Under-explored Dimensions in Spatial Sound

Michael Cohen∗
University of Aizu

Abstract

An introduction to spatial sound in the context of hypermedia, interactive multimedia, and virtual reality is presented. Basic principals of relevant physics and psychophysics are reviewed (ITDs: interaural time differences, IID: interaural intensity differences, and frequency-dependent attenuation capturable by transfer functions). Modeling of sources and sinks (listeners) elaborates such models to include such as intensity, radiation, distance attenuation & filtering, and reflections & reverberation. Display systems—headphones and headsets, loudspeakers, nearphones, stereo, home theater and other surround systems, discrete speaker systems, speaker arrays, WFS (wave field synthesis), and spatially immersive displays—are described. Distributed applications are surveyed, including stereotelephony, chat-spaces, and massively multiplayer online role-playing games (MMORPGs), with references to immersive virtual environments.

CR Categories: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities; H.5.2 [Information Interfaces and Presentation]: User Interfaces—Auditory (non-speech) feedback; H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing—Methodologies and techniques; H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing—Modeling; H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing—Signal analysis, synthesis, and processing; H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing—Systems

Keywords: ambient media, awareware, impulse response, narrowcasting, pervasive computing, transfer function, ubicomp (ubiquitous computing), virtual auditory display, wearware, whereware

Figure 1: Synthetic Spatial Sound, Source → Sink.
1 Introduction

On a stereo reproduction system, sound basically comes from only the left and right transducers, whether headphones or loudspeakers. Simple audio systems project only a one-dimensional arrangement of the (stereo captured or mixed) sources. In traditional sound reproduction, the apparent direction from which a sound emanates is typically controlled by panning, shifting the balance of the unmodified sound source between left and right channels [Everest and Streicher 1998]. But this technique yields images that are diffuse, and located only between the speakers.

Spatial sound involves technology that allows sound sources to have not only a left–right (sway) attribute (as in a conventional stereo mix), but up–down (heave) and back–forth (surge) qualities as well. It is related, but goes beyond, systems like quadrophonics and surround sound [Holman 2008]. Augmenting a sound system with spatial attributes opens new dimensions for audio; spatial sound is a rich audio analog of 3D graphics.

Spatial sound projects audio media into space by manipulating sound sources so that they assume virtual positions, mapping them from “zero-space” (the source channel) into multidimensional-space (the perceptual envelope around the listener). These virtual positions allow auditory localization, a listener’s psychological separation in space of the channels, via space-domain multiplexing.

By creating psychoacoustic effects with DSP, scientists are developing ways of generating this 3D sound imagery [Chowning 1977] [Blauert 1997] as “spatial sound.” Spatial sound takes theatre-in-the-round and turns it inside-out, embedding the listener in a soundscape, a landscape of sound.

2 Spatial Sound

A direct approach to spatial sound, termed “acoustic space synthesis,” distributes sources by physically locating loudspeakers in the place where each source is located, relative to the listener. These loudspeakers can be statically placed, or perhaps moved around by mechanical means. Alternatively, spatial cues can be directly captured by a dummy head, gimbaled with respect to a fixed speaker. But such implementations are awkwardly static and not portable. Therefore, the rest of this description concentrates on digital signal processing (DSP) synthesis of spatial cues.

A sound spatializer creates the impression that the sound is coming from different sources and different places, just like one would hear “in person.” Spatial hearing can be stimulated by assigning each “phantom source” a virtual position with respect to the listener and simulating the auditory positional cues. A display based on this technology exploits human ability to quickly and subconsciously localize sound sources.

Figure 2: Woodworth’s Formula for Interaural Time Delay (ITD): This model is a simplification: The actual radius of a typical adult head is nominally 8.75 cm, and the ears are not symmetrically astride the diameter, but slightly behind, at around 100° and 260°. The difference in the arrival time of a planar wavefront can be estimated as $\tau = r(\theta + \sin \theta)/C$, where $r$ is the assumed radius of the head, $\theta$ is the bearing of a far-field source, and $C$ is the speed of sound. Time difference cues are registered at starts and ends of sounds (onsets and offsets). Primarily based on low frequency content of a sound signal, ITD is usable for frequencies of wavelength up to twice the distance between ears, about 1700 Hz, (340 m/s)/(2 \times .1 m/cycle), above which phase extraction from an early-arriving signal pair is ambiguous.

