Spatial soundscape superposition, Part II: Signals and systems

Michael Cohen¹,* and William L. Martens²,†

¹Spatial Media Group, Computer Arts Laboratory, University of Aizu, Tsuruga, Iki-machi, Aizu-Wakamatsu, 965–8580 Japan
²Discipline of Physiology, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, NSW 2006, Australia

Abstract: We present an analytical framework for a cognitively informed organization of signals involved in computational representations of spatial soundscape superposition, defined here as “procedural superposition,” building on the accompanying article Part I, where we discussed physical (acoustical) and perceptual (subjective and psychological) frameworks for soundscape representations in virtual auditory displays. Exploiting multimodal sensation and mental models of situations and environments, convention and idiom can tighten listeners’ apprehension of an auditory scene, using metaphor and relaxed expectation of sonorealism to enrich communication. Besides physical and psychological combinations, procedural (logical and cognitive) superposition considers metaphorical mappings between audio sources and virtual location, including such aspects as separation of visual and auditory perspectives; separation of direction and distance; parameterized binaural and spatial effects, including directionality; range-compression and indifference; layering of soundscapes; “audio windowing” (analogous to graphical user interface windows), narrowcasting, and multipresence as a strategies for managing privacy; and rotation as revolution. These auditory display strategies leverage virtual relaxations of sonorealism to enable enhanced soundscape representation.

Keywords: Cognitive models, Computational soundscapes, Metaphorical mixing, Nonsonorealistic rendering (NSR)

PACS number: 43.60.Dh, 43.60.Qv, 43.25.Lj, 43.38.Tj, 43.38.Vk, 43.60.Sx
[doi:10.1250/ast.41.297]

1. INTRODUCTION: STAGES OF COMPOSITION

Auditory displays are broadly and richly embedded in modern life. We are positively assailed by communication sounds, competing with each other for attention. Spatial soundscape superposition can be usefully characterized by where the combination actually occurs. As outlined by Table 1, superposition of spatial soundscapes occurs in various ways: physically, perceptually, and procedurally — sound, sensation, and signals, following [1]’s titular description of the auditory process (albeit in rotated order). Physical (acoustic) superposition describes such aspects as configuration of personal sound transducers, cross-fading among multiple sources, speaker arrays, and the modern challenge of how to integrate and exploit mobile devices and “smart speakers.” Perceptual (psychological) superposition describes such aspects as binaural image formation [2], subjective attributes of auditory objects and spaces, and multimodal sensory interpretation. Procedural (logical and cognitive) superposition describes higher-level relaxation of insistence upon literal auralization that leverages idiom and convention to enhance practical expressiveness.

Acoustic superposition (sound) is reviewed in a forthcoming chapter [3], and perceptual (sensation) considerations are reviewed in a companion paper [4]. This paper focuses on procedural (signals) signal-level considerations, metaphorical mixing of spatial soundscapes.

2. PROCEDURAL SUPERPOSITION (LOGICAL & COGNITIVE CONVENTIONS): SIGNALS AND SYSTEMS

We consider procedural models of signals that inform spatial soundscap superposition and cognitive apprehension, metaphorical associations with which listeners decode sound fields. When interacting with virtual displays, explicit models consciously reinterpret perceptual impressions of physical phenomena. In graphics, non-photorealistic rendering (NPR) describes deliberate distortion or remapping of imagery, for the purposes of art or information visualization, subsuming realism to some superseding
goal, such as visual interest or perspicuity, ease in appreciation or understanding. Similarly, auditory displays also admit such relaxation of literal, “sonorealistic” renderings (analogous to “photorealistic” graphical renderings). Shared assumptions, social conventions, and learned idioms compress communication expression. The following subsections describe some “NSR” (non-sonorealistic renderings), used to enhance or enable designation, in the semiotic sense of consensual understanding.

2.1. Separation of Visual and Auditory Perspectives: Displaced Display

Normally, personal audition and vision are thought of as concentric, the respective sensory organs embodied together as they are in one’s head. When pressed to identify a single center of identity and personal consciousness, most people would indicate their brain, conveniently nestled as it is behind the eyes and between the ears. However, telesensory instrumentation allows and encourages independence of modalities. For simple example, movies, video games, and TV shows present audiovisual scenes that resemble what one might plausibly see and hear if one were at the position of the camera and its assumedly coincident microphone.

