Spatial Soundscape Superposition and Multimodal Interaction

Michael Cohen and William L. Martens

Abstract Contemporary listeners are exposed to overlaid cacophonies of sonic sources, both intentional and incidental. Such soundscape superposition can be usefully characterized by where such combination actually occurs: in the air, at the ears of listeners, in the auditory imagery subjectively evoked by such events, or in whatever audio equipment is used to mix, transmit, and display such signals. This chapter regards superposition of spatial soundscapes: physically, perceptually, and procedurally. *Physical* (acoustic) superposition describes such aspects as configuration of personal sound transducers, panning among multiple sources, speaker arrays, and the modern challenge of how to integrate and exploit mobile devices and "smart speakers." *Perceptual* (subjective and psychological) superposition describes such aspects as binaural image formation, auditory objects and spaces, and multimodal sensory interpretation. *Procedural* (logical and cognitive) superposition describes higher-level relaxation of insistence upon literal auralization, leveraging idiom and convention to enhance practical expressiveness, metaphorical mappings between real objects and virtual position such as separation of direction and distance; range -compression and -indifference; layering of soundscapes; audio windowing, narrowcasting, and multipresence as strategies for managing privacy; and mixed reality deployments.

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

M. Cohen (✉)
Spatial Media Group, Computer Arts Laboratory, University of Aizu,
Aizu-Wakamatsu, Fukushima 965-8580, Japan
e-mail: mcohen@u-aizu.ac.jp

W. L. Martens
Discipline of Physiology, School of Medical Sciences, Faculty of Medicine and Health,
University of Sydney, Sydney, NSW 2006, Australia

J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_13

Table 1 Spatial soundscape superposition

Stage	Domain	Realm	Practice	Considerations
Sound	Acoustics	*Physics*	*Transmission*: air mixing plus bone-conduction	personal and public transducers, panning, speaker arrays, smart speakers, mobile-ambient interfaces
Transduction	Biophysics, biochemistry	Physiology	Cochlear implants	Critical bands and ERBs, auditory or loudness recruitment
Sensation	Psychology, psychoacoustics	*Perception*: sensorineural processes	*Apprehension*: subjective composition, vection	Auditory objects, binaural imagery, multimodal stimulation
Signals	Cognition	*Procedure*: central auditory process	*Interpretation*: logical convention, metaphorical mapping and mixing, culture and semiotics	Parameterized directionalization and spatialization, layering and audio windowing, mental models, practical interpretation

actually occurs. As anticipated by Table 1, this chapter regards superposition of spatial soundscapes: physically, perceptually, and procedurally—sound, sensation, and signal, following Hartmann's titular description (1999) of the auditory process (albeit in rotated order).

2 Physical Superposition (Air Mixing): Sound

Normal circumstances combine sound in air, as when ordinary sources such as voices naturally add. The air acts as a linear mixer, superposing respective pressure disturbances. Modern instances of such superposition involve electroacoustics, using speakers to display some organized diffusion, such as sound distribution and panning. Physical combination leverages installed speakers as well as mobile devices such as cell phones, laptop computers, and smart speakers. Stereo speakers, sound bars, and home theater systems are common installations. In environments such as automobiles, ordinary loudspeakers can be replaced with novel actuator systems, such as distributed mode actuators (DMAs), distributed mode loudspeakers (DMLs), and multiactuator panels (MAPs). For example, the Ac2ated Sound (https://continental-automotive.com/en-gl/Passenger-Cars/Information-Management/Multimedia-Systems/Ac2ated-Sound) system attaches transducing drivers to car interior elements, using the pillars and dashboard for high- and mid-frequency reproduction, and large components—such as the ceiling, back covers of seats, and rear shelf—for low frequencies. More specialized spaces have super-directional (sound beam) loudspeakers and phased arrays. Some speakers, such as the Nexo CDD (configurable directivity device) (https://nexo-sa.com/systems/geo-s12/technology/), even feature adjustable directivity.

Directly connecting speakers to microphones, as in simple channel-based telepresence installations—dating back to the Théâtrophone exhibited at the Paris Electrical Exhibition in 1881, and formalized by Alan Blumlein (https://en.wikipedia.org/wiki/Alan_Blumlein) in the 1930s—can recreate sound fields. More active manipulations process audio streams by time delay and filtering, which can be implemented in digital signal processing (DSP) systems via recursive delay-and-add networks, or as time-domain convolution or equivalent frequency-domain multiplication.

2.1 Speaker Arrays

Besides theatrical spatial sound systems and the *sui generis* Audium (http://www.audium.org)—shown in Fig. 1—various institutions maintain polyphonic media art centers and concert halls, including such high-density loudspeaker arrays (HDLAs) (Lyon 2016, 2017) as the AlloSphere in Santa Barbara, California (http://www.allosphere.ucsb.edu), the BEAST (Birmingham ElectroAcoustic Sound Theatre) (http://www.beast.bham.ac.uk) project (Birmingham, UK), CCRMA at Stanford University (https://ccrma.stanford.edu) Stanford University, Palo Alto, California; Espace de PROJECTION ("Espro") (http://web4.ircam.fr/1039.html?&L=1) at IRCAM (Paris), "The Cube" (http://icat.vt.edu/studios.html) at Virginia Tech's Institute for Creativity, Arts, and Technology (ICAT), and the Spatial Sound Institute, Budapest, Hungary (https://spatialsoundinstitute.com). Annual festivals highlight multichannel sound, including Berlin's Club Transmediale (https://transmediale.de), Edmonton's Sea of Sound (http://www.beams.ca/SeaofSound.htm), and Ontario's New Adventures in Sound Art (http://naisa.ca).

Audio diffusers such as sound field renderers can pantophonically (horizontally) and periphonically (horizontally and vertically) distribute parallel inputs across speaker arrays using a mixer as a crossbar directionalizer. Such architecture scales up to arbitrary degrees of polyphony: multichannel songs, conference chat-spaces, and immersive soundscapes can be dynamically displayed via such controllers. For instance, a dynamic map interface, like that shown in Fig. 2a, can control distribution of multiple channels across a ring of speakers, like that in Fig. 2b, panning signals across an adjustable spread (or "aperture") of speakers.

Spatial sound display is receptive to any number of modulations. Perception of situated sources includes impression not only of position and emission characteristics (relative location and orientation directivity), but also environmental effects, such as reflection, occlusion, obstruction, echo, and reverberation, as measured by such related metrics as Reverberation Time RT_{60}, Definition D_{50}, Clarity C_{80}, and interaural cross-correlation (IACC; http://asastandards.org/Terms/interaural-cross-correlation/).

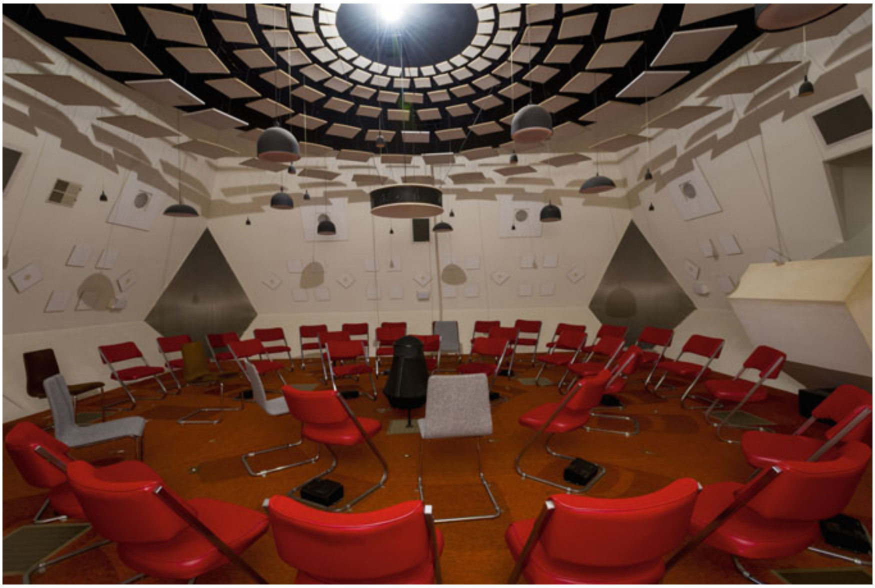
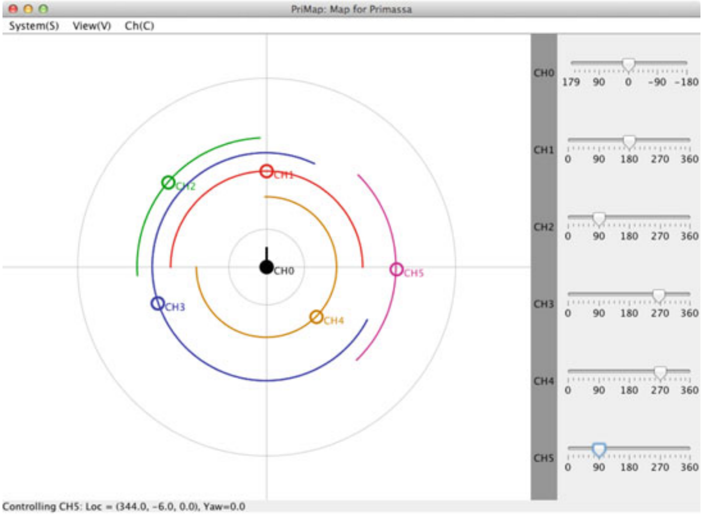


Fig. 1 The Audium (San Francisco)

Adjustment of the virtual position of sources can be independent of the underlying audio streams or somehow related. For simple example, a virtual musical source might encode harmony by moving around a space to signal chord progression (Herder and Cohen 2002) as a pedagogical tool. Dynamic gestures, auditory vectors comprising moving sources, can be used not only for spatial music but for "acoustic arrows," animated sonic beacons for way-finding and multimodal displays, accompanying such verbal directions as "come hither and proceed thither." Likewise, a simulated environment can be adjusted and parameterized by such variables as spatial dimensions, geometry, and liveness (absorption and diffusion material characteristics).