2.1 Directionalization and Localization

Binaural (stereo) localization cues include interaural time (phase) delay (ITD), a function of the interaural distance, as illustrated by Figure 2; interaural intensity difference (IID), a consequence of the acoustic head shadow; and notches in the binaural spectral response. These perceptual cues are captured by head-related transfer functions (HRTFs), impulse responses measured for the head and pinna (outer ear) of human or dummy heads [Hartmann 1999]. For each spherical direction, a left–right pair of these transfer functions is measured and stored. Spatial sound can be implemented by driving input signals, sampled with an A/D, through these HRTFs in a convolution engine, creating psychoacoustic localization effects by introducing binaural spatial cues into an original monaural signal.

Thus, such a (hardware- or software-based) convolution engine implements finite impulse response (FIR, also known as tapped delay) digital audio filters whose output signals are presented to stereo loudspeakers or headphones. The device performs a mathematical convolution of arbitrary audio signals, such as voices, with the filter coefficients that “place” the signals within the perceptual 3-space of...
2.2 Spatial Reverberation

The implementation of spatial sound just described is called “dry”; it includes no notion of a virtual room, and hence no echoes. Spatial reverberation is a “wet” technique for simulating the acoustic information used by people listening to sounds in interior environments. A spatial reverberation system, as illustrated by Figure 3, creates an artificial ambient acoustic environment, by simulating echoes consistent with the placement of a source in a virtual room.

There are two classes of generated echoes: early reflections, which are discretely generated (delayed, with frequency-dependent amplitude), and late-field reverberation, which are continuous and statistically averaged. The early reflections are the particular echoes generated by the source, and the late field reverberation “tail” is the non-specific ambience of the listening environment.

2.2.1 Early Reflections

Representing the particular echoes generated by the source, the early reflections (echoes), off the floor, walls, and ceiling, provide indirect sound to the listener. The energy ratio of direct to indirect sound is an important distance cue.

A ray-tracing algorithm can be used to simulate the timing and direction of individual reflections, which are then spatialized, as if they were separate sources. (Each separately spatialized audio source, incident or reflected, requires its own DSP channel and transfer function.)

2.2.2 Late Field Reverberation

Late field reverberation reflects the ambience of the virtual auditorium or listening room. Reverberant implementations of spatial sound—as discussed or instantiated by employ a recursive, or IIR (infinite impulse response) section to yield dense global reverberation effects.

2.2.3 Combined Effect

A filter that combines early reflections with late field reverberation is sometimes called TDR, for tapped-delay-plus-recirculation. The simulation captures cues to perceived sound direction, sound distance, and room characteristics.

This combination of high-order recursive and non-recursive filters enables a spatial reverberation system to accept descriptions of such characteristics as room dimensions and wall absorption, as well as time-varying source and listener positions. Thus, given a monophonic sound source and a specification of source and listener motion in a model room, the spatial reverberator captures the sound field arriving at the model listener’s ear-drums.

2.3 Dynamic Responsiveness

Spatial sound and spatial reverberation systems have varying degrees of responsiveness. This hierarchy is reflected by Table 1, whose items are ordered in increasing responsiveness.

- static non-realtime (ordinary stereo mix)
- binaural recording
- dynamic non-realtime (enhanced stereo mix)
- static realtime (ordinary teleconference)
- dynamic realtime
- dynamic realtime with head-tracking

Table 1: Hierarchy of Spatial Sound Responsiveness.

Binaural recordings, through a dummy head, can provide convincing, immersive, auditory experience, albeit purely sampled, and not directly reconfigurable. The most primitive sort of synthetic, a static non-realtime system, must be programmed a priori, and the convolutions performed asynchronously. Such a system might be
used to mix music [Moore 1983], for instance, but would be somewhat inadequate for modeling a conversation of ambulatory (mobile or moving) participants. An improvement over such an asynchronous system would admit dynamic position vectors [Chowning 1977], allowing the virtual positions of the sources to change over time, creating a more salient positional percept.