For architectural walk- (or fly-)throughs and auralizations [5], visual and auditory perspectives should match, as if cameras and microphones were integrally deployed. For a concert, an auditory display might be “with perspective” (i.e., aligned with visual display), either directly (via coincident microphone capture) or coherently simulated. However, performed electroacoustic music can be captured by a variety of overhead, on-stage, and spot (accent) microphones, mixed down and distributed for monitoring (realtime self-audition by musicians), sound reinforcement (for live audience), and recording and/or transmission (for archive or distribution). Mediated concert experiences such as music videos separate visual and auditory perspectives, not insisting that capture, render, or simulation of aural perspective match optical position. Such conventions extend to spatial media, as cameras might be binocular, visual displays stereographic, microphones stereophonic, and auditory displays binaural.

Cinematic and gaming idioms also relax literal associations, freely exercising liberty to set aside assumptions of alignment of auditory and visual perspectives. For example, BGM (background music) is non-diegetic (conceptually outside a visual scene), accommodated by such independence. One auditionally attends multiple spaces at once, apprehending not only a story scene, but also, implicitly, its musical accompaniment. Displacement can reflect temporal offset as well as spatial. For instance, in sort of the same way that a panning camera leads a moving character by framing comfortably ahead, sound of a subsequent scene is often introduced before corresponding visuals.

A viewing audience’s or gamer’s perspective is privileged, enjoying not only extraordinary visual situation and arrangement (cinematography, including various lenses, cameras, and display formats; editing, including shot transitions, split-screening, montage, etc.), but also artificial auditory access, with flexible correspondence among display modalities. The 2nd-person perspective popular in RPGs (role-playing games) is characterized by such displacement, as the auditory perspective, through which one listens and speaks, is that of an associated avatar, not that of its tethered viewpoint. That is, the human gamer is projected into an avatar, typically viewed from slightly behind and above, through a loosely attached virtual camera. Likewise, projected location of sound associated with the player (generated by a game engine or voice-chat captured from the human pilot) is that of the avatar, not the lagging visual perspective.

2.2. Separation of Orientation and Location: Directionalization vs. Localization

“Localization” can describe both projection and ap-
prehension of spatial position—as by an audio renderer that compiles soundscapes, as well as by a human listener estimating position of auditory objects during audition. Practically, localization comprises directionalization plus distance effects or judgement. For subjective displays, parameterized explicitly or implicitly by standpoint and direction, polar or spherical coordinates are more convenient than rectilinear Cartesian coordinates since polar or spherical coordinates are non-homogeneous, in that range is dimensionally different from azimuth and elevation (length vs. direction). Representation or apprehension of distance (ρ) is different from that for horizontal (θ) or vertical (ϕ) direction, and is measured with different units (such as meters vs. radians) and can be decoupled.

Contemporary spatial sound systems handle three kinds of audio encodings:

channel-based, associated with fixed (“bed”) display configurations (headphones, stereo speakers, home theater layouts, theatrical arrangements, etc.), including matrix encodings,

scene-based, such as Ambisonics recordings and streams that capture sound fields at particular locations, and

object-based, associating streams with particular objects in a scene (human speakers, musical instruments, auditory events, etc.) and assuming that an audio renderer will directionalize or spatialize these tracks for a given display.

Directionality is usually more important than distance expression, but a fully featured 3D display should allow spatialization into the intimate “whisper space” near-field [6]. Standard 5.1 configurations, using channel-based encodings such as those deployed in home theater arrangements, do not usually exploit elevational cues. However, preconditioning signals with anatomical transfer functions before display through loudspeakers [7,8] can simulate height cues.

For object-based encodings, sound objects are most simply directionalized by intensity panning to loudspeakers near a phantom source, but amplitude- or gain-based techniques [9–12] cannot realistically convey spatial effects such as early reflections (echoes), modal resonances (standing waves), and late reverberation. Monaural audio streams can be directionalized or spatialized for binaural display with usual ITD (interaural time difference), IID (interaural intensity difference, or head shadowing, a.k.a. ILD, interaural level difference), and HRIR-based (or HRIR-based) filtering.