Besides articulated sound directionalization and spatialization, distributed display allows extended diffusion. Spatial extent can be suggested by multiple virtual sources and/or loudspeakers driven together, which resultant auditory (or apparent) source width (ASW) is interpretable as line, area, or volume sources. Much the same way that vibrato makes a note seem louder (Wolfe 2018), or aural exciters (which add high-order harmonic extensions to a signal) enhance conspicuity, wiggling a source or pulsating its size can make it "shimmer" to stand out. To draw attention to a virtual source, an aware agent (a software component that monitors, confirms, and sharpens user focus) can modulate various aspects, including perturbing its position, and dilating and contracting it (adjusting its spatial volume). Such highlighting can push a track to prominence in a mix, like a musical warble, trill or quaver.

Perceptual rivalry—such as contradictory IID (interaural intensity difference, or head shadowing, a.k.a. ILD, interaural level difference) and ITD (interaural time delay)



Controlling CH5: Loc = (344.0, -6.0, 0.0), Yaw=0.0

(a) Asw pantophony: visualization and control interface for sink azimuth and concentric auditory source widths (asws). Arc angles correspond to diffusion aperture across an annularly arranged (circumferential) speaker array. (Map and diffuser developed by Yoshiyuki Yokomatsu.)

"Come out with your hands up—you're surrounded by men with megaphones."

(b) Megaphonic pantophony. (©2020 The New Yorker Collection from `cartoonbank.com`. All rights reserved.)

Fig. 2 Pantophonic perimeter

cues—leads to diffuse source imagery, which a listener describes as a "fuzzy" region. Such localization blur can be thought of as the resolution of spatial hearing.

Because of the horizontal arrangement of our ears, paralleling the gravity-oriented arrangement of our limbs and eyes, IID and ITD panning affect virtual source azimuthal direction (but not elevation or range); it is easier to create auditory "*bokeh*" (out-of-focus blurring) laterally than vertically or longitudinally.

2.2 Panning

A panoramic potentiometer (or "pan-pot") can control distribution of audio power across multiple speakers. To avoid panned signal coherence from disturbing broadened display (via such artifacts as the precedence or Haas effect), image dispersion and signal decorrelation (via such DSP techniques as all-pass filtering, vibrato, and chorusing) can be used to scramble the phase of frequency components (Kendall 2010). DBAP: Distance-based Amplitude Panning (Lossius et al. 2009), MIAP: Manifold-Interface Amplitude Panning (Seldess 2014), VBAP: Vector Base Amplitude Panning (http://legacy.spa.aalto.fi/research/cat/vbap/) (Pulkki 1997) and DirAC: Directional Audio Coding (Pulkki et al. 2011, http://legacy.spa.aalto.fi/research/cat/DirAC/) can be considered generalizations of pan-pots. Phased array and beam-forming soundfield synthesis (SFS) techniques—such as wavefield (or wavefront) synthesis (WFS) and boundary surface control (BoSC) or boundary element methods (BEMs)—which modulate not only gain but also the delay and spectra of multiple signals, require active signal-processing, more aggressive DSP than just amplitude modulation and frequency filtering.

2.3 Personal Listening Systems, PSAPs, Hearables, Hearing Aids, and XR (Extended Reality—AR: Augmented Reality, MR: Mixed Reality, and VR: Virtual Reality)

The contemporary panoply of personal listening devices is surveyed and summarized by Fig. 3. Besides ordinary speakers, personal sound amplification products (PSAPs) and hearing aids are increasingly popular and important, performing vigorous DSP, including directional capture, filtering and active noise control (or cancellation, ANC), ameliorating sensorineural and conductive hearing loss such as presbycusis, age-related hearing loss, as well as compensating for loudness recruitment, rapid increase with amplitude of perceived loudness. By detecting characteristics of an environment, audio processing can automatically change parameters to accommodate various situations (conversation, restaurant, TV, cinema, driving, telephone, concert, etc.). Equalization can be done monaurally using ipsilaterally embedded processors, but binaural processing can be performed on a smartphone, including "diminished

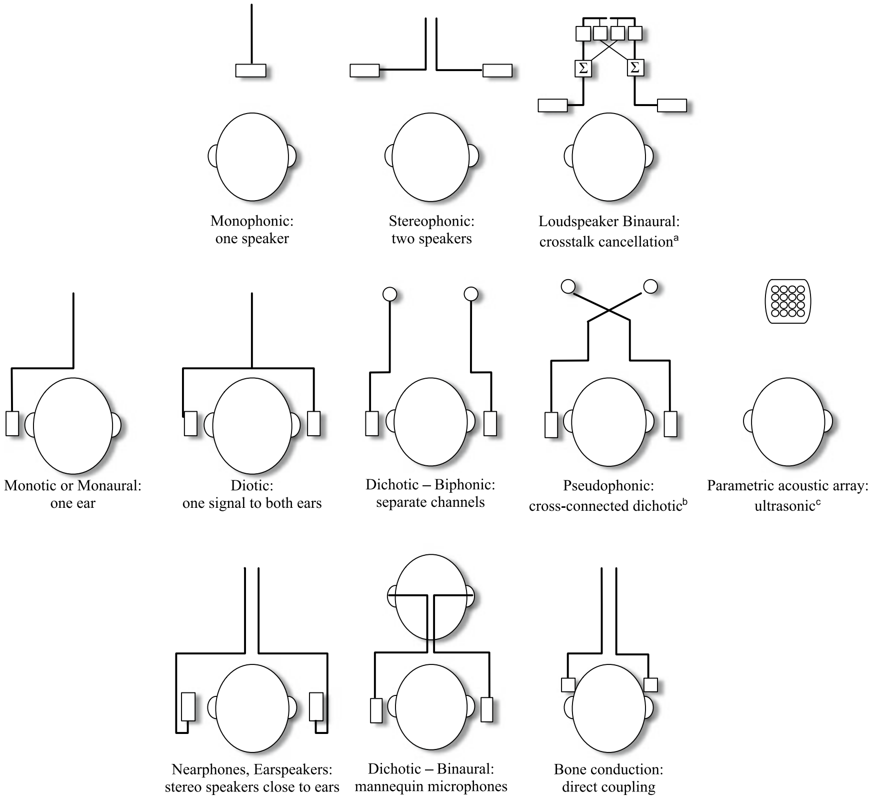


Fig. 3 Personal sound displays: A variety of form factors for personal audition, arranged in order of intimacy. The drivers, symbolized by rectangles, may be wireless, such as Bluetooth earbuds and headsets. (Extended by Cohen 2016 from Streicher and Everest 2006 and Marui and Martens 2006.) **a** By preconditioning stereo signals, speaker crosstalk can be controlled and significantly cancelled or compensated for (cross-talk cancellation, CTC). A special implementation of this technique is called “transaural” (Bauck and Cooper 1996; Choueiri 2018). **b** Pseudophonic arrangements allow dramatic demonstration of the importance of active, head-turning directionalization, as front–back and up–down disambiguations are subverted, even if a subject can see the source (Martens et al. 2011). **c** Ultrasonic displays (a.k.a. parametric loudspeakers), such as that described by Ochiai et al. (2017), represent a special case: inaudible ultrasonic signals demodulate in the air, so the audible source is the air itself, not the driver. Somewhat similarly, some new displays exploit the photoacoustic effect (Sullenberger et al. 2019), by which sound is formed as a result of material absorbing light, such as a laser beam

reality” off-axis rejection for focused hearing. Such architecture also supports disintermediation, eliminating unnecessary dataflow stages: hearing assistance transmission systems using induction looping (telecoils) or FM radio can be replaced by beaming stereo streams directly to earphones, avoiding cumbersome reconstruction, transduction by external speakers, and recapture by hearing aid microphones before resynthesis by in-ear drivers. Hearing aids feature “superhearing” processing (hyperacuity: hypersensitivity and hyperselectivity) informed by auditory scene analysis (Bregman 1990) and deep learning (Wang 2017), enhancing sound segregation and isolating speech. "Human hacking"—bionic augmentation such as prostheses, cochlear implants, and bone-anchored hearing aids—invites extended auditory displays.

Contemporary personal audition systems also include virtual reality (VR) and augmented reality (AR) auditory displays, typically using head-mounted displays (HMDs). VR and AR are generally considered mixed reality (MR), 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). XR can be applied to visualization of sound fields by overlaying visual intensity indication upon actual acoustic spaces (Inoue et al. 2017), but more relevantly and importantly, it can leverage environmental or ambient resources for richer soundscapes. This subject is revisited below in Sect. 2.5.

Some exotic headphones highlight innovative capability, calibrating for anatomy or featuring head-tracking and multidriver arrays to emulate directional sources. For example, the Sennheiser Ambeo headphones (https://en-us.sennheiser.com/in-ear-headphones-3d-audio-ambeo-smart-headset) feature ANC, "hear-through" acoustic transparency, and binaural recording. Bose Frames (https://www.bose.com/en_us/products/frames.html) sunglasses have earstem-embedded, personal back-firing speakers, a microphone for voice control and conferencing, Bluetooth connectivity, and head-tracking for AR applications such as audio tour guides. The Panasonic Wear Space (https://panasonic.net/design/flf/works/wear-space/) features ANC wireless headphones extended with head-wrapping fabric, enhancing concentration by blocking noise and peripheral visual distractions. Nura headphones (https://www.nuraphone.com) have a circumaural body combined with earbuds; its set-up calibration analyzes otoacoustic emissions (OAEs), weak sound generated by the cochlea, to adjust equalization; and tactile bass is delivered through "immersion mode" earcup drivers. Sony 360 Reality Audio (https://www.sony.com/electronics/360-reality-audio) headphones, calibrated by probe microphones, are part of a larger system dedicated to flexible display of 3D audio. The "Aware" headphone (http://www.unitedsciences.com/the-aware-kickstart-the-hearable-revolution/) or "hearable" (https://www.everydayhearing.com/hearing-technology/articles/hearables/) has integrated EEG (electroencephology) sensors, allowing estimation of a wearer's mental state (as reviewed below in Sect. 5.1).