Out-of-realtme calculations can be accelerated by special purpose hardware, such as a DSP or GPU. A system quick enough to do real-time convolutions allows synchronous applications, as might be required for a simple teleconferencing system. If such a system were further enhanced to allow the transfer functions to be updated “on-the-fly,” then the spatial parameters could vary along with the source signals. With such capability, dynamically selected filter parameters can yield the sensation of sound sources moving with respect to a sink, the listener.

2.4 Head Tracking

An important effect in stereophonic localization is the ability to “triangulate” the position of a sound source by moving one’s head back and forth. Since the virtual sources don’t actually exist, but are simulated by multiple channels of audio, the system should be aware of the position of the listener’s head in order to faithfully emulate the desired sound image. One way to do this is to track a user’s head, for instance by means of a sensor attached to the user’s head. A head-tracker, which could be attached to a set of stereo headphones or a head-mounted display, can strobe the orientation of the listener’s head, adjusting the selection of the HRTFs convolving the input signals. This feature is important for robust localization and soundscape stabilization [Thurlow and Runge 1967; Thurlow et al. 1967], since a source will stay stationary or stabilized in (user-relative) motion trajectory when a user turns her head, effectively triangulating its location. A sophisticated system synthesizes a 3D distribution of stationary or moving sound sources, whose acoustics reflect the user’s head orientation at each moment.

Besides the personal display systems illustrated by Figure 4, an ultimate consequence of which would be talking clothing, like that humorously imagined by Figure 6, speaker array systems allow multiple users to enjoy a shared environmental display. Wavefield (or wavefront) synthesis (WFS) systems use densely distributed DSP-driven speaker arrays to reconstruct, by Huygens’ Principle, complex wavefronts for which the localization of virtual sources does not depend upon or change with a listener’s position. An end-to-end outline of the entire source→sink virtual acoustics model is shown in Figure 1.

3 Fresh Directions: Wearware, Whereware, and Awareware

A panoramic potentiometer controls the placement of a channel in a conventional (left–right stereo) mix. By using an audio window system as a mixing board, a multidimensional pan pot, the user could set parameters reflecting these positions. A listener altering these parameters could experience the sensation of wandering around a conference room, among the conferees. Members of the audience at a (live or recorded) concert could actively focus on a particular channel by sonically hovering over the shoulder of a musician in a virtual concert hall. Minglers at a virtual party might freely circulate. Sound presented in this dynamically spatial fashion is as different from conventional mixes as sculpture is from painting. Entities in cyberspace may be imbued with sound qualities as flexibly as they are assigned visual attributes.

Spatial sound is mature enough [Begault 1994] [Carlile 1996] [Gilkey and Anderson 1997] [Jot 1999] [Rumsey 2001] [Begault 2004] [Rabenstein and Spors 2008] to start to strain at ordinary applications. Researchers have heretofore understandably focused on “low hanging fruit,” but the time is ripe for stretching the capabilities of spatial audio, exploring more difficult and exotic systems [Suzuki et al. 2010].

As network bandwidth allows increasingly polyphonic soundscapes, users might be overwhelmed with stimuli, as suggested by Figure 7. We have been working on a way to prioritize such potentially overwhelming sources with narrowcasting, a simplification of which is expressed by Figure 8. Narrowcasting [Fernando et al. 2009] [Alam et al. 2009a; Alam et al. 2009b], by way of analogy with broad- and multicastring, is an idiom for limiting media streams for privacy, both protecting secrets (“deafen” & “attend”) and preventing interruptions (“mute” & “select”). These narrow-casting attributes can be crossed with spatialization in a stereotelephonic system, and used for “polite calling,” or what can be thought of as “awareware,” reflecting a sensitivity to one’s availability.

Spatial sound can be used to enhance driving experiences [Villegas and Cohen 2010a]. As illustrated by Figure 9, localized audio sources can be aligned with real-world locations for increased driver situation awareness and safety.

Spatial sound can be used to enhance musical experiences. Speciality music publishers include spatial music in their catalog, and spatial music is an under-exploited capability of Mr3 [Breebaart and Faller 2007]. Interactive music browsing can also leverage such capability. As seen in Figure 10, we are using “Open Wonderland” as

2http://www.openwonderland.org
Figure 4: Binaural (Personal) Sound Displays.