For scene-based encodings such as Ambisonics, each loudspeaker receives its own weighted sum of all channels, spatially sampling spherical harmonic coefficients. An Ambisonic microphone array captures a sound field and encodes a multichannel signal for flexible redirectionalization. Of the three affine transformations (scaling, rotation, translation), only rotation is modeled, so such soundfield recordings can be thought of as “prebaked,” forgoing flexibility (such as standpoint excursion or interaural baseline adjustability, which scales anatomical cues such as ITD and IID and changes binaural disparity) for optimized rendering.

2.3. Directionality Processing

Head motion is like “antenna pointing,” but is also “body language,” a kind of display. Situational context, voice intonation, facial expression, eye-gaze, and gesture all inform decoding of proxemic cues. Eye-gaze, which can be approximated from head orientation, is used for social signaling and can trigger computer-mediated events. Individually apprehended spatial sound tells the eyes where to look, but “gaze direction” (understanding where someone else is looking), awareness of directed or projected visual attention, alerts conversants about objects of regard.

Mouth-emitted sounds are non-isotropic, and speech is directional. In general, a speaker might not directly face a listener, and projected direction of speech is an important social signal. As illustrated by Fig. 1, a listener estimates not only direction but also orientation of a speaker. Using hints such as ratio of direct:indirect energy and darkening (via low-pass filtering) of utterances, listeners recognize which way a speaker is facing or talking, inferring targets of directed address [13]. Symmetrically, talkers are aware of orientation of listeners, and may modulate their voices according to appreciation of the listening difficulty of those facing away (akin to the Lombard effect, in which talkers strengthen vocalizations in the presence of ambient noise).

An aware interface detects user receptivity and automatically adjusts displays of all types to complement user state by modulating the environment. An aware renderer is parameterized not only by direction but also orientation of...
sources relative to sinks, modulating delivered audio streams to convey such fine cues. (A “sink” is the dual of a source, used instead of “listener” to distinguish it from an actual human, including allowing designation of multiple sinks for a single user, as explained in Sect. 2.8 below.) Smart speakers should be aware speakers.

Source and sink directivity can both be modeled by emulating idealizations of microphone receptivity patterns, combinations of omnidirectional (unipolar, \( \Theta \)) and directional (dipolar, figure-of-eight, \( \Theta \Theta \)) radiation as well as sensitivity \[1\]. For representative instance, the Google VR Audio and Resonance Audio Unity plug-ins model directionality by “alpha” \((0 \leq \alpha \leq 1)\) and “sharpness” \((1 \leq \text{sharpness})\). Normalized gain profiles are calculated as \(|(1 - \alpha + \alpha \cos(\theta))^{\text{sharpness}}|\), where \(\theta\) is the relative direction of (for projection or emission) a sink with respect to a source or (for reception or sensitivity) a source w.r.t. a sink, bilinear dipole weighting coefficient \(\alpha\) scales directionality, dipole power \(\text{sharpness}\) exaggerates such non-isotropy, and the absolute value function rectifies polarity inversion \((-\Theta)\). When \(\alpha\) is zero, the pattern is isotropic (and the sharpness is irrelevant); as \(\alpha\) approaches unity, directivity becomes increasingly lobed. “Earshot,” combined radia-
tion and reception, is the product of these for each source \(\rightarrow\) sink combination.

### 2.4. Nonrealistic Range-based Attenuation

Just as with computer graphics, it is common to introduce both approximate and more complicated models for sound propagation (diffusion, reflection, reverberation, refraction & diffraction in presence of obstacles or occluders, dispersion, absorption and scattering) to realize both improved performance and also expressive control. Intensity of a point source spherically radiating sound waves naturally observes a inverse square relation with distance, because bounding volumes in 3-space are 2D hulls, so amplitude gain, a root power quantity proportional to RMS pressure and square root of intensity, has a reciprocal relation with range. Localization of sound sources becomes even sharper if volume control is driven by models that roll off more rapidly than this physical \(1/\rho\) law (where \(\rho\) is the distance between source and sink). In contrast, it is sometimes assumed that in small spaces the amplitude of a reverberant signal changes little with range, and that in large spaces it is roughly proportional to \(1/\sqrt{\rho}\) \[10\].