2.4 Panic in the Anech: Extending Live Direct Sound with Environmental Indirect Sound

Another type of physical superposition does not usually employ binaural technology, but becomes very interesting when it does. If a violin is played under anechoic conditions, or captured in a non-reverberant practice room, the performer will typically dislike the unnatural character of the sound—that is, "panic in the anech[oic chamber].” A commonplace non-binaural solution is to submit the ‘dry’ input source to reverberation processing and loudspeaker reproduction to create the more musically familiar ‘wet’ sound signal, so that the performer can hear the sound of their violin in a manner more typical of an acoustically live performance space. Now imagine the binaural counterpart to this, where the direct sound of the violin is captured by a closely-placed, instrument-mounted (“spot”) microphone, and this signal is processed for binaural display such that the indirect sound of a reverberant space responding to the instrumental sound is realistically reproduced via ear-speakers (drivers positioned near but not on the auricles, without circumaural cushions or contact with the pinnae), deployed to allow direct sound from the violin to enter the ears without interference. The performer hears the direct sound from the violin as usual, but with plausibly realistic binaural information in the reproduced indirect sound superposed upon it. This can be valuable for a performer during rehearsal, as the enriched reproduction can mimic the acoustics of the performance space for which they would like to be prepared.

Similarly, when speaking or singing, one’s voice returns to one’s ears with information about the room and its interaction with the voice, yielding an impression of the space. The room acoustical contribution to the sound of one’s voice can be represented via the Oral-Binaural Room Impulse Response (OBRIR), so that self-generated ‘direct’ sounds can be combined in the air (i.e., air-mixed, including the ever-present, bone-conducted, vocal sound: the “human sidetone”) with environmental ‘indirect’ sound that has been electroacoustically introduced (Cabrera et al. 2009). In one such deployment, indirect sound associated with a sound source was reproduced via a pair of ear-speakers, so that binaurally recreated indirect sound could be added to unobstructed ‘live’ sound propagating directly from mouth to ear.

A converse arrangement that also relies upon acoustically transparent ear-speakers is that which might be used to superimpose virtual sound sources upon ‘live’ environmental sound so as to minimize interference of the ear-speakers with natural spatial hearing. For example, in augmented spatial auditory displays providing navigational aid to visually challenged users, minimized interference from a binaural auditory display system is required, since navigation by the blind can be enabled through use of available sonic information, often with refined skills using sound alone. Removing this “open-ear” channel by covering the pinna or plugging the ear canal with insert earphones would disable a needed sensory system, causing drastic reduction in the considerable acuity such users exhibit with their own natural spatial hearing for navigation.

Clear directional imagery was demonstrated for speech signals using such an open-ear binaural superposition system, developed with the commercially available “TOPlay” Open Guided Sound (OGS) earphones (Pereira and Martens 2018; http://www.toplay-ogs.com). Speech signal localization performance using OGS earphones, featuring so-called “TrueOpen” technology to deliver sound directly to the ear-canal entrance with minimal obstruction of the pinnae, was comparable to that assessed using a 196-channel loudspeaker array. Additional “mobile-ambient” systems are discussed in the following subsection.

2.5 Mobile-Ambient Systems: Combination of Personal and Public Displays

Table 2 shows a variety of audio and visual output devices, ordered by intimacy. In analogy to laptop and desktop computing, "eartop" and "eyetop" form factors describe closely attached personal displays. Eartop transducers featuring sound displays for individuals can be integrated with public loudspeaker systems. Even closed-back or circumaural headphones are not completely acoustically opaque, leaking sound in both directions. That is, ambient speakers can be used to complement headphone-displayed soundscapes.

In situations where public and private resources are both available, combinations can leverage advantages of each. As suggested by Fig. 4, hybrid configurations will emerge, such as loudspeaker arrays in conjunction with eartop displays (Satongar et al. 2015) and arrangements of mobile phone speakers. A cinema could feature individual binaural channels, like those served by SoundFi (http://soundfi.me), as well as the theatrical multichannel system, for personalized auditory display, including localized dialog and multilingual narration. Bass management might route low-frequency effects (LFEs) to shared subwoofers whilst sending higher frequency bands to per-

Table 2 Audio and visual displays along private ↔ public continua

Proxemic context	Architecture	Display	
		Audio	Visual
Intimate, Personal, Private	Headset, XR, wearable computer	*Eartop* (earwear), headphone, earbud, earphone, hearing aid, PSAP, hearable, in-ear monitor, bone-conduction (cheekbone, neckband, collarbone, …)	*Eyetop* (eyewear), HWD (head-worn display), HMD (head-mounted display)
Individual	Chair	Smartphone, nearphone, ear-speaker, "sound shower" isolation directional display	Smartphone, tablet, *laptop* display, *desktop* monitor
Interpersonal	Couch or bench	Loudspeaker (e.g., stereo dipole, transaural™)	HDTV, "fishtank VR"
Multipersonal, Familiar	Home theater, vehicle, spatially immersive display (e.g., Cave,™ Cabin)	Surround sound, soundbar, ITU 5.1, 7.1.4, NHK 22.2, etc.	Projection, 4K, 8K
Social	Club, theater	Speaker array (e.g., VBAP, DirAC, DBAP, WFS)	Large-screen display (e.g., IMAX)
Public	Stadium, concert arena	Public address system, (additional sound reinforcement, with delay towers for distant listeners), siren, klaxon	Multiple screens (additional image display to reach distant viewers)

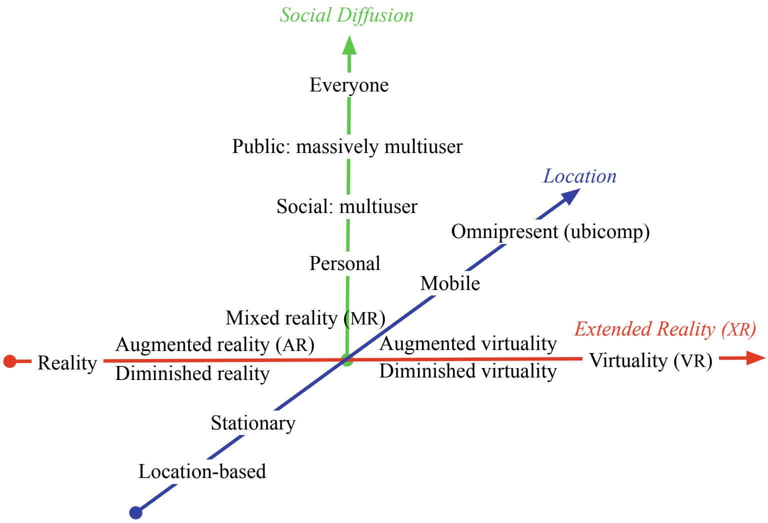


Fig. 4 Extended reality (XR), location, and social diffusion taxonomy—The horizontal Extended Reality (XR) axis is the original AR–VR continuum (Milgram and Coquhoun, Jr. 1999); ***Location*** (longitudinal axis) refers to where such XR systems are used; Social Diffusion (vertical axis) refers to degree of concurrent usage. Adapted and extended from (Broll et al. 2008)

sonal transducers. The dichotomy between mobile computing and site-specific LBS (location-based services) is resolved with "mobile-ambient" transmedial interfaces that span both personal, mobile devices and public, shared resources (Cohen 2016).

2.6 Implications: IoT and Ubicomp

Global popularity of mobile computing creates opportunities for new kinds of computer-human interaction, including democratized control and distributed display. For instance, even technophobes uncomfortable with personal computers can enjoy rich interaction with smartphones. The social diffusion of wireless devices has been paralleled by a separate development of networked appliances: internet of things ("IoT"), ubicomp (ubiquitous computing), and pervasive computing. Sensors and displays will increasingly find their way into everyday circumstances, allowing exploitation by roomware media managers, software for smart buildings.

In computer graphics, "projection mapping" refers to adjusting presentation for display on irregular surfaces, preconditioning contents to anticipate a physical space into which a scene is projected. Auditorily, flexible sound renderers encourage such display context-sensitivity. A simple example is a loudspeaker crossover circuit, which frequency-band filtering matches spectral responses of a multidriver speaker. A more novel example is an opportunistic mixer that routes channels among available

Table 3 Saturated: distributed and pervasive, continuous and networked, transparent or invisible—spatial hierarchy of ubicomp or ambient intimacy

Smart spaces, smart cities, urban (or street) computing
Cooperative or intelligent buildings and smart homes
Roomware and reactive rooms
Spatially immersive displays
Information furniture
Networked appliances, smart displays
Handheld, mobile, nomadic, portable, and wireless (unplugged) devices
Wearable computers, smart watches, smart glasses, hearables, XR HMDs
Computational clothing (smart clothes), hearing aids, PSAPs

resources, discovered and managed by smart homes, intelligent building controllers, "urban (or street) computing," and "smart city" infrastructure. As outlined by Table 3, displays should collaborate across all scales. (These ideas are revisited below in Sects. 5.1 and 5.3.)

3 Perceptual Superposition (Subjective Compositing): Sensation

Whereas the previous section of this chapter dealt with the great variety of physical soundscape superposition to which listeners are exposed, this section addresses perceptual experiences associated with such exposure. The treatment recognizes the complexity of binaural image formation when listeners move relative to sound reproduction systems whilst simultaneously receiving sensory input through multiple modalities, including not only auditory, but also visual and vestibular systems (Martens and Cohen 2020).

Perceptual superposition depends, of course, upon binaural stimuli presented via physical superposition (appearing as afferent signals), but spatial hearing also depends on observers being aware of their own motion in the world (perhaps through efferent signals associated with motor commands, but also though cognitive factors that exert top-down influences on operations such as binaural image formation).

Because perception can be influenced as much by cognitive factors as by stimulus parameters, purely bottom-up (signal-driven, or afferent) models of spatial perception sometimes yield poor predictions of human experience. This is particularly evident in results of studies that include listening conditions allowing listener movement, such as listening while walking (Martens et al. 2011). Although it is difficult to experimentally determine the role of binaural cognition, as scientific studies focus predominantly upon overt behavior, it is reasonable to suppose that cognitive factors (based, for example, upon expectations) operate during listener movement by disambiguating raw sensations through implicit hypothesis testing, such as that associated with “symbol processing” (Blauert et al. 2013; Blauert 2017). Auditory scenes are mentally constructed in the context of potentially abstract thoughts and concepts associated with procedural superposition, which is taken up in the next section of this chapter. Before delving into that topic, the process of binaural image formation shall be discussed, and the complexity of this process, appearing superficially simple, will be revealed.

3.1 Binaural Image Formation: Perceptual Fusion (Integrated Superposition) and Fission (Segregation)

Binaural image formation is the process by which acoustic events to which listeners are exposed lead to the experience of associated auditory events. These auditory events comprise auditory objects that are heard to be located in auditory spaces. While this seems straightforward enough, the process of auditory image formation is neither simple nor well understood. Indeed, there is not always a one-to-one relationship between acoustic events and auditory events. Single acoustic events may give rise to multiple auditory events: perceptual fission (segregation) has occurred. Multiple acoustic events may give rise to only one auditory event: perceptual fusion (integrated superposition) of incoming energy into a coherent entity has occurred. Superposition of sonic events that are presented with the intention of creating an integrated unitary percept will not necessarily be successful, so principles of fusion and fission are examined here. Under typical binaural listening conditions, when the sounds of an external acoustic event impinge upon ears of a human listener, an auditory image of a sounding object typically results. This auditory image may or may not be heard as externalized, i.e., heard as occurring outside the listener’s head. If externalized, the auditory image may be described as an *auditory object*, a mental representation associated with an acoustic event resulting from perceptual fusion of the incoming sound energy into a single, coherent entity. In discussion of binaural image formation, this distinction between acoustical and auditory events should be clearly defined: sounding objects associated with acoustic events have *actual* positions in the physical space surrounding the listener; associated auditory objects have *apparent* positions in auditory space, a mentally constructed space in which auditory events can occur. Acoustic events that occur in reverberant environments are usually heard as occurring outside a listener’s head (i.e., as externalized auditory objects), and yet it is important to recognize these auditory objects as mental projections into psychological constructions of those reverberant environments as they are perceived.

In the context of this discussion on soundscape superposition, understanding principles underlying binaural image formation is key to linking physical superposition and perceptual superposition. This is not a new idea. Plenge (1974) proposed that a sound stimulus should form a coherent auditory image if and only if natural processes

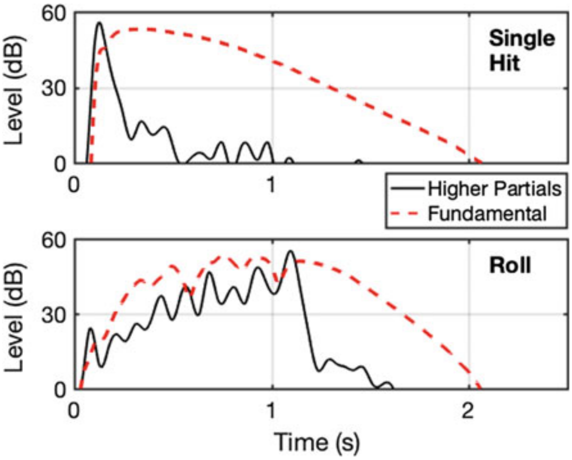


Fig. 5 Examples of temporal envelopes of frequency components for two types of marimba performance, where the dashed curves show the envelope for the fundamental frequency and the solid curves show the sum of the higher-frequency overtones

of spatial hearing are engaged. His model stressed that sound localization has as its first condition…

> […] the ability, learned in early childhood, to classify [auditory] events as sound events. This ability may comprise, besides the perception of direction and distance, the ontogenetic earlier fusion of the information coming through both ears into one general acoustic image.

In free-field sound localization research, asking a listener to report the location of a sound stimulus is reasonable, even when the sound stimulus is as simple as a gated sinusoid. But when a listener uses headphones, such simple stimuli are often heard as within the listener's head ("IHL": inside-the-head-locatedness (Wenzel et al. 2018)), under which conditions Plenge (1974) would term the task *lateralization* rather than *localization*. Even when broadband binaural stimuli are employed, there is no guarantee of externalization and coherent auditory imagery (Toole 1969).

Consider the auditory imagery associated with the binaural presentation of a musical note played on a marimba. Even when a high-quality microphone captures a dry but realistic sounding marimba performance, and then that signal is transformed for headphone presentation through a listener's own measured head-related or anatomical transfer functions (HRTFs or ATFs), the fundamental frequency component of the marimba note typically segregates spatially from the higher-frequency partials of the note which decay more rapidly (and correspond to the brief "strike tone," rather than the more slowly decaying resonance corresponding to the nominal pitch (Perrott et al. 1987)).

For the "single hit" marimba performance shown in the upper panel of Fig. 5, it is easy to see how there might be segregation based on the difference in the temporal envelope of the fundamental frequency component versus that of the higher-frequency partials, which are summed to produce the single solid curve. If, however, a series of rapid marimba notes is performed as in the "roll" performance in the lower panel, listeners have the opportunity to rotate their heads while listening. The two temporal envelopes, while not strictly correlated, nonetheless rise and fall together, so that coordinated lateral shifts in the tone's fundamental and higher partials accompany head-turning or "idling" postural sway. Whether listeners use their natural head acoustics, or use a headphone-based binaural display incorporating active head-tracking, there is an increased likelihood of perceptual fusion of all these frequency components in this dynamic case. Then, if the presentation includes an effective (i.e., spatially realistic) binaural simulation of indirect sound, the binaural image of the marimba tones will likely be heard as both unified and externalized. It is tempting to propose that a Gestalt principle could be operating, where the fundamental frequency that normally segregates from the strike tone of each note might be integrated based upon the 'common fate' of all partials as they shift in lateral angle in response to head-turning.

Whereas in free-field conditions it would be reasonable to elicit a report of the direction and distance of the marimba as a sounding object in physical space, without head-tracking, headphone presentation of a spatially static and dry marimba tone creates a complex percept that cannot be assigned a single direction or position in space. For many years, much of the spatial hearing literature considering headphone presentation has obscured this issue by using the term "localization judgments" to identify such estimates of the position of auditory objects.

Decades ago, Shaw (1982) argued for the importance of a distinction between performance in localizing sound objects and the ability to report the direction and distance of an auditory object experienced during headphone listening. He proposed that headphone studies of auditory spatial imagery be referred to as *space perception* rather than *sound localization*. If this sage advice had been heeded, considerable misunderstanding in the literature might have been avoided. Coupled with an emphasis on spatially static sources and listeners, many reported research results have contributed less to practical applications of binaural technology than desired. Philosophical underpinnings of the above issues are well addressed in a paper by Blauert (2012) that introduces into this discussion the concept of "Perceptionism." A perceptionist's approach to psychoacoustics is also a perspective on methods used in evaluating effectiveness of binaural technology, emphasizing methods that should benefit those engaged in optimizing spatial auditory display technology for real-world applications rather than artificial arrangements in research laboratories.

3.2 Moving Listeners: Dynamic Multimodal Sensory Integration

Much recent research regarding multimodal sensory integration in spatial hearing relates to the importance of voluntary motion in allowing listeners to understand changes in binaural stimuli coupled with changes in the orientation and position of those listeners (Pastore et al. 2020, this volume). Particularly telling in this regard are the results of studies using pseudophonic displays that swap signals between the left and right ears—as shown in this chapter's Fig. 3 and described by its caption b. For example, when listeners are fitted with pseudophonic displays that afford a "live" interchange between left and right ear signals, and are then instructed to walk through an environment attempting to localize sources such as speech sounds, the naturally occuring head-motion-coupled variation in interaural directional cues dominates other localization cues (Martens et al. 2011). If, however, sources with emphasis on higher-frequency content are presented from stable "world-centric" positions, there is less dominance of head-motion-coupled changes in low-frequency interaural cues over spectral cues associated with the pinna. In fact, directional ambiguities can result from the cue conflict that results from such pseudophonic displays when broadband noise bursts are localized (Martens et al. 2013). However, when speech is the stimulus, continuous changes in orientation of the head during walking (such as head-turning) contribute to the creation of strong auditory illusions that are hard to suppress, even when the mouth of the talker is clearly visible. That so-called "Phantom Walker" study showed that when listeners with swapped left and right ear signals were asked to walk past a continuously viewed speech source emanating from a fixed spatial position, the source was heard to be moving through space at twice the listener's rate, and arriving from a spatial region that was reversed with regard to all three spatial axes: left for right, front for back, and above for below. For example, despite having the stationary talker producing the speech sound in clear view as listeners walked toward that talker (where the "ventriloquism" effect might operate), the sound was invariably heard to be approaching from behind, and the voice of this illusory Phantom Walker overtook listeners as they passed by the physically stationary source. These head-coupled interaural cues are so strong that they defeat the contradictory "pinna-based" directional cues, as well as the visual cues (anchored on the actual talker).

Such observations have also been made in studies in which listeners were asked to turn their heads in a constrained fashion while dorsally located loudspeakers presented sources that shifted laterally across the rear hemifield, creating illusions of frontward incidence (Macpherson 2013), through a reversal of interaural cues accompanying head-turning. While these results replicate those of the classic study by Wallach (1940), a related, but possibly surprising result emerged when walking listeners rolled their heads while listening to speech sounds arriving from elevated loudspeakers in an analogous reversal of interaural cues accompanying head-rolling (Martens et al. 2011). Just as front–back reversals are associated with pseudophonic treatment during head-turning (Brimijoin and Akeroyd 2012), above–below reversals were shown to be associated with pseudophonic treatment during head-rolling (with cueing of source elevation depending on the resulting lateral shifts of source images). As have results of other related studies, Kawaura et al. (1991) suggest the dominance of dynamic interaural cues over spectral directional cues, at least for speech sounds containing energy mostly below 5 kHz. When sources containing more high-frequency energy are presented, presumably allowing pinna-based spectral cues greater influence on binaural image formation, the rate of these illusory reversals is greatly reduced (Martens et al. 2013).

To be clear, such head-motion-coupled directional cues do not require or depend upon gross listener movements. Indeed, even when listeners are asked to remain still during a sound localization task, they still move their heads by small but measurable amounts (Wersényi and Wilson 2015), and they seem to move their heads just as much when engaged in natural listening activities, such as watching movies (Kim et al. 2013). Again, these recent studies of vestibular and other motion-based influences on binaural perception of auditory direction are preceded by important earlier studies. In introducing the topic of such non-acoustic influences on binaural perception, Lackner (1983) noted that studies of directional hearing conducted with a fixed head position and orientation clarify only part of the human capacity for spatial hearing:

> Ordinarily a person is freely moving about and his head and trunk position vary both respect to each other and to external objects. Under these conditions the auditory cues at the ears from a stationary sound source change continuously. [...] In localizing an external sound source a person thus must monitor not only the auditory cues he receives from the sound source, but also his own body movements and ongoing position.

Some classic papers on the role of head movement in the context of other non-acoustic cues in sound localization provide a wealth of observations on this topic. (The accompanying chapter by Suzuki et al. (2020) also explores such concerns.) Most notable was early work by Wallach (1940), who observed that head-turning during presentation of a sound stimulus made it possible to distinguish whether a sound arrived from in front or in back of a listener. He noted that when the head was turned to the left, the auditory image associated with a frontal sound source would shift towards the right ear, whereas a dorsal source would shift towards the left. This enables front/rearward distinctions to be made on the basis of head-motion-coupled changes in interaural cues producing variation in the lateral angle of the auditory image. Under conditions in which pinna cues and movement cues indicated incidence from contrasting hemifields, these dynamic interaural cues dominated pinna cues to direction. Wallach also presented such dynamic sound stimuli under conditions in which an illusion of self-rotation was induced by placing stationary subjects inside a revolving screen that filled the visual field. Since their heads were not actually rotating, vestibular cues were absent, and yet listeners experienced self-motion due to these visual cues, and experienced front-to-back reversal when the lateral angle of a frontal sound stimulus was made to shift with head movement as it would were it arriving from the rear.

In another relatively early study, Thurlow and Runge (1967) also investigated the influence of head-rotation on directional hearing, again manually inducing head movements rather than allowing the listener to perform them actively. They examined errors in both azimuth and elevation judgments for a number of types of angular head movement. Without belaboring specifics of the experiments, general results can be summarized as follows: Relative to a condition in which no head movement was allowed, rotation of the head reduced errors in azimuth judgment as expected. However, head-rotation did not significantly reduce errors in elevation judgments. If, alternatively, a subject's head was rolled from side to side while listening (which, in the terminology of the original paper, was called 'pivoted,' as tabulated by Table 4), elevation errors were reduced and azimuth errors were not. This makes sense when considering what happens to the lateral angle of an elevated stationary source when first one ear is dropped closer to the ipsilateral shoulder, and then the other is dropped towards its adjacent shoulder: the lateral shift is the opposite of what is experienced

Table 4 Angular motions of the head ("cocking")

Euler rotation	Plane	Active semicircular canal	Informal designation	Gesture	Expression
Pitch	Median	Superior, anterior	Tip	Nod	Affirmation, concurrence: "yes"
Yaw	Horizontal	Horizontal, lateral	Rotate	Turn, shake	Denial, contradiction: "no"
Roll	Frontal	Posterior	Pivot	Roll, rock, wag, tilt	Uncertainty, questioning: "maybe"

for stationary sources located well below ear level. When the head was tipped forward and back (facing down then up), neither error rate was reduced significantly, as might be expected from the above analysis, since no lateral shifts would occur.

A more recent study of the relative influence of tipping and pivoting considered perceptual attributes associated with many simultaneous sources, rather than the single source studied in (Thurlow et al. 1967). In a study of immersive spatial impression by Martens and Han (2018), multichannel program material—presented via a 10-channel array of loudspeakers distributed about a hemispherical array that included 'height channels')—produced a sense of auditory spatial diffuseness comparable to more truly diffuse stimuli presented using twice as many loudspeakers. In contrast, the spatial impression was noticeably less diffuse when the same 10-channel program was reproduced via a more conventional "without-height" loudspeaker array (i.e., employing loudspeakers located only on a single plane near the listener's ear level). However, this with- versus without-height discrimination in auditory spatial diffuseness was possible in only one of the three head-movement conditions that were tested, and that was the condition in which head-rolling was active.

Considering the geometry involved, it should be clear that above-below disambiguation is enabled by head-rolling-coupled lateral shifts of auditory images along the interaural axis, as demonstrated by Martens et al. (2011). Head-pitching cannot produce analogous disambiguating changes in lateralization for sources that are stable from the world-centric standpoint. For example, if sources are stabilized to remain within the median or even an offset sagittal plane, no lateral shifts occur with head-pitching, but only variation in the HRTF (or ATF) occurs at each ear. Studies have also investigated whether vestibular sensations are strictly required for head-rotation to disambiguate source incidence angles (whether head-turning or -rolling). For example, Lackner (1977) found that illusory self-rotation could be induced by a rotating sound field. He rotated six loudspeakers mounted on a circular frame around the heads of subjects in the dark. Not only did the subjects report that they themselves were rotating and that the sound field was stationary, but they also exhibited compensatory nystagmoid eye movements like those that would occur if they were actually being rotated. More recent studies have examined the compression of auditory space during rapid head-turns (Leung et al. 2008), confirming that self-motion can have strong effect on auditory scene analysis (Kondo et al. 2012).

3.3 Implications: Multisensory Interfaces

Results of these classic experiments indicate bidirectional interaction between perception of head and body orientation and auditory spatial perception. Such characteristics can be exploited by modern communication systems. For example, besides smartphone-embedded IMUs (inertial measurement units), mobile devices feature various techniques for position sensing. SLAM (simultaneous localization and mapping) techniques—including depth perception, motion tracking, markerless feature tracking, depth from stereo, structure from motion, and area learning—are used in visual position sensing/systems (VPS). Head- and eye-tracking can refine positional awareness. Rich models of both internal and external spaces inform rendering of multichannel, multimodal displays that leverage "sensor fusion" among various sensory modalities. These observations are elaborated in the conclusion to this chapter, which follows the survey of idiomatic soundscape conventions presented in the next section.

4 Procedural Superposition (Logical and Cognitive Conventions): Signals

Having reviewed in previous sections combinations of spatial soundscapes regarding physical (*sound*) and perceptual (*sensation*) considerations, we finally consider procedural models of *signals* that inform soundscape composition and cognitive apprehension, higher-level metaphorical associations with which listeners decode sound fields (Cohen and Martens 2020).

When interacting with virtual displays, explicit mental models aid in the conscious reinterpretation of perceptual impressions. In graphics, non-photorealistic rendering (NPR) describes deliberately expressive 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. Analogously, auditory displays also admit such relaxation of literal, "sonorealistic" renderings. Shared assumptions, social conventions, and learned idioms compress communication expression. The following subsections describe some "nonsonorealistic renderings" (NSR) used to enhance or enable de**sign**ation, in the semiotic sense of consensual understanding (Jekosch 2005; Sodnik and Tomažič 2015).

4.1 Separation of Visual and Auditory Perspectives

Normally, personal audition and vision are thought of as concentric, the respective sensory organs embodied together as they are in one's head. 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. Such conventions extend to spatial media, as cameras might be binocular, visual displays stereographic, microphones stereophonic, and auditory displays binaural. However, telesensory instrumentation allows and encourages independence of modalities.

For architectural walk- (or fly-)through and auralizations (Kleiner et al. 1993), visual and auditory perspectives should match, as if cameras and microphones were integrally deployed. For a concert, an auditory display might be presented "with perspective" (i.e., aligned with visual display), either directly (acquired via coincident microphone) or coherently simulated. However, performed electroacoustic music can be captured by a variety of overhead, on-stage, and spot (accent) microphones, mixed and distributed for monitoring (realtime self-audition by the musicians), sound reinforcement (for live audience), and recording or transmission (for archive or distribution). Mediated concert experiences such as music videos separate visual and auditory perspectives, not insisting that capture, rendering, or simulation of aural perspective match optical position.

Cinematic and gaming idioms also relax literal associations, freely exercising liberty to set aside assumptions of alignment of auditory and visual perspectives. For example, background "score" music (BGM) is non-diegetic (conceptually outside a story space, like narration) and accommodated by such independence. One auditionally attends multiple spaces at once, apprehending not only a narrative 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 optical perspective (cinematography, montage, etc.), but also artificial auditory access, with flexible correspondence among display modalities. The "2nd-person perspective" popular in role-playing games (RPGs) 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 a puppet or "vactor" (virtual actor), typically viewed from slightly behind and above, through a loosely attached virtual camera. Likewise, projected location of sound associated with such an avatar (generated by a game engine or voice-chat captured from the human pilot) is that of the avatar, not the lagging virtual camera.

4.2 Separation of Orientation and Location—Directionalization Versus Localization

Table 5 juxtaposes location, orientation, and position as well as static posture versus dynamic gesture. Space, at least at sensible levels of apprehension, is 3-dimensional, and location is most simply represented numerically by Cartesian triplets (x, y, z). For example, CAD models are usually represented as vertices, edges, faces, and solids. Such subject-independence is allocentric (Roginska and Geluso 2018) or exocentric, independent of listeners or observers.

For subjective displays, parameterized explicitly or implicitly by standpoint and egocentric direction, polar or spherical coordinates are more convenient than rectilinear coordinates since they are non-homogeneous, in that range (ρ) is dimensionally different from azimuth (θ) and elevation (ϕ). That is, representation or projection of distance is different from horizontal and vertical direction, and can be decoupled.

For object-based encodings, monaural audio streams can be localized for binaural display with ITD, IID, and HRTF-based filtering. Sound objects are most simply directionalized by intensity panning to loudspeakers near a phantom source, but such amplitude- or gain-based techniques cannot realistically convey spatial effects such as early reflections (echoes), modal resonances (standing waves), and late reverberation.

Ordinary surround sound and 5.1 configurations, using channel-based encodings such as those deployed in home theater arrangements, do not usually exploit elevational cues, such as those deliverable via height or overhead ("voice of god") channels. However, preconditioning signals with ATFs before display through loudspeakers can simulate height cues (Jo et al. 2010; Tanno et al. 2014).

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 re-directionalization.

Of the three affine transformations (scaling, rotation, and translation), Ambisonics accommodates only rotation, so such soundfield recordings can be thought of as "prebaked," forgoing "remixing" flexibility (such as standpoint excursion or interaural baseline adjustability, which scales anatomical signals such as ITD and IID and changes binaural disparity) for optimized rendering.

4.3 Directionality Processing

Head motion, such as was discussed in the last subsection, is not only like "antenna pointing," but also "body language," a kind of display. Situational context, voice intonation, facial expression, gaze and gesture all inform exquisite decoding of proxemic cues. Head gestures as shown earlier in Table 4 are just 1st-order conventions; such communication is rich and subtle. Eye-gaze, which can be approximated from head

Table 5 Physically spatial dimensions: taxonomy of positional degrees of freedom, including cinematographic gestures (assuming right-handed coordinate system, with xy horizontal plane and z gravitational up)

Position (Pose)					
Static (Posture)		**Dynamic (Gesture)**			
Location (Displacement)		**Translation (Perturbation)**		**Along**	**Perpendicular**
(Extent)	**Scalar**	**Camera Motion**	**Directionality (Force)**	**Axis**	**to Plane**
lateral, horizontal (breadth, width)	abscissa x	sway track (“crab”)	left ↔ right	x	median (sagittal)
frontal, longitudinal (depth, length)	ordinate y	surge dolly	back (aft): retreat (drag) ⤡ front, forth (fore): advance (thrust)	y	frontal (coronal)
vertical (height)	altitude z	heave boom (crane, “ped”)	up: ascend (lift) ↕ down: descend (weight)	z	horizontal (transverse)
Orientation, Direction, Attitude		**Rotation (Spin)**		**About Axis**	**In Plane**
elevation, inclination	ϕ	pitch (tumble, flip) tilt	climb/dive	x	median (sagittal)
(“barrel roll”)	ψ	roll (tilt, bank, flop, “Dutch”)	left/right	y	frontal (coronal)
azimuth	θ	yaw (whirl, twist) pan	CCW/CW	z	horizontal (transverse)
Location and Orientation		**Revolution & Rotation**		**Axis**	**In Plane**
focal pivot	x, y, θ	phase-locked orbit “spin-around” or inspection	(CCW/CW)	(z)	(horizontal)

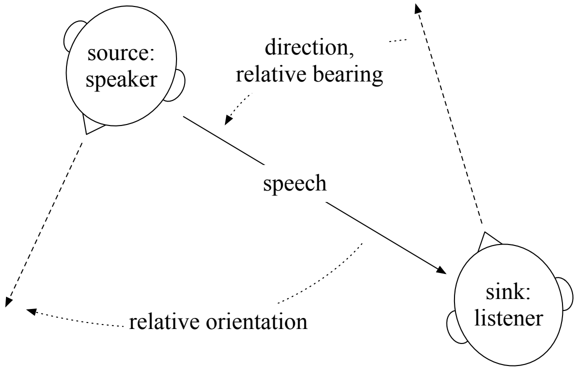


Fig. 6 Direction and orientation: psychoacoustic cues as proxemic social signals. (By "direction" we mean here the relative bearing of a source with respect to a sink, independent of its egocentric rotation; by "orientation" we mean the direction a source is facing.)

orientation, is used for social signaling and can trigger computer-mediated events. Individually apprehended spatial sound tells the eyes where to look, but "gaze indirection" (understanding where someone else is looking), awareness of directed or projected visual attention, alerts conversants about objects of regard. Mouth-emitted sounds are anisotropic, and speech is directional.

As illustrated by Fig. 6, a listener estimates not only direction but also orientation of a talker. Using hints such as ratio of direct-to-indirect intensity and darkening (via low-pass filtering) of utterances, listeners recognize which way a talker is facing, inferring targets of directed address. Symmetrically, talkers are aware of orientation of listeners, and modulate their voices according to appreciation of the listening difficulty of those facing away from them (akin to the Lombard effect, in which talkers strengthen vocalizations in the presence of ambient noise). An aware renderer such as a roomware auditory display is parameterized not only by direction but also orientation of sources relative to sinks, modulating delivered audio streams to convey such fine cues.[1]

Sink and source directivity can be modeled by emulating idealizations of microphone receptivity patterns, combinations of omnidirectional (unipolar) and directional (dipolar) radiation as well as sensitivity (Hugonnet and Walder 1998). For typical instance, the Google VR Audio (https://developers.google.com/vr/ios/spatial-audio) and Resonance Audio (https://resonance-audio.github.io/resonance-audio/) Unity plug-ins model directionality by "`alpha`" ($0 \leq \alpha \leq 1$) and "`sharpness`" ($1 \leq$ `sharpness`). Normalized gain fields are calculated as $|(1-\alpha)+\alpha\cos(\theta)|^{\text{sharpness}}$, where θ 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 weighting coefficient α scales directionality, dipole power `sharpness` exaggerates such non-isotropy, and the absolute value function rectifies

[1] 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. 4.8 below.

polarity inversion.[2] When α is zero, the pattern is isotropic (and the `sharpness` is irrelevant); as α approaches unity, directivity becomes increasingly lobed. "Earshot," combined radiation and reception, is the product of these for each source → sink combination.

Such sensitivity directivity patterns are analogous to clipping frusta of computer graphics rendering. Such hyper-acuity of apprehension or heightened directionality of projection are best suited for AR applications embedded in real world contexts, since purely virtual exposure and receptivity are not constrained by such coarse models as lobed directivity. These are generalized by narrowcasting, described below in Sect. 4.7.

4.4 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, and diffraction in the presence of obstacles or occluders, dispersion, absorption and scattering) to realize both improved performance and expressive control. Intensity of a point source spherically radiating sound waves naturally observes an inverse square relation with distance, so amplitude gain, a root power quantity proportional to RMS pressure and the square root of intensity, observes a reciprocal (inverse-proportional) relation with range. Distance modulation and estimation of virtual sound sources becomes even sharper if volume control is driven by models that roll-off more rapidly than this physical gain $\propto 1/\rho$ law, where ρ is the distance between source and sink. In contrast, it is sometimes assumed that, in small spaces, amplitude of a reverberant signal changes little with range, and that in large spaces it is roughly proportional to $1/\sqrt{\rho}$ (Pulkki et al. 2011).

Excepting extreme circumstances in spatial sound teleconferencing, such as when a virtual source approaches antipodal position, geotagged 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. 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.

For example, realistic display would attenuate most sources below audibility. In everyday experience, even very loud sources are rarely heard beyond a few kilometers, and conversational intensities are normally inaudible beyond tens of meters. With the usual -6 dB/range doubling attenuation, the level 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 auditory threshold, and practical inaudi-

[2] Similar plug-ins are also offered by other companies, including Facebook (https://facebookincubator.github.io/facebook-360-spatial-workstation/), Microsoft (https://docs.microsoft.com/en-us/azure/cognitive-services/acoustics/what-is-acoustics), and Yamaha (https://research.yamaha.com/ja/technologies/vireal/).

bility occurs even closer because of background noise. Fortunately, utilities for way-finding (such as Microsoft Soundscape (https://www.microsoft.com/en-us/research/product/soundscape/)), direction-giving, and conferencing do not need to render sonorealistic range cues.

Besides intensity-controlled loudness, other cues to simulate or suggest distance can be separately modulated (Jot 1999), including initial time-delay gap, the interval between a direct sound and its first reflection; the previously mentioned direct:indirect ratio of the power 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. Direction is usually more important than distance expression, but a fully featured display should allow localization into one's "whisper space" (Villegas and Cohen 2010) to convey such near-field intimacy, such as that evoked by autonomous sensory meridian response (ASMR) programs.

Relatedly, a rendering engine might perform "spotlight mixing," exaggerating loudness of frontal objects assumed to be foci of attention, analogous to foveal rendering in computer graphics. Alternatively, as frontal objects could be assumed to be visible and therefore already conspicuous, rearward objects might be particularly amplified (Bailey 2007), or their auditory position or timbre animated to "catch one's ear." Such "gaze mixing" (https://docs.microsoft.com/en-us/windows/mixed-reality/spatial-sound) is a sensory substitution kind of multimodal coordination, which also includes "audio haptics," reactive sounds for touchless interactions, compensating for a lack of force-feedback in virtual displays.

4.5 Extreme Dynamic Range Compression: Location-Indifferent Intensity

Dynamic range is the ratio of the intensities of the strongest and weakest parts of a signal, and range in the sense of source → sink distance can be used to attenuate level, distance fall-off. In the limit, compression of dynamic range associated with distance-dependent attenuation approaches range-insensitivity. Separation of orientation and location, including distance independence, allows directionalization without localization. In spatial user interfaces, compass bearings such as "North" are obviously 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 characters (such as icons or avatars 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 immersive photospherical or volumetric visual and stereophonic auditory display represent a popularization of VR-style interfaces. Google Cardboard (https://arvr.google.com/cardboard/), the Merge Headset (https://merge edu.com/headset), Oculus Quest (https://www.oculus.com/quest), and Samsung Gear VR (https://www.samsung.com/global/galaxy/gear-vr/) use sensors for head-tracked binocular display of stereoscopic contents and stereophonic display of spatial audio. Orientation can be tracked by a micro-electro-mechanical system (MEMS) IMU—including gyroscope, accelerometer, and magnetometer—estimating bearing via aggregating sensor fusion, but if location is not tracked (as via GPS or optical tracking), user virtual standpoint is not directly adjusted.

Some scene-based interfaces ignore location and use only orientation. Spatial sound sources can be directionalized without range-parameterized gain modulation. With head-tracking, a subjective soundscape can be counter-rotated, panned to stabilize a scene, but not perturbed. Orientation sensitivity supports location-based sound fields. For example, fields captured or encoded into Ambisonics B-format (with 4 channels) are easily rendered at runtime, down-mixed to a panned stereo pair heard through head-tracked headphones or up-mixed to a real or virtual speaker array.

4.6 Layered Listening and Audio Windowing

Procedural mixing allows user interfaces to algorithmically combine and distribute audio signals. Networked and object-based articulated sources invite audio-level (as opposed to acoustic-level) modulation, and logical layers are a natural model for such composition. Cinema and electronic gaming encourage richly textured soundscapes, including music, sound effects (SFX), 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. Such a cascade is equivalent to a tree of metamixers (Cohen 2015), a dataflow arrangement in which compositing operations are modeled as routing matrix switches with effects applied at each crosspoint—"programmable shaders" fanning-out into amplifiers for a combination of personal and public transducers, headphones and loudspeakers. Multichannel Audio Digital Interface (MADI) (http://www.aes.org/publications/standards/search.cfm?docID=17) and Dante (https:www.//audinate.com) are popular standards for multichannel audio networking and interfaces. Audio middleware and engines such as CSound (https://csound.com/), Faust (http://faust.grame.fr/), FMOD (https://fmod.com), JUCE (https://juce.com), Max/MSP (https://cycling74.com/products/max/), Pure Data (http://puredata.info), Reaktor (https://www.native-instruments.com/en/products/komplete/synths/reaktor-6/), SuperCollider (https://supercollider.github.io), and Wwise (https://www.audiokinetic.com/products/wwise/) can render auditory scenes.

Audio windowing (Cohen 2016), 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 (Begault 1994). Spatial soundscapes, like layers in graphical applications or tracks in musical compositions, can be combined simply by summing, although some scaling (amplification or attenuation), equalization, and other conditioning yields 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 muted or muzzled and some sinks might be deafened or muffled.

As was illustrated by Fig. 4, mixed reality can not only add information to naturally captured scenes, but can also remove information. Interpretation of "XR" can include "Diminished Reality." Diminished audio reality can be thought of as hiding or masking otherwise apparent auditory scene components, such as engine sounds (as in ANC), 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" (Middlebrooks et al. 2017), whereby particular objects are "heard out" of a cacophonous mix. The are generalized together by auditory source separation, auditory scene analysis (Bregman 1990), and blind source separation.

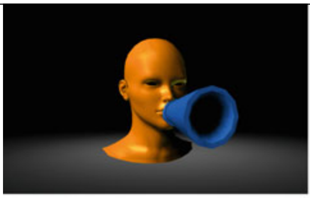
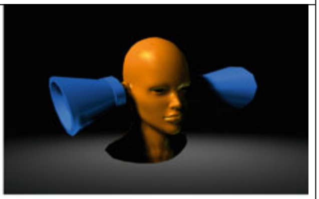
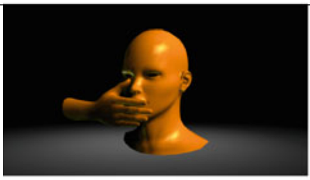
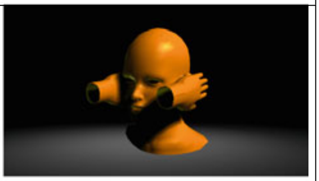
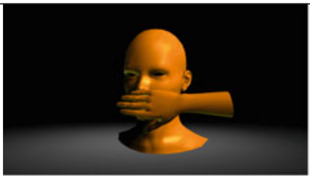
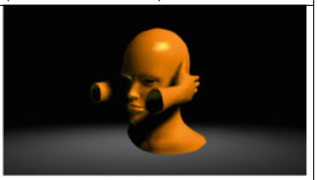
4.7 Narrowcasting: Privacy and Attention Management

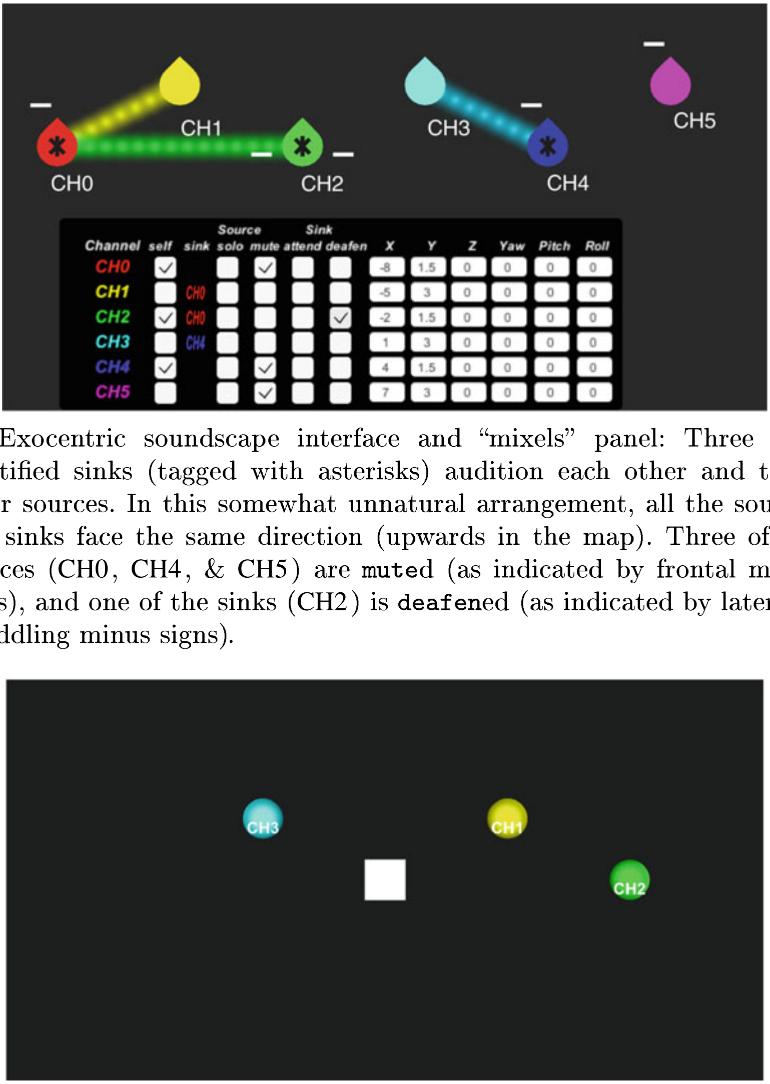
"Privacy" has two interpretations. The first association is that of avoiding "leaks" of confidential information, protecting secrets. The second association is "freedom from disturbance," not being bothered by interruption. Narrowcasting operations manage privacy in both senses, filtering duplex information through an articulated communication model. In analogy to any-, broad-, multi-, and unicasting, narrowcasting is an idiom for limiting and focusing media streams. Sources and sinks are symmetric duals in virtual spaces. 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 being `muted`, 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`" (Cohen 2000).

Formalized by the permission scheme expressions shown in Fig. 8, narrowcasting (Alam et al. 2009; Cohen et al. 2009) exposure and distributes attention. Advanced floor control symbology—for chat-spaces, concerts, and conferences—is outlined by Table 6. Modulation of source exposure or sink attention needn't be "all or nothing"—nimbus (projection) and focus (receptivity) can be respectively partially softened with muzzling and muffling (Cohen 1993)—see Fig. 7.

That is, nuanced operations can soften state transition, allowing non-binary control—not just `on`–`off` but intermediate 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`" switch of a conferencing service.

Table 6 Narrowcasting for $^{s}\mathrm{OU}^{\mathrm{rce}}_{\mathrm{Tput}}$ and $^{s}\mathrm{IN}^{\mathrm{k}}_{\mathrm{put}}$. (Figurative avatars by Julián Villegas.)

		Source (♂)	Sink (♀)
Function		Radiation (effector), emission	Reception (sensor), collection
Level Adjustment		Amplification, Attenuation	Sensitization, Desensitization
Media Direction		OUTput (display), production, push	INput (control), consumption, pull
Perspective		Object	Subject
Presence Locus		Nimbus (projection, exposure)	Focus (receptivity, sensitivity)
Auditory	Instance	Speaker	Listener
	Transducer	Loudspeaker	Microphone array or dummy-head
	Organs	Mouth	Ears
Enable (spur)	Include	`solo`, `select`	`attend` (harken)
	Metaphorical Device	Megaphone, loud-hailer, bull-horn	Ear trumpets
	Icon	+ ○	+○+
	Avatar		
Disable (spurn)	Exclude	`mute`	`deafen`
	Suppress	Muzzle	Muffle
	Icon	− ○	-○-
	Avatar, own *reflexive*	 (Thumb up)	 (Thumbs down)
	Avatar, other *transitive*	 (Thumb down)	 (Thumbs up)



			Source		Sink							
Channel	self	sink	solo	mute	attend	deafen	X	Y	Z	Yaw	Pitch	Roll
CH0	☑		□	☑	□	□	-8	1.5	0	0	0	0
CH1	□	CH0	□	□	□	□	-5	3	0	0	0	0
CH2	☑	CH0	□	□	□	☑	-2	1.5	0	0	0	0
CH3	□	CH4	□	□	□	□	1	3	0	0	0	0
CH4	☑		□	☑	□	□	4	1.5	0	0	0	0
CH5	□		□	☑	□	□	7	3	0	0	0	0

(a) Exocentric soundscape interface and "mixels" panel: Three self-identified sinks (tagged with asterisks) audition each other and three other sources. In this somewhat unnatural arrangement, all the sources and sinks face the same direction (upwards in the map). Three of the sources (CH0, CH4, & CH5) are `muted` (as indicated by frontal minus signs), and one of the sinks (CH2) is `deafen`ed (as indicated by laterally straddling minus signs).

(b) "Flattened" soundscape, collapsed around a single listener perspective: Active sources and deafened sinks are partitioned across active sinks, which necessarily coalesce into a singular perspective. Since the autofocus algorithm assigns sources CH1 & CH2 to CH0 and CH3 to CH4, as indicated in the mixels panel above, the soundcape reduces to this subjective arrangement around the notional human listener, iconified by the central white square.

Fig. 7 Dynamic map featuring display and control of spatial sound sources and sinks, including narrowcasting, multipresence, and autofocus (Cohen and Kojima 2018), with contributions by Akane Takeshige, Peter Larson, and Koki Tsuda with Rintarō Satō

The general expression of inclusive selection is

$$\begin{aligned}\texttt{active}(x) = \neg\texttt{exclude}(x)\wedge \\ ((\exists\, y\ (\texttt{include}(y) \wedge (\texttt{self}(x) \Leftrightarrow \texttt{self}(y)))) \ \Rightarrow\ \texttt{include}(x)).\end{aligned} \tag{1}$$

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

$$\begin{aligned}\texttt{active}(\mathrm{source}_x) = \neg\texttt{mute}(\mathrm{source}_x)\wedge \\ ((\exists\, y\ (\texttt{solo}(\mathrm{source}_y) \wedge (\texttt{self}(x) \Leftrightarrow \texttt{self}(y)))) \ \Rightarrow\ \texttt{solo}(\mathrm{source}_x)),\end{aligned} \tag{2a}$$

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 auditory sink relation is

$$\begin{aligned}\texttt{active}(\mathrm{sink}_x) = \neg\texttt{deafen}(\mathrm{sink}_x)\wedge \\ ((\exists\, y\ (\texttt{attend}(\mathrm{sink}_y) \wedge (\texttt{self}(x) \Leftrightarrow \texttt{self}(y)))) \ \Rightarrow\ \texttt{attend}(\mathrm{sink}_x)).\end{aligned} \tag{2b}$$

Fig. 8 Formalization of narrowcasting functions in predicate calculus notation, where '¬' means "not," '∧' means conjunction (logical "and"), '∃' means "there exists," '⇒' means "implies," and '⇔' means "is equal to" (mutual implication, "if and only if"). Duality between source and sink operations is strong, and the semantics are analogous: an auditory object is inclusively enabled by default unless, (i) it is explicitly excluded with mute (for sources) or deafened (for sinks), or, (ii) peers in the same self/non-self class are explicitly included with solo/select (for sources) or attend (for sinks) when the considered object is not

4.8 Multipresence and "Anyware"

Ordinary correspondence between inhabited bodily apprehension and consciousness is one-to-one, but telexistence (Tachi 2015) can soften such rigidly focused subjectivity, relaxing the singularity of human experience. Multitasking users want to have presence in several locations at once. For instance, a telephone exemplifies auditory telepresence, projecting conversants to other places besides their corporeal "meat-space" base.

Enriched user interfaces, especially with position-tracking systems or real-time locating systems, encourage multipresence, the inhabiting by representative sources and sinks of multiple locations simultaneously, allowing a human user to designate doppelgänger delegates in distributed domains. Exocentric interfaces supporting "out-of-body" experience enable parallel spaces, across which can be designated multiple instances of self-identified avatars (Cohen 1998; Ranaweera et al. 2015) as shown in Fig. 7. "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 vision, which cannot naturally overlay separate scenes. 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 (Cohen and Fernando 2009), like a “snap-to grid.” 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.

4.9 Implications: Nonsonorealistic Rendering and Multimodal Cognition

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. Communication culture is not innate but learned. Listening is not a one-off event, but continuous experience. Sound displays use acquired associations, rather than direct emulation of natural phenomena. An assumed sophistication of listeners decoding nonliteral displays admits an acceptance of plausible but nonveridical cues.

Many situations do not call for an auralization-style re-creation of a particular soundscape but instead are best served by some kind of metaphorical space. Practical auditory conventions such as those described by this section refine expression. For instance, by using an audio windowing system as a mixing console, a multidimensional pan-pot, users and applications determine rich parameters to compile source and sink positions and their environments, rendering as a distributed diffuser or spatial sound stager. Presence is more important than fidelity, audiophilic predilection for “absolute sound” or perceived need for Master Quality Authenticated (MQA; http://mqa.co.uk) streaming notwithstanding.

Purely auditory displays hardly exist. Normal physical environments ensure that ordinary events are perceived multimodally. Spatial sound cues are aspects of a rich ecology of environment-embedded signs. Almost always, “in the wild,” visual cues and other context complement projected auditory scenes. Soundscapes are not apprehended “in a vacuum”: some map, conventional understanding, or at least situation awareness aids decoding. Multimodal interfaces empower overlapping displays.

Cognitive processes can resolve otherwise confusing soundscapes. For instance, a flashing light (as on an active smart speaker, or the “Lyric Speaker,” (https://lyric-speaker.com) which animates words in *karaoke*-style sync with songs) can disambiguate conflicting cues. Listeners are inclined to be forgiving, suspending not only disbelief but also insistence on sonorealism, so sonic situations can be efficiently communicated. Mental models are used to interpret multimodal events, including those generated by non-literal displays. For instance, independence of location and orientation can flatter and “flatten” multipresent auditory localization. An advantage of separating translation and rotation is that directionalizability can be preserved even across multiple frames of reference. Such distributed presence can be coupled with vehicle or position tracking. Moving can twist (but deliberately not shift) multiple representations, maintaining consistent proprioceptive alignment of overlaid sound sources.

5 Crowds and Clouds: Final Thoughts and Conclusions

5.1 Ubicomp and IoT: Extreme Sound Reinforcement

Ordinary rooms often host electronic appliances such as TVs, desktop and laptop computers, game consoles and controllers, smart speakers, as well as tablets, and smartphones of “second screening” (multitasking) occupants, who might also have HMDs or smart glasses for XR, wearable computers (such as smart watches and hearables), PSAPs and hearing aids. These multitudes of speakers and microphones, displays and sensors, can be integrated by roomware.

In ubicomp environments, generally multiple users must be accommodated. Urban computing offers even broader challenges and opportunities: public signage and auditory displays can serve AR messages to tracked users. A distributed ecosystem of electronic devices defies top-down management but invites bottom-up coordination. Privacy, attention, and sensitivity parameterize rendering of soundscapes. Delegated by human users, software agents and intelligent assistants will negotiate private and collective access to resources. Transducers of AI-infused networked appliances can work in concert with personal “awareable” devices to optimize personal and public experience. Syndicates of groupware interfaces will pool crowd-sourced data and share displays: mediated social sensing and signaling.

In an “ABC” (always best connected) world, persistent chat-spaces are expected: selectively continual connectivity with one’s family, friends, and colleagues. Aware interfaces infer user receptivity, tuning an environment by automatically adjusting displays of all types to reward attention. Activity sensors, position trackers, and monitors cooperate to optimize comfort, efficiency, and productivity. IoT-style smart speakers should be situationally aware, using amalgamated sensing—microphones, cameras (including thermal and infrared sensors), mo-cap, EEG, and fitness trackers and biosensors (capturing microexpressions of voice, gaze, body language, pupil dilation, heartbeat and pulse variability, galvanic skin response, body heat, etc.)—to gauge mood, empathetically adjusting soundscapes to support users (Crum 2019).

Compiling a heterogeneous display, for listeners in arbitrary positions, across speakers of various sizes, orientations, directivities, spectra, acoustic intensities, and irregular and dynamic arrangement is endlessly challenging: extreme sound reinforcement. However, opportunistic networked managers (Choi et al. 2016) can exploit disparate devices for enriched presentation, carving out “sound zones.” Reflexive display-and-capture systems can be used to calibrate diffusion in a “closed loop,” like that used by structured light sensing. For instance, roomware might arrange to ‘borrow’ or ‘lease’ nearby sensors and effectors to adjust parameters. Representative contemporary applications demonstrate such cyberphysical cooper-ation between speakers and microphones and suggest the potential of such symbiosis: