal filters:

$$\mathbf{C}_{\mathrm{rt}}(z) = z^{-m}\mathbf{C}(z). \tag{4.37}$$

The ear signals are thus merely delayed versions of the inputs. For simplicity, we ignore this in the following discussion. We should also note that this treatment assumes anechoic conditions and ignores the effects of room reflections.

We can rewrite the inverse head transfer matrix as:

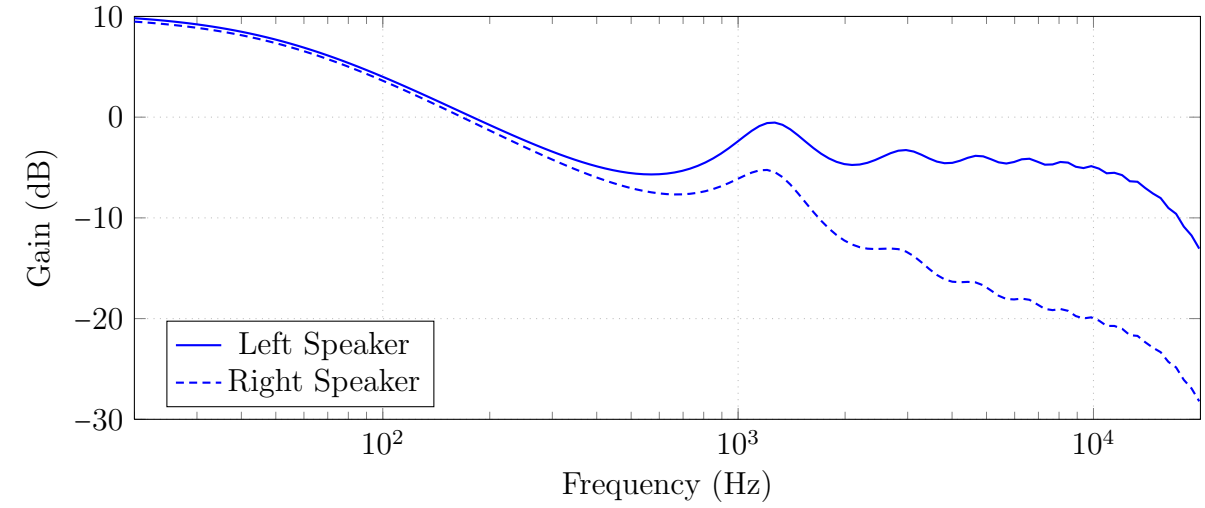
$$\mathbf{H}_{(0)}^{-1} = \begin{bmatrix} \frac{1}{H_{LL}^{(0)}} & 0 \\ 0 & \frac{1}{H_{RR}^{(0)}} \end{bmatrix} \begin{bmatrix} 1 & -\mathrm{ITF}_R^{(0)} \\ -\mathrm{ITF}_L^{(0)} & 1 \end{bmatrix} \frac{1}{1 - \mathrm{ITF}_L^{(0)} \mathrm{ITF}_R^{(0)}}. \tag{4.38}$$

The inverse HRTFs compensate for the speaker locations; the off-diagonal –ITF terms generate the crosstalk cancellation signals; and the scalar term compensates for higher-order crosstalks (crosstalk from the crosstalk cancellation signals, etc.). Figure 4.15 illustrates an implementation for a crosstalk canceller based on this factorization. Figure 4.16 shows the response when the left input is an impulse and the right input is silence.

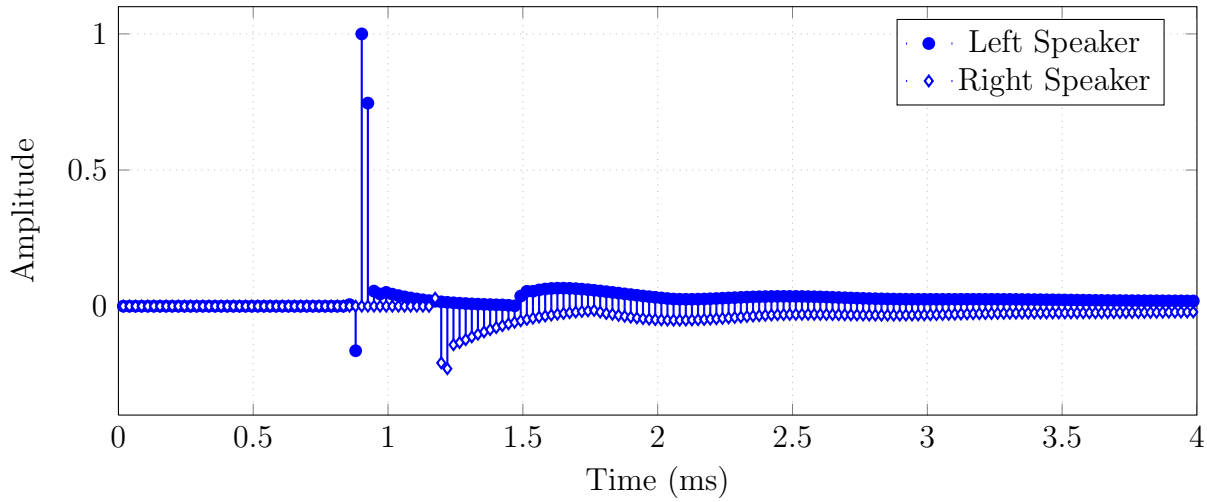
4.2.4. Spectral Equalization

The recursive section at the end of figure 4.15 implements overall spectral equalization after crosstalk cancellation. Since it is common to both channels, it cannot affect interaural differences, though it can obviously affect spectral spatial cues. In practice, this equalization amounts primarily to a bass boost. This is because there is little head-shadowing at low frequencies, and so significant bass energy must be cancelled at the contralateral ear.

The impulse response for this loop is fairly long. While it has not been rigorously demonstrated, there is strong evidence that a short time response is desirable for crosstalk cancellation [103, 104]. This makes intuitive sense because a longer time response could cause parts of the cancellation signal to arrive after the 1 ms integration time for summing localization (see figure 3.6 on page 67). Naïvely, we might consider varying the phase response to find another filter with the same magnitude response but a shorter time response, since overall "phase equalization" is not particularly important. Unfortunately, this is not possible.



(A) Magnitude Response



(B) Impulse Response

FIGURE 4.16. Magnitude and (normalized) impulse response of the crosstalk cancellation circuit. The left channel input is an impulse, and the right channel is silence. The speaker geometry is that of overhead speaker pair A in table 5.4 (page 148). The target is a centered listener directly under the line of the speakers. The impulse response decays for roughly another 4 ms.

The loop is an all-pole filter, and we assume causality and stability (discussed below). Every causal stable all-pole filter is minimum-phase.[2] Minimum-phase implies minimum energy delay: no other function with the same magnitude response will have a shorter time response.[3] However, we propose shortening the impulse response while maintaining almost the same magnitude response by the simple expedient of including a constant multiplier g_{EQ} in series with the ITF functions. The loop transfer function is then:

$$\mathrm{EQ}(z) = \frac{1}{1 - g_{\mathrm{EQ}}\mathrm{ITF}_L(z)\mathrm{ITF}_R(z)} \tag{4.39}$$

(omitting the (0) notation for simplicity). The magnitude of the frequency response is [100]:

$$\left|\mathrm{EQ}(e^{j\omega})\right| = \frac{1}{\sqrt{1 + g_{\mathrm{EQ}}^2\left|\mathrm{ITF}_L\mathrm{ITF}_R\right|^2 - 2g_{\mathrm{EQ}}\left|\mathrm{ITF}_L\mathrm{ITF}_R\right|\cos\left(\measuredangle\left(\mathrm{ITF}_L\mathrm{ITF}_R\right) - T_d^{\mathrm{EQ}}\omega F_s\right)}} \tag{4.40}$$

(where each ITF function is evaluated at $z = e^{j\omega}$). T_d^{EQ} is the net time delay (in seconds) of $\mathrm{ITF}_L\mathrm{ITF}_R$. Figure 4.17 shows the magnitude response for three values of g_{EQ}. It also shows the accumulated energy over time, normalized by the total energy of the response. That is:

$$E_{\mathrm{EQ}}[n] = \frac{\sum_{m=0}^{n}\left|h_{\mathrm{EQ}}\right|^2}{\sum_{m=0}^{\infty}\left|h_{\mathrm{EQ}}\right|^2}. \tag{4.41}$$

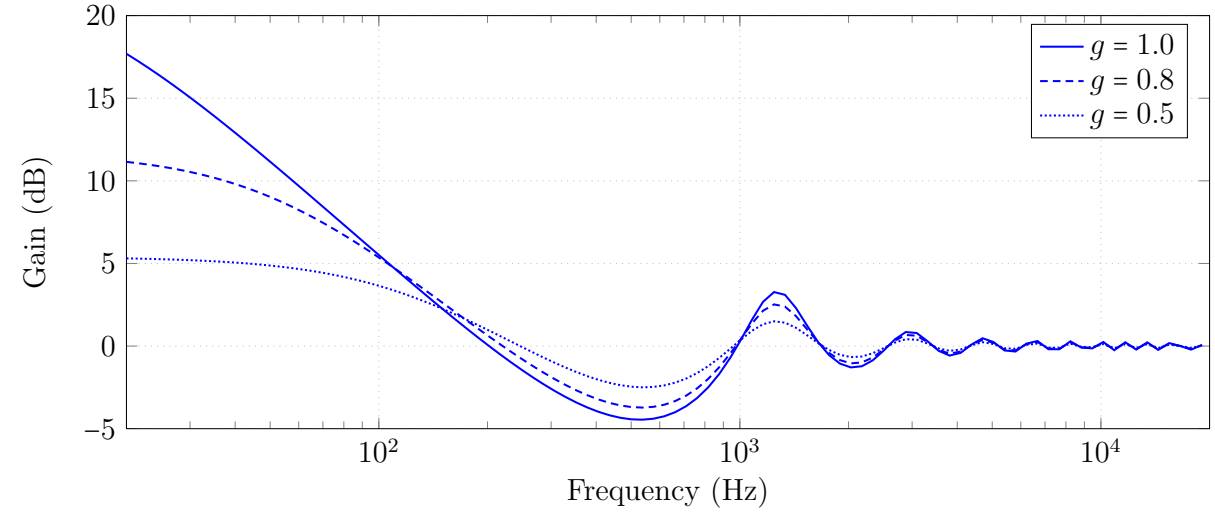
This is the effective length of the response [82]. The evaluations in chapter 5 all use $g_{\mathrm{EQ}} = 0.8$.

[2] The Region of Convergence (ROC) of the z-transform cannot contain any poles. Stability implies that the ROC includes the unit circle. Causality implies that the ROC extends outward from the outermost pole; hence all poles are inside the unit circle. There are no zeros (except possibly at $z = 0$), so all poles and zeros are inside the unit circle, the definition of minimum-phase [105].

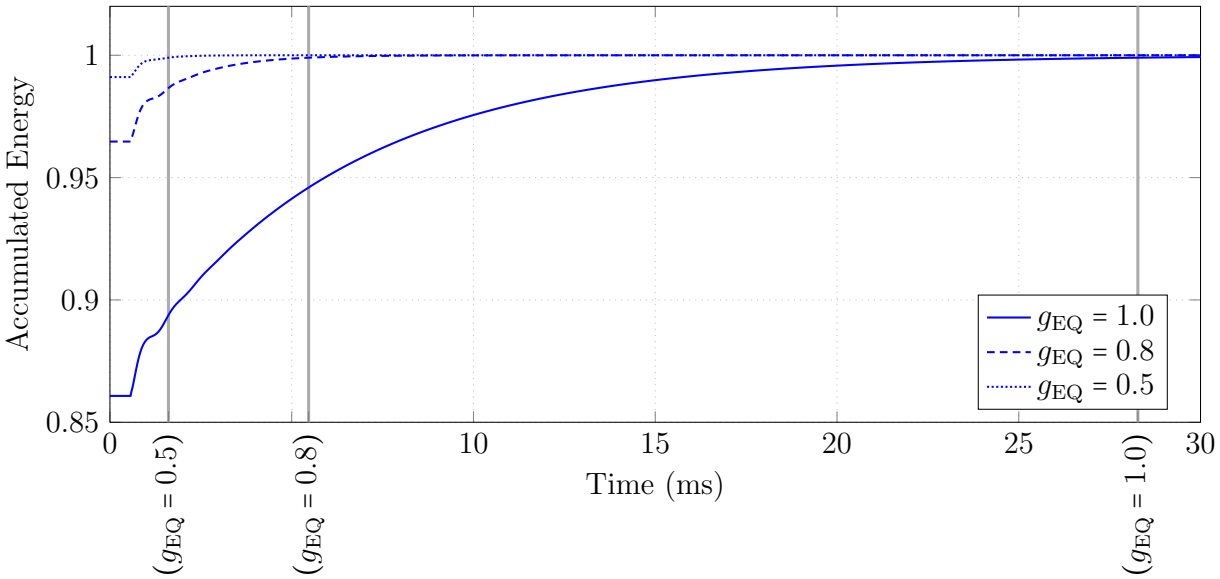
[3] Formally: consider all sequences $h_{\mathrm{EQ}}[n]$ having the same magnitude response $\left|H_{\mathrm{EQ}}(e^{j\omega})\right|$. Then the function

$$E_{\mathrm{EQ}}[n] = \sum_{m=0}^{n}\left|h_{\mathrm{EQ}}[m]\right|^2$$

is maximized when $h_{\mathrm{EQ}}[n]$ is minimum-phase, $\forall n \geq 0$ [105].



(A) Magnitude Response



(B) Accumulated Energy. The three vertical lines indicate the times at which 99.9% of the energy has arrived.

FIGURE 4.17. Magnitude response and normalized accumulated energy of the recursive EQ loop for three values of the gain factor g_{EQ}. The speaker geometry is that of pair A in table 5.4 (page 148). The target location is centered and directly under the line of the speakers.

To ensure a realizable circuit, the loop must contain at least one sample delay. This is true in almost all practical configurations. However, it can become false when the angle between the loudspeakers is very small from the perspective of the listener, due to a small speaker span or very distant listener.[4] In this case, a single sample delay must be inserted and the equalization will not be theoretically exact. The practical effect is usually small and unlikely to make a noticeable difference with a spherical head model. To ensure stability, the poles of the loop z-transform must have magnitude less than 1. This is true when

$$g_{\mathrm{EQ}} \left| \mathrm{ITF}_L \left(e^{j\omega} \right) \mathrm{ITF}_R \left(e^{j\omega} \right) \right| < 1 \qquad \forall \omega. \tag{4.42}$$

While we do not prove it, this is always the case for forward-facing listeners. However, the magnitude can come arbitrarily close to unity, particularly for loudspeakers with a narrow span. For numerical stability, it is advisable to choose $g_{\mathrm{EQ}} \leq 0.98$ or so. Further discussion of realizability and stability is found in [48].

While we have ensured stability, the system gain must not exceed 0 dB at any frequency to avoid clipping. Clearly the loop transfer curve does exceed this threshold, so we must limit the gain of the input, reducing the dynamic range of the system. The maximum loop gain occurs at DC, so we must reduce the input gain by that amount. A lower value of g_{EQ} therefore requires less dynamic range reduction. In practice, complex source signals will usually not attain the theoretical worst-case, in large part because they contain little energy at very low frequencies. We can quickly find the required gain reduction by processing

[4] As a practical reference, the delay was never less than four samples for any of the configurations used in chapter 5.

representative source material, monitoring the output of the loop, and reducing the input gain until there is no clipping.

4.2.5. Energy Compensation

It is beneficial to improve on direct path compensation without requiring crosstalk cancellation. For example, it is well-known that crosstalk cancellation is difficult and prone to artifacts at high frequencies [29, 48]. We use the method described by Gardner [48], generalized for arbitrary sources. We first divide each input channel into low and high frequency bands at a specified crossover frequency. Gain scaling is applied to the high frequencies so that the sum of the direct and crosstalk signals produces the correct high-frequency power at each ear. The bands are then summed and processed with a bandlimited crosstalk canceller: crosstalk cancellation is bypassed at high frequencies, leaving only direct path compensation. This is done by placing a lowpass filter LP in series with each ITF term in the circuit. We refer to this overall process as *energy compensation* (with implied direct path compensation). Conceptually, it is important to note that we can declare all frequencies "high frequencies" and use energy compensation in lieu of crosstalk cancellation. Figure 4.18 shows the complete processing chain of the system.

In the following discussion, it is understood that we are considering only the high-frequency portion of the signals and transfer functions. To compute the energy scaling gains, we model the left and right channels as incoherent white noise. This has two consequences: first, the high-frequency energy in each channel is evenly distributed across frequencies; second, the direct and crosstalk signals combine incoherently at the ears, so that the powers add (see section A.3).

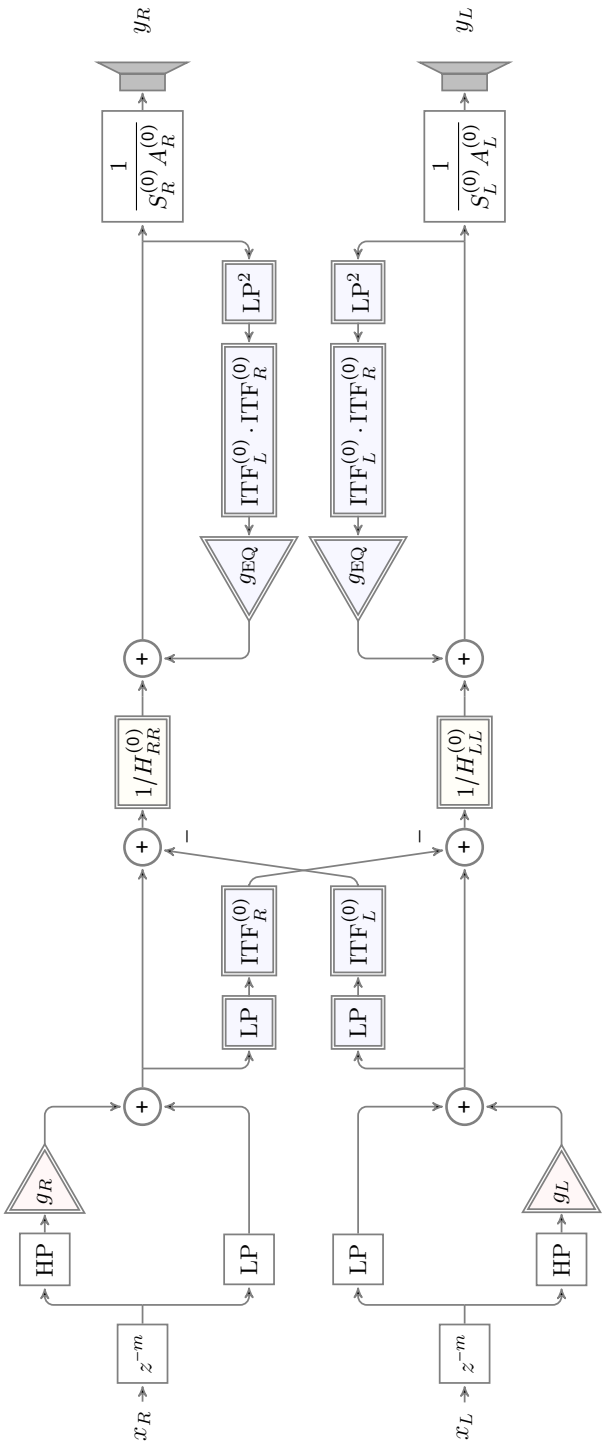


FIGURE 4.18. Complete processing chain of the system. A double border indicates a component that may be enabled or disabled.

We begin by adding the scaling gains to the direct path compensation circuit:

$$\begin{bmatrix} e_L \\ e_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{bmatrix} \begin{bmatrix} g_L H_{LL}^{-1} & 0 \\ 0 & g_R H_{RR}^{-1} \end{bmatrix} \begin{bmatrix} x_L \\ x_R \end{bmatrix}. \tag{4.43}$$

Next we replace signals by their power, and transfer functions by their energy:

$$\begin{aligned} \begin{bmatrix} \sigma_{e_L}^2 \\ \sigma_{e_R}^2 \end{bmatrix} &= \begin{bmatrix} 1 & E_{\mathrm{ITF}_R} \\ E_{\mathrm{ITF}_L} & 1 \end{bmatrix} \begin{bmatrix} g_L^2 & 0 \\ 0 & g_R^2 \end{bmatrix} \begin{bmatrix} \sigma_{x_L}^2 \\ \sigma_{x_R}^2 \end{bmatrix} \\ &= \begin{bmatrix} g_L^2 \sigma_{x_L}^2 + g_R^2 \sigma_{x_R}^2 E_{\mathrm{ITF}_R} \\ g_L^2 \sigma_{x_L}^2 E_{\mathrm{ITF}_L} + g_R^2 \sigma_{x_R}^2 \end{bmatrix}. \end{aligned} \tag{4.44}$$

Finally, we can solve for the gains that will make the power at the ears equal to the input power:

$$\begin{aligned} \begin{bmatrix} g_L^2 \\ g_R^2 \end{bmatrix} &= \left(\begin{bmatrix} 1 & E_{\mathrm{ITF}_R} \\ E_{\mathrm{ITF}_L} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{x_L}^2 \\ \sigma_{x_R}^2 \end{bmatrix} \right) \begin{bmatrix} 1/\sigma_{x_L}^2 & 0 \\ 0 & 1/\sigma_{x_R}^2 \end{bmatrix} \\ &= \frac{1}{1 - E_{\mathrm{ITF}_L} E_{\mathrm{ITF}_R}} \begin{bmatrix} 1 - E_{\mathrm{ITF}_R} \left(\frac{\sigma_{x_R}^2}{\sigma_{x_L}^2} \right) \\ 1 - E_{\mathrm{ITF}_L} \left(\frac{\sigma_{x_L}^2}{\sigma_{x_R}^2} \right) \end{bmatrix} \end{aligned} \tag{4.45}$$

(note that the matrix-vector multiplication is not associative, because of the asymmetric dimensionality of the column vector). If either row on the right-hand side is negative, there is no real solution. This will happen, for example, using binaural synthesis to project a virtual source that is more lateral than the loudspeakers. The desired IID is greater than what results when only one speaker is emitting a signal; any power radiated from the other speaker can only reduce the IID. In this case, we set the unrealizable gain equal to 0, and scale the

other gain so that the total power at both ears combined equals the total desired power. The total power and total desired power can be found by adding the two rows of equation (4.44):

$$\sigma_{e_L}^2 + \sigma_{e_R}^2 = g_L^2 \sigma_{x_L}^2 \left(1 + \mathrm{ITF}_L\right) + g_R^2 \sigma_{x_R}^2 \left(1 + \mathrm{ITF}_R\right). \tag{4.46}$$

We set the nonzero gain so that

$$\sigma_{e_L}^2 + \sigma_{e_R}^2 = \sigma_{x_L}^2 + \sigma_{x_R}^2. \tag{4.47}$$

Signal power is computed by taking a running mean-square average over a specified time window. The evaluations in chapter 5 used a window of 120 ms, which is roughly comparable to the loudness integration time of the auditory system [143]. For a crossover frequency of ω_X (in radians/sample), the high-frequency energy of the ITF transfer function is found by integating (see section A.3):[5]

$$\begin{aligned} E_{\mathrm{ITF}} &= \frac{1}{\pi - \omega_X} \int_{\omega_X}^{\pi} \left|\mathrm{ITF}\left(e^{j\omega}\right)\right|^2 \mathrm{d}\omega \\ &= \frac{1}{\pi - \omega_X} \left(\frac{\gamma_{\mathrm{contra}}}{\gamma_{\mathrm{ips}}}\right)^2 \\ &\quad \times \left(\frac{2\omega_0 \left((\pi/2) - \lambda_x\right)\left(\alpha_{\mathrm{contra}}^2 - \alpha_{\mathrm{ips}}^2\right) + \alpha_{\mathrm{ips}} \left(\pi - \omega_X\right)\left(\omega_0^2 - \alpha_{\mathrm{contra}}^2\right)}{\alpha_{\mathrm{ips}}\left(\omega_0^2 - \alpha_{\mathrm{ips}}^2\right)}\right), \end{aligned} \tag{4.48}$$

where

$$\lambda_X = \tan^{-1}\left(\frac{\tan\left(\omega_X/2\right)\alpha_{\mathrm{ips}}}{\omega_0}\right). \tag{4.49}$$

[5] A word of caution:

$$E_{\mathrm{ITF}} \neq \frac{E_{H_{\mathrm{contra}}}}{E_{H_{\mathrm{ips}}}}.$$

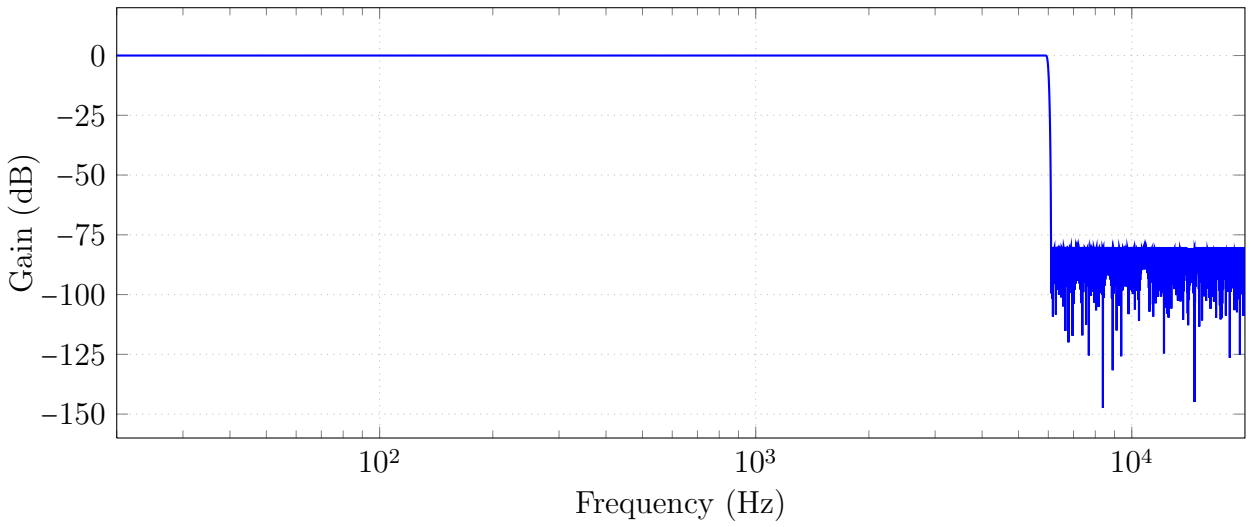


FIGURE 4.19. Lowpass FIR filter with a cutoff of 6 kHz.

The crossover filters must be linear phase, to avoid alignment errors when the bands are recombined. The lowpass filters LP were designed in MATLAB. They are 1022-order equiripple filters with a transition band of 200 Hz, designed with the Parks-McClellan algorithm (see [105] for an overview of this method and further references). Stopband rejection is better than 80 dB. Filters were designed for cutoffs of 2, 4, 6, 8, 10 and 14 kHz. It is also trivial to apply energy compensation to all or no frequencies, equivalent to cutoffs at 0 Hz or the Nyquist frequency, respectively. Highpass filters are derived simply as the complement, HP = 1 – LP. Figure 4.19 shows the lowpass filter for a cutoff of 6 kHz.