Excepting extreme circumstances in spatial sound teleconferencing, such as when a virtual source approaches antipodal position, on opposite side of the planet, geo-tagged sources can be rendered basically horizontally, but with elevation — ignoring spherical curvature of the earth, but allowing relative altitude effects such as mountains and valleys. Rendering distance effects literally is not always useful. For many applications, such as conferencing and navigation, it is convenient to separate direction and range, rendering the former faithfully but the later metaphorically or not at all.

In everyday experience, even very loud sources are rarely audible beyond a few kilometers, and conversational intensities are normally inaudible beyond 10s of meters. With the usual \(-6\,\text{dB/range doubling attenuation},\) intensity of a typical conversational human speaker, measuring, say, 60 dB SPL at 1 m, weakens a millionfold at 1 km to 0 dB, a nominal threshold of audibility, and practical inaudibility occurs even closer because of background noise. Fortunately, utilities for way-finding (such as Microsoft Soundscape), direction-giving, and conferencing needn’t render sonorealistc cues.

Besides intensity-controlled loudness, other cues to simulate or suggest distance can be separately modulated \[15\], including initial time-gap delay, the interval between a direct sound and its first reflection; the ratio of the energy of direct sound to that of reverberation; motion parallax, subjective shift of a source when the head is moved; and high-frequency attenuation. Nature, including air, is a low-pass filter, and receding sources naturally manifest darkening, thinning of higher frequency components.

Relatedly, a rendering engine might perform “spotlight mixing,” exaggerating loudness of frontal objects assumed to be of interest. Alternatively, as frontal objects can be assumed to be visible and therefore already conspicuous, rearward objects might be particularly amplified \[16\], or their auditory position or timbre animated to “catch one’s ear.” Such “gaze mixing” or “attention focus” is a sensory substitution kind of multimodal coordination, acoustically highlighting objects, which also includes “audio haptics,” reactive sounds for synaesthetically touchless interaction, compensating for lack of force-feedback in virtual displays.

### 2.5. Extreme Range Compression: Location-indiffer-
ent Intensity

Dynamic range is the ratio of the power of the loudest and quietest parts of a signal, and range in the sense of source \(\rightarrow\) sink distance can be used to attenuate intensity, distance fall-off. In the limit, compression of dynamic range associated with distance-dependent attenuation approaches range-insensitivity, by raising the floor of the transfer function describing range-related attenuation. Separation of orientation and location, including distance-independent, allows directionalization without localization, as suggested by Fig. 2. In spatial user interfaces, compass bearings such as “North” are purely directional (like computer graphics directional lights, as opposed to area-, point-, or spot-lights), but even grounded objects with specific locations (such as one’s home or office) or

---

*Acoust. Sci. & Tech. 41, 1 (2020)*
characters (such as avatars or icons representing conversants) can project as range-indifferent sources, by normalizing or compressing range-dependent intensities. Sound spatialization can preserve direction but collapse distance.

Affordable systems for projecting immersive photospherical or volumetric visual displays and stereophonic auditory displays represent democratization of VR-style interfaces. Google Cardboard, the Merge AR/VR Headset, Samsung (Galaxy) Gear VR, and Oculus Quest use intrinsic sensors and monitors for head-tracked binocular display of stereoscopic contents and stereophonic display of spatial audio. Orientation can be tracked by a MEMS (microelectromechanical system) IMU (inertial motion unit) — including gyroscope, accelerometer, and magnetometer — estimating bearings via aggregating sensor fusion, but if location is not tracked, virtual standpoint is not adjusted.

Such scene-based interfaces ignore location and use only orientation. Spatial sound sources can therefore be directionalyzed without range-parameterized gain modulation. The subjective soundscape is counter-rotated by headpose, panned to stabilize a scene, but not perturbed. Rotation sensitivity supports location-based captured sound fields such as those handled by Ambisonics surround sound. Ambisons features independent coding and decoding, so, for example, capture or encoding into B-format (with 4 channels) is easily rendered at runtime, downmixed to a panned stereo pair heard through head-tracked headphones or up-mixed to a speaker array.

2.6. Layering and Audio Windowing

Procedural mixing allows user interfaces to combine and distribute audio signals. Cinema and electronic gaming encourage richly textured soundscapes, including music, SFX (sound effects), narration and dialog channels. Room effects such as echo and reverb can be added by ambience processors.

Graphical compositing, à la Photoshop-style layers, allows various blending modes, articulated effects applied at each phase of the “bit bucket brigade,” a chain of filters like a sequence of guitar effects pedals or a composition of digital effects to enrich expression. (SFX are audio samples or streams; digital effects are processes applied to such streams.) Such a cascade is equivalent to a tree of metamixers [17], a data-flow arrangement in which compositing operations are modeled as crossbar routing matrix switches with effects at each crosspoint: “programmable shaders” finally fanning-out into loudspeaker amplifiers.

Audio windowing [18], in analogy to graphically windowing user interfaces (and not to be confused with signal processing data sequence extraction), treats soundscapes as articulated elements in a composite display [19]. Spatial soundscapes, like layers in graphical applications or tracks in musical compositions, can be combined simply by summing, although in practice some scaling (amplification or attenuation), equalization, and other conditioning can yield better results. For instance, interior soundscapes might be reverberated, to distinguish them from outdoor scenes. To make a composited soundscape manageable, some sources might be situationally muted or muzzled and some sinks might be deafened or muffled.

2.6.1. Ducking and automatic rearrangement

Ducking is used by DJs to attenuate music during voice-over, including side-chain compression. Important “phone calls” (realtime voice-chats) can soften less critical channels, and driver advisories in vehicular navigation systems can lower music intensity. More complicated interaction is possible. For instance, a voice call might not only reduce intensity of music, but also change its directionization, moving it from a frontal stage-like position to a less conspicuous rearward location.

2.6.2. Diminished reality: sensor censor

Contemporary personal audition systems also include virtual reality (VR) and augmented reality (AR) auditory displays, usually featuring HMDs (head-mounted displays). As shown in by Fig. 3, VR and AR are generally considered mixed reality (MR) [23], so the abstraction of all of these in current parlance is XR, where the ‘X’ stands not only for “eXtended” but also for “Augmented,” “Mixed,” and “Virtual” (as at the end of this subsection’s title). Mixed reality can not only add information to naturally captured scenes, but can also remove information, so interpretation of “XR” can include “diminished reality” (extending the initialized elaboration). A cinemagraph, or “living picture,”
is an interpolation between static and dynamic displays, modelable as applying a transparency–opacity map to a still image overlaid upon a coincident video. Similarly, diminished audio reality can be thought of as hiding or masking otherwise apparent auditory scene components, such as engine sounds (as in active noise cancellation), objectionable ambient “room tone,” or an unwelcome voice (such as that of a boring interloper). Such “unmixing” suppression of particular sources is the opposite of the “cocktail party effect” [24], whereby particular objects are “heard out” of a cacophonous mix. The are generalized together by auditory source separation, auditory scene analysis [25], and blind source separation (BSS).

2.7. Narrowcasting: Privacy Management

“Privacy” has two interpretations. The first association is that of avoiding “leaks” of confidential information, protecting secrets. The second association is “freedom from disturbance,” in the sense of not being bothered. Narrowcasting operations manage privacy in both senses, filtering duplex information through an articulated communication model. In analogy to any-, broad-, multi-, and uni-casting, narrowcasting is an idiom for limiting and focusing media streams. Sources and sinks are symmetric duals in virtual spaces, respectively representing emitters and collectors. A human user might be represented by both a source and a sink in a groupware environment, or perhaps by multiple instances of such delegates, and both one’s own and others’ sources and sinks can be adjusted for privacy. Sound sources can be explicitly “turned off” by muting, or implicitly ignored by selecting some others. Similarly, audibility of a soundscape is controlled by embedded sinks, which can be explicitly deafened or implicitly desensitized if other sinks are “attended” [26].

Formalized by the permission scheme expressions shown in Fig. 4, narrowcasting [27,28] controls exposure and distributes attention. Advanced floor control symbolology — for chat-spaces, concerts, and conferences — is outlined by Table 2. Modulation of source exposure or sink attention [29] needn’t be “all or nothing”: nimbus (projection) and focus (receptivity) can be respectively partially softened with muzzling and muffling [30]. Nuanced operations can soften state transition, allowing non-binary control — not just on–off but intermediate

The general expression of inclusive selection is

\[
\text{active}(x) = \neg\text{exclude}(x) \land ((\exists y (\text{include}(y) \land (\text{self}(x) \equiv \text{self}(y)))) \Rightarrow \text{include}(x)).
\] (1)