Figure 5: 5.1 Home Theater Bass Management. 5.1 systems have five full-range channels—three in the front and two surround—and a 200 Hz low frequency effects (LFE) channel. Before the bass management system there is the LFE channel. After bass management, there is a subwoofer signal. The LFE isn’t exactly the “subwoofer channel,” since a bass management system may direct bass to one or more subwoofers (if present) from any channel, not just from the LFE.
The general expression of inclusive selection is

\[ \text{active}(x) = \neg \text{exclude}(x) \land (\exists y \text{include}(y) \Rightarrow \text{include}(x)). \]  

(1)

So, for \text{mute} and \text{select} (solo or cue), the relation is

\[ \text{active(source}_x) = \neg \text{mute(source}_x) \land (\exists y \text{select(source}_y) \Rightarrow \text{select(source}_x)), \]  

(2a)

\text{mute} explicitly turning off a source, and \text{select} disabling the collocated (same room/window) complement of the selection (in the spirit of “anything not mandatory is forbidden”). For \text{deafen} and \text{attend}, the relation is

\[ \text{active(sink}_x) = \neg \text{deafen(sink}_x) \land (\exists y \text{attend(sink}_y) \Rightarrow \text{attend(sink}_x)). \]  

(2b)

Figure 8: Simplified formalization of narrowcasting and selection functions in predicate calculus notation, where ‘\neg’ means “not,” ‘\land’ means conjunction (logical “and”), ‘\exists’ means “there exists,” and ‘\Rightarrow’ means “implies.” The suite of inclusion and exclusion narrowcast commands for sources and sinks are like analogs of burning and dodging (shading) in photographic processing. The duality between source and sink operations is tight, and the semantics are identical: an object is inclusively enabled by default unless, a) it explicitly excluded (with \text{mute} || \text{deafen}), or, b) peers are explicitly included (with \text{select} || \text{attend: confide or harken}) when the respective avatar is not. Narrowcasting attributes are not mutually exclusive, and the dimensions are orthogonal. Because a source or a sink is active by default, invoking \text{exclude} and \text{include} operations simultaneously on an object results in its being disabled. For instance, a sink might be first \text{attended}, perhaps as a member of some non-singleton subset of a space’s sinks, then later \text{deafened}, so that both attributes are simultaneously applied. (As audibility is assumed to be a revocable privilege, such a seemingly conflicted attribute state disables the respective sink, whose attention would be restored upon resetting its \text{deafen} flag.) Symmetrically, a source might be \text{selected} and then \text{muted}, akin to making a “short list” but relegated to backup.

Figure 9: “Back Seat Driver”: Localized Beacons and Directionalization for Vehicular Wayfinding—Spatial Sound Whereware
Collaborative virtual environments (CVEs), like Second Life, offer immersive experiential network interfaces to online worlds and media. The Smithsonian Institution, the national museum of the U.S.A., curates a world music archive, the “Folkways Collection,” and is engaged in a partnership with the folkwaysAlive! Project at the U. of Alberta (Edmonton, Canada), whose mandate is to develop the Folkways Collection in new ways. We have developed an immersive browser for this collection, “Folkways in Wonderland.” Avatar-represented users can navigate freely in this distributed groupware-enabled “Second Life”-like environment, auditioning recorded world music displayed as album art positioned near the origin of the music, and enjoying spatialized music and voice chat.

For instance, directionality is perhaps more important than range simulation, but a full 3D display should allow spatialization into the intimate “whisper space” near-field, as explored by an accompanying paper by Villegas & Cohen [2010b]. Standard 5.1 configurations, as often used in home theater (since they are supported by DVD and Blu-Ray), and which bass management is illustrated in Figure 5, don’t usually exploit elevational cues, but an accompanying paper by Jo, Martens, & Park [2010] considers how to simulate elevation, height cues. Finally, most models of spatial sound assume simple radiation patterns, but real sources are not uniformly directional. Maki, Kimura, & Katsumoto [2010] describe their experiments on articulated directivity with a special spherical loudspeaker with multiple transducers.

References