In section 5.4.1 we suggest a subband approach. For reference, the ITF energy in a band from ω_1 to ω_2 is given by:

$$E_{\mathrm{ITF}_{\mathrm{band}}} = \frac{1}{\omega_2 - \omega_1} \left(\frac{\gamma_{\mathrm{contra}}}{\gamma_{\mathrm{ips}}} \right)^2 \times \left(\frac{2\omega_0 \left(\lambda_2 - \lambda_1\right) \left(\alpha_{\mathrm{contra}}^2 - \alpha_{\mathrm{ips}}^2\right) + \alpha_{\mathrm{ips}} (\omega_2 - \omega_1) \left(\omega_0^2 - \alpha_{\mathrm{contra}}^2\right)}{\alpha_{\mathrm{ips}} \left(\omega_0^2 - \alpha_{\mathrm{ips}}^2\right)} \right). \tag{4.50}$$

4.3. Software Realization

The system is implemented in SuperCollider, an audio processing language [90, 130]. A number of factors make SuperCollider a natural choice:

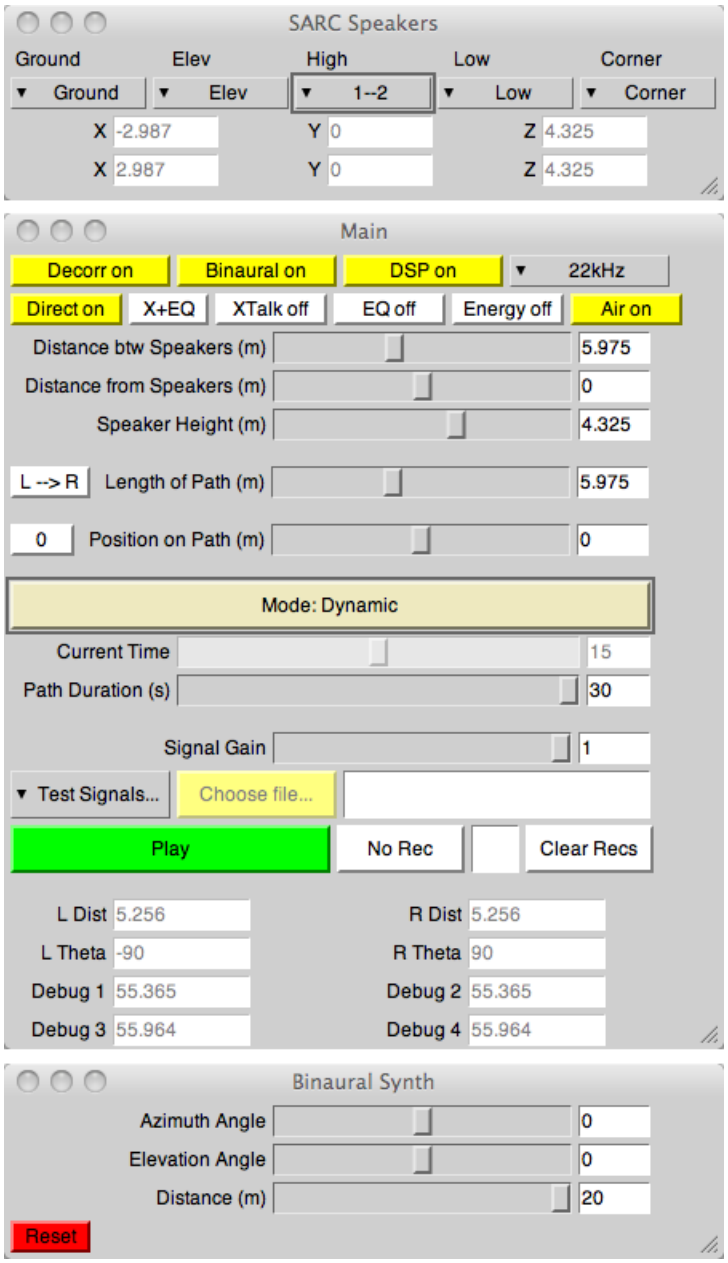
- It is highly optimized for real-time operation, but can also render audio off-line
- Internal processing uses 32-bit floats for high numerical accuracy
- There are numerous built-in tools and extension libraries for spectral processing and audio stream manipulation
- It supports arbitrary numbers of input and output channels
- It is based on Smalltalk, and hence allows very general routing and DSP
- It has built-in tools for GUI design
- It is free open-source software (FOSS)

The primary drawback is that SuperCollider is best supported under Mac OS X. Ports do exist for Linux and Windows, but they are substantially more difficult to install and maintain. Nevertheless, they are usable and rapidly maturing. SuperCollider is legitimately considered cross-platform.

SuperCollider uses a block-based synthesis architecture, meaning that each unit generator computes a fixed number of output samples before the next unit is called. This is very efficient but makes single-sample feedback extremely difficult (though not technically impossible). During the listening evaluations (see section 5.3), the block size was set to 4 samples, the smallest practical size for realtime operation. This means that the EQ loop in figure 4.15 possessed a minimum 4-sample delay. For theoretical delays shorter than this, the actual equalization will deviate, though interaural differences will not be affected. Fortunately, very short delays are required only for very narrow speaker pairs, or when the target is very distant

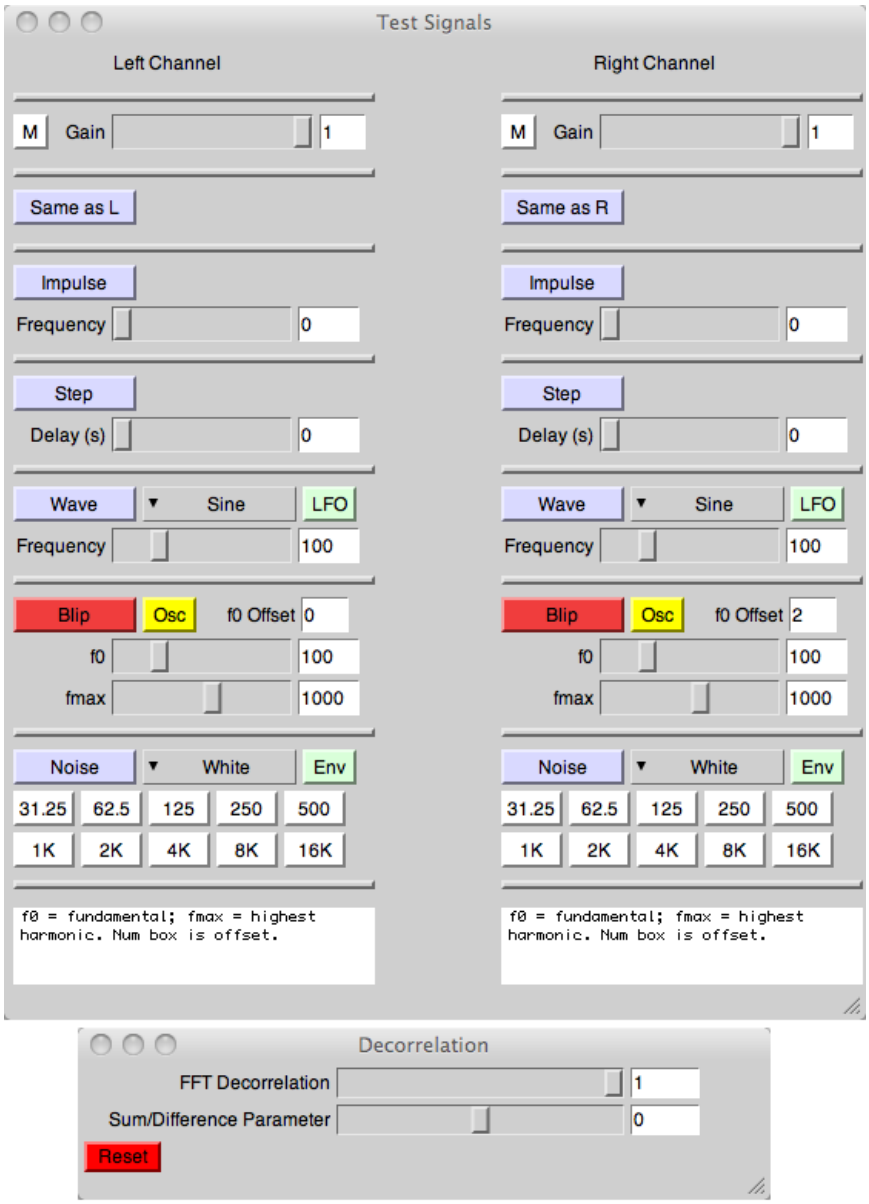
from the speakers and highly off-center. Neither situation occurred during the listening. For non-realtime operation, the block size, and hence the minimum feedback delay, can be set to 1 sample.

Realtime operation is controlled through five GUI windows, shown in figure 4.20. In the primary window, users specify which components of the system's signal processing to use (these can be enabled or disabled dynamically). They also define the basic geometry of the listening environment (an auxiliary window provides shortcuts for the SARC layout described in section 5.3.1), and set the duration and spatial extent of the target location's path. A static mode maintains a fixed target location for an indefinite duration. Speaker geometry is input using three parameters: distance between the speakers; distance from their projection into the xy-plane to the target line; and height above ear level. Source signals are selected in a second window. These include impulse and step functions for numerical testing, as well as the noise and oscillator signals described in section 5.3.2. File input is also supported. Live input from an audio interface is not currently supported but could be added quite easily. A fourth window specifies the correlation of the source channels (see appendix B); the final window synthesizes binaural cues using the spherical head model. Binaural synthesis using the CIPIC HRIR database (see section 5.1.1) is also implemented but was not used for this dissertation.



(A) Shortcuts for SARC speaker layout; primary system controls; binaural synthesis controls.

FIGURE 4.20. System GUI.



(B) Test signals; decorrelation controls.

FIGURE 4.20 (CONT'D). System GUI.

CHAPTER 5

System Evaluation

This chapter provides evidence to evaluate the performance of the system in meeting the stated goals. Primary assessment is done through simulations of the soundfield that results for various configurations of the system. Relevant interaural differences are computed and displayed as a function of listener location. This data is used to support assertions made about the perceptual impact of the system. Additional confirmation comes from the results of listening conducted to determine whether the system behaves as expected.

5.1. Numerical Modelling

Simulating ear signals requires recorded HRTF data. In this section we describe the data used and the processing needed for application to the present context.

5.1.1. HRTF Database

The HRTFs used to simulate the soundfield were taken from the CIPIC HRTF database [2, 26]. The CIPIC database is a public-domain collection of HRTFs (technically, time-domain HRIRs) recorded for 45 subjects. The impulse responses are sampled at 44.1 kHz and windowed to 200 samples. They are blocked-meatus, free-field equalized recordings. Briefly, this means that ear canal resonances and the transfer function of the electroacoustic playback/recording chain are not included in the responses. This only affects overall spectral equalization, and does not affect interaural differences. The CIPIC measurements were made using speakers 1 m

from the subject. At this distance or beyond, HRTFs can be considered distance-independent with only slight error (see section 4.1.4). Hence, assuming no listeners are closer than 1 m to a speaker, we do not need range-dependent HRTFs to accurately model the soundfield. Due to the difficulty of obtaining reliable low-frequency data, the HRIRs cannot be considered accurate below about 200 Hz. The database download includes full documentation on the recording procedure; general discussion of HRTF recording and equalization methods can be found in [98]. The only subject used in this dissertation is "Subject 021," a KEMAR head-and-torso mannequin with large pinnae. Background information on KEMAR is provided in [23]. Coordinates for the CIPIC HRIRs are given in "interaural polar coordinates" shown in figure 5.1. Conversion factors between this coordinate system and the spherical coordinates described in section 4.2.1 are given in table 5.1. Recorded HRIRs are available at every combination of the following angles (interaural polar coordinates, in degrees):

$$\begin{aligned} \alpha &\in [-80, -65, -55, -45 : 5 : 45, 55, 65, 80] \\ \beta &\in (5.625[0 : 1 : 49] - 45) \end{aligned} \tag{5.1}$$

(where the MATLAB notation $a : n : b$ means "a sequence beginning at a and ending at b, counting by an increment of n").

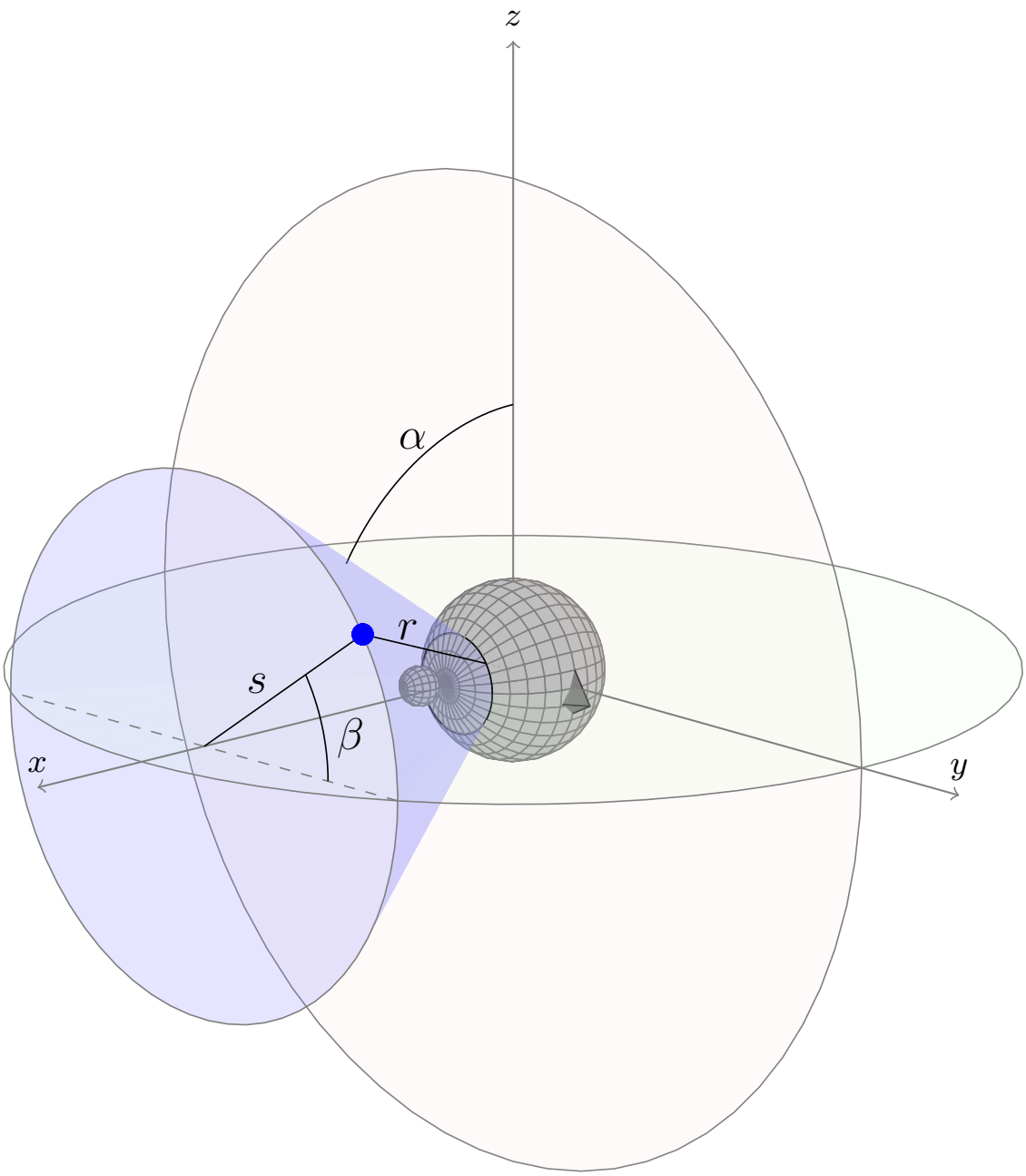


FIGURE 5.1. Interaural polar coordinates used for CIPIC HRIRs (after [101]). Notice that α defines a "cone of confusion."

		To		
		Interaural	Cartesian	Spherical
From	Interaural		$s = r\sqrt{1-\sin^2(\alpha)}$ $x = r\sin(\alpha)$ $y = s\cos(\beta)$ $z = s\sin(\beta)$	$s = r\sqrt{1-\sin^2(\alpha)}$ $\theta = \tan^{-1}\left(\frac{r\sin(\alpha)}{s\cos(\beta)}\right)$ $\phi = \frac{\pi}{2} - \tan^{-1}\left(\sqrt{\frac{1}{\cos^2(\alpha)\sin^2(\beta)} - 1}\right)$ $r = r$
	Cartesian	$\alpha = \sin^{-1}(x/r)$ $\beta = \tan^{-1}(z/y)$ $r = \sqrt{x^2+y^2+z^2}$		$\theta = \tan^{-1}(x/y)$ $\phi = \frac{\pi}{2} - \tan^{-1}\left(\sqrt{x^2+y^2}/z\right)$ $r = \sqrt{x^2+y^2+z^2}$
	Spherical	$\alpha = \sin^{-1}\left(\sin(\theta)\cos(\phi)\right)$ $\beta = \tan^{-1}\left(\frac{\sin(\phi)}{\cos(\theta)\cos(\phi)}\right)$ $r = r$	$x = r\sin(\theta)\cos(\phi)$ $y = r\cos(\theta)\cos(\phi)$ $z = r\sin(\phi)$	

TABLE 5.1. Coordinate conversions. Note that we are measuring ϕ upward from the xy-plane, not downward from the z-axis (i.e., we use latitude, not colatitude). Note also that we measure azimuth angles from the y-axis, not the x-axis. This leads to $\tan^{-1}(x/y)$ rather than the usual $\tan^{-1}(y/x)$ (for example, when converting from cartesian to spherical coordinates). The four-quadrant version of the $\tan^{-1}$ function (`atan2`) must be used. These equations are valid in the upper hemisphere; slight modifications are necessary for the lower hemisphere.

5.1.2. HRTF Interpolation

HRIRs for angles between the recorded locations were obtained through bilinear interpolation in the time domain.[1] Although this method is not optimum, it works quite well for a data set that is densely spatially sampled [13, 119]. Dense spatial sampling is a major design feature of the CIPIC database, so the approach is appropriate.[2]

The HRIR for arbitrary angles α, β is formed as the weighted sum of the four surrounding recorded HRIRs (see figure 5.2):

$$h_{\alpha\beta} = w_{11}h_{11} + w_{12}h_{12} + w_{21}h_{21} + w_{22}h_{22}, \tag{5.2}$$

where the weights are computed from the distance to the recorded HRIR as follows:

$$w_{mn} = \left(1 - \frac{|\alpha - \alpha_m|}{|\alpha_2 - \alpha_1|}\right)\left(1 - \frac{|\beta - \beta_n|}{|\beta_2 - \beta 1|}\right). \tag{5.3}$$

Direct application of this method to the raw HRIRs is not satisfactory, because they contain varying amounts of delay, and hence are not aligned in time. We can compensate for this via a well-established three-step process [77]:

(1) Separate each HRIR into minimum-phase and allpass portions.

(2) Substitute a pure (frequency-independent) delay for the allpass portion.

(3) Bilinearly interpolate the minimum-phase filter and delay values separately.

[1] Frequency-domain interpolation requires accurate and reliable phase unwrapping. An excellent discussion on this surprisingly obscure topic is found in [119].

[2] Although not used for this research, a binaural synth was constructed using the CIPIC data and the interpolation method described here. The perceptual effect is quite convincing and serves as additional informal confirmation of the validity of this implementation.

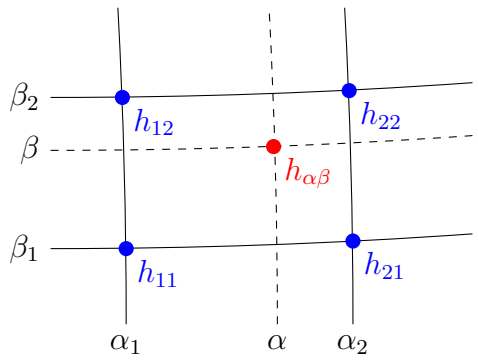


FIGURE 5.2. Bilinear interpolation: the HRIR $h_{\alpha\beta}$ for an arbitrary location is constructed as the weighted sum of the four surrounding recorded HRIRs. See text for details.

Minimum-phase decomposition can be achieved through the Hilbert transform, a relationship between the real and complex portions of the Fourier transform [105]. The allpass component, although not actually required for our calculations, is obtained by dividing in the frequency domain. The full theory of Hilbert transforms is well outside the scope of this document, but the MATLAB function in listing 5.1 performs the needed computations. Figure 5.3 shows an example of minimum-phase interpolation.

We now reduce the allpass filter to a single delay value. This "minimum-phase-plus-delay" model does involve the loss of some phase information, but this is justified because HRTFs are approximately minimum-phase-plus-delay, at least up to 10 kHz [92, 102]. In addition, the discarded phase information has little or no perceptual impact [77]. The pure delay can be found by computing the cross-correlation φ of the original measured HRIR with the minimum-phase reconstruction; the delay is the lag τ which maximizes this function. We can obtain subsample accuracy by modelling the cross-correlation as a parabola $\varphi(\tau) = A\tau^2 + B\tau + C$ near the maximum [70]. Call the integer lag which maximizes cross-correlation τ_0, with corresponding φ_0. We have similar points for lags ± 1 sample from τ_0. The unique parabola

LISTING 5.1. MATLAB code to compute a minimum-phase/allpass decomposition. After [62].

```
function [hmin, hall] = minalldecomp(h)

na = 3; %For numerical accuracy. 0 gives good
        %results while 3 gives excellent tolerance.
fftsize = 2^(nextpow2(h)+na);

fh = fft(h, fftsize);
fmin = exp(conj(hilbert(log(abs(fh)))));
fhall = fh ./ fhmin;

hmin = real(ifft(fhmin, fftsize);
hall = real(ifft(fhall, fftsize);
```

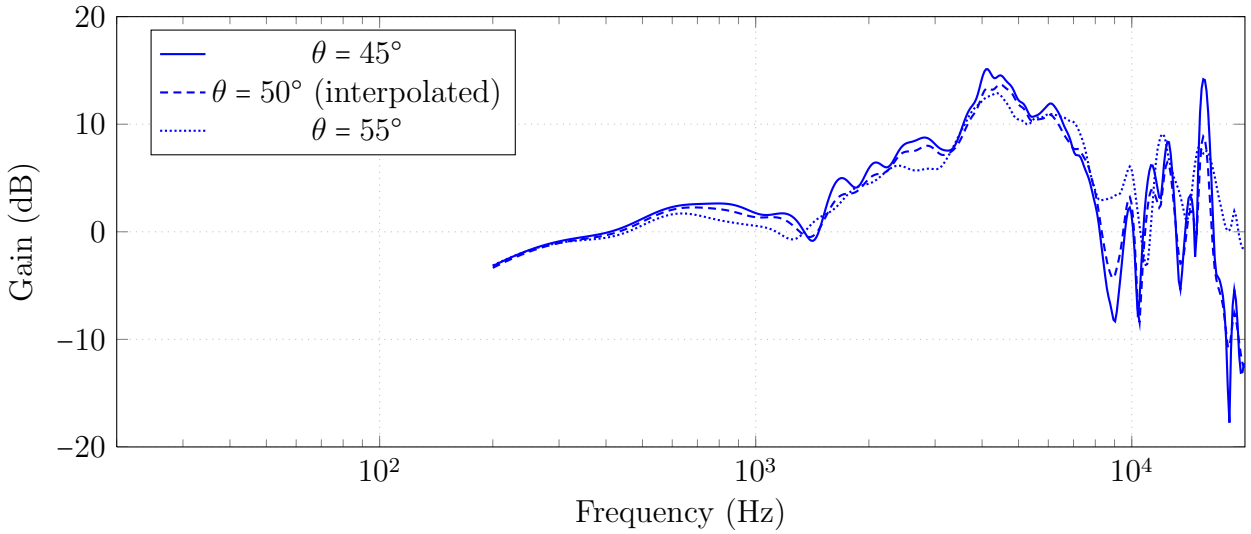


FIGURE 5.3. Recorded KEMAR HRTFs for $\theta = 45°$ (solid line) and $\theta = 55°$ (dotted line); interpolated HRTF for $\theta = 50°$ (dashed line). All three are for $\phi = 0°$, and show the response at the right (ipsilateral) ear.

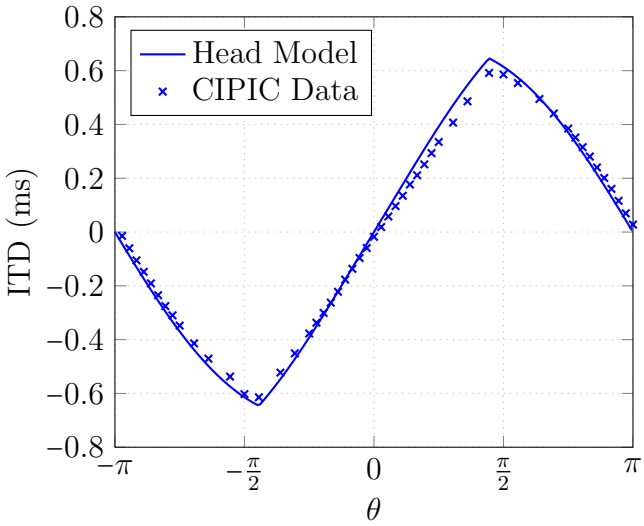


FIGURE 5.4. ITD as a function of source angle, for a source in the horizontal plane. The crosses indicate the computed ITD values for the CIPIC data. The solid line shows the modelled ITD values.

which passes through these three points can be found by solving the matrix equation:

$$\begin{bmatrix} \tau_{-1}^2 & \tau_{-1} & 1 \\ \tau_0^2 & \tau_0 & 1 \\ \tau_{+1}^2 & \tau_{+1} & 1 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} \varphi_{-1} \\ \varphi_0 \\ \varphi_{+1} \end{bmatrix}. \tag{5.4}$$

The extremum of this parabola occurs at $\tau = -B/2A$. The same method can be used to compute the maximum or minimum cross-correlation with subsample accuracy. Figure 5.4 compares the delays found with this method to the predictions of the spherical head model.

The angles $\alpha = \pm 90°$ are unique because the indicated direction is independent of β. To facilitate implementation and avoid special cases, HRTFs for $\alpha = 90°$ were formed by averaging the minimum-phase impulse responses for $\alpha = 80°$ over all stored elevations. The equivalent process was done for $\alpha = -90°$. We attempted to calculate the delay values by

linear extrapolation, but this tended to produce an ITD that was too high by about 0.1 ms, so the delays for $\alpha = \pm 90°$ were manually adjusted to fit the pattern shown in figure 5.4.

5.2. Soundfield Simulations

This section presents the results of simulations which verify the effectiveness of the system. For a particular set of system parameters, we can model the ear signals that arrive at a given location relative to the loudspeakers, using the interpolated HRTFs described above. We can then compute the distribution of spatial cues across the audience, to verify that the result is as designed.

5.2.1. Description of Soundfield Plots

The soundfield plots depict a "bird's-eye-view" of the simulated audience area. The x- and y-coordinates are always understood to be given in meters (m). The layout is based on the SARC facility described in section 5.3.1. For perspective, figure 5.5 depicts the location of all speaker pairs simultaneously. Table 5.2 describes the precise speaker geometry used. Finally, table 5.3 provides a key for the symbols used in the soundfields.

The plots show equal-value contour lines of various interaural difference measures, evaluated as a function of listener location. We adopt the convention that IID and ITD are calculated as "left ear minus right ear," so that a source on the left side of the head produces a positive IID and a negative ITD.[3] Consistent with this, dashed contours indicate either a negative gain or a positive time delay. All intensity differences are measured in decibels (dB); all timing differences are measured in milliseconds (ms). Often, we assume a desired target of

[3] If this seems counterintuitive, recall that sound arrives *louder* (positive gain) and *earlier* (negative time delay) at the ipsilateral ear.

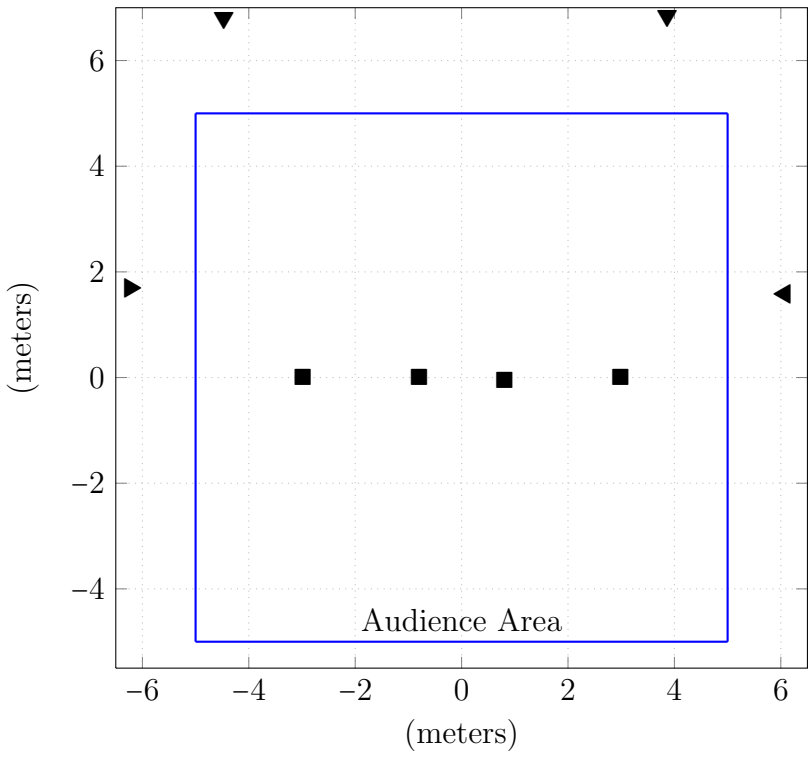


FIGURE 5.5. Overhead scale view of the simulated audience area and four speaker pairs.