Therefore, for mute and solo (or select), the source relation is

\[
\text{active(source}_x) = \neg\text{mute}(\text{source}_x) \land ((\exists y (\text{select}(y) \land (\text{self}(x) \equiv \text{self}(y)))) \Rightarrow \text{select}(\text{source}_x)), \quad (2a)
\]

mute explicitly turning off a source, and solo disabling the collocated complement of the selection (in the spirit of “anything not mandatory is forbidden”). For deafen and attend, the sink relation is

\[
\text{active(sink}_x) = \neg\text{deafen}(\text{sink}_x) \land ((\exists y (\text{attend}(y) \land (\text{self}(x) \equiv \text{self}(y)))) \Rightarrow \text{attend}(\text{sink}_x)). \quad (2b)
\]

Fig. 4 Formalization of narrowcasting and selection functions in predicate calculus notation, where ‘\(\neg\)’ means “not,” ‘\(\land\)’ means conjunction (logical “and”), ‘\(\exists\)’ means “there exists,” ‘\(\Rightarrow\)’ means “implies,” and ‘\(\equiv\)’ means “is equal to” (mutual implication, “if and only if”). Inclusion and exclusion narrowcasting commands are like analogs of burning and dodging (shading) in photographic processing. Duality between source and sink operations is strong, and the semantics are analogous: an auditory object is inclusively enabled by default unless, a) it is explicitly excluded with mute (for sources) or deafened (for sinks), or, b) peers in the same self/non-self class are explicitly included with solo/select (for sources) or attend (for sinks) when the object is not.
gains as well — and also signal-processing filter cascades at each opportunity. Narrowcasting attributes can be integrated with spatialization and used for “polite calling” or “awareware,” reflecting sensitivity to one’s availability, like the “online–offline” status switch of a conferencing service.

### 2.8. Multipresence and “Anyware”

Ordinary correspondence between inhabited bodily apprehension and consciousness is one-to-one, but telexistence [31] can soften such rigidly focused subjectivity, relaxing the singularity of human experience. A telephone, for simple example, exemplifies auditory telepresence,
projecting conversants to other places besides their corporeal “meat-space” base.

Multitasking users want to have presence in several locations at once. Enriched user interfaces, especially with position-tracking systems, encourage multipresence, the inhabiting by sources and sinks of multiple spaces simultaneously, allowing a human user to designate doppelgänger delegates in distributed domains. Exocentric interfaces supporting “out-of-body” experience enable parallel spaces, across which one designates multiple instances of “self,” self-identified avatars [32,33], as illustrated by Fig. 5. “Anyware” multipresence models separate but combinable scenes, allowing users to enjoy selectively distributed attendance.

Direct superposability of soundscapes makes audition especially open to multipresence (unlike, for particular instance, vision, which can’t easily overlay separate scenes). A soundscape interpreter can resolve source → sink correspondences, directionalizing, localizing, or spatializing each source to its best sink, a function of respective and mutual direction and orientation, directionality, and range. The apparent paradoxes of auditory multipresence can be resolved by an “autofocus” technique that uses Helmholtz reciprocity (exchangeability of sources and sinks) and simulated precedence effect (perceptual fusion) to disambiguate soundscapes [35], like a “snap-to” grid. Auditory interfaces can simulate “superhearing” including hyperacuity: hyperselectivity as well as hypersensitivity.

2.9. Rotation as Revolution

Even more aggressively forced relaxations of sonorealism are useful. For example, rotation can be remapped as revolution. The width of many sound-producing objects is too narrow to allow reliable auditory perception of their turning. Binaural disparity of an object’s rotation is limited by apparent girth of the object, which is partly determined by the distance at which it is inspected. As seen in Fig. 6, a musical tuning fork has directionality [36,37]. The variation can be heard by twisting the tines close to one’s ear, but is not otherwise audible.

However, reprojecting central twisting as peripheral orbit, with sources revolving in an “inspection gesture” instead of rotating, can auditorily express orientation. By attaching to an object under subjective inspection a virtual strut holding a virtual source, as shown in Fig. 7, rotation can be displayed as orbiting revolution [38], making obvious an otherwise subtle aspect, a kind of sonification.
Experience is not only realtime, but an accumulation of history, and constancy of reality suggests continuity of existence: objects are not intermittent, but persistent, and normal physical environments ensure that ordinary events can be perceived multimodally. Spatial sound cues are aspects of a rich ecology of environment-embedded signs.