Pair	*Elevation*	*Symbol*	*Distance Between Speakers* (m)	*Distance in Front of Target* (m)	*Height Above Ear Level* (m)
A	High	■	5.975	0.000	4.325
B	High	■	1.607	−0.022	4.903
C	Floor	▼	8.332	6.827	0.107
D	Floor	▶◀	12.305	1.645	0.067

TABLE 5.2. Speaker pairs used for the soundfield simulations. The target location was always in a line under the two ceiling pairs. "Distance in front of target" is the distance forward of this line in the xy-plane.

Symbol	*Meaning*
▼	Ear-level front speakers
▶◀	Ear-level side speakers
■	Ceiling speakers
★	Target location

TABLE 5.3. Symbols used for the bird's-eye-view contour plots of the soundfield.

no interaural differences. This is purely for convenience; section 5.2.5 gives an example where a large IID is desired.

The soundfields were generated using SuperCollider's non-realtime rendering mode. For given source signals and processing parameters, the speaker signals were computed and saved. Iterating over (at least) a 21×21-point grid of listener coordinates, ear signals were computed using the air-propagation model of equation (4.29) and the CIPIC HRTF data. Care was taken to use identical speaker signals for all iterations, even in the case of pseudorandom noise signals. This data was then imported into MATLAB and visualized using the `contour` plotting function. We make no attempt to smooth the contours.

5.2.2. Soundfield Dependence on Speaker Location

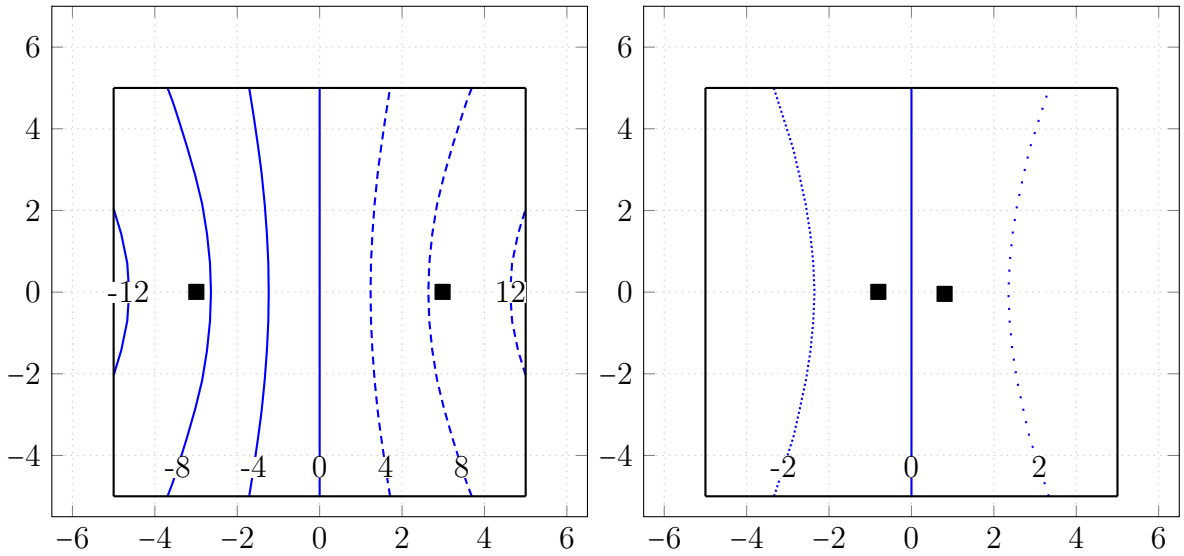
The arrangement of loudspeakers relative to the audience has a strong influence on the resulting soundfield. We discuss this quantitatively with respect to arrival differences and channel separation.

Arrival Differences

Recall from section 3.2 that a lead and a lag sound which arrive within a certain amount of

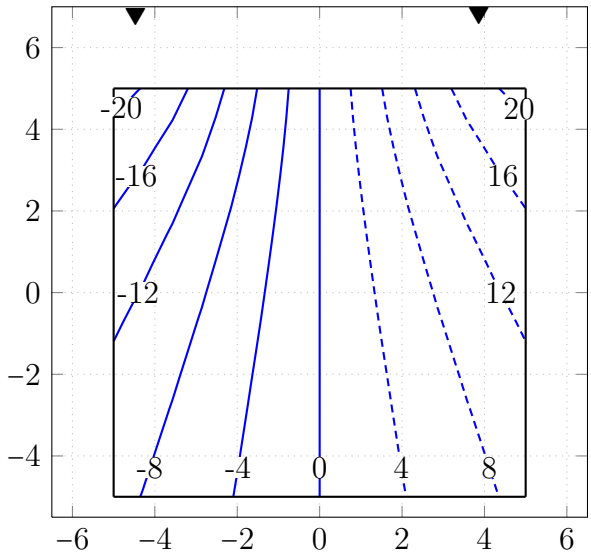
time will be perceived as a single sound. Lag delays greater than the echo threshold create a disruptive effect. The precise point at which this happens varies greatly with the source material, and estimates in the literature disagree significantly [84]. As a very rough guide, we consider an ITD of ±4 ms to define the boundary of the zone of effective integration. Effective imaging with delays at least this long is possible with decorrelated signals [74]. Similar statements hold for the IID, but arrival time is almost always the limiting factor for two-loudspeaker reproduction.

Given this, it makes sense to look for a speaker geometry that minimizes timing differences. For several speaker pairs, figures 5.6 and 5.7 show the differences in arrival time and intensity respectively that result purely from path length differences (i.e., to a point in free space, without considering the listener's head). Two patterns emerge. First, overhead pairs afford far more opportunity to create a symmetric soundfield. A symmetric soundfield is important both to prevent extremely bad listening positions, and to facilitate consistency across the soundfield. In order to achieve overall symmetry with two speakers placed symmetrically on the left and right, they must be in the plane perpindicular to the audience area and bisecting it into front to back halves. The only points on the ground that fall in this plane are to the extreme left and right (pair D is just slightly forward of this). However, elevated speakers allow the rest of this plane to be utilized. Second, it is clear that closer loudspeakers result in smaller absolute differences across the soundfield. This has been noted for ear-level loudspeakers, usually in the context not of creating a more uniform audience area, but of making crosstalk cancellation for a single listener more robust to head motion [11, 76, 131]. Again, speakers overhead can be both close and symmetric.

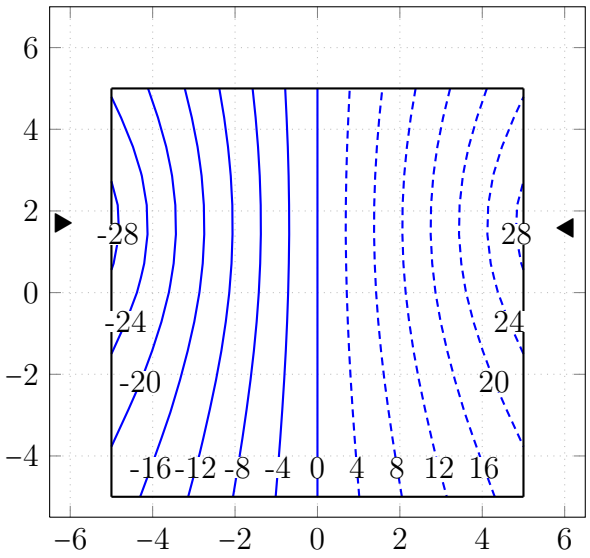


(A) Time of arrival difference (ms), speaker pair A

(B) Time of arrival difference (ms), speaker pair B

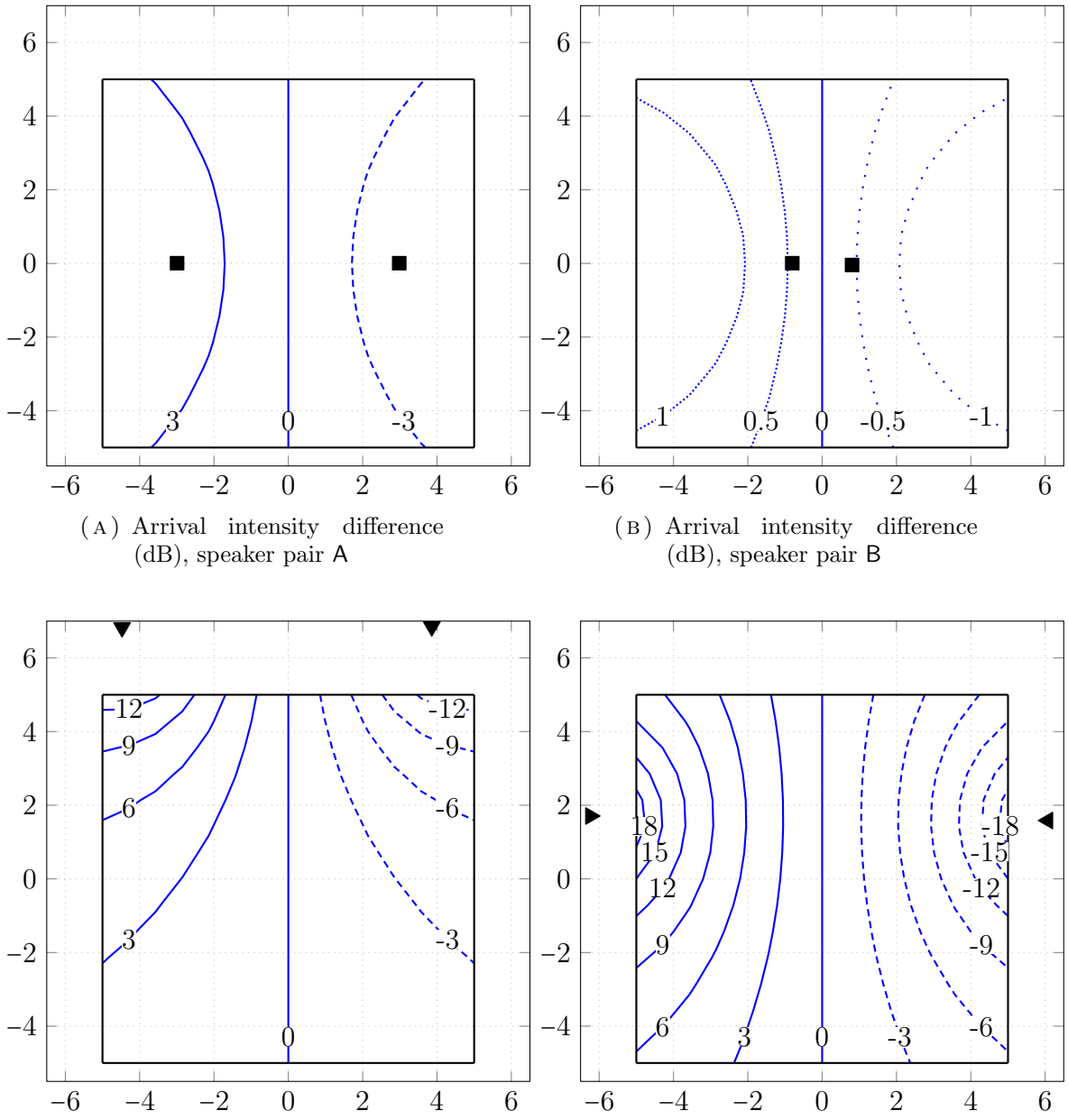


(C) Time of arrival difference (ms), speaker pair C



(D) Time of arrival difference (ms), speaker pair D

FIGURE 5.6. Difference in arrival time for a dual-mono signal, using speaker pairs A, B, C and D. Note the different scale for pair B; using the same scale as the other speakers, only the center line of 0 ms difference appears.



(A) Arrival intensity difference (dB), speaker pair A

(B) Arrival intensity difference (dB), speaker pair B

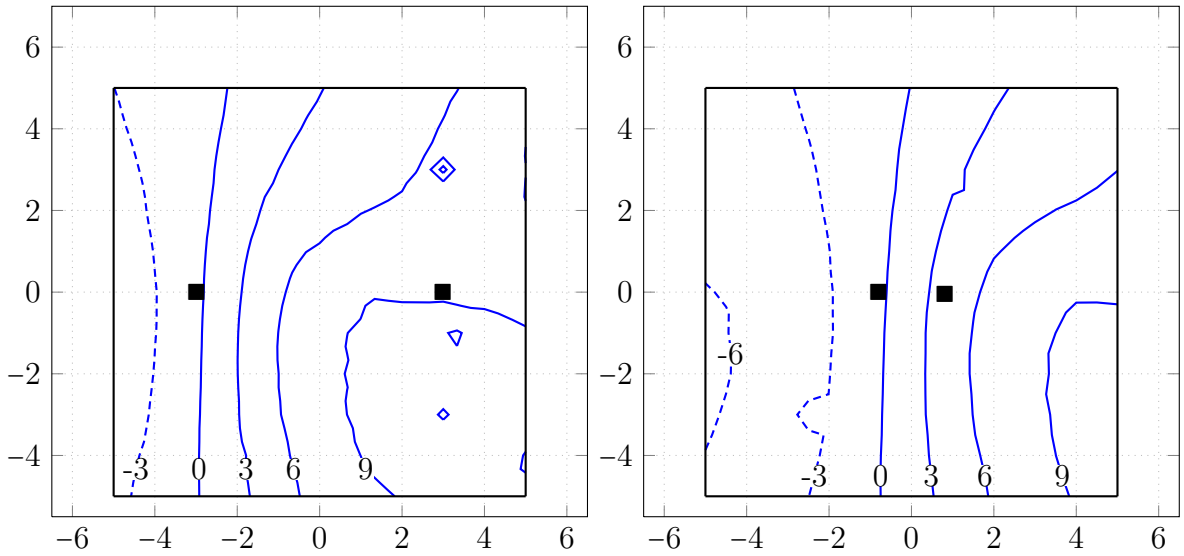
(C) Arrival intensity difference (dB), speaker pair C

(D) Arrival intensity difference (dB), speaker pair D

FIGURE 5.7. Difference in arrival intensity for a dual-mono signal, using speaker pairs A, B, C and D. Note the different scale for pair B; using the same scale as the other speakers, only the center line of 0 dB difference appears.

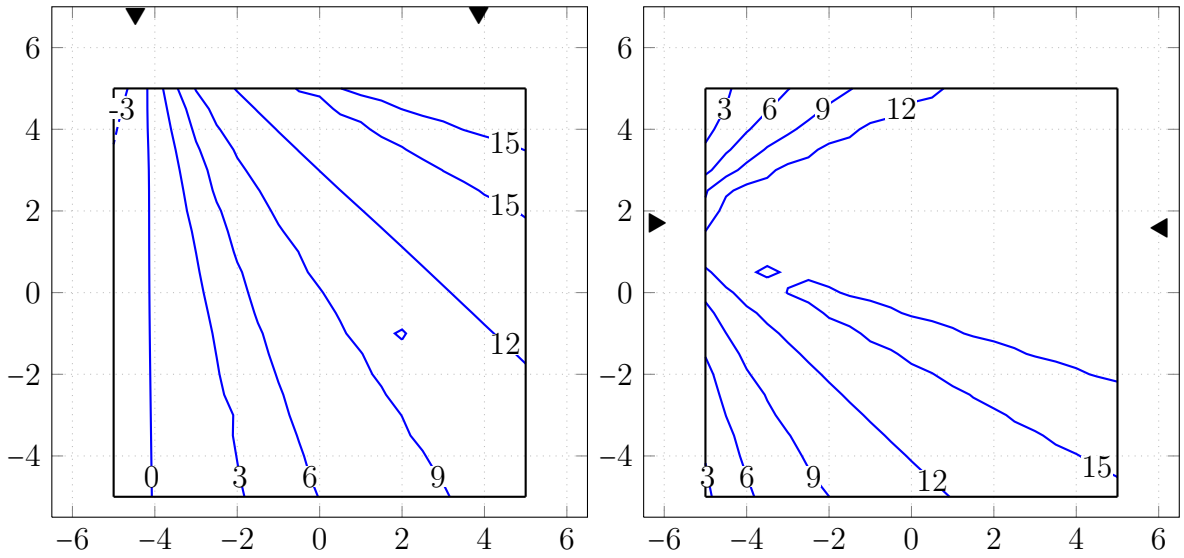
Channel Separation

It is natural to conclude from the above discussion that the narrow overhead pair is superior, but this is not necessarily so. Figure 5.8 illustrates another property of the soundfield, that of *channel separation*. This is essentially a measure of our ability to control the signals at the ears independently. To compute channel separation, we choose an impulse as the input for the left channel, and silence for the right channel. The resulting IID (with the listener's head now in place) is the left-to-right channel separation. The right-to-left channel separation will not usually be equal, because of asymmetric listeners and the asymmetric CIPIC data. Figure 5.8 shows the *natural* left-to-right channel separation, averaged from 200 Hz–10 kHz, that results solely from physical head-shadowing (i.e., no processing is applied to the inputs). Negative values outside of the speaker span mean the sound is actually arriving with more energy at the unintended ear. This quantifies the intuition that it is difficult to control spatial cues for listeners outside the span of the speakers. Narrower pairs clearly result in a larger area of negative channel separation. Just as importantly, the average channel separation with a narrow pair is lower even for the area between the speakers. Taken in conjunction, figures 5.6, 5.7 and 5.8 suggest that speaker pair A is a good compromise.



(A) L-R channel separation (dB), speaker pair A

(B) L-R channel separation (dB), speaker pair B



(C) L-R channel separation (dB), speaker pair C

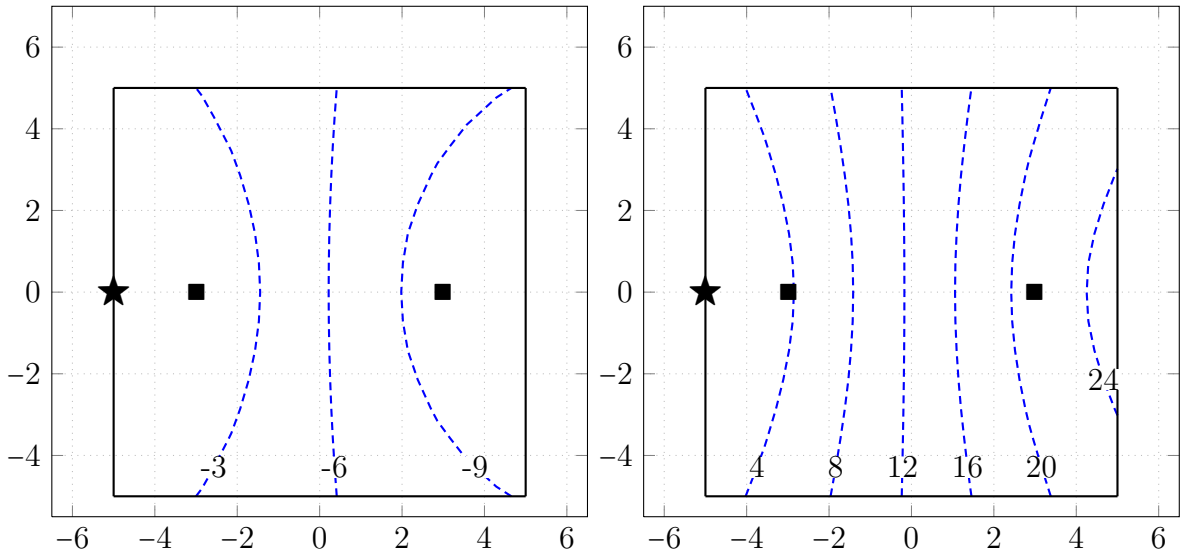
(D) L-R channel separation (dB), speaker pair D

FIGURE 5.8. Natural left-to-right channel separation, averaged from 200 Hz–10 kHz, for speaker pairs A, B, C and D. The contour lines skew to the right because the ears are set back on the head, so there is more channel separation when the speakers are in front of the listener.

5.2.3. Moving Target Location

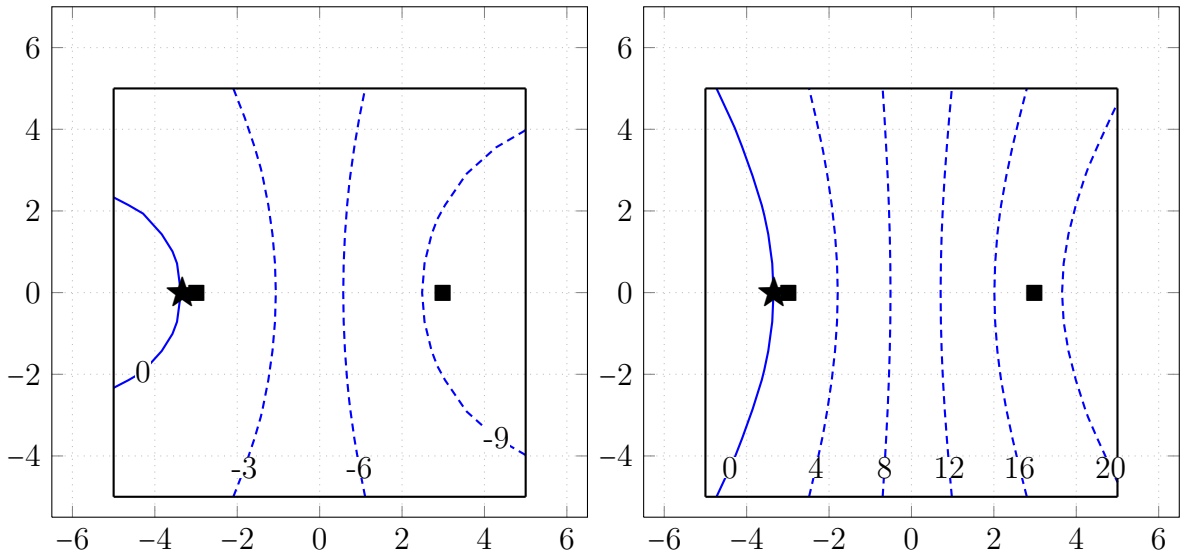
This section briefly discusses the effect of a moving target location. Figure 5.9 shows differences in arrival time and intensity for four frames of a moving target location, using speaker pair A. The target moves from the far left of the audience area to the origin (motion to the right is symmetric). Considering a fixed audience position, the sound image moves antiparallel to the path motion, because the interaural cues first favor the right, then transition to favoring the left. In a given time span, positions at the front or rear experience a greater change than centered positions. At all target locations, time of arrival is more limiting than arrival intensity for delivering spatial cues.

For comparison, figure 5.10 shows the result when speaker pair C is used, for the first two target positions shown in figure 5.9. While the zone of valid interaural cues is roughly the same size, the maximum arrival differences are greater. This actually becomes more apparent as the target area moves toward the center, as can be seen in figures 5.6 and 5.7.



(A) Arrival intensity difference (dB), Frame 1/4

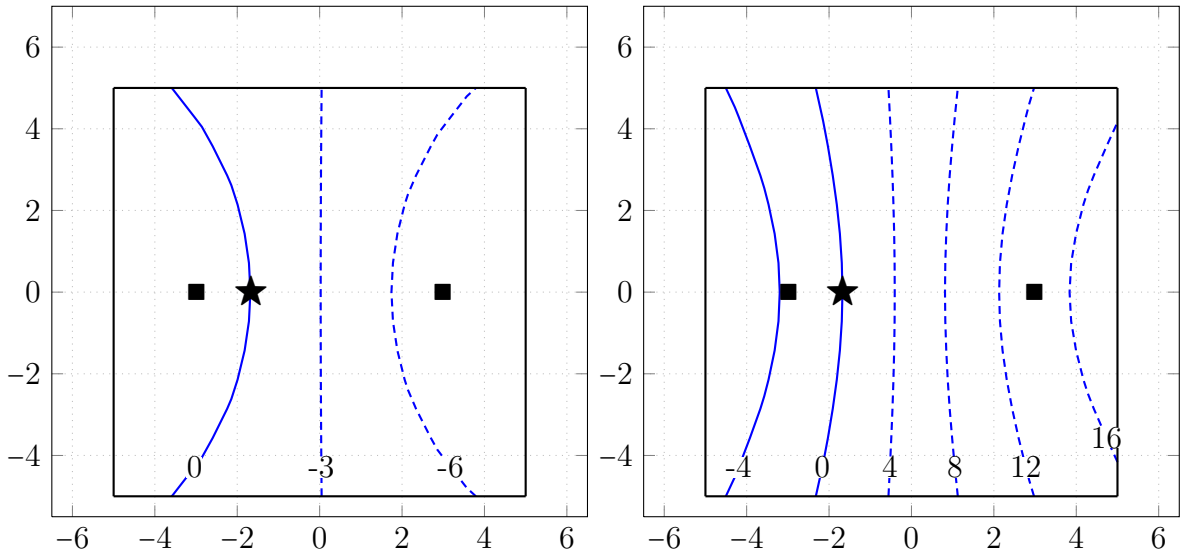
(B) Time of arrival difference (dB), Frame 1/4



(C) Arrival intensity difference (dB), Frame 2/4

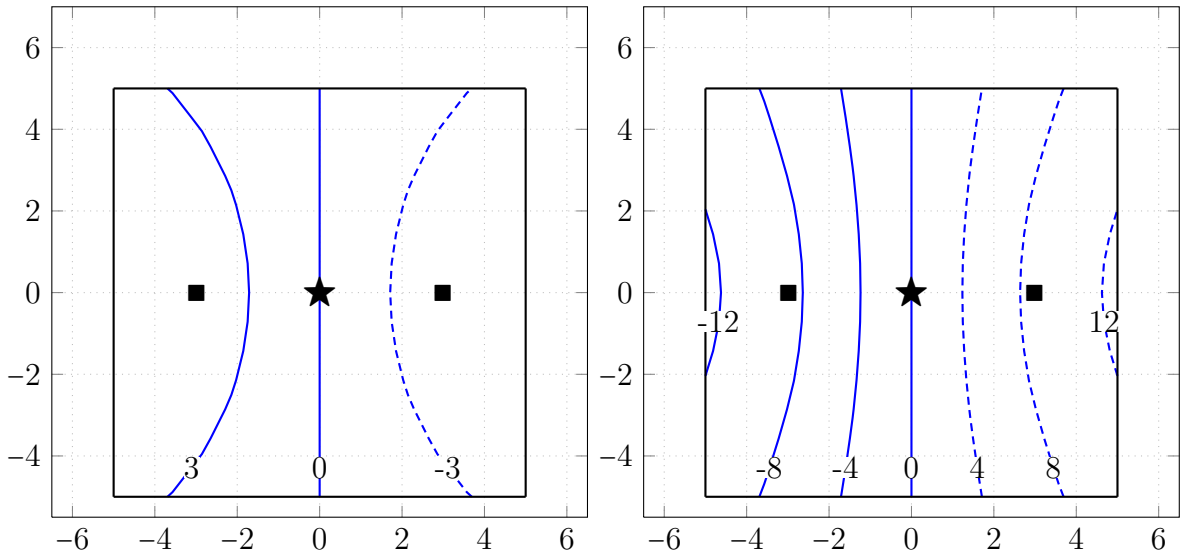
(D) Time of arrival difference (dB), Frame 2/4

FIGURE 5.9. Differences in arrival time and intensity for a moving target path using speaker pair A. The target location moves from the far left to the origin.



(E) Arrival intensity difference (dB), Frame 3/4

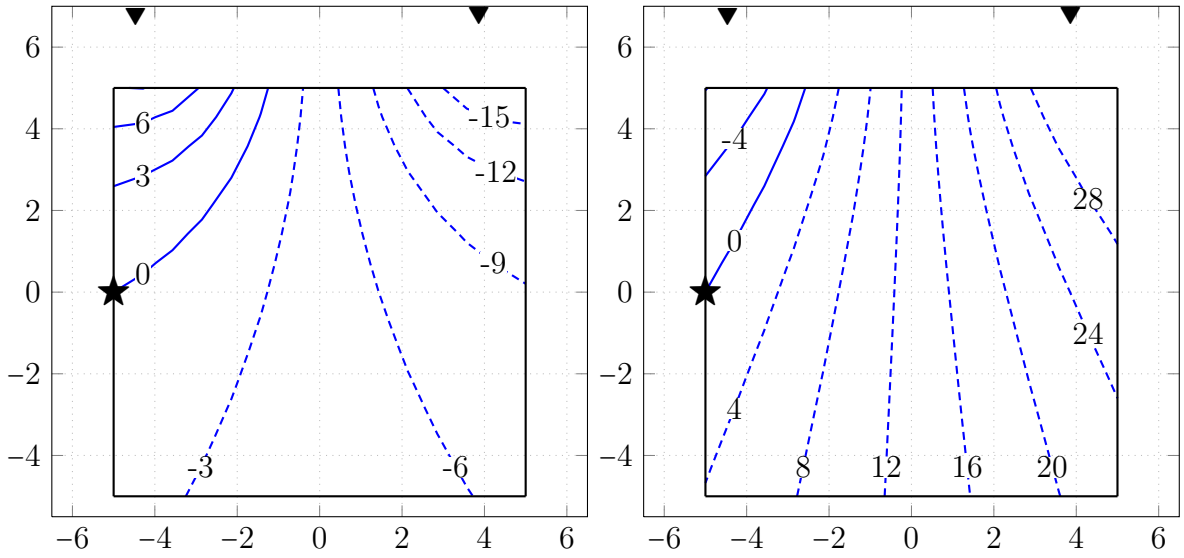
(F) Time of arrival difference (dB), Frame 3/4



(G) Arrival intensity difference (dB), Frame 4/4

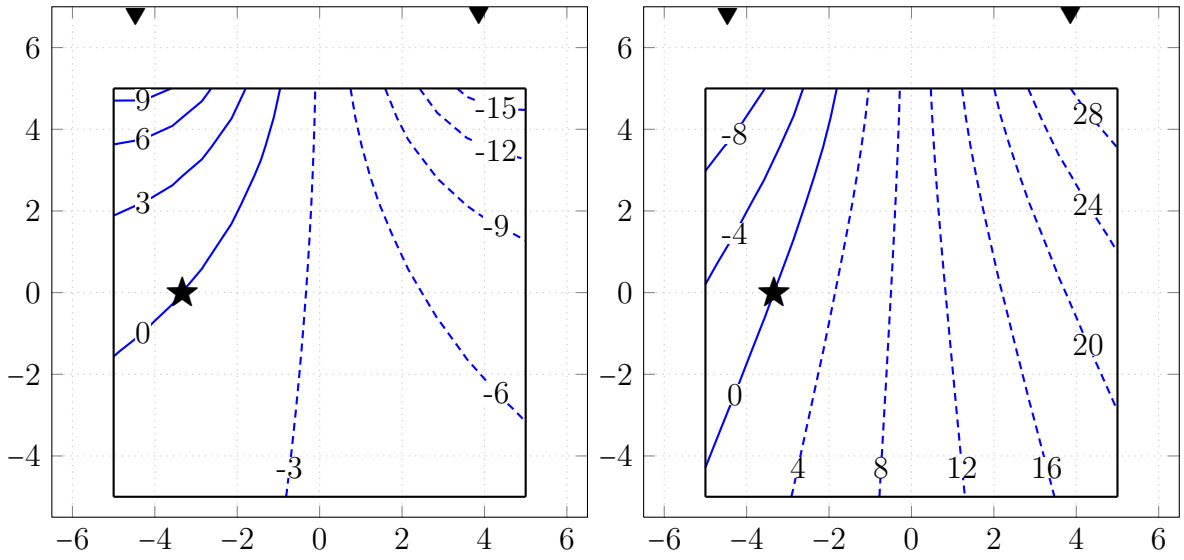
(H) Time of arrival difference (dB), Frame 4/4

FIGURE 5.9 (CONT'D). Differences in arrival time and intensity for a moving target path using speaker pair A. The target location moves from the far left to the origin.



(A) Arrival intensity difference (dB), Frame 1/2

(B) Time of arrival difference (dB), Frame 1/2



(C) Arrival intensity difference (dB), Frame 2/2

(D) Time of arrival difference (dB), Frame 2/2

FIGURE 5.10. Differences in arrival time and intensity for a moving target path using speaker pair C. The target location moves from the far left partway toward the origin.

5.2.4. Direct Path Compensation and Energy Compensation

In this section we consider the effect of direct path compensation and energy compensation on IID, ITD and IAC values at different listener locations.

Direct Path Compensation

The main purpose of direct path compensation is to equalize spectral features that relate to the loudspeaker position rather than the desired image position. Blauert describes a series of "boosted bands," frequency regions which are prominent when a source is behind, overhead or in front, and which cause sources to localize to those regions [17]. For example, a peak at 8 kHz is a strong cue for localization overhead. This will make it difficult to project broadband frontal images via overhead loudspeakers. Direct path compensation using true HRTFs would invert this peak, creating (ideally) a neutral spectrum. The spherical head model is not sufficiently detailed for this (but see the suggestion in section 6.2). However, it can equalize the overall balance of high and low frequencies.

Another important function of direct path compensation is to account for the unequal time delays due to head-related effects for off-centered listeners. Figure 5.11 shows the ITD at 21 target locations ranging from $x = -5$ m to $x = 0$ m (with $y = 0$ m). The source is a dual-mono impulse, so the ideal result is an ITD of 0 ms. The dotted line shows the result with path length compensation only, and the dashed line shows the effect of direct path compensation. Beyond the speaker position (left of the solid vertical line), neither case gives 0 ms ITD. Even so, direct path compensation is partly successful. Within the speaker span, direct path compensation creates approximately the correct result, while path length compensation only succeeds for centered listeners.

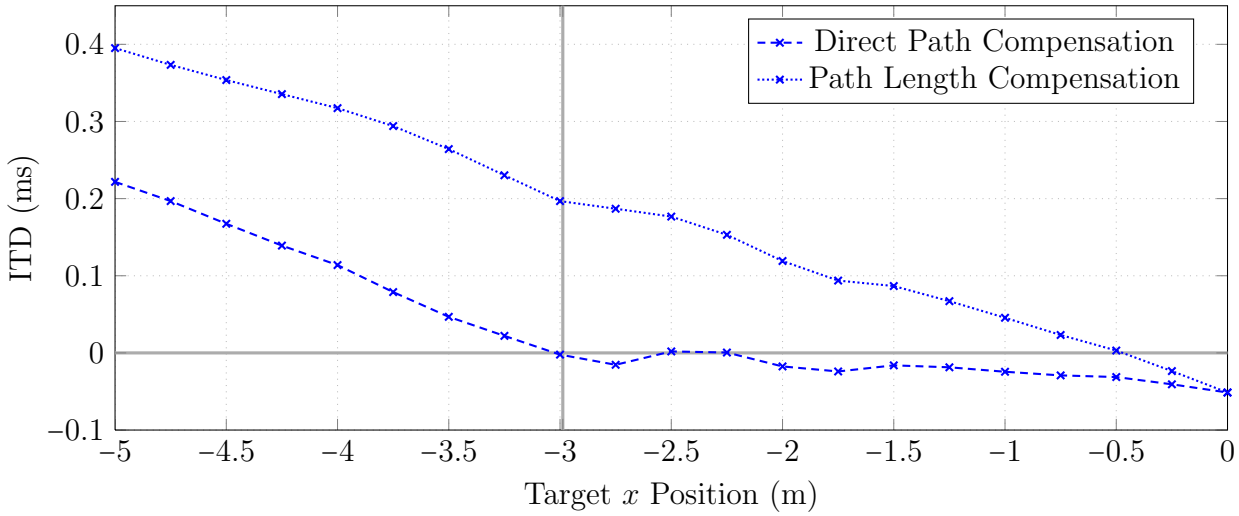


FIGURE 5.11. ITD for a target position from $x = -5\,\text{m}$ to $x = 0\,\text{m}$, with $y = 0\,\text{m}$. The dashed line shows direct path compensation; the dotted line shows path length compensation only. The solid horizontal line shows the optimal result, 0 ms. The solid vertical line shows the point at which the listener is directly underneath the left loudspeaker.

Energy Compensation

Direct path compensation also permits some control over the IID at the target, but this can be further improved upon using energy compensation. Figure 5.12 shows the IID at 21 target locations ranging from $x = -5\,\text{m}$ to $x = 0\,\text{m}$ (with $y = 0\,\text{m}$). The source is decorrelated white noise, bandlimited around 2 kHz (using the method described in section 5.3.2). The dotted line shows path length compensation only; the dashed line shows direct path compensation; the solid line shows energy compensation. Again, none of these methods succeeds outside the span of the speakers, though with energy compensation the total power will be approximately correct. Results with broadband noise are similar but somewhat more irregular. Results with dual-mono noise are unsatisfactory. Section 5.4.1 discusses the assumptions of the energy compensation model and suggests possible refinements.

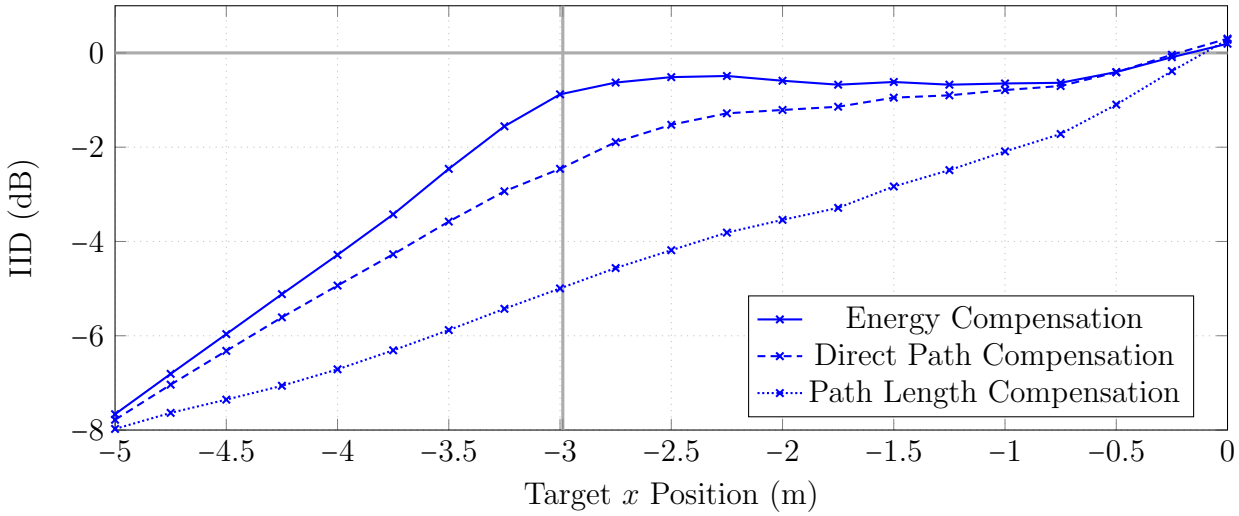
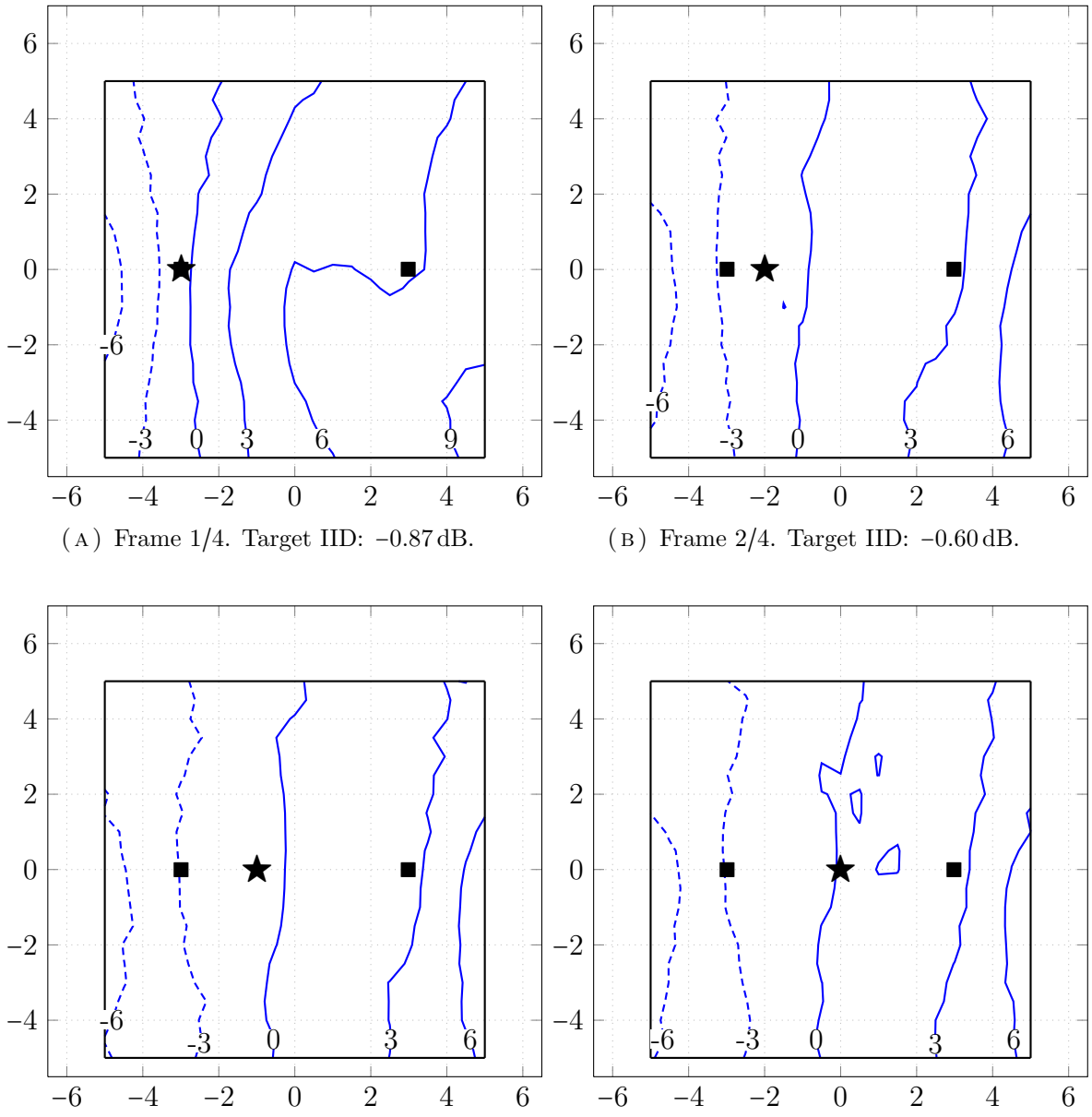


FIGURE 5.12. IID for a target position from $x = -5\,\text{m}$ to $x = 0\,\text{m}$, with $y = 0\,\text{m}$. The solid line shows energy compensation; the dashed line shows direct path compensation; the dotted line shows path length compensation only. The solid horizontal line shows the optimal result, 0 dB. The solid vertical line shows the point at which the listener is directly underneath the left loudspeaker. See also figure 5.24 on page 167, and the accompanying discussion.

Figure 5.13 shows the progression of a target path using energy compensation. The source is decorrelated noise, bandlimited around 2 kHz, with equal energy in both channels. The plots show overall IID, calculated by comparing the signal power arriving at each ear. Energy compensation appears quite successful at maintaining the desired IID at the target. We would not expect this to be true for a target outside the span of the speakers. The contour lines extend the entire length of the listening area, albeit with quite a bit of irregularity.



(A) Frame 1/4. Target IID: −0.87 dB.

(B) Frame 2/4. Target IID: −0.60 dB.

(C) Frame 3/4. Target IID: −0.65 dB.

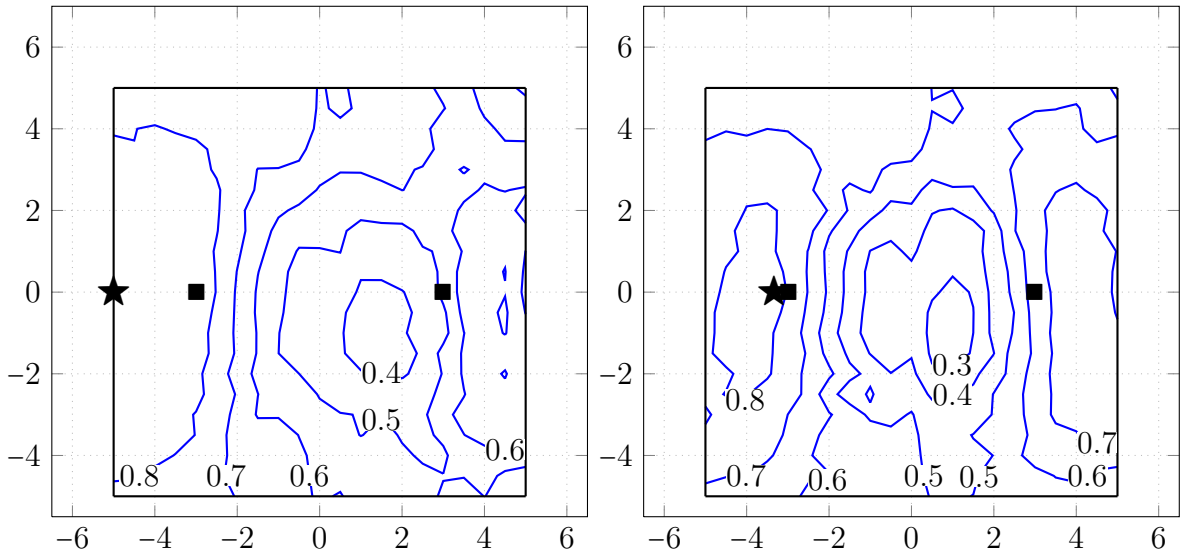
(D) Frame 4/4. Target IID: 0.19 dB.

FIGURE 5.13. IID using energy compensation and speaker pair A. Source IID is 0 dB. The target location moves from directly under the left speaker to the origin.

Decorrelation

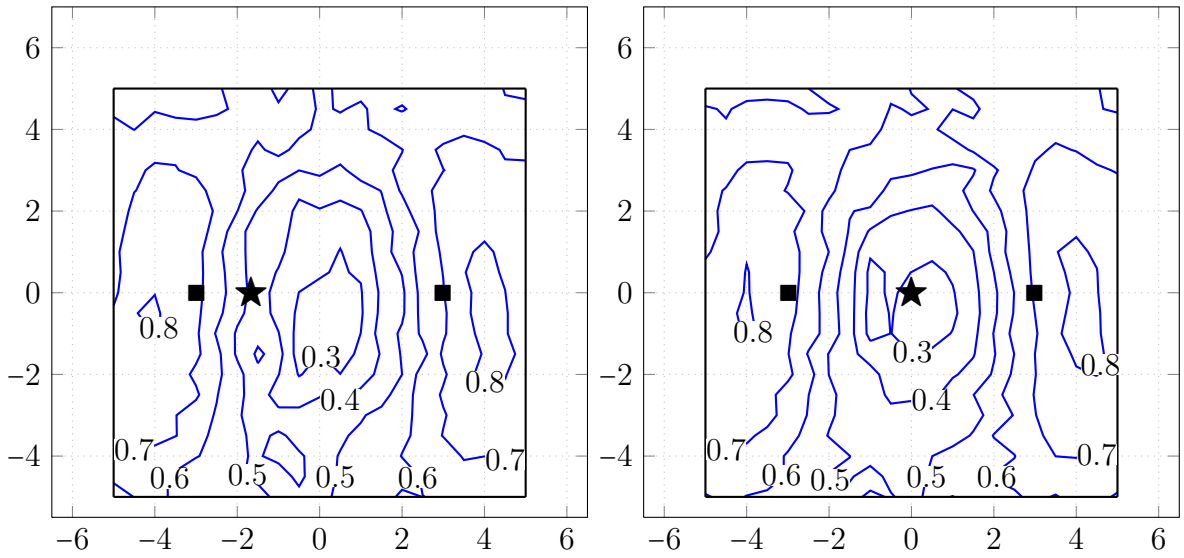
Figure 5.14 shows the effect of using direct path compensation with decorrelated white noise, bandlimited around 2 kHz. The target moves from the far left of the audience area to the origin. In this case, the area of lowest decorrelation actually moves in the opposite direction from the target path. The region of lowest correlation is relatively small. Fortunately, the ear is much less discriminating for low absolute values of IAC than for high absolute values [3, 74]. Values less than 0.5 or so are sufficient to create diffuseness, if not as substantial as complete decorrelation. Without using crosstalk cancellation, it appears quite difficult to produce low correlation outside of the speaker span. Sources with a negative correlation might be useful, though we do not pursue the idea here.

Energy compensation interferes with decorrelation when the target position is beyond the span of the speakers. This is because it disables one output completely to minimize IID. A possible solution is to crossfade between energy compensation and direct path compensation only, based on the target position relative to the loudspeaker span.



(A) Frame 1/4. IAC at target: 0.82.

(B) Frame 2/4. IAC at target: 0.81.



(C) Frame 3/4. IAC at target: 0.46.

(D) Frame 4/4. IAC at target: 0.28.

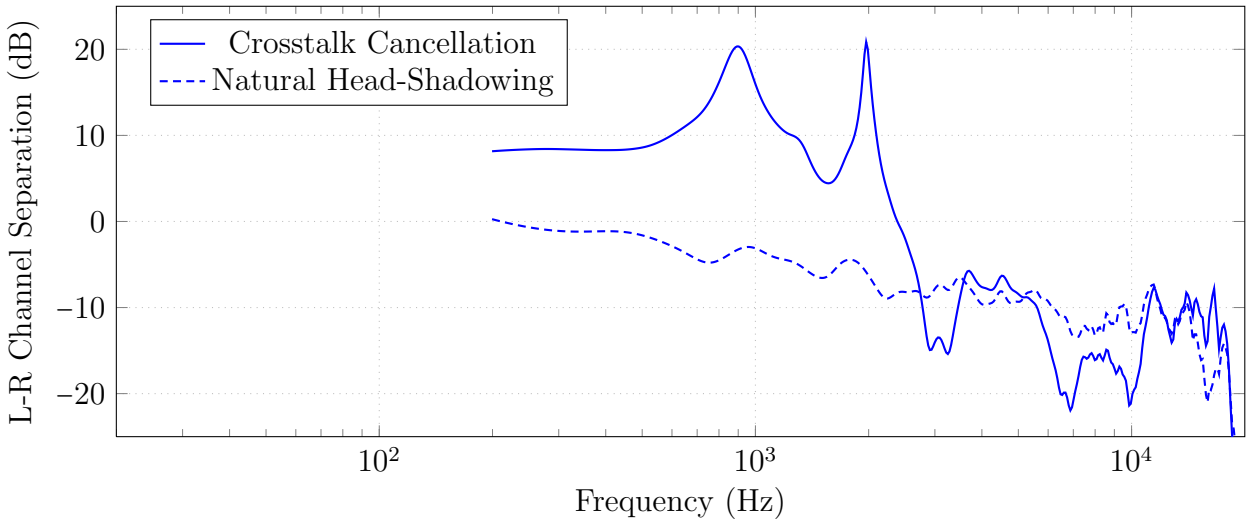
FIGURE 5.14. IAC using direct path compensation over speaker pair A. The source is decorrelated white noise (ICC = 0.16), bandlimited around 2 kHz. The target location moves from the far left of the audience area to the origin.

5.2.5. Crosstalk Cancellation

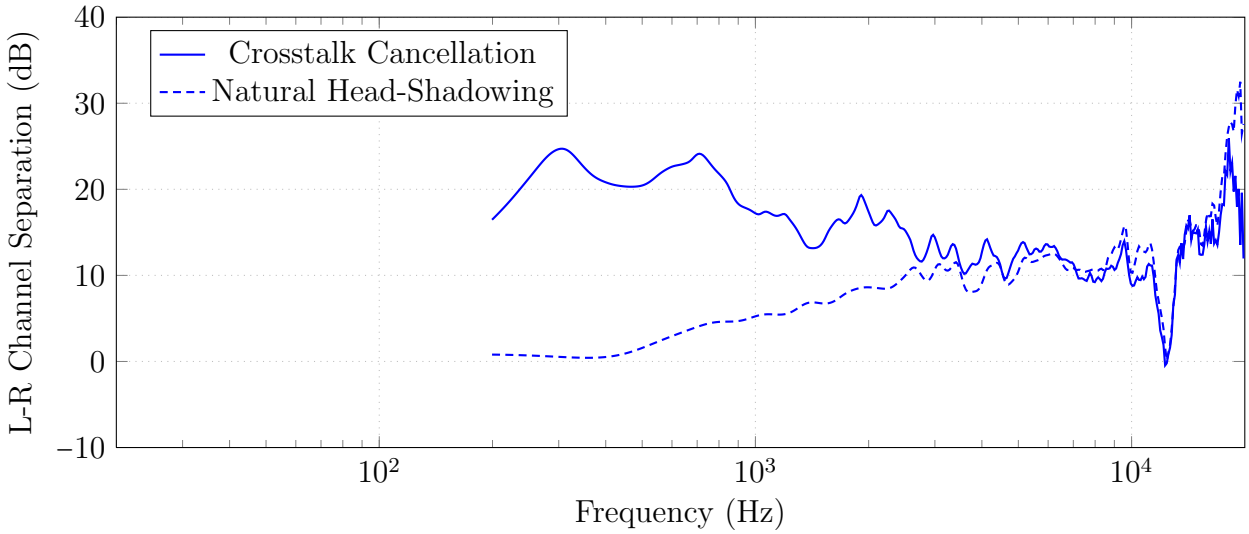
This section investigates the performance of crosstalk cancellation. Frequency-dependent channel separation is shown, as well as soundfields for both decorrelation effects and binaural cue delivery.

Channel Separation

Before considering soundfields, we first discuss the frequency-dependent behavior of crosstalk cancellation. Figure 5.15 shows L-R channel separation that results from using crosstalk cancellation over speaker pair A. Separation is shown at two listener locations, both along the line under the speaker pair. Natural head-shadowing separation is shown for comparison; performance of the crosstalk canceller is measured by the increase in separation relative to this baseline. The first location is to the far left, $x = -5\,\text{m}$. This is far outside the speaker span; both the left and right speakers are to the right of the head. Despite this, crosstalk cancellation succeeds up to about 2.5 kHz with a roughly 10 dB improvement. While not ideal, this is sufficient for many spatial effects, particularly because low-frequency ITD cues dominate perception in some circumstances [138]. Performance is even better for a centered listener. Crosstalk cancellation approaches natural separation around 3.5 kHz but makes a slight improvement to about 7 kHz. Note, however, that there is still significant high-frequency separation above this point. The two plots together suggest that crosstalk cancellation is best limited to below about 3 kHz.



(A) L-R channel separation at $x = -5\,\text{m}$.



(B) L-R channel separation at $x = 0\,\text{m}$.

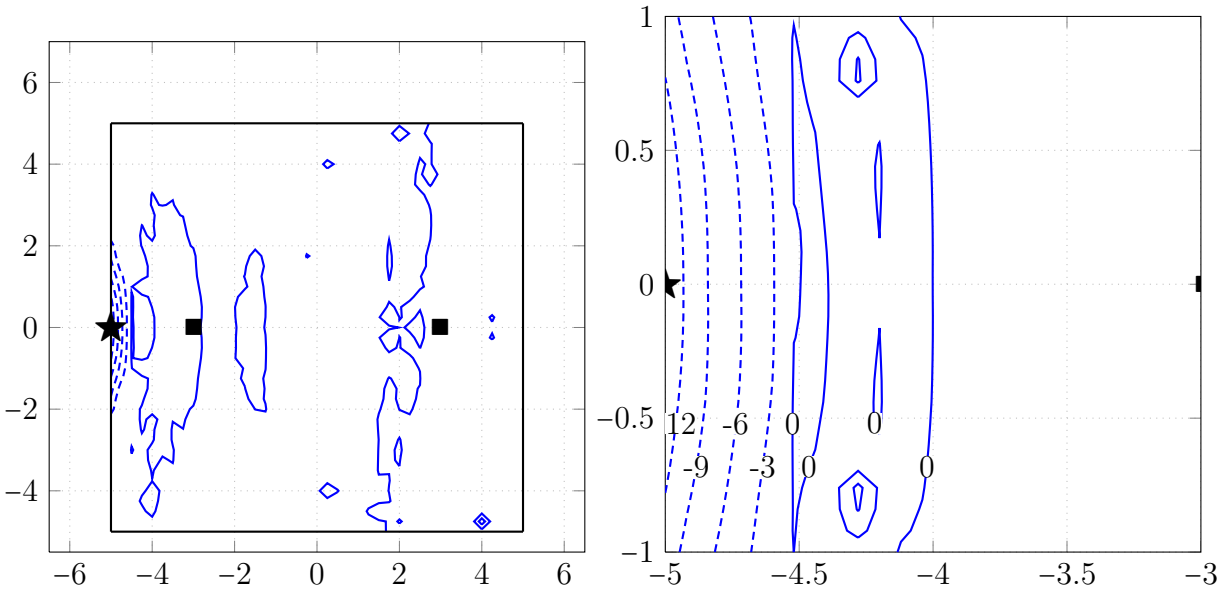
FIGURE 5.15. L-R channel separation using speaker pair A for two listener positions, both at $y = 0$. The solid line shows separation with crosstalk cancellation; the dashed line shows natural separation due to head shadowing.