Communication culture is not innate but learned. Listening isn’t a one-off event, but continuous experience. Sound displays use acquired associations, metaphorical signs, rather than direct emulation of natural phenomena. Cultural background admits an awareness of expectation, an acceptance of plausible but nonveridical cues, an assumed sophistication of listeners to decode nonliteral displays.

Many situations do not call for an auralization-style recreation of a particular soundscape but some kind of metaphorical space. Practical auditory conventions, short-hand expressions, and idioms such as those described by this survey refine expression. By using an audio windowing system as a mixing board, a multidimensional pan-pot, users and applications determine rich parameters to compile source and sink positions and their environments, rendering as a distributed diffuser (spatial sound stager or spatializer). Presence is more important than fidelity, audiophilic predilection for “absolute sound” or perceived need for Master Quality Authenticated streaming notwithstanding.

Purely auditory displays hardly exist. Almost always “in the wild” appear visual cues that complement projected aural scenes. Soundscape are not apprehended “in a vacuum”: some map, conventional understanding, or at least situational awareness aids decoding. Multimodal interfaces flatter overlapping displays. Media device orchestration [39] uses ad hoc arrays of devices to deliver or augment media experience. Articulation and comodulation of display parameters can coordinate audio and visual displays to accommodate attention, mood, and circumstances. Confederation of information appliances can enhance awareness, expressiveness, and experience.

Cognitive processes can resolve otherwise confusing soundscapes. For instance, moving lips or a flashing light (as on active “smart speaker”) can disambiguate conflicting cues. Listeners are inclined to be forgiving, suspending not only disbelief but also insistence on natural fidelity, so sonic situations can be efficiently communicated. Exploiting multimodal sensation and mental models of situations and environments, convention and idiom can tighten apprehension of a scene, using metaphor and relaxed expectation of sonorealism to enrich communication.

REFERENCES

[1] W. M. Hartmann, Signals, Sound, and Sensation (AIP Press, New York, 1999).
[2] E. M. Wenzel, D. R. Begault and M. Godfroy-Cooper, “Perception of spatial sound,” in Immersive Sound: The Art and Science of Binaural and Multi-channel Audio, A. Roginska and P. Geluso, Eds. (Routledge, Focal Press, Taylor & Francis, New York, 2018), Chap. 1, pp. 5–39.
[3] M. Cohen and W. L. Martens, “Spatial soundscape superposition and multimodal interaction,” in The Technology of Binaural Understanding, J. Blauert and J. Braasch, Eds. (Springer, Berlin, Heidelberg, 2019), Chap. 13, in press.
[4] W. L. Martens and M. Cohen, “Spatial soundscape superposition, Part I: Subject motion and scene sensibility,” Acoust. Sci. & Tech., 41, 288–296 (2020).
[5] M. Kleiner, B.-I. Dalenbäck and P. Svensson, “Auralization—
An overview,” *J. Aud. Eng. Soc.*, 41, 861–875 (1993).

[6] J. Villegas and M. Cohen, “frr: Modulating range in headphone-reproduced spatial audio,” VRCAI: Proc. Int. Conf. Virtual-Reality Continuum and Its Applications in Industry, Seoul, December (2010).

[7] H. Jo, W. L. Martens, Y. Park and S. Kim, “Confirming the perception of virtual source elevation effects created using 5.1 channel surround sound playback,” VRCAI: Proc. Int. Conf. Virtual-Reality Continuum and Its Applications in Industry, pp. 103–110, Seoul, December, ACM (2010).

[8] K. Tanno, A. Saji and J. Huang, “A 3D sound generation system with horizontally arranged five-channel loudspeakers,” *IEICE Trans. Inf. Syst.*, J97-D, 1044–1052 (2014).

[9] T. Lossius, P. Baltazar and T. de la Hogue, “Dhap—Distance-based-amplitude panning,” *ICMC: Proc. Int. Computer Music Conf.*, August (2009).

[10] V. Pulkkki, T. Lokki and D. Rocchesso, “Spatial effects,” in DAFX: Digital Audio Effects, 2nd ed., U. Zölzer, Ed. (John Wiley & Sons, Ltd., West Sussex, U.K., 2011), Chap. 5, pp. 139–184.

[11] V. Pulkkki, “Virtual source positioning using vector base amplitude panning,” *J. Aud. Eng. Soc.*, 45, 456–466 (1997).

[12] Z. Seldess, “MIAF: Manifold-interface amplitude panning in max/msp and pure data,” *Audio Eng. Soc. Conv. 137*, Los Angeles, October (2014).

[13] C. Rascon and I. Meza, “Localization of sound sources in robotics: A review,” *Rob. Auton. Syst.*, 96, 184–210 (2017).

[14] C. Hugonnet and P. Walder, *Spatial Effects*,” ’in Presence: Teleoperators Virtual Environ., 4, 364–386 (1995).

[15] V. Pulkki, T. Kokki and D. Rocchesso, “Spatial effects,” in DAFX: Digital Audio Effects, 2nd ed., U. Zölzer, Ed. (John Wiley & Sons, Ltd., West Sussex, U.K., 2011), Chap. 5, pp. 139–184.

[16] M. Cohen, “Throwing, pitching, and catching sound: Audio windowing models and modes,” *IJMMS: J. Pers.-Comput. Interact.*, 39, 269–304 (1993).

[17] S. Tachi, *Teleexistence*, 2nd ed. (World Scientific Publishing Company, Singapore, 2015).

[18] M. Cohen, “Quantity of presence: Beyond person, number, and pronouns,” in *Cyberworlds*, T. L. Kunii and A. Luciani, Eds. (Springer-Verlag, Tokyo, 1998), Chap. 19, pp. 289–308.

[19] R. Ranaweera, M. Cohen and M. Frishkopf, “Narrowcasting and multipresence for music auditioning and conferencing in social cyberworlds,” *Presence: Teleoperators Virtual Environ.*, 24, 220–242 (2015).

[20] M. Cohen and H. Kojima, “Multipresence and autofocus for interpreted narrowing,” *AES Int. Conf. Spatial Reproduction—Aesthetics and Science*, Tokyo, August (2018).

[21] M. Cohen and O. N. N. Fernando, “Awareware: Narrowcasting attributes for selective attention, privacy, and multipresence,” in *Awareness Systems: Advances in Theory, Methodology and Design*, P. Markopoulos and W. Mackay, Eds. (Springer-Verlag, London, 2009), Chap. 11, pp. 259–289.

[22] D. A. Russell, “On the sound field radiated by a tuning fork,” *Am. J. Phys.*, 68, 1139–1145 (2000).

[23] E. J. Heller, *Why You Hear What You Hear* (Princeton University Press, Princeton, 2013).

[24] M. Cohen and I. Jayasingha, “Auditory and haptic disambiguation of browsing models: Turnornamis and virtual viewpoints,” in *HAID: Int. Workshop Haptic and Auditory Interaction Design*, I. Oakley and S. A. Brewster, Eds., Seoul, November, Springer Lecture Notes in Computer Science 4813 (2007).

[25] J. Francombe, J. Woodcock, R. J. Hughes, R. Mason, A. Franck, C. Pike, T. Brooks, W. J. Davies, P. J. B. Jackson, T. J. Cox, F. M. Fazi and A. Hilton, “Qualitative evaluation of media device orchestration for immersive spatial audio reproduction,” *J. Aud. Eng. Soc.*, 66, 414–429 (2018).
Michael Cohen is a Prof. at the University of Aizu in Japan, where he heads the Spatial Media Group in the Computer Arts Lab, teaching computer music, game design, 3D modeling and printing, and audio interfaces. He received an Sc.B. in Electrical Engineering from Brown University in 1980, M.S. in Computer Science from the University of Washington in 1988, and Ph.D. in EECS from Northwestern University in 1991. He is a member of the VRSJ and ACM, and a Senior Member of the IEEE.

William L. Martens is a perceptual psychologist specializing in spatial hearing research and the simulation of the acoustical cues used in human sound localization. Currently an Assoc. Prof. in Audio and Acoustics at the University of Sydney, he has contributed to development of commercial spatial sound processing technologies, such as the 3D Positional Audio software for controlling the EMU-8000 DSP chip on the AWE-32 SoundBlaster sound card for Creative Labs (U.S. patent awarded in 1999).