Binaural Cues

Figure 5.16 demonstrates the delivery of binaural cues using crosstalk cancellation. The spherical head model was used to impose DTFs for a virtual source at 90° and only 2.2 cm from the surface of the head. This close range should result in a very large negative IID at low frequencies. The figure shows IID averaged from 200 Hz to 1 kHz. The source IID for this virtual range is −21.7 dB, while the IID at the target locations is only around −14 dB. Nevertheless, this is a substantial IID for low frequencies. The figure also shows that cues can be delivered beyond the span of the loudspeakers, at least for low frequencies. The region has quite a large front to back spatial extent; sufficient to sweep across the entire audience at some point. However it is extremely narrow left to right. As the target area passes over, an area of positive IID follows, which would tend to move the source quickly to the left, while still possibly appearing somewhat close to the head.

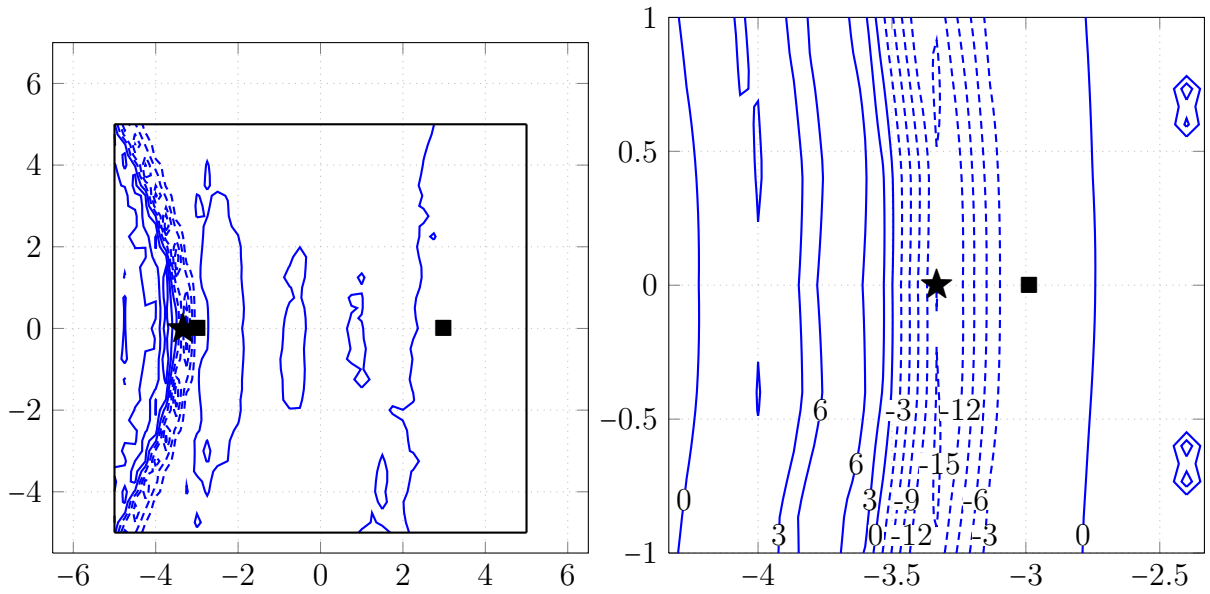
Decorrelation

Figure 5.17 illustrates the resulting IAC when crosstalk cancellation is applied to decorrelated white noise, bandlimited to the octave around 1 kHz. The ear signals at the target are indeed almost decorrelated. Compared to the IID cues in the example above, the region of near-optimal cues is more irregular and does not have as great a spatial extent. As mentioned above, an IAC below about 0.5 will produce some diffusion. The primary advantage here is that low correlation can be produced outside of the span of the speakers.



(A) Frame 1/4. IID at target: −14.0 dB.

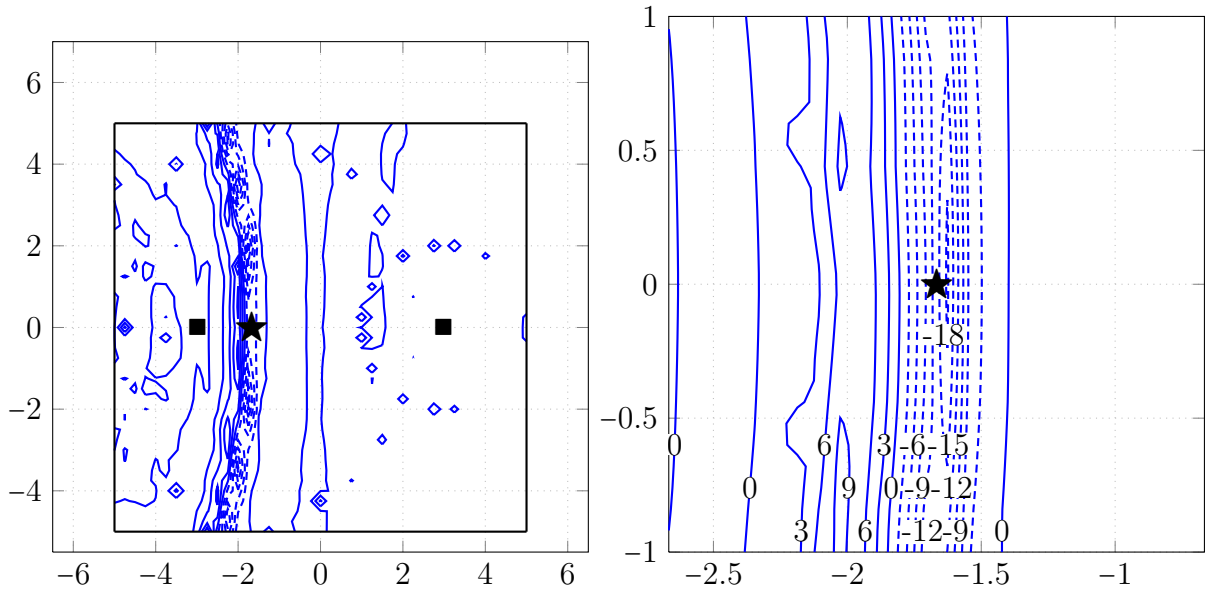
(B) Frame 1/4 (Magnified)



(C) Frame 2/4. IID at target: −15.0 dB.

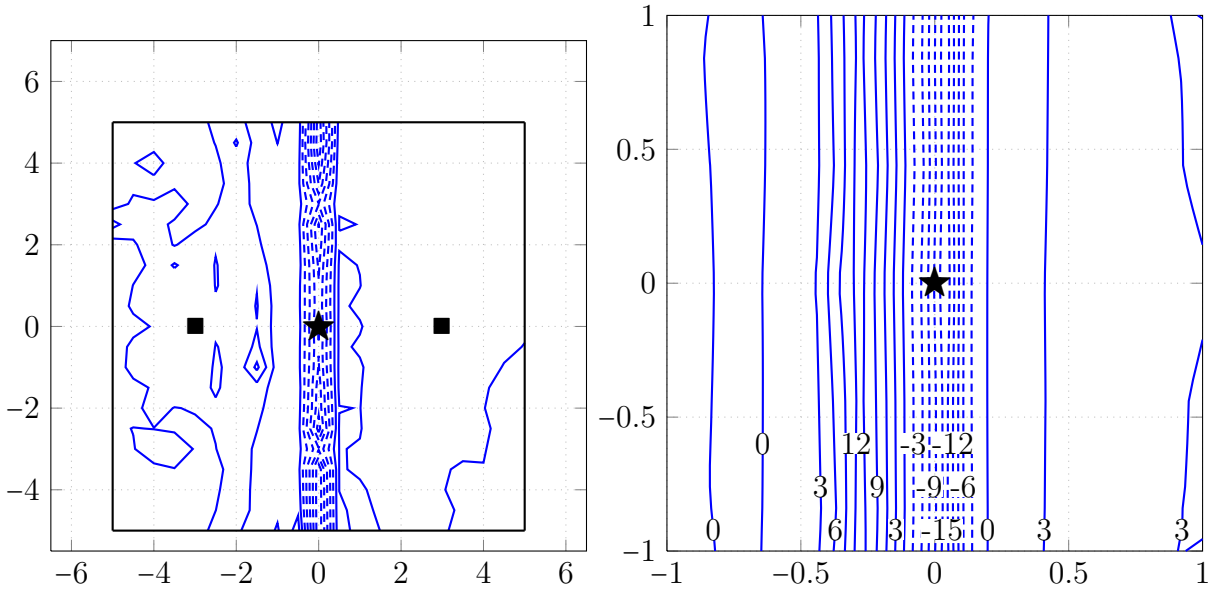
(D) Frame 2/4 (Magnified)

FIGURE 5.16. IID from 200 Hz to 1 kHz using crosstalk cancellation over speaker pair A. The source is a binaural signal synthesized at $\theta = 90°$, 2.2 cm from the surface of the head. Source IID is −21.7 dB. The target location moves from the far left of the audience area to the origin.



(E) Frame 3/4. IID at target: −14.6 dB.

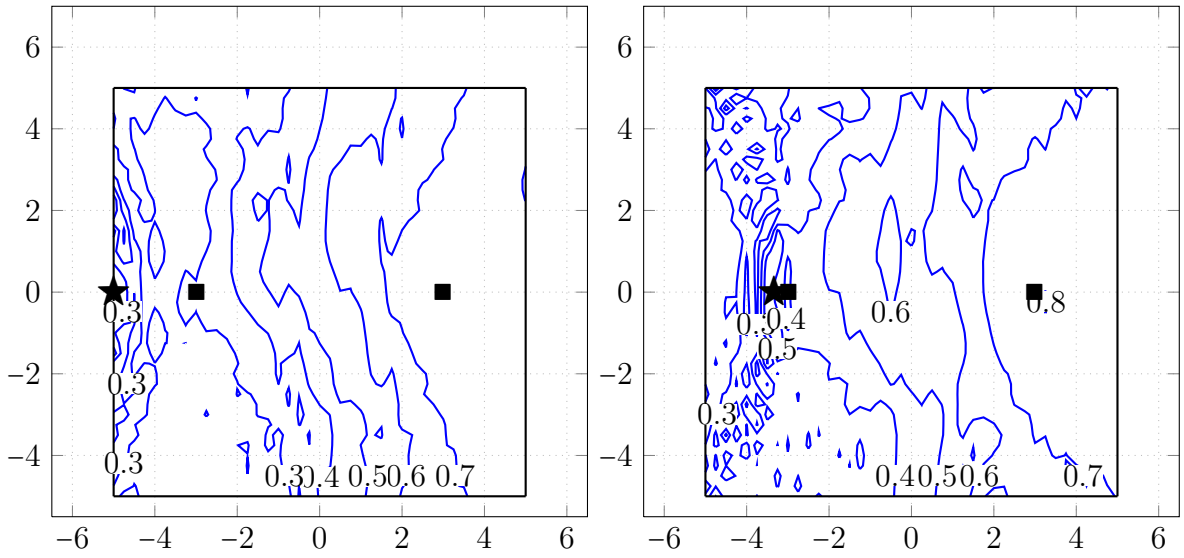
(F) Frame 3/4 (Magnified)



(G) Frame 4/4. IID at target: −12.4 dB.

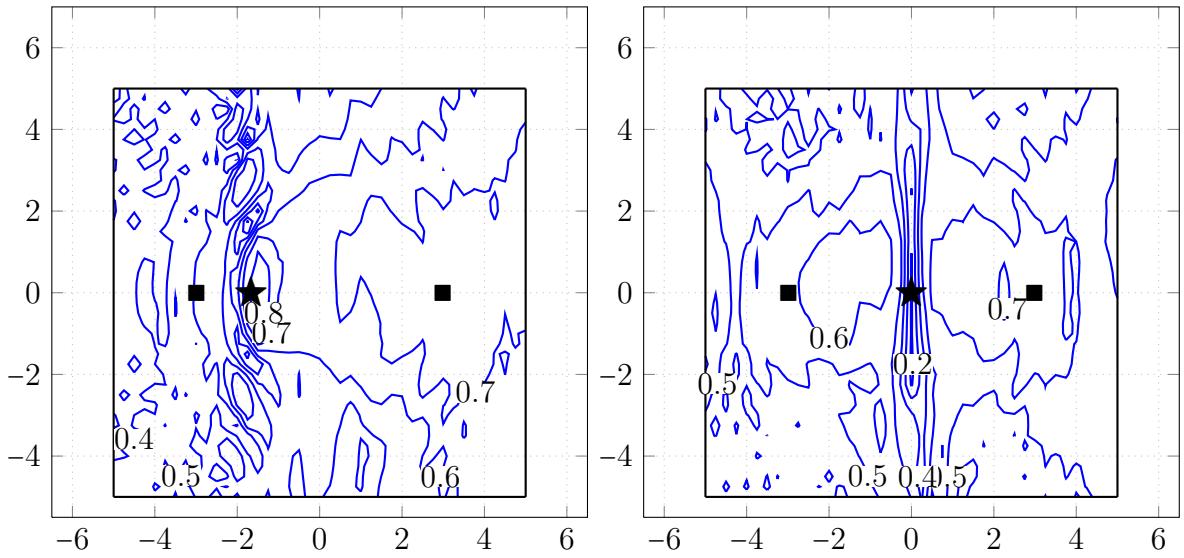
(H) Frame 4/4 (Magnified)

FIGURE 5.16 (CONT'D). IID from 200 Hz to 1 kHz using crosstalk cancellation over speaker pair A. The source is a binaural signal synthesized at $\theta = 90°$, 2.2 cm from the surface of the head. Source IID is −21.7 dB. The target location moves from the far left of the audience area to the origin.



(A) Frame 1/4. IAC at target: 0.23.

(B) Frame 2/4. IAC at target: 0.17.



(C) Frame 3/4. IAC at target: 0.18.

(D) Frame 4/4. IAC at target: 0.18.

FIGURE 5.17. IAC using crosstalk cancellation over speaker pair A. The source is decorrelated white noise (ICC 0.16), bandlimited around 1 kHz. The target location moves from the far left of the audience area to the origin.

5.3. Listening Demonstrations

To confirm accuracy of implementation, listening demonstrations were carried out using a variety of parameter configurations. It must be emphasized that these listening tests were not quantitative perceptual trials or controlled experiments, which are outside the scope of the present project. The observations presented here should be considered as confirmation of the system's functional correctness.

5.3.1. Listening Conditions

The demonstrations were conducted in the Sonic Laboratory of the Sonic Arts Research Centre (SARC) at Queen's University, Belfast [118]. The Sonic Lab is a purpose-built facility designed for loudspeaker experimentation and electroacoustic diffusion. Figure 5.18 shows the lab; figure 5.19 shows a diagrammatic representation. The lab measures 17 m L × 13 m W × 14 m H. The main loudspeakers are arranged in four groups of eight: one at ear level; one roughly two meters above ear level; one about five meters above ear level; and one about five meters below ear level. The "floor" at audience level is an acoustically transparent grate, allowing use of the lower group on the structural floor. In addition, there are eight corner speakers and six subwoofers. All of these speakers are moveable with varying degrees of effort. Table 5.4 gives data on the speaker pairs that were employed. For these tests, the front pair of ceiling speakers was relocated and placed on the same line as the rear ceiling pair, but with a larger separation. This resulted in both a wide and a narrow pair directly over the audience area. The outer pair was lowered slightly, so that both pairs were approximately equally distant from a centered listener (see figure 5.20). During the listening, the movable acoustic

TABLE 5.4. Speaker pairs used for the listening demonstrations. The target location was in a line under the two ceiling pairs. "Distance in front of target" is the distance forward of this line in the *xy*-plane.

Pair	*Speaker Type*	*Group*	*Distance Between Speakers* (m)	*Distance in Front of Target* (m)	*Height Above Ear Level* (m)
A	Meyer UPM-1P	High	5.975	0.000	4.325
B	Meyer UPM-1P	High	1.607	−0.022	4.903
C	Genelec 1038B	Floor	8.332	6.827	0.107
D	Genelec 1037B	Floor	12.305	1.645	0.067
A (alt.)*	Meyer UPM-1P	High	6.753	0.765	4.903
E†	Meyer UPJ-1P	Mid-high	11.404	3.150	2.116

* Used only in section 5.3.5. In this case only, the target line was at the mixing desk.
† Rarely used due to buzzing in the left loudspeaker.

absorption on each wall was lowered. This nominally brings the 1 kHz reverberation time to 0.4 s, although this was not verified.

The software was running on an Apple MacBook Pro (early 2008 model; dual-core 2.5 GHz processor; 2 GB RAM). Audio was taken from the analog output (unfortunately, using the digital output was not possible) and run into the lab's Digidesign Control|24 mixing desk. From there the sound could be routed dynamically to any speaker pair, via Digidesign 192 I/O converters.

FIGURE 5.18. The SARC Sonic Lab, facing toward the front. Floor-height, mid-high and ceiling-level speakers are visible, as is the stage, audience area, absorbent wall panels and mixing desk.

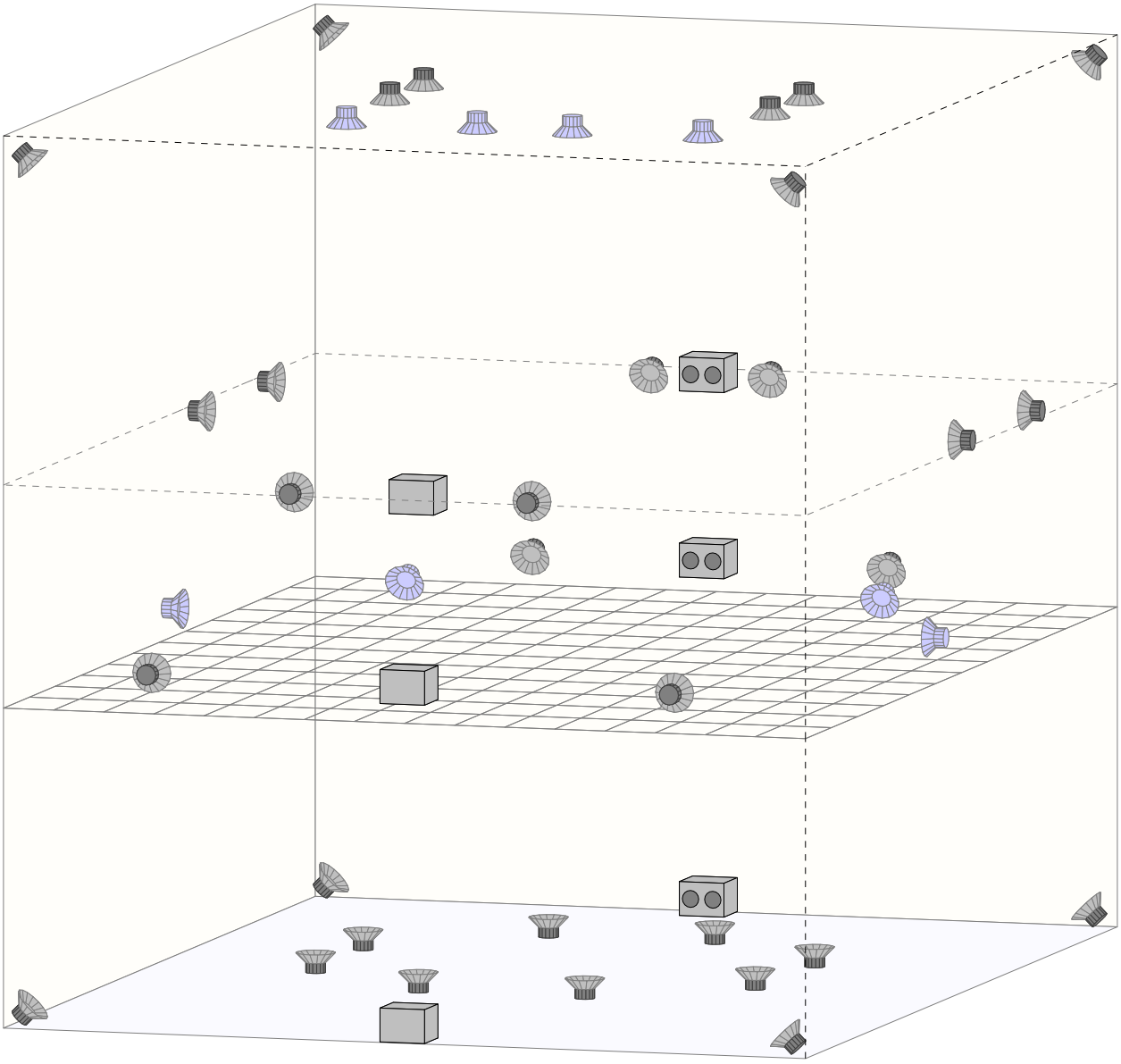


FIGURE 5.19. Schematic diagram of the Sonic Lab. Note that for these tests the front two ceiling speakers were relocated to the same line over the audience as the rear ceiling pair, with a wider separation.

FIGURE 5.20. The wide and narrow overhead speaker pairs employed during the listening. This view is facing toward the rear of the lab. Notice that the panels supporting the outer pair have been slightly lowered.

$\frac{2}{3}f_0$ (Hz)	f_0 (Hz)	$\frac{4}{3}f_0$ (Hz)
166.67	250	333.33
333.33	500	666.67
666.67	1000	1333.33
1333.33	2000	2666.67
2666.67	4000	5333.33

TABLE 5.5. Center and −6 dB cutoff frequencies for the bandpass filter used during the listening.

5.3.2. Source Stimuli

Two broad types of signals served as initial sources for further processing. First, several variations of white noise were used. In some instances, identical noise was used for both the left and right source channels. This is referred to as "dual-mono" noise. In other cases, independent noise generators were used. This results in a source signal with an Inter-Channel Coherence (ICC) of 0. For some tests the noise covered the full audible bandwidth. Frequently however, the noise was bandlimited (identically on both channels). This was done using a series of two biquadratic bandpass filters [20], implemented by the "BEQSuite" in SuperCollider. The overall filter results in a 1-octave bandwidth between the cutoff frequencies 6 dB below the peak. If the center frequency is f_0, these cutoff points will happen at $(2/3)f_0$ and $(4/3)f_0$. Table 5.5 lists the filter values that were used during the listening. Figure 5.21 shows the transfer function of this filter, centered at 1 kHz. Note that there is significant energy outside the nominal bandwidth.

In addition to noise, source signals were created by adding several sine waves between a minimum and maximum frequency, at intervals of 100 Hz. This allows for perfectly bandlimited stimuli. The frequency of the sine waves were modulated randomly by about ±1%, either

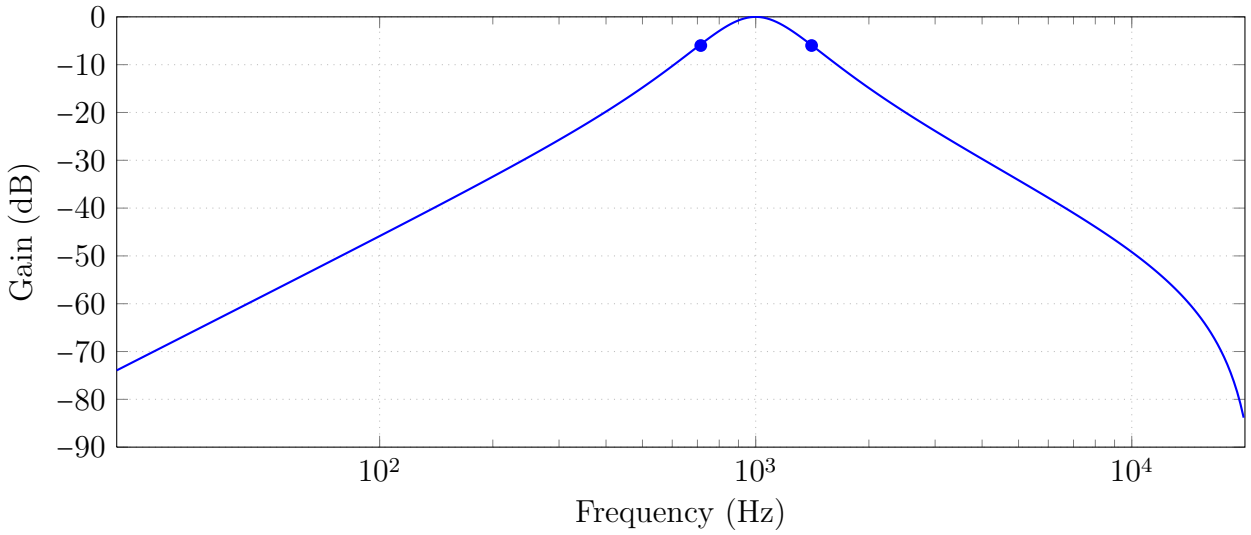


FIGURE 5.21. Magnitude response of the filter used to bandlimit noise, centered at 1 kHz. The marks indicate the −6 dB points.

identically for both channels (dual-mono) or independently. In both cases all sines in a given channel followed the same modulation pattern. This "common-fate" modulation encourages the auditory system to perceive the sound as a single source, rather than as independent partials [19]. We call this kind of stimulus "summed sines," with the understanding that this implies frequency modulation. To obtain decorrelated signals, the right channel was offset by 2 Hz. We refer to this as "frequency-based decorrelation" (see section B.2). In a few cases, we kept the frequency of both channels the same and used the methods described in section B.1 to produce decorrelated channels. We refer to this as "phase decorrelation." To demonstrate crosstalk cancellation, simple binaural synthesis based on the spherical head model was used to process dual-mono summed sines.

5.3.3. Overhead Speakers and Decorrelated Sources

The initial listening had two primary goals: first, to gain familiarity with the behavior of sounds presented via overhead loudspeakers; second, to compare dual-mono and decorrelated sources. Target location processing was not used.

Dual-Mono Sines

To begin, dual-mono summed sines with partials beginning at 100 Hz were played over both overhead pairs, A and B.

Sound Image. It was observed that the perceived elevation of the image depended strongly on the maximum frequency, with higher frequencies appearing more elevated. Up to about 3 kHz, the image spanned a range from horizontal to about 30°. As the maximum frequency increased from 3 kHz to 6 kHz, the image split up into several bands, despite the common-fate modulation. The highest band seemed elevated approximately at the level of the ceiling speakers. Once the maximum reached about 9 kHz, the entire sound image was at ceiling level. The width of the dual-mono image was fairly narrow (as would be expected). With speaker pair B, the image was far more narrow than the speakers themselves.

Effect of Listener Location. Moving from directly under the speakers to 1 m in front or behind widened the image slightly, and also prevented most front/back ambiguity. Lateral head motion perceptibly affected image azimuth. With speaker pair A, moving left or right approximately 2 m from center would place the sound completely in the near speaker. This is consistent with summing localization theory.

Other Observations. We also noted a dependence of elevation on signal level. As the level was gradually raised from silence, the image began in the horizontal plane and moved upward until it obtained its maximum elevation. We hypothesized that this was due to the well-known Fletcher-Munson equal-loudness curves (or their revised modern equivalent [69]). These indicate that at low levels, human hearing has reduced sensitivity to low and high frequencies. As the level is increased, sensitivity becomes more flat (though not completely so). This effect is asymmetric, so that as level increases, the spectral centroid (see section A.3) will also increase.

Decorrelated Sines

Compared to a dual-mono signal, summed sines processed with frequency-based decorrelation behaved quite differently.

Sound Image. The image was much broader and more difficult to localize. The perceived sound energy was not distributed evenly from left to right, but rather was strongly bimodally distributed, with a weaker (but still present) center fill. We noted that phase decorrelation created a more even distribution; however the remainder of our listening was conducted with frequency-based decorrelation. With respect to perceived height, decorrelated sines maintained a lower elevation through at least 6 kHz. Only once the maximum frequency was raised beyond that did the image move toward the ceiling. Finally, it seemed that frequencies in the 3–6 kHz range localized quite close to the head.

Effect of Listener Location. Consistent with expectation, decorrelated signals exhibited far less sensitivity to listener position. Head motion in a fixed seating location had virtually no impact on the auditory image. For speaker pair A, moving up to 2 m left or right of center had little effect on the perceived image, except perhaps a modest shift of overall energy distribution toward the near speaker. With speaker pair B, moving 1 m left or right of center (virtually directly underneath one speaker) produced a very asymmetric image, extending vertically toward the near speaker and horizontally toward the far speaker. Nonetheless, the sound image was continuous and had significant perceived width. Moving ±2 m off center (outside the physical span of the speaker pair) was sufficient to collapse the image to approximately the actual speaker location. For both pairs, it seemed that sitting 1 m in front of or behind the speakers decreased the ability to localize the speakers somewhat, compared to sitting directly underneath the line of the speakers.

5.3.4. Direct Path Compensation and Energy Compensation

The listening described in this section focused on the effects of direct path compensation and energy compensation. A brief first demonstration was followed by a more systematic evaluation.

Preliminary Listening

Stimuli. Initial assessment of direct path compensation used summed sines and frequency-based decorrelation. We first used speaker pair A. The target listener position was moved from directly under the left speaker to directly under the right speaker over 30 seconds.

Observations. For partials from 100–500 Hz, the diffusion was so total that moving the target position had little effect, and no comparison could be made. With partials from 500–1500 Hz, we observed that direct path compensation created a somewhat smoother transition as the target location passed by. Results with partials from 1.5–3 kHz and 3–6 kHz were similar, but with more evident motion of the target area and a more noticeable improvement with direct path compensation.

Non-Elevated Speakers. The procedure was then repeated using speaker pair C, maintaining the same target path. At low frequencies, the image was broad but clearly out front of the listener, with the result that there was no envelopment. Immersion did increase steadily as the frequency content rose. This contrasts with the overhead speakers, which produced an enveloping effect at all frequencies. The lateral distribution became very bimodal at higher frequencies.

Primary Listening

Stimuli. More extensive listening was done with bandlimited noise, primarily using speaker pair A, the wide overhead speakers. The same target line and path duration as the earlier tests were used. The tests directly compared five stimuli/processing combinations: dual-mono noise without and then with direct path compensation; and decorrelated noise without and with direct path compensation, as well as with energy compensation. The dual-mono source served as a baseline for comparison with the decorrelated signals.

Dual-Mono Noise. For dual-mono noise bandlimited around 4 kHz, the image was very elevated, and largely in the speakers when outside the target location. Strong flanging effects could be heard, especially moving into and out of the target. Direct path compensation slightly reduced this. Results for the other bands were similar, though lower frequencies affected perceived elevation (see below).

Decorrelated Noise. With bandlimiting around 4 kHz and no energy compensation, the image had a bimodal distribution. There was no "panning" sensation as the target moved, but rather a gradual shift in the balance of energy. The image was distant and elevated at speaker level. There was a modest but perceptible increase in envelopment when positioned in the target location. With energy compensation, there was a significantly more even left-to-right distribution. As a result, though the absolute amount of envelopment was roughly similar, it had a more pleasing subjective quality. Figure 5.12 on page 137 suggests that energy compensation produces a lower IID, which would reinforce a center image. There was also a clearer perception of motion due to the moving target location, which might be desirable or not, depending on the aesthetic context. Results for bandlimiting around 2 kHz were similar.

For bandlimiting around 1 kHz, direct path compensation made a significant difference, creating a more continuous left-to-right distribution of energy in the sound image. However, energy compensation somewhat restored the bimodal distribution. The assumptions of this processing likely become invalid at this frequency range (see section 5.4.1), which is reasonable as it was designed for high frequencies.

Finally, centered around 500 Hz, decorrelation created a very wide image, though the distribution was bimodal without direct path compensation. With this compensation, there

was an even left-to-right distribution, and a very gradual transition as the target moved. Energy compensation was not tested for this frequency band.

Other Observations. With bandlimiting around 1 kHz, the image began diffusing vertically, with lower frequencies at the bottom and higher frequencies more elevated. Bandlimiting around 500 Hz, there was substantial vertical diffusion, as low frequencies were directly in front, and higher frequencies appeared toward the ceiling. With dual-mono noise, there were somewhat distinct vertical bands, while decorrelation created a more continuous field.

Processing with direct path and energy compensation was also evaluated for bandlimiting around 2 kHz over speaker pairs C and D. The image was no more enveloping than with the overhead pair. For pair C, the image was actually narrower and more distant than the speakers themselves, which is not encouraging for the typical stereo layout. Although the apparent source width was roughly comparable to speaker pair A, the image's location clearly out front of the listener created substantially less envelopment. With pair D, the image was a bit more enveloping, but not nearly as much as might be suggested by the extreme lateral positioning of this pair.

5.3.5. Crosstalk Cancellation

Initial attempts at crosstalk cancellation were disappointing. It was eventually determined that the crosstalk cancellation signal was delayed approximately 1 ms too much relative to the direct path compensation signal (a redundant modelling delay had been applied to the crosstalk signal; the bug was obscured by the routing to allow dynamic modification of the processing). This seemingly small alignment error drastically reduces the resulting channel

separation and effectively invalidates the crosstalk cancellation. We were later able to return to SARC to conduct a basic listening session with corrected code.

Decorrelation

We listened first over the narrow ceiling pair B. Using decorrelated noise bandlimited around 1 kHz, crosstalk cancellation created a much wider apparent source width when in the target zone, consistent with the lower IAC predicted by theory. A moving target location was equally as effective, but there were strong flanging artifacts when listening outside of the target area. This is inherent to noise stimuli and would likely be reduced, though perhaps not eliminated, with other kinds of sources.

Binaural Cues

Next, the spherical head model was used to impose binaural cues on dual-mono summed sines, with partials from 100–1000 Hz. Targeting was set for a centered listener. For a virtual source located straight ahead, the sound image appeared overhead. However, for virtual sources at ±90°, the image was convincingly to the side. A virtual source at ±130° appeared to the side and somewhat behind. Lateralization beyond the loudspeaker span is a strong indication that crosstalk cancellation was functioning properly. Attempts to use the range-dependent model to produce a sound image close to the head were unsuccessful however. The spherical head model can create this effect quite convincingly when listening over headphones, so this points to the difficulty of producing large interaural level differences at low frequencies using loudspeakers.

Other Observations

The listening was repeated using a wide ceiling pair. Speaker pair A had been returned to its original position, so pair A (alt.) was used as a substitute. Since this pair is almost directly over the mixing desk, that was used as the target line, rather than the line in the audience area used for all other tests. Results were essentially identical to those of pair B.

Curiously, binaural localization was somewhat improved when the spectral equalization loop was disabled. Since the loop should not affect interaural differences, this is unexpected. One possibility is that SARC is more reverberant at low frequencies, and so crosstalk cancellation is less successful when this range is prominent, but this is purely speculative.

Clearly, extensive further study is needed, but these initial trials are quite encouraging. Two observations from nearfield monitoring suggest possible future directions. First, synthesizing virtual sources close to the head is reasonably effective in this configuration, an additional tentative hint that room properties are involved. Second, logical but sometimes unexpected source motion can occur with moving targets. For example, consider a virtual source 90° to the right, and a target moving left to right. At the beginning of the path, the sound appears to the right, since the right speaker is emitting its signal earlier and louder. As the target reaches the listener and the binaural cues become effective, the image becomes more lateralized to the right. Finally, the image then moves to the left, as the left speaker becomes dominant.

5.4. Sources of Uncertainty or Error

This section discusses the known sources of uncertainty which affect the performance of the system. Many of the following are not "errors" per se, but rather basic design choices which

reflect the incomplete information available in real-world listening situations. For example, the spherical head model is an imperfect match for the unknown HRTFs of an arbitrary listener, leading to actual ear signals which are uncertain.

5.4.1. Accuracy of the Model

We begin by investigating the validity of the assumptions made by the target location processing. In some cases, we suggest possible improvements (see also section 6.2).

Spherical Head Model

Figure 5.22 shows the magnitude response of three filters. First, the KEMAR HRTF for a source at $\theta = 90°$, $\phi = 55°$ (the geometry for a listener centered under speaker pair A) to the ipsilateral (right) ear. Second, the inverse spherical head model filter for the same angles (at a distance $r = 5\,\text{m}$). Finally, the overall response when cascading the two filters in series is given. In other words, the figure shows the net result when attempting to invert this HRTF using the spherical head model. Visually, this is not immediately encouraging. The "inverted" response is far from flat. However, the head model is successful at creating a response with a narrower dynamic range through at least 10 kHz.[4] Different geometry or a different listener might lead to better or worse performance. Importantly, we are interested not in the response at each individual ear, but with the interaural differences. For example, 5 dB inversion at each ear could lead to as much as a 10 dB reduction in interaural differences. Evidence for

[4]The CIPIC data contains an unknown overall gain offset. The plots for the KEMAR HRTF and the net result were raised by 2.5 dB, the estimated gain offset based on the low-frequency behavior. This way, the success of the inversion can be judged qualitatively by examining how much the net result deviates from 0 dB. Quantitatively, one would measure the decrease in dynamic range over the relevant frequency band.

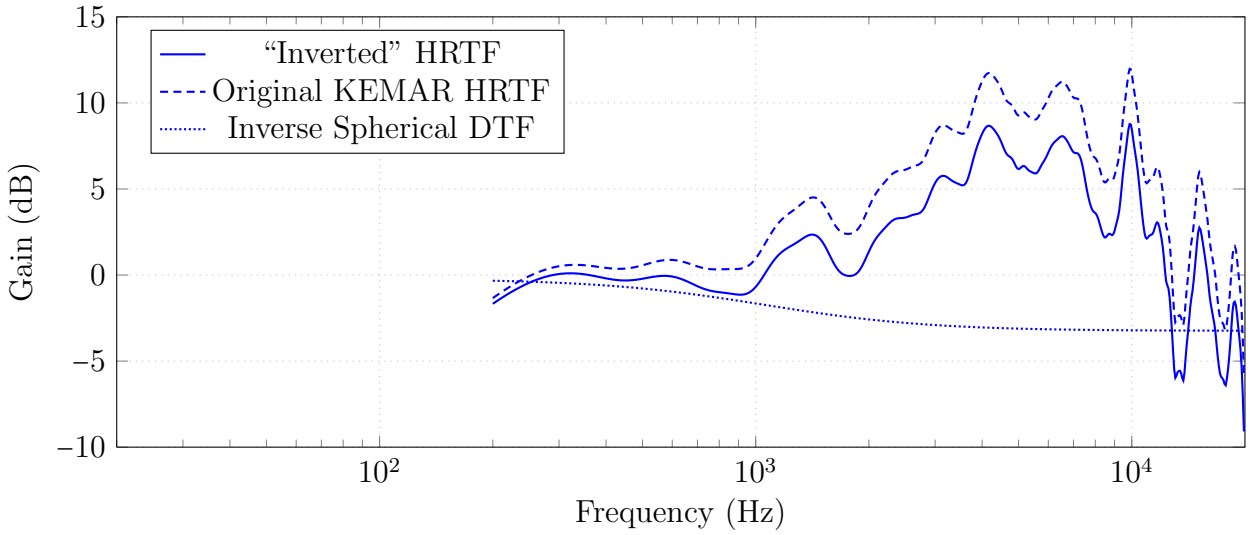


FIGURE 5.22. Head-model error: the solid line shows the spherical head-model inverse filter from a source at $\theta = 90°$, $\phi = 55°, r = 5\,\text{m}$ to the ipsilateral (right) ear; in series with the KEMAR HRTF for the same angles. Deviation from 0 dB gain indicates imperfect inversion. The dashed line shows the KEMAR HRTF only. Frequencies below 200 Hz are omitted because the CIPIC data is not reliable; see section 5.1.1.

the system's ability to control interaural differences was presented earlier in this chapter. Nonetheless, figure 5.22 suggests that improvements to the model would be productive.

Figure 5.4 on page 122 shows the ITD values for the CIPIC data for a source in the horizontal plane, as computed using the method in section 5.1.2. The solid line gives the modelled ITD for a source at 1 m using equation (4.23). RMS deviation is 0.03 ms. This data indicates that the KEMAR head is just slightly smaller than the spherical head model.

Energy Compensation Model

The energy compensation model makes several assumptions that are not always justified. First, it assumes that source power is evenly distributed across frequency (i.e., the source is white noise). This is often a reasonable approximation for the high-frequency region, but not always,

especially with synthesized signals such as bandlimited noise. For example, suppose we are using one-octave bands of pink noise (since pink noise has equal power per octave), applying energy compensation at all frequencies. With a band centered around 1 kHz, much of the energy might reach the contralateral ear; with a band centered around 4 kHz, comparatively little would. Nonetheless, both inputs would result in the same energy compensation gains. This could be solved by using a subband approach, which is straightforward in theory. However, it does require that source power be analyzed in the frequency domain, rather than using a simple running mean-square average in the time domain.

Second, the model assumes that the direct and crosstalk signals add incoherently at each ear, so that the power of their sum is the sum of their individual powers, but this is not always true. At high frequencies, the crosstalk signal will be delayed by more than one period of the relevant frequency range, and hence we expect that the assumption is valid. At lower frequencies, the delay will be less than one period and the assumption breaks down. The source ICCC also affects the coherence at the ears. The ICCC could be computed dynamically, and used in conjunction with the transmission model to estimate coherence at the ears, but this creates considerable conceptual and computational complexity. A compromise could assume a convenient ICCC for the source and use the transmission model to estimate the coherence of the direct and crosstalk signals at the ears. For further discussion of these assumptions in a somewhat different context, see [48].

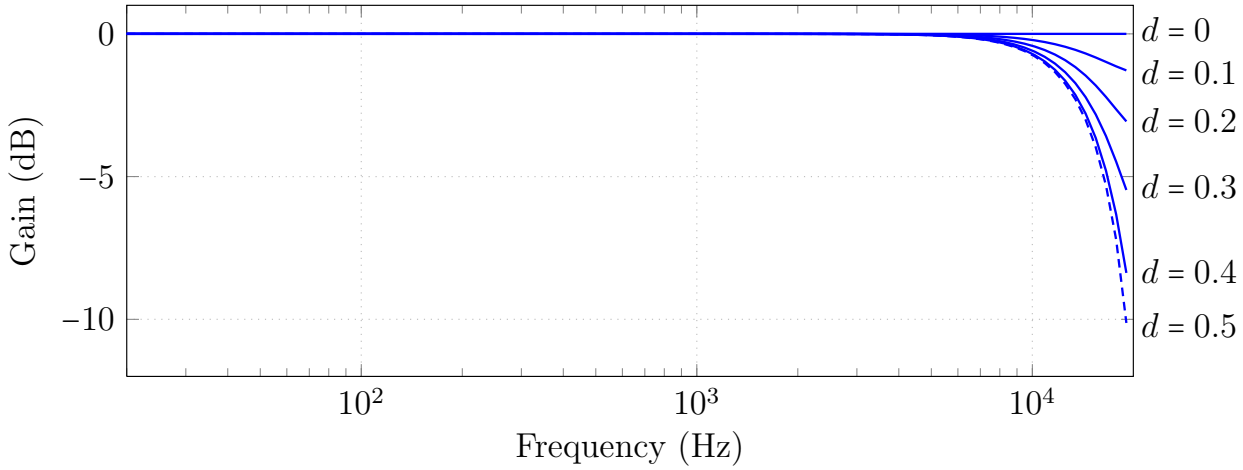
Anechoic Assumption

The model considers only direct and crosstalk transmission from the speakers to the ears. In non-anechoic conditions, sound also reaches listeners via reflections from room surfaces, which

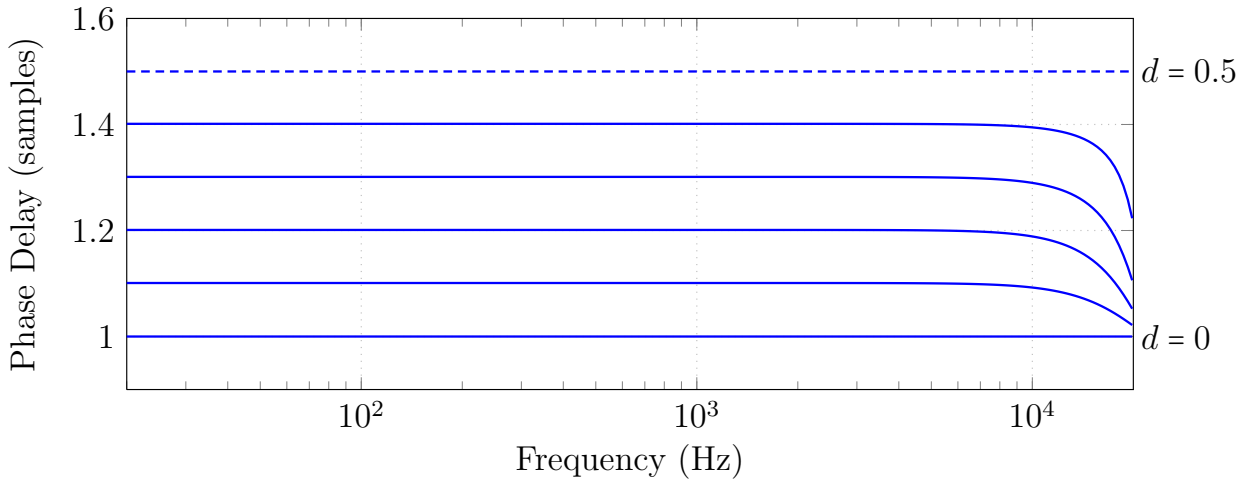
will obviously influence the soundfield. Quantitative analysis of the effect room reflections on the performance of spatial audio systems is difficult and there are few references. For crosstalk cancellation, Cooper and Bauck assert that reflections arriving after 1–2 ms are fairly benign and contribute only to a general room impression. However, crosstalk cancellation in the presence of a single reflection is studied in [117], which finds that reflections arriving as late as 10 ms have a significant impact on localization performance. While absolute localization accuracy is not a goal of the present system, this does suggest that reflections create a less ordered soundfield. Griesinger has studied the relative influence of room reflections and ICC on the resulting IAC, albeit in the context of small rooms [58]. He found that for very reflective spaces, room reflections are the dominant factor, but for more typical conditions, source coherence was quite significant. At SARC, the extensive wall absorption means that the strongest reflections probably come from the cement structural floor, approximately 5 m below ear level. At this distance, reflections arrive roughly 15 ms after the direct sound. It is certainly intuitive that the low reverberation time and distant reflective surfaces at SARC permit greater control, but the precise effect of the room remains unknown.

Other Modelling Error

The fractional delay approximation also introduces error into the model. SuperCollider uses a third-order Lagrange interpolating filter for fractional delays. While this is generally considered an excellent approximation, it nonetheless produces error at high frequencies, particularly above 10 kHz. Figure 5.23 shows the magnitude response and group delay for several fractions of a sample. Extensive further discussion of fractional delays is found in [83].



(A) Magnitude Response



(B) Phase Delay

FIGURE 5.23. Magnitude response and phase delay of the third-order Lagrangian interpolating fractional delay filter. Shown for fractions of a sample d, from $d = 0$ to the worst case $d = 0.5$ (dashed line).

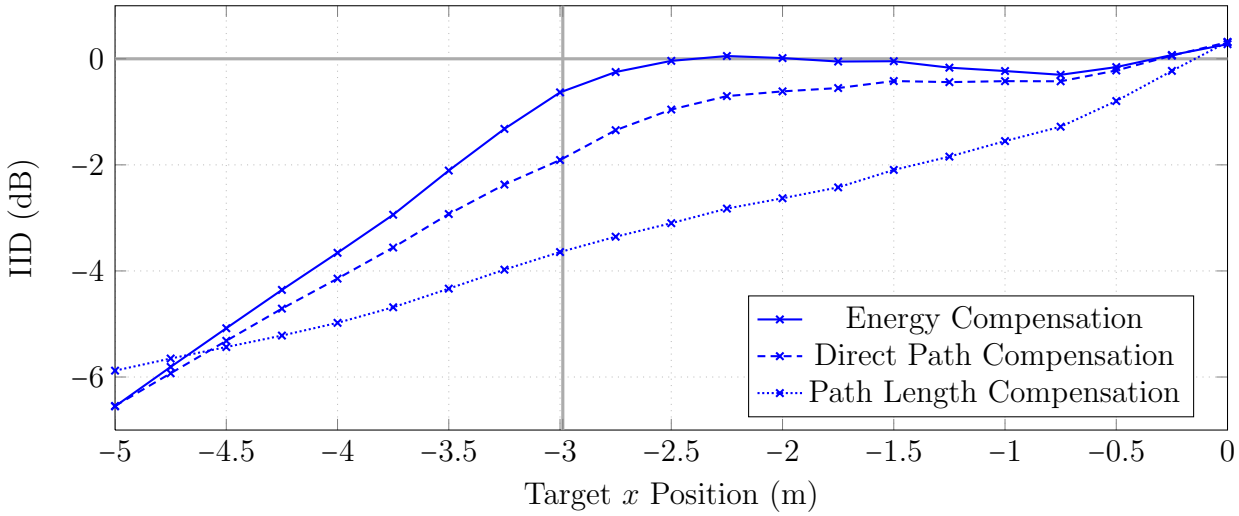


FIGURE 5.24. IID at several target positions. The γ function was used to compensate for range-dependent effects in the modelling HRTFs. Cf. figure 5.12 on page 137.

The accuracy of the HRTF database will obviously affect the numerical simulations. Reliability at low frequencies and range-dependent effects were discussed in section 5.1.1. The inverse γ function from the head model could be used to quickly compensate for first-order range dependency in the recorded data. In preliminary tests this seemed to improve some simulation results. For example, figure 5.24 shows that this method improves the computed IID results when using energy compensation. However, more rigorous verification is needed. The CIPIC documentation also mentions potentially low signal-to-noise ratios at the contralateral ear. Beyond that, the process of recording HRTFs is well outside the scope of this document. The quality of the CIPIC database is generally considered excellent.

Finally, the model of air propagation is only a first approximation. We assume free-field propagation, meaning that the sound level decreases by 6 dB for each doubling of the distance. However, in an enclosed space, the sound level will drop somewhat less due to the presence of reflections. This could be incorporated as part of a room model, or accounted for separately.

We have also ignored the well-known effect of air absorption. This is a frequency-dependent (as well as humidity- and atmospheric pressure-dependent) phenomenon which results in greater attenuation of high frequencies [10, 66]. The correction is quite small for most listening scenarios: at a distance of 10 m, there is almost no effect at 1 kHz. Frequencies of 10 kHz are attenuated roughly 3 dB more than the free-field model without air absorption. The effect becomes important when modelling room reflections, which might easily travel 40 m or more through the air before reaching the listener. For direct sounds, the simple model we have used is only a minor source of error.

5.4.2. Loudspeakers

The loudspeakers are a major source of uncertainty. Typically, loudspeaker transfer functions are unknown, across both frequency and spatial position. Failing to compensate for loudspeaker characteristics can substantially degrade spatial audio performance [114]. The datasheets published for the loudspeakers at SARC do provide some information; Genelec offers fairly detailed specifications [50, 51], while the Meyer data is inadequate to construct even a simple model [94, 95]. Individual speakers can vary significantly from the published figures. The speakers at SARC seemed well-matched within each model, but this was not verified. Spatial response is likely the largest cause of non-unity speaker transfer functions. The Meyer UPM-1Ps at the ceiling level exhibited a particularly noticeable variation in on- and off-axis high-frequency response. Lacking measurement equipment adequate to quantify these variations, all loudspeakers were modelled as ideal point sources with a unity transfer function.

Second, loudspeaker positions are not typically measured precisely. Accurate measurements of the speaker locations at SARC had been made at one point [65], but this data was clearly

no longer completely valid at the time of these trials. Reasonable corrections were made but without access to a laser distance meter, precise data on the ceiling speakers in particular was impossible. As a very rough estimate, the x- and y-coordinates of the speakers were probably accurate to a few centimeters; the height of the elevated speakers, to 10 cm.

Finally, the left loudspeaker in pair E, the mid-high side pair, produced a buzzing sound which strongly cued localization to that speaker location. Consequently, we were unable to make extensive use of that pair.

CHAPTER 6

Conclusions and Future Directions

This chapter summarizes the preceding discussion and the contributions of this research. It then provides suggestions for possible next steps.

6.1. Summary

A new approach to spatial audio for large venues was presented. After a historical review of prior approaches and the necessary perceptual background, the primary components of the system were outlined. Spatial cues are targeted to a specific audience location, which is swept over time. Because the cues tend to extend to a region in front of and behind the target, the majority of listeners will receive them at some point. Targeting relies on estimates of the ear signals based on a spherical head model. This provides good overall results while avoiding strong spectral features which might create audible artifacts.

We also made the novel observation that overhead loudspeakers are advantageous for spatialization. This configuration permits maximal symmetry of the soundfield, minimizing path-length differences. Further, it allows practical access to a range of loudspeaker placements which can favor either a broader region of desirable timing cues, or better channel separation.

The spherical head model was thoroughly analyzed. Analytical expressions were found for many properties of the model. Using the exact solution for a rigid sphere, the model was extended to include range dependency. Several means of controlling the target location were detailed, with new extensions described for the equalization section of the crosstalk canceller,

and for the energy compensation model. The complete system was implemented in software, and runs in real time on standard computer hardware.

The system was evaluated by visualizing properties of the soundfield over the listening area for different conditions of operation. This supported our claim for the benefits of overhead loudspeakers. The processing was found to successfully control parameters relating to spatial cues over a wide area. These include interaural intensity and delay, decorrelation, and frequency-dependent binaural cues. In addition, listening was conducted at the SARC facility to confirm the functional accuracy of the implementation. The subjective responses to this listening were largely consistent with expectations based on the theoretical analysis. Finally, we discussed sources of uncertainty that arise in practical deployment.

6.2. Future Directions

There are numerous avenues available to extend the scope of this work. We suggest here several possibilities which offer interesting potential.

Controlled Perceptual Experiments

An obvious next step is to conduct precise controlled experiments to quantify the perceptual performance of the system. This is a far from trivial task. There are general difficulties common to all such trials; in particular, ecological validity is certainly a consideration. More importantly, the success of the system should not be judged by absolute localization accuracy, which is (relatively) easy to quantify. Rather, responses must capture in some way a subjective sense of quality of spatial experience. Rumsey has proposed an extensive taxonomy of descriptors for spatial scenes [116], and details the use of several experimental

methodologies to quantify subjective characterizations of spatial attributes [115; 14, with Berg]. These papers also contain excellent guidance on auditory perceptual experiments in general.

Improvements of the Model

Section 5.4.1 pointed out several approximations made by the model, and suggested some possible improvements. One refinement that might be particularly relevant to overhead loudspeakers is an ellipsoidal rather than spherical head model. Duda et al. show that this improves ITD accuracy by up to about 0.1 ms, particularly at high elevations and large azimuth angles [43]. Unfortunately, there is no analytic expression for the ITD of this model, and heuristic algorithms are required.

It is almost certainly worthwhile to incorporate some information on the physical characteristics of the transducers. Unlike nearfield monitoring, reproduction in large spaces means that the majority of the audience is *not* on-axis with respect to the speakers, so their spatial response becomes very important. It is hardly reasonable to expect the average user to have access to detailed transfer function measurements. Perhaps a simple low-order filter model is all that is required. A few parameter sets could be predetermined for the most common loudspeaker design types. If the user has access to a Sound Pressure Level (SPL) meter, readings of bandlimited noise at a few polar angles could further calibrate the filter.

Modifications must carefully balance increased target accuracy with potential artifacts. The above improvements maintain a smoothly varying spatial and frequency-domain response. Processing based in the time domain requires particular caution. For example, it is natural to consider room modelling for the purposes of echo cancellation. Away from the target location

though, the cancellation will fail and this would likely have the opposite effect, introducing an additional perceived reflection. Perhaps an attenuated cancellation signal could mitigate this while still improving target accuracy. Along similar lines, a crosstalk canceller with a variable gain control on the cancellation signals is potentially quite interesting. This might allow a continuous transition from a large target area to a narrow but more accurate one. Although bandlimited crosstalk cancellers obviously reduce gain above the cutoff, a broadband gain control does not seem to be considered in the literature.

One final suggestion that also appears unexplored is to use different head models for different components of the compensation circuit. It was mentioned that certain "boosted" frequency bands correspond to sound sources located behind, in front or overhead [17]. Direct path compensation could employ a more detailed head model, though one still extremely smooth compared to actual HRTF data. An inverse head model with a broad valley in the overhead band (for overhead speakers of course) might inhibit perception of the speaker location and facilitate imaging at lower elevations. Crosstalk cancellation should still use the original model, since it relies on phase alignment and is therefore more prone to artifacts.

Optimal Speaker Locations and Multiple Loudspeakers

In section 5.2.2 we discussed the interaction between speaker placement and the resulting soundfield, and considered some of the possible tradeoffs between different factors. The optimal placement under various conditions remains unknown, however. This almost certainly depends on the absolute size of the room, as well as the intended spatial effects. A reasonable guess is that localization effects, which require good control of ITD, should utilize narrower speakers. Decorrelation effects or virtual sources close to the head (because of the great

difficulty in creating a large low-frequency IID) should use a wider pair, to take advantage of natural head-shadowing.

While this system explicitly avoids the need for multiple loudspeakers, it is always interesting to consider their use. Certain applications are natural; if reliable rear localization is needed, rear loudspeakers are appropriate. We mention three more interesting possibilities, considering both the narrow and wide overhead pairs together. First, the speakers could be used pairwise, following the target position. Second, the four speakers could all be used simultaneously for crosstalk cancellation at a single target. This forms an overdetermined linear system, which theoretically permits a reduction in the spatial gradient of the sound field [4]. In other words, spatial cues would change more slowly over space, widening the target area. However this analysis assumes perfect crosstalk cancellation—it is possible that, using an imperfect head model or considering non-targeted listeners, the additional cancellation signals would actually cause *more* artifacts. Finally, and perhaps most in the spirit of the present work, the outer and inner pair could be used in a source-dependent way, favoring either minimized path-length differences or maximized head-shadowing, as we have discussed. The best way to determine this balance in real time is another open question.

Parameters of the Source Signals and Target Path

The source stimuli play a critical role in the kinds of spatial attributes experienced by the listener. Basic properties such as bandwidth and time-varying correlation are likely candidates for further study. More exotic stimuli are also intriguing. For instance, Faller and Merimaa suggest that ITD and IID cues are only utilized when the IAC is above a certain threshold [46, 93]. It would be interesting to create stimuli with a different coherence in each critical

band [18]. Perhaps this would simply segregate the sound image into separate components, but conceivably this could yield a source both diffuse and tightly localized. Clearly, this is only one possibility out of many. Finally of course, exploration of actual musical sources would be valuable.

The motion of the target path is also relevant. The precise effect of such basic parameters as the duration of path motion remains unknown. As just one example of how this might be utilized, we mention an asymmetry observed during the listening demonstrations. It seemed that listeners who were toward the beginning of the path would experience a longer duration of spatial effect. We speculated that there was a "persistence effect," causing the ear to retain the impression once established. This is conceptually, if not perceptually, similar to the well-known Franssen effect: a sound is played from one loudspeaker, then faded out while simultaneously faded in to a second speaker. The sound remains localized in the first speaker for many seconds [17].

Musical Application

As a closing thought, we believe that this work should not remain used solely in a research context. Composers face fewer constraints and so find possibilities that are difficult to quantify. The complexities of source and target parameters become further magnified when considering multiple simultaneous inputs, perhaps with independent target motion. It is possible to imagine a situation where every audience member is receiving cues for at least one source at all times. Perhaps sources are split into multiple components, in a process similar to granular synthesis, with the target for each component following rapidly in succession. Hopefully this study has created the foundation for new avenues of musical expression.

APPENDIX A

Standard Signal Processing Definitions

In this appendix we give a few definitions of basic functions and methods of quantifying various signal properties. A standard reference text for discrete-time signal processing is [105]. Statistical signal processing is covered in [108, 110]. An excellent and very comprehensive reference for audio signal metrics is [111].

A.1. General Definitions

The discrete-time delta sequence is given by

$$\delta[n] = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \neq 0. \end{cases} \tag{A.1}$$

The discrete-time unit step sequence is given by

$$u[n] = \begin{cases} 1 & \text{if } n \geq 0 \\ 0 & \text{if } n < 0. \end{cases} \tag{A.2}$$

The discrete-time Fourier transform and inverse Fourier transform are given by

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n} \tag{A.3a}$$

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}\,\mathrm{d}\omega. \tag{A.3b}$$

Since the Fourier transform is in general complex, we often refer to its magnitude and phase. It is sometimes convenient to use the group delay instead of the phase. Group delay is defined as

$$\tau_g(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\Omega} \measuredangle X(e^{j\omega}), \tag{A.4}$$

where we use the unwrapped (continuous) phase.

A.2. Basic Statistical Measures

We take a very informal approach to statistical signal processing, omitting all development of logical prerequisites such as random processes and wide-sense stationarity. In most cases involving audio, we must compute using short-time approximations, which are guaranteed to exist.

The *mean*, *average* or *expected value* is defined as:

$$\begin{aligned} \mu_{x[n]} &= \mathcal{E}\{x[n]\} \\ &= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x[n]. \end{aligned} \tag{A.5}$$

A useful result is that expectation is linear; given constants a and b, we have

$$\mathcal{E}\{ax[n] + by[n]\} = a\,\mathcal{E}\{x[n]\} + b\,\mathcal{E}\{y[n]\}\,. \tag{A.6}$$

The *mean-square* or *average power* is:

$$\begin{aligned} \mathcal{P}\{x[n]\} &= \mathcal{E}\left\{|x[n]|^2\right\} \\ &= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} |x[n]|^2. \end{aligned} \tag{A.7}$$

The *variance*, denoted σ^2, is a measure of deviation from the mean, given by:

$$\begin{aligned}\sigma^2_{x[n]} &= \mathcal{E}\left\{\left|x[n] - \mu_{x[n]}\right|^2\right\} \\ &= \mathcal{E}\left\{\left|x[n]\right|^2\right\} - \left|\mu_{x[n]}\right|^2 .\end{aligned} \tag{A.8}$$

For signals with no DC power, i.e., 0 mean, the variance and average power are equal. For simplicity, we assume that all audio signals have no DC offset. For uncorrelated signals only (see section A.4), it can be shown with a bit of algebra that the power of the sum is the sum of the power:

$$\mathcal{P}\{x_1[n] + x_2[n]\} = \mathcal{P}\{x_1[n]\} + \mathcal{P}\{x_2[n]\} . \tag{A.9}$$

Definitions for continuous-time signals are similar:

$$\mathcal{E}\{x(t)\} = \lim_{T\to\infty} \frac{1}{2T} \int_{t=-T}^{T} x(t)\,\mathrm{d}t \tag{A.10}$$

$$\mathcal{E}\left\{\left|x(t)\right|^2\right\} = \lim_{T\to\infty} \frac{1}{2T} \int_{t=-T}^{T} \left|x(t)\right|^2\,\mathrm{d}t \tag{A.11}$$

$$\sigma^2_{x(t)} = \mathcal{E}\left\{\left|x(t)\right|^2\right\} - \left|\mu_{x(t)}\right|^2 . \tag{A.12}$$

A.3. Energy and Power

Over a specified period, a signal's energy is simply the sum of the squares of its samples:

$$E_x = \sum_{n=0}^{N-1} \left|x[n]\right|^2 . \tag{A.13}$$

Consistent with the definition already given, power is the average energy per sample:

$$\mathcal{P}\{x\} = \frac{1}{N} \sum_{n=0}^{N-1} \left|x[n]\right|^2 . \tag{A.14}$$

Parseval's theorem gives a frequency-domain expression for a signal's total energy:

$$\sum_{n=\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left|X(e^{j\omega})\right|^2 \, \mathrm{d}\omega. \tag{A.15}$$

The energy in an arbitrary frequency band from ω_1 to ω_2 is given by (using the fact that, for $x[n]$ real, $|X(e^{j\omega})| = |X(e^{-j\omega})|$):

$$E_{\text{band}} = \frac{1}{\omega_2 - \omega_1} \int_{\omega_1}^{\omega_2} \left|X(e^{j\omega})\right|^2 \, \mathrm{d}\omega. \tag{A.16}$$

The spectral centroid of a signal is the "center of mass" of its energy distribution. It is the frequency such that, if we divided the signal into two bands at that point, the upper and lower bands would have equal energy. It is found by taking the average of all frequencies (or frequency bins), weighted by each frequency's amplitude [111]:

$$C = \frac{\sum_{n=0}^{N-1} A_n f_n}{\sum_{n=0}^{N-1} f_n}. \tag{A.17}$$

A.4. Correlation

Correlation is one kind of similarity measure. Intuitively, the correlation between two signals measures how well the value of one predicts the value of the other (a bit more precisely, correlation only captures linear relationships). Correlation, properly normalized, ranges from −1 to +1. A correlation of +1 indicates a perfect linear relationship, while a correlation of −1 indicates a perfect inverse linear relationship. A correlation of 0 indicates that the values of one signal permit no prediction about the values of the other. Correlation is a function of the *lag*, denoted τ, which is simply a delay value applied to one signal.

Often, we are concerned with the correlation between signals arriving at the two ears. The *interaural cross-correlation* is defined by:

$$\begin{aligned} \text{IACC} &= \varphi_{LR}(\tau) \\ &= \frac{\mathcal{E}\{e_L(t)e_R(t+\tau)\}}{\sqrt{\mathcal{E}\{e_L(t)^2\}\,\mathcal{E}\{e_R(t)^2\}}} \\ &= \frac{\int_{t1}^{t2} e_L(t)e_R(t+\tau)\,\mathrm{d}t}{\sqrt{\int_{t1}^{t2} e_L(t)^2\,\mathrm{d}t \int_{t1}^{t2} e_R(t)^2\,\mathrm{d}t}}. \end{aligned} \tag{A.18}$$

The *interaural coherence* is given by:

$$\text{IAC} = \underset{-t\le\tau\le+t}{\text{extr}}\, \varphi_{LR}(\tau), \tag{A.19}$$

where "extr" denotes the extremum; i.e., the minimum or maximum with the greatest absolute value. (In practice, the term "interaural cross-correlation" frequently refers to this measure). Often the value of t is set to 1 ms [46]. Strictly speaking, this measure applies only to ear signals; when referring to input channels we should speak of the *inter-channel cross-correlation* (ICCC) and *inter-channel coherence* (ICC). Similar definitions for cross-correlation and coherence of arbitrary signals x and y are readily obtained. The special case $x = y$ yields the signal's *autocorrelation*. The autocorrelation at $\tau = 0$ is always 1, hence "autocoherence" is not a useful concept.

APPENDIX B

Techniques for Creating Decorrelated Signals

This appendix presents three methods for controlling the correlation between signals.

B.1. Phase Decorrelation

Kendall describes a method for obtaining two signals with any desired correlation φ from a mono source by manipulating their phase [74]. The version here is very slightly adapted. First, two independent random sequences are scaled so that their values are between $-\pi$ and $+\pi$; call the scaled sequences A and B. Next, two allpass filters are created by treating linear combinations of the random sequences as phase specifications and taking the IFFT to find the impulse responses. The phase for the left channel is specified as

$$\measuredangle L = \begin{cases} A & \text{if } \varphi = 0 \\ A + B & \text{otherwise.} \end{cases} \tag{B.1}$$

The right channel is given by

$$\measuredangle R = \begin{cases} |\varphi|\, A + B + \pi & \text{if } \varphi < 0 \\ \varphi A + B & \text{if } \varphi \geq 0. \end{cases} \tag{B.2}$$

Both the left and right phases are wrapped to stay between $\pm\pi$. Finally, the source is convolved separately with each allpass filter, resulting in two outputs with the desired correlation.

This concept is particularly simple to implement when frequency-domain processing is available. The sequences A and B are created as above. Next the source is duplicated and the FFT of both copies is taken. Finally, the phase of each bin is set according to the random sequences, and the IFFT is used to convert back to the time domain.[1]

An extension of this method based on critical bands and some additional references are found in [18].

B.2. Frequency Decorrelation

Two orthogonal signals (i.e., signals with no frequencies in common) will have a cross-correlation of 0. This derives from the frequency-domain interpretation of cross-correlation. Consider two signals $x[n]$ and $y[n]$, with Fourier transforms $X(e^{j\omega})$ and $Y(e^{j\omega})$. We can define the *cross-spectral density* by:

$$\Phi_{xy}(e^{j\omega}) = X(e^{j\omega})Y^*(e^{j\omega}) \tag{B.3}$$

(where the asterisk denotes complex conjugation). By a slight generalization of the Wiener-Khintchine theorem, cross-correlation and cross-spectral density form a Fourier transform pair:

$$\varphi_{xy}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_{xy}(e^{j\omega})e^{j\omega n}\,\mathrm{d}\omega \tag{B.4a}$$

$$\Phi_{xy}(e^{j\omega}) = \sum_{n=-\infty}^{\infty} \varphi_{xy}[n]e^{-j\omega n}. \tag{B.4b}$$

[1] My thanks to Scott Wilson, of the University of Birmingham, for his SuperCollider implementation.

By direct substitution,

$$\varphi_{xy}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})Y^*(e^{j\omega})e^{j\omega n}\,\mathrm{d}\omega. \tag{B.5}$$

Clearly then, if X and Y are orthogonal over ω, the cross-correlation $\varphi_{xy}[n]$ will be 0. A more rigorous development is found in [108, 110]. By creating a dual-mono signal and slightly shifting the frequencies in one channel, we can obtain decorrelated signals. In practice of course, we must account for the frequency resolution both of the short-time Fourier transform and of the ear. Generally speaking, shifting by $1-2\%$ is sufficient.

B.3. Sum and Difference Processing

Unlike the prior two methods, this technique requires a stereo signal as input. We begin by defining the so-called "shuffler matrix" [29]:

$$\mathbf{U} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{B.6}$$

Note that $\mathbf{U} = \mathbf{U}^{-1}$. When applied to the input $\mathbf{x}$, we obtain sum and difference signals:

$$\begin{aligned} \mathbf{U}\mathbf{x} &= \begin{bmatrix} (x_L + x_R)/\sqrt{2} \\ (x_L - x_R)/\sqrt{2} \end{bmatrix} \\ &\equiv \begin{bmatrix} x_S \\ x_D \end{bmatrix}. \end{aligned} \tag{B.7}$$

Next, the sum and difference signals are processed by a matrix $\mathbf{M}$:

$$\mathbf{M}(\mathbf{Ux}) = \begin{bmatrix} \Sigma & 0 \\ 0 & \Delta \end{bmatrix} \begin{bmatrix} x_S \\ x_D \end{bmatrix} = \begin{bmatrix} \Sigma x_S \\ \Delta x_D \end{bmatrix}. \tag{B.8}$$

Finally, the shuffler matrix is applied again to obtain the new (processed) left and right channels:

$$\begin{aligned} \mathbf{x}^{\mathbf{SD}} &= \mathbf{U}^{-1}(\mathbf{MUx}) \\ \begin{bmatrix} x_L^{\mathrm{SD}} \\ x_R^{\mathrm{SD}} \end{bmatrix} &= \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma x_S + \Delta x_D \\ \Sigma x_S - \Delta x_D \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} (\Sigma + \Delta) x_L + (\Sigma - \Delta) x_R \\ (\Sigma - \Delta) x_L + (\Sigma + \Delta) x_R \end{bmatrix}. \end{aligned} \tag{B.9}$$

The last form suggests that increasing Σ has the effect of adding crosstalk between the two channels, increasing the inter-channel cross-correlation. Increasing Δ adds out-of-phase crosstalk, decreasing the inter-channel cross-correlation. This method will of course fail if $x_D = 0$ (the signal is "dual-mono") or $x_D = \sqrt{2}\, x_L$ (dual-mono with the polarity reversed in one channel); however we ignore these degenerate cases.

To maintain roughly constant power, we relate Σ and Δ such that the sum of their squares is always 1. For $\phi \in [0, \frac{\pi}{2}]$, we can write:

$$\begin{aligned} \Sigma &= \sin(\phi) \\ \Delta &= \cos(\phi). \end{aligned} \tag{B.10}$$

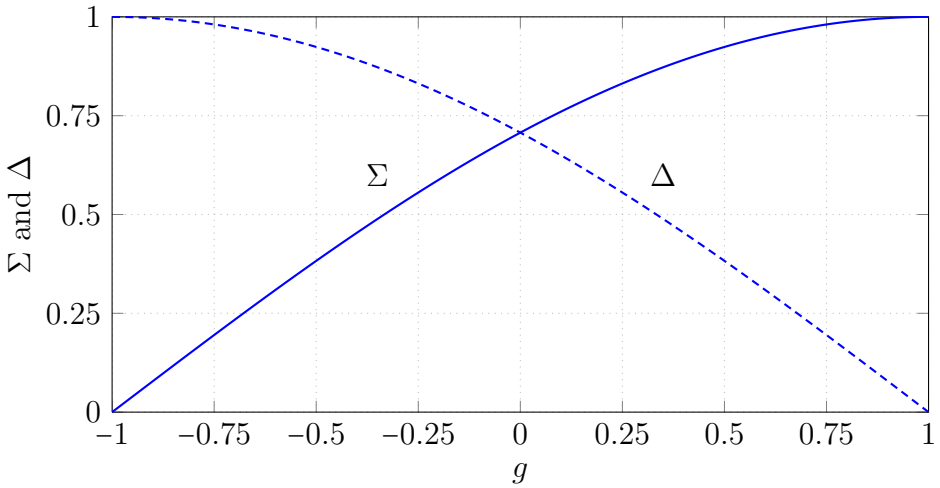


FIGURE B.1. Variation of Σ and Δ with respect to the parameter g.

Scaling this to a more convenient interval, we get

$$\begin{aligned} \Sigma &= \sin\left(\frac{\pi}{4}(g+1)\right) \\ \Delta &= \cos\left(\frac{\pi}{4}(g+1)\right) \end{aligned} \tag{B.11}$$

for $g \in [-1, 1]$. When $g = +1$, we have $x_L^{\mathrm{SD}} = x_R^{\mathrm{SD}}$ and the channels are perfectly correlated. Similarly, when $g = -1$ we have $x_L^{\mathrm{SD}} = -x_R^{\mathrm{SD}}$, and the channels are perfectly negatively correlated. (However note that in general, the inter-channel coherence does not equal g). Figure B.1 shows Σ and Δ as functions of g.

If we assume that the sum and difference signals are uncorrelated white noise signals of equal power P, the total power in each ear (and hence the total power at both ears) is

constant, independent of the value of g:

$$\begin{aligned}\mathcal{P}\left\{x_L^{\mathrm{SD}}\right\} &= \mathcal{E}\left\{\left(\Sigma x_S + \Delta x_D\right)^2/2\right\} \\ &= \frac{\Sigma^2}{2}\mathcal{E}\left\{x_S^2\right\} + \frac{\Delta^2}{2}\mathcal{E}\left\{x_D^2\right\} \\ &= \left(\Sigma^2 + \Delta^2\right)\frac{P}{2} \\ &= \frac{P}{2},\end{aligned} \tag{B.12}$$

and similarly for the right channel.

We can also calculate the inter-channel cross-correlation. We rely on the fact that the autocorrelation of white noise is 0 for $\tau \neq 0$ [108].

$$\begin{aligned}\mathrm{ICCC} &= \frac{\mathcal{E}\{x_L^{\mathrm{SD}}(t)x_R^{\mathrm{SD}}(t+\tau)\}}{\sqrt{\mathcal{E}\{x_L^{\mathrm{SD}}(t)^2\}\,\mathcal{E}\{x_R^{\mathrm{SD}}(t)^2\}}} \\ &= \frac{\mathcal{E}\left\{\left(\Sigma x_S(t) + \Delta x_D(t)\right)\left(\Sigma x_S(t+\tau) - \Delta x_D(t+\tau)\right)\right\}/2}{\sqrt{\mathcal{E}\left\{\left(\Sigma x_S(t) + \Delta x_D(t)\right)^2\right\}\mathcal{E}\left\{\left(\Sigma x_S(t) - \Delta x_D(t)\right)^2\right\}/4}} \\ &= \frac{\mathcal{E}\{\Sigma^2 x_S(t)x_S(t+\tau)\} - \mathcal{E}\{\Delta^2 x_D(t)x_D(t+\tau)\}}{\sqrt{\left(\mathcal{E}\{\Sigma^2 x_S(t)^2\} + \mathcal{E}\{\Delta^2 x_D(t)^2\}\right)^2}} \\ &= \frac{(\Sigma^2 - \Delta^2)\sigma^2\delta(\tau)}{(\Sigma^2 + \Delta^2)\sigma^2} \\ &= (\Sigma^2 - \Delta^2)\delta(\tau).\end{aligned} \tag{B.13}$$

The ICC then readily follows:

$$\begin{aligned}\mathrm{ICC} &= \operatorname*{extr}_{\tau}(\Sigma^2 - \Delta^2)\delta(\tau) \\ &= \Sigma^2 - \Delta^2.\end{aligned} \tag{B.14}$$

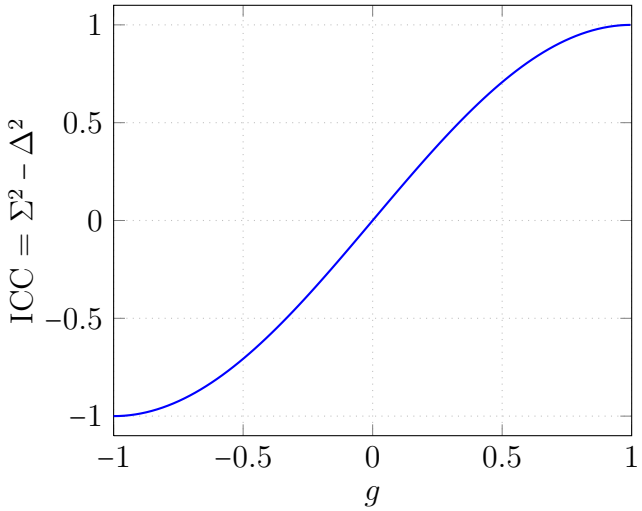


FIGURE B.2. ICC when the left and right channels are linear mixtures of uncorrelated white noise, controlled by the parameter g. See text for details.

Figure B.2 shows the ICC as it varies with the parameter g.

Bibliography

[1] Algazi, V. Ralph, Carlos Avendano, and Richard O. Duda. "Estimation of a Spherical-Head Model from Anthropometry." *Journal of the Audio Engineering Society* 49:6 (2001), pp. 472–479.
URL: http://www.aes.org/e-lib/browse.cfm?elib=10188.
URL: http://interface.cipic.ucdavis.edu/PAPERS/AES_ITD.pdf.

[2] Algazi, V. Ralph, Richard O. Duda, D. M. Thompson, and Carlos Avendano. "The CIPIC HRTF Database." In proceedings: *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY 2001, pp. 99–102.
ISBN: 978-0-7803-7126-2.
DOI: 10.1109/ASPAA.2001.969552.
URL: http://interface.cipic.ucdavis.edu/data/doc/CIPIC_HRTF_Database.pdf.

[3] Ando, Yoichi and Yoshitaka Kurihara. "Nonlinear Response in Evaluating the Subjective Diffuseness of Sound Fields." *Journal of the Acoustical Society of America* 80:3 (1965), pp. 833–836.
DOI: 10.1121/1.393906.
URL: http://link.aip.org/link/?JAS/80/833/1.

[4] Asano, F., Y. Suzuki, and T. Sone. "Sound Equalization Using Derivative Constraints." *Acta Acustica united with Acustica* 82:2 (1996), pp. 311–321.

[5] "Audio Engineering Society Events."
URL: http://www.aes.org/events/ (retrieved on 12/04/2008).

[6] Austin, Larry. "Sound Diffusion in Composition and Performance Practice. II: An Interview with Ambrose Field." *Computer Music Journal* 25:4 (2001), pp. 21–30.
DOI: 10.1162/01489260152815260.

[7] Bamford, Jeffrey Stephen. "An Analysis of Ambisonic Systems of First and Second Order." Master's thesis, University of Waterloo, 1995.
URL: http://audiolab.uwaterloo.ca/~jeffb/thesis/thesis.html.

[8] Barbour, James L. "Subjective Consumer Evaluation of Multi-Channel Audio Codecs." Presented at the AES 119th Convention, New York, NY 2005.
URL: http://www.aes.org/e-lib/browse.cfm?elib=13279.

[9] Barron, Michael. "The Subjective Effects of First Reflections in Concert Halls: The Need for Lateral Reflections." *Journal of Sound and Vibration* 15:4 (1971), pp. 475–494.
DOI: 10.1016/0022-460X(71)90406-8.
URL: http://www.sciencedirect.com/science/article/B6WM3-4951JM7-FN/2/3f496bf6546e926be082cef002852d46.

[10] Bass, H. E., H.-J. Bauer, and L. B. Evans. "Atmospheric Absorption of Sound: Analytical Expressions." *Journal of the Acoustical Society of America* 52:3B (1972), pp. 821–825.
DOI: 10.1121/1.1913183.
URL: http://link.aip.org/link/?JAS/52/821/1.

[11] Bauck, Jerry. "A Simple Loudspeaker Array and Associated Crosstalk Canceller for Improved 3D Audio." *Journal of the Audio Engineering Society* 49:1/2 (2001), pp. 3–13.
URL: http://www.aes.org/e-lib/browse.cfm?elib=10205.

[12] Bauck, Jerry and Duane H. Cooper. "Generalized Transaural Stereo and Applications." *Journal of the Audio Engineering Society* 44:9 (1996), pp. 683–705.
URL: http://www.aes.org/e-lib/browse.cfm?elib=7888.

[13] Begault, Durand R. *3-D Sound for Virtual Reality and Multimedia*. Cambridge, MA: Academic Press Professional, 1994.
ISBN: 978-0-12-084735-8.
URL: http://human-factors.arc.nasa.gov/publibrary/Begault_2000_3d_Sound_Multimedia.pdf.

[14] Berg, Jan and Francis Rumsey. "Spatial Attribute Identification and Scaling by Repertory Grid Technique and Other Methods." In proceedings: *AES 16th International Conference: Spatial Sound Reproduction*, Rovaniemi, Finland 1999, pp. 51–66.
URL: http://www.aes.org/e-lib/browse.cfm?elib=8049.
URL: http://epubs.surrey.ac.uk/recording/43/.

[15] Berger, Ivan. "Binaural Sound Hits the Big Screen." *Audio: The Equipment Authority* 81 (Oct. 1997), pp. 32–33.

[16] Berkhout, Augustinus J., Diemer de Vries, and Peter Vogel. "Acoustic Control by Wave Field Synthesis." *Journal of the Acoustical Society of America* 93:5 (1993), pp. 2764–2778.
DOI: 10.1121/1.405852.
URL: http://link.aip.org/link/?JAS/93/2764/1.

[17] Blauert, Jens. *Spatial Hearing: The Psychophysics of Human Sound Localization.* Rev. ed. Cambridge, MA: MIT Press, 1997.
ISBN: 978-0-262-02413-6.

[18] Bouéri, Maurice and Chris Kyriakakis. "Audio Signal Decorrelation Based on a Critical Band Approach." Presented at the AES 117th Convention, San Francisco, CA 2004.
URL: http://www.aes.org/e-lib/browse.cfm?elib=12948.

[19] Bregman, Albert S. *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press, 1994.
ISBN: 978-0-262-52195-6.

[20] Bristow-Johnson, Robert. "Cookbook Formulae for Audio EQ Biquad Filter Coefficients."
URL: http://www.musicdsp.org/files/Audio-EQ-Cookbook.txt (retrieved on 04/22/2009).

[21] Brown, C. Phillip and Richard O. Duda. "An Efficient HRTF Model for 3-D Sound." In proceedings: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY 1997.
ISBN: 978-0-7803-3908-8.
DOI: 10.1109/ASPAA.1997.625596.
URL: http://interface.cipic.ucdavis.edu/PAPERS/Brown1997(Efficient3dHRTFModels).pdf.

[22] Brown, C. Phillip and Richard O. Duda. "A Structural Model for Binaural Sound Synthesis." *IEEE Transactions on Speech and Audio Processing* 6:5 (1998), pp. 476–488.
DOI: 10.1109/89.709673.
URL: http://interface.cipic.ucdavis.edu/PAPERS/Brown_Duda98.pdf.

[23] Burkhard, M. D. and R. M. Sachs. "Anthropometric Manikin for Acoustic Research." *Journal of the Acoustical Society of America* 58:1 (1975), pp. 214–222.
DOI: 10.1121/1.380648.
URL: http://link.aip.org/link/?JAS/58/214/1.

[24] Busby, Kevin. "BEAST Home Page."
URL: http://www.ea-studios.bham.ac.uk/BEAST/ (retrieved on 12/03/2006).

[25] Chion, Michel and Walter Murch. *Audio-Vision: Sound on Screen.* New York: Columbia University Press, 1994.
ISBN: 978-0-231-07898-6.

[26] "CIPIC Interface Laboratory: Database."
URL: http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm (retrieved on 04/19/2009).

[27] Clark, Keith. "A New Day at the Colosseum." *Live Sound International* (June 2003).
URL: http://www.livesoundint.com/archives/2003/june/celine/celine.php.

[28] Clozier, Christian. "The Gmebaphone Concept and the Cybernéphone Instrument." *Computer Music Journal* 25:4 (2001), pp. 81–90.
DOI: 10.1162/01489260152815305.

[29] Cooper, Duane H. and Jerald L. Bauck. "Prospects for Transaural Recording." *Journal of the Audio Engineering Society* 37:1/2 (1989), pp. 3–19.
URL: http://www.aes.org/e-lib/browse.cfm?elib=6108.

[30] Cooper, Duane H. and Jerald L. Bauck. "Head Diffraction Compensated Stereo System." Pat. 4,893,342. 1990.
URL: http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=4893342.

[31] Cunningham, Mark. "Welcome to the Machine: The Story of Pink Floyd's Live Sound." Pts. 1–4. *Sound on Stage* 5-8 (Mar.–Jun. 1997).
URL: http://www.pinkfloyd-co.com/band/interviews/art-rev/art-sos1.html (continues with art-sos2.html; art-sos3.html; art-sos4.html).

[32] Davis, Mark F. "History of Spatial Coding." *Journal of the Audio Engineering Society* 51:6 (2003), pp. 554–569.
URL: http://www.aes.org/e-lib/browse.cfm?elib=12218.

[33] De Lancie, Philip. "Big Enough for Everest: Inside IMAX Sound." *Millimeter* (June 1, 1999).
URL: http://digitalcontentproducer.com/mag/video_big_enough_everest/.

[34] Desantos, Sandra. "Acousmatic Morphology: An Interview with François Bayle." *Computer Music Journal* 21:3 (1997), pp. 11–19.

[35] "Digital Theater System." *Wikipedia*.
URL: http://en.wikipedia.org/wiki/Digital_Theater_System (retrieved on 12/04/2006).

[36] "Dolby Digital." *Wikipedia*.
URL: http://en.wikipedia.org/wiki/Dolby_Digital (retrieved on 12/04/2006).

[37] Dolby Laboratories. "Dolby B, C and S Noise Reduction Systems: Making Cassettes Sound Better."
URL: http://www.dolby.com/uploadedFiles/zz-_Shared_Assets/English_PDFs/Professional/212_Dolby_B,_C_and_S_Noise_Reduction_Systems.pdf (retrieved on 12/04/2006).

[38] Dolby Laboratories. "Surround Sound Past, Present and Future: A History of Multichannel Audio from Mag Stripe to Dolby Digital." 1999.
URL: http://www.dolby.com/uploadedFiles/zz-_Shared_Assets/English_PDFs/Professional/2_Surround_Past.Present.pdf (retrieved on 12/04/2006).

[39] Dolby Laboratories. "5.1-Channel Production Guidelines." (Issue 1, part no. S00/12957). 2000.
URL: http://www.dolby.com/uploadedFiles/zz-_Shared_Assets/English_PDFs/Professional/L.mn.0002.5.1guide.pdf (retrieved on 12/04/2006).

[40] Dolby Laboratories. "Dolby Surround Mixing Manual." (Issue 2, part no. 91536). 2005.
URL: http://www.dolby.com/uploadedFiles/zz-_Shared_Assets/English_PDFs/Professional/44_SuroundMixing.pdf (retrieved on 12/04/2006).

[41] Dolby, Ray. "An Audio Noise Reduction System." *Journal of the Audio Engineering Society* 15:4 (1967), pp. 383–388.
URL: http://www.aes.org/e-lib/browse.cfm?elib=1083.

[42] Dolby, Ray. "The Spectral Recording Process." *Journal of the Audio Engineering Society* 35:3 (1987), pp. 99–118.
URL: http://www.aes.org/e-lib/browse.cfm?elib=5222.

[43] Duda, Richard O., Carlos Avendano, and V. Ralph Algazi. "An Adaptable Ellipsoidal Head Model for the Interaural Time Difference." In proceedings: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASPP)*, Phoenix, AZ 1999, vol. 2 pp. 965–968.
ISBN: 978-0-7803-5041-0.
DOI: 10.1109/ICASSP.1999.759855.
URL: http://interface.cipic.ucdavis.edu/PAPERS/Icassp99.pdf.

[44] Duda, Richard O. and William L. Martens. "Range Dependence of the Response of a Spherical Head Model." *Journal of the Acoustical Society of America* 104:5 (1998), pp. 3048–3058.
DOI: 10.1121/1.423886.
URL: http://link.aip.org/link/?JAS/104/3048/1.

[45] Everest, F. Alton. *Master Handbook of Acoustics*. 4th ed. New York: McGraw-Hill, 2001.
ISBN: 978-0-07-136097-5.

[46] Faller, Christof and Juha Merimaa. "Source Localization in Complex Listening Situations: Selection of Binaural Cues Based on Interaural Coherence." *Journal of the Acoustical Society of America* 116:5 (2004), pp. 3075–3089.
DOI: 10.1121/1.1791872.
URL: http://link.aip.org/link/?JAS/116/3075/1.

[47] Fetterman, William. *John Cage's Theatre Pieces: Notations and Performances*. Amsterdam, Netherlands: Harwood Academic Publishers, 1996.
ISBN: 978-3-7186-5642-4.

[48] Gardner, William G. "3-D Audio Using Loudspeakers." PhD dissertation, Massachusetts Institute of Technology, 1997.

Gardner, William G. *3-D Audio Using Loudspeakers*. Norwell, MA: Kluwer Academic Publishers, 1998.
ISBN: 978-0-7923-8156-3.

[49] Gellert, W., S. Gottwald, M. Hellwich, H. Kästner, and H. Künstner. *The VNR Concise Encyclopedia of Mathematics*. 2nd ed. New York: Van Nostrand Reinhold, 1989.
ISBN: 978-0-442-20590-4.

[50] Genelec. "1037B Data Sheet." (Document no. BBA37001). 2000.
URL: http://www.genelec.com/documents/datasheets/DS1037b.pdf.

[51] Genelec. "1038B Data Sheet." (Document no. BBA0055001a). 2005.
URL: http://www.genelec.com/documents/datasheets/DS1038b.pdf.

[52] Gerzon, Michael. "Periphony: With-Height Sound Reproduction." *Journal of the Audio Engineering Society* 21:1 (1973), pp. 2–10.
URL: http://www.aes.org/e-lib/browse.cfm?elib=2012.

[53] Gerzon, Michael. "Surround-Sound Psychoacoustics." *Wireless World* 80 (Dec. 1974), pp. 483–486.
URL: http://www.audiosignal.co.uk/Resources/Surround_sound_psychoacoustics_USL.pdf.

[54] Gerzon, Michael. "Papers on Ambisonics and Related Topics." 1995.
URL: http://www.york.ac.uk/inst/mustech/3d_audio/gerzonrf.htm (retrieved on 12/04/2006).

[55] Gray, Henry. *Anatomy of the Human Body*. 20th ed. Philadelphia: Lea and Febiger, 1918.
URL: http://www.bartleby.com/107/.

[56] Griesinger, David. "Spaciousness and Envelopment in Musical Acoustics." Presented at the AES 101st Convention, Los Angeles, CA 1996.
URL: http://www.aes.org/e-lib/browse.cfm?elib=7378.
URL: http://www.davidgriesinger.com/spac4.pdf.

[57] Griesinger, David. "The Psychoacoustics of Apparent Source Width, Spaciousness and Envelopment in Performance Spaces." *Acta Acustica united with Acustica* 83:4 (1997), pp. 721–731.
URL: http://www.davidgriesinger.com/SPAC7A.DOC.

[58] Griesinger, David. "Multichannel Sound Systems and their Interaction with the Room." In proceedings: *AES 15th International Conference: Audio, Acoustics and Small Spaces*, Copenhagen, Denmark 1998.
URL: http://www.aes.org/e-lib/browse.cfm?elib=10207.
URL: http://www.davidgriesinger.com/multichan.pdf.

[59] Griesinger, David. "Objective Measures of Spaciousness and Envelopment." In proceedings: *AES 16th International Conference: Spatial Sound Reproduction*, Rovaniemi, Finland 1999, pp. 27–41.
URL: http://www.aes.org/e-lib/browse.cfm?elib=8051.
URL: http://www.davidgriesinger.com/objmeas.pdf.

[60] Gundry, Ken. "An Introduction to Noise Reduction."
URL: http://www.dolby.com/uploadedFiles/zz-_Shared_Assets/English_PDFs/Professional/1000_kens_corner.pdf (retrieved on 12/04/2006).

[61] Harrison, Jonty. "Sound, Space, Sculpture: Some Thoughts on the 'What', 'How' and 'Why' of Sound Diffusion." *Organised Sound* 3:2 (1999), pp. 117–127.
DOI: 10.1017/S1355771898002040.

[62] Hawksford, Malcolm Omar. "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design." *Journal of the Audio Engineering Society* 45:1/2 (1997), pp. 37–62.
URL: http://www.aes.org/e-lib/browse.cfm?elib=7869.
URL: http://www.essex.ac.uk/csee/research/audio_lab/malcolmspubdocs/J37%20Digital%20signal%20processing%20tools,%20crossover%20design.pdf.

Hawksford, Malcolm Omar. "Corrections to 'Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design'." *Journal of the Audio Engineering Society* 45:6 (1997), p. 497.
URL: http://www.aes.org/e-lib/browse.cfm?elib=7849.

[63] Hertzog, Christian. "Morton Subotnick." In: *Contemporary Composers*. Ed. by Morton, Brian and Pamela Collins. Chicago: St. James Press, 1992.
ISBN: 978-1-55862-085-8.
URL: http://www.mortonsubotnick.com/about.html.

[64] "Historique du GRM."
URL: http://www.ina.fr/entreprise/activites/recherches-musicales/historique.html (retrieved on 12/04/2006).

[65] Hollerweger, Florian. "Sonic Lab Measurements." Unpublished e-mail. Dec. 7, 2007.

[66] Huopaniemi, Jyri, Lauri Savioja, and Matti Karjalainen. "Modeling of Reflections and Air Absorption in Acoustical Spaces: a Digital Filter Design Approach." In proceedings: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY 1997.
ISBN: 978-0-7803-3908-8.
DOI: 10.1109/ASPAA.1997.625594.
URL: http://www.acoustics.hut.fi/~ruba/pubs/refair.ps.

[67] Hurwitz, Matt. "The Sound of Love." *Live Design* (Aug. 1, 2006).
URL: http://livedesignonline.com/mag/sound_love/.

[68] IMAX Corp. "IMAX: The 15/70 Filmmaker's Manual." 1999.
URL: http://www.imax.com/images/corporate/pdfs/filmmaker.pdf (retrieved on 12/04/2006).

[69] International Organization for Standardization. "ISO 226:2003—Normal Equal-Loudness-Level Contours." 2003.
URL: http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=34222.

[70] Jacovitti, Giovanni and Gaetano Scarano. "Discrete Time Techniques for Time Delay Estimation." *IEEE Transactions on Signal Processing* 41:2 (1993), pp. 525–533.
DOI: 10.1109/78.193195.

[71] Johnson, David. "Chasing the Blues." *Live Design* (Oct. 1, 2000).
URL: http://livedesignonline.com/mag/show_business_chasing_blues_3/index.html.

[72] Kay, Jonathan, Kimber Ghent, Brian Chumney, and Erik Lutkins. "Film Sound History: Film Sound Formats."
URL: http://www.mtsu.edu/~smpte/table.html (retrieved on 12/03/2006).

[73] Kendall, Gary S. "A 3-D Sound Primer: Directional Hearing and Stereo Reproduction." *Computer Music Journal* 19:4 (1995), pp. 23–46.
URL: http://www.garykendall.net/papers/3-DPrimer1995.pdf.

[74] Kendall, Gary S. "The Decorrelation of Audio Signals and its Impact on Spatial Imagery." *Computer Music Journal* 19:4 (1995), pp. 71–87.
URL: http://www.garykendall.net/papers/Decorrelation1995.pdf.

[75] Kendall, Gary S. and Mauricio Ardila. "The Artistic Play of Spatial Organization: Spatial Attributes, Scene Analysis and Auditory Spatial Schemata." In proceedings: *Computer Music Modeling and Retrieval (CMMR): Sense of Sounds*, Copenhagen, Denmark 2007, pp. 125–138.
ISBN: 978-3-540-85034-2.
DOI: 10.1007/978-3-540-85035-9.
URL: http://www.garykendall.net/papers/KendallArdila2008.pdf.

[76] Kirkeby, Ole, Philip A. Nelson, and Hareo Hamada. "Local Sound Field Reproduction Using Two Closely Spaced Loudspeakers." *Journal of the Acoustical Society of America* 104:4 (1998), pp. 1973–1981.
DOI: 10.1121/1.423763.
URL: http://link.aip.org/link/?JAS/104/1973/1.

[77] Kistler, Doris J. and Frederic L. Wightman. "A Model of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-Phase Reconstruction." *Journal of the Acoustical Society of America* 91:3 (1992), pp. 1637–1647.
DOI: 10.1121/1.402444.
URL: http://link.aip.org/link/?JAS/91/1637/1.

[78] Klapholz, Jesse. "Fantasia: Innovations in Sound." *Journal of the Audio Engineering Society* 39:1/2 (1991), pp. 66–70.
URL: http://www.aes.org/e-lib/browse.cfm?elib=5994.

[79] Kramer, Lorr. "DTS: Brief History and Technical Overview."
URL: http://www.dts.com/media/uploads/pdfs/history,whitepapers,downloads.pdf (no longer available).

[80] Kuhn, George F. "Model for the Interaural Time Differences in the Azimuthal Plane." *Journal of the Acoustical Society of America* 62:1 (1977), pp. 157–167.
DOI: 10.1121/1.381498.
URL: http://link.aip.org/link/?JAS/62/157/1.

[81] Kurozumi, Kohichi and Kengo Ohgushi. "The Relationship Between the Cross-Correlation Coefficient of Two-Channel Acoustic Signals and Sound Image Quality." *Journal of the Acoustical Society of America* 74:6 (1983), pp. 1726–1733.
DOI: 10.1121/1.390281.
URL: http://link.aip.org/link/?JAS/74/1726/1.

[82] Laakso, Timo I. and Vesa Välimäki. "Energy-Based Effective Length of the Impulse Response of a Recursive Filter." *IEEE Transactions on Instrumentation and Measurement* 48:1 (1999), pp. 7–17.
DOI: 10.1109/19.755042.
URL: http://www.acoustics.hut.fi/~vpv/publications/icassp98-impl.pdf.

[83] Laakso, Timo I., Vesa Välimäki, Matti Karjalainen, and Unto K. Laine. "Splitting the Unit Delay: Tools for Fractional Delay Filter Design." *IEEE Signal Processing Magazine* 13:1 (1996), pp. 30–60.
DOI: 10.1109/79.482137.
URL: http://www.acoustics.hut.fi/software/fdtools/ (MATLAB tools).

[84] Litovsky, Ruth Y., H. Steven Colburn, William A. Yost, and Sandra J. Guzman. "The Precedence Effect." *Journal of the Acoustical Society of America* 106:4 (1999), pp. 1633–1654.
DOI: 10.1121/1.427914.
URL: http://link.aip.org/link/?JAS/106/1633/1.

[85] Maconie, Robin. *The Works of Karlheinz Stockhausen*. 2nd ed. Oxford and New York: Clarendon Press and Oxford University Press, 1990.
ISBN: 978-0-19-315477-3.

[86] Malham, David G. and Anthony Myatt. "3-D Sound Spatialization Using Ambisonic Techniques." *Computer Music Journal* 19:4 (1996), pp. 58–70.

[87] Manning, Peter. *Electronic and Computer Music*. Rev. ed. New York: Oxford University Press, 2004.
ISBN: 978-0-19-514484-0.

[88] Martens, William L. "The Impact of Decorrelated Low-Frequency Reproduction on Auditory Spatial Imagery: Are Two Subwoofers Better Than One?" In proceedings: *AES 16th International Conference: Spatial Sound Reproduction*, Rovaniemi, Finland 1999.
URL: http://www.aes.org/e-lib/browse.cfm?elib=8048.

[89] Martens, William L. "Perceptual Evaluation of Filters Controlling Source Direction: Customized and Generalized HRTFs for Binaural Synthesis." *Acoustical Science and Technology* 24:5 (2003), pp. 220–232.
DOI: 10.1250/ast.24.220.
URL: http://www.jstage.jst.go.jp/article/ast/24/5/24_220/_article.

[90] McCartney, James. "Rethinking the Computer Music Language: SuperCollider." *Computer Music Journal* 26:4 (2002), pp. 61–68.
DOI: 10.1162/014892602320991383.

[91] McLean, Anthony. "Digital Promises Realized: Boldly Going Where No One Has Been." *Live Sound International* (Jan./Feb. 2002).
URL: http://www.livesoundint.com/archives/2002/janfeb/promises/promises.php.

[92] Mehrgardt, S. and V. Mellert. "Transformation Characteristics of the External Human Ear." *Journal of the Acoustical Society of America* 61:6 (1977), pp. 1567–1576.
DOI: 10.1121/1.381470.
URL: http://link.aip.org/link/?JAS/61/1567/1.

[93] Merimaa, Juha. "Analysis, Synthesis, and Perception of Spatial Sound: Binaural Localization Modeling and Multichannel Loudspeaker Reproduction." PhD dissertation, Helsinki University of Technology, 2006.
URL: http://lib.tkk.fi/Diss/2006/isbn9512282917/.

[94] Meyer Sound Laboratories. "UPJ-1P Data Sheet." (Part no. 04.134.097.01 B). 2005.
URL: http://www.meyersound.com/pdf/products/ultraseries/upj-1p_ds.pdf.

[95] Meyer Sound Laboratories. "UPM-1P Data Sheet." (Part no. 04.084.004.01 C). 2005.
URL: http://www.meyersound.com/pdf/products/ultraseries/upm-1p_ds.pdf.

[96] Miles, Michael. "New Horizons: Live Concerts In Surround?" *Live Sound International* (Dec. 2003).
URL: http://www.livesoundint.com/archives/2003/dec/surround/surround.php.

[97] Miles, Michael. "New Horizons: Approaches to Live Surround." *Live Sound International* (Mar. 2004).
URL: http://www.livesoundint.com/archives/2004/march/horizons.pdf.

[98] Møller, Henrik. "Fundamentals of Binaural Technology." *Applied Acoustics* 36:3-4 (1992), pp. 171–218.
DOI: 10.1016/0003-682X(92)90046-U.
URL: http://www.sciencedirect.com/science/article/B6V1S-480V04C-5P/2/65a0163685c0317ad4b23d218dbda871.

[99] Mooney, James R. "Sound Diffusion Systems for the Live Performance of Electroacoustic Music." PhD dissertation, University of Sheffield, 2005.
URL: http://www.james-mooney.co.uk/research/sound_diffusion/JM_Sound_Diffusion_Systems.pdf.

[100] Moorer, James A. "About This Reverberation Business." *Computer Music Journal* 3:2 (1979), pp. 13–28.
URL: http://www.jamminpower.com:8080/jp/PDF/reverberation%20business.pdf.

[101] Morimoto, Masayuki, Kazuhiro Iida, and Motokuni Itoh. "Upper Hemisphere Sound Localization Using Head-Related Transfer Functions in the Median Plane and Interaural Differences." *Acoustical Science and Technology* 24:5 (2003), pp. 267–275.
DOI: 10.1250/ast.24.267.
URL: http://www.jstage.jst.go.jp/article/ast/24/5/24_267/_article.

[102] Nam, Juhan, Miriam A. Kolar, and Jonathan S. Abel. "On the Minimum-Phase Nature of Head-Related Transfer Functions." Presented at the AES 125th Convention, San Francisco, CA 2008.
URL: http://www.aes.org/e-lib/browse.cfm?elib=14698.

[103] Nelson, Philip A. and J. F. W. Rose. "Errors in Two-Point Sound Reproduction." *Journal of the Acoustical Society of America* 118:1 (2005), pp. 193–204.
DOI: 10.1121/1.1928787.
URL: http://link.aip.org/link/?JAS/118/193/1.

[104] Nelson, Philip A. and J. F. W. Rose. "The Time Domain Response of Some Systems for Sound Reproduction." *Journal of Sound and Vibration* 296 (2006), pp. 461–493.
DOI: 10.1016/j.jsv.2005.12.051.

[105] Oppenheim, Alan V., Ronald W. Schafer, and John R. Buck. *Discrete-Time Signal Processing*. 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1999.
ISBN: 978-0-13-754920-7.

[106] Oppenheim, Alan V., Alan S. Willsky, and Syed Hamid Nawab. *Signals and Systems*. 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1996.
ISBN: 978-0-13-814757-0.

[107] Owinski, Bobby. "Frequently Asked Questions about Surround Sound."
URL: http://www.surroundassociates.com/fqmain.html#8.1 (retrieved on 12/05/2006).

[108] Papoulis, Athanasios and S. Unnikrishna Pillai. *Probability, Random Variables and Stochastic Processes*. Boston: McGraw-Hill Higher Education, 2002.
ISBN: 978-0-07-122661-5.

[109] Parsons, Alan. "Four Sides of the Moon." *Studio Sound Magazine* (June 1975).
URL: http://www.stereosociety.com/FourSides.html.

[110] Peebles Peyton, Jr. *Probability, Random Variables, and Random Signal Principles*. 4th ed. New York: McGraw-Hill, 2000.
ISBN: 978-0-07-366007-3.

[111] Peeters, Geoffroy. "A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project." IRCAM technical report, v1.0 (4/23/2004).
URL: http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf.

[112] Pope, Stephen. "Research on Spatial and Surround Sound at CREATE." 2001.
URL: http://www.create.ucsb.edu/wp/SpatialSnd.2.pdf (retrieved on 12/05/2006).

[113] Pritchett, James. *The Music of John Cage*. Cambridge and New York: Cambridge University Press, 1993.
ISBN: 978-0-521-41621-4.

[114] Qiu, X. and M. Vorlaender. "Effects of Practical Loudspeaker Characteristics on Virtual Acoustic Imaging Systems." In proceedings: *Acoustics '08*, Paris 2008, pp. 1719–1724.
ISBN: 978-2-9521105-4-9.
URL: http://intellagence.eu.com/acoustics2008/acoustics2008/cd1/data/articles/001250.pdf.

[115] Rumsey, Francis. "Subjective Assessment of the Spatial Attributes of Reproduced Sound." In proceedings: *AES 15th International Conference: Audio, Acoustics and Small Spaces*, Copenhagen, Denmark 1998, pp. 122–135.
URL: http://www.aes.org/e-lib/browse.cfm?elib=8096.
URL: http://epubs.surrey.ac.uk/recording/44/.

[116] Rumsey, Francis. "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm." *Journal of the Audio Engineering Society* 50:9 (2002).
URL: http://www.aes.org/e-lib/browse.cfm?elib=11067.

[117] Sæbø, Asbørn. "Influence of Reflections on Crosstalk Cancelled Playback of Binaural Sound." PhD dissertation, Norwegian University of Science and Technology, 2001.
URL: http://urn.kb.se/resolve?urn=urn:nbn:no:ntnu:diva-2004.

[118] "SARC / Sonic Arts Research Centre / Queen's University Belfast."
URL: http://www.sarc.qub.ac.uk/main.php (retrieved on 04/17/2009).

[119] Scarpaci, Jacob William. "Creation of a System for Real Time Virtual Auditory Space and its Application to Dynamic Sound Localization." PhD dissertation, Boston University, 2006.
URL: http://www.bu.edu/dbin/binaural/pubs/Sca06.pdf.

[120] Scheiber, Dave. "Sound Recognition." *St. Petersburg Times* (Nov. 2, 2003).
URL: http://www.sptimes.com/2003/11/02/Floridian/Sound_recognition.shtml.

[121] Schroeder, Manfred Robert and Bishnu S. Atal. "Computer Simulation of Sound Transmission in Rooms." *IEEE International Convention Record* 11:7 (1963), pp. 150–155.

[122] Schroeder, Manfred Robert, D. Gottlob, and K. F. Siebrasse. "Comparative Study of European Concert Halls: Correlation of Subjective Preference with Geometric and Acoustic Parameters." *Journal of the Acoustical Society of America* 56:4 (1974), pp. 1195–1201.
DOI: 10.1121/1.1903408.
URL: http://link.aip.org/link/?JAS/56/1195/1.

[123] "SDDS Movies."
URL: http://www.sdds.com/news_movies.cfm (retrieved on 12/04/2006).

[124] Smyth, Mike and Stephen Smyth. "APT-X100: A Low-Delay, Low Bit-Rate, Sub-Band ADPCM Audio Coder for Broadcasting." In proceedings: *AES 10th International Conference: Images of Audio*, London 1991, pp. 41–56.
URL: http://www.aes.org/e-lib/browse.cfm?elib=5400.

[125] Sonic Arts Network. "Cut and Splice: Acousmonium." 2006.
URL: http://www.sonicartsnetwork.org/press/Cut_&_Splice_20%25c9ess_release.doc (retrieved on 12/05/2006).

[126] "Sound-on-Film." *Wikipedia*.
URL: http://en.wikipedia.org/wiki/Sound-on-film (retrieved on 12/04/2006).

[127] Stockhausen, Karlheinz. "Message from Stockhausen: LICHT-Timing."
URL: http://www.stockhausen.org/LICHT_timing.html (retrieved on 01/08/2007).

[128] Stockhausen, Karlheinz. *Spiral, für einen Solisten, Nr. 27.* (UE14957). Vienna: Universal Edition, 1973.
ISMN: 979-0-008-00866-5.
ISBN: 978-3-7024-1971-4.

[129] Stockhausen, Karlheinz. "Electroacoustic Performance Practice." *Perspectives of New Music* 34:1 (1996), pp. 74–105.
URL: http://www.jstor.org/stable/833486.

[130] "Supercollider Hub."
URL: http://supercollider.sourceforge.net/ (retrieved on 02/04/2007).

[131] Takeuchi, Takashi, Philip A. Nelson, and Hareo Hamada. "Robustness to Head Misalignment of Virtual Sound Imaging Systems." *Journal of the Acoustical Society of America* 109:3 (2001), pp. 958–971.
DOI: 10.1121/1.1349539.
URL: http://link.aip.org/link/?JAS/109/958/1.

[132] Todd, Craig C., G. A. Davidson, M. F. Davis, L. D. Fielder, B. D. Link, and S. Vernon. "AC-3: Flexible Perceptual Coding for Audio Transmission and Storage." Presented at the AES 96th Convention, Amsterdam, Netherlands 1994.
URL: http://www.aes.org/e-lib/browse.cfm?elib=6436.

[133] Treib, Marc. *Space Calculated in Seconds: The Philips Pavilion, Le Corbusier, Edgard Varèse*. Princeton, NJ: Princeton University Press, 1996.
ISBN: 978-0-691-02137-9.

[134] Truax, Barry. "Barry Truax Home Page."
URL: http://www.sfu.ca/~truax/ (retrieved on 12/04/2006).

[135] Truax, Barry. "Composition and Diffusion: Space in Sound in Space." *Organised Sound* 3:2 (1998), pp. 141–146.
DOI: 10.1017/S1355771898002076.

[136] University of Southern California. "Immersive Audio Lab Home Page."
URL: http://audiolab.usc.edu/.
URL: http://imsc.usc.edu/research/project/immersiveaudio/index.html.

[137] *U.S. Standard Atmosphere, 1976*. Washington, DC: U.S. Government Printing Office, 1976.
URL: http://hdl.handle.net/2060/19770009539.
URL: http://ntrs.nasa.gov/search.jsp (search document ID 19770009539).

[138] Wightman, Frederic L. and Doris J. Kistler. "The Dominant Role of Low-Frequency Interaural Time Differences in Sound Localization." *Journal of the Acoustical Society of America* 91:3 (1992), pp. 1648–1661.
DOI: 10.1121/1.402445.
URL: http://link.aip.org/link/?JAS/91/1648/1.

[139] Wilson, Kim. "Yes Live in DTS." *Audio Video Revolution* (Dec. 5, 2006).
URL: http://www.avrev.com/music/revs/yes/index.html (no longer available).

[140] Wilson, Kim. "Tomlinson Holman's Latest Experiment." *Audio Video Revolution* (Dec. 13, 2007).
URL: http://www.avrev.com/equip/tomholman/.

[141] Wörner, Karl Heinrich. *Stockhausen: Life and Work*. Rev. ed. Berkeley: University of California Press, 1973.
ISBN: 978-0-520-02143-3.

[142] Xenakis, Iannis. "Xenakis on Xenakis." *Perspectives of New Music* 25:1/2 (1987), pp. 16–63.
URL: http://www.jstor.org/stable/833091.

[143] Yabe, Hirooki, Mari Tervaniemi, Janne Sinkkonen, Minna Huotilainen, Risto J. Ilmoniemi, and Risto Näätänen. "Temporal Window of Integration of Auditory Information in the Human Brain." *Psychophysiology* 35:5 (1998), pp. 615–619.
DOI: 10.1017/S0048577298000183.
URL: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=28803&fulltextType=SC&fileId=S0048577298000183.

[144] Zouhar, Vit, Rainer Lorenz, Thomas Musil, Johannes Zmölnig, and Robert Höldrich. "Hearing Varèse's Poème Électronique Inside a Virtual Philips Pavilion." In proceedings: *11th International Conference on Auditory Display (ICAD)*, Limerick, Ireland 2005, pp. 247–252.
URL: http://www.idc.ul.ie/icad2005/downloads/f76.pdf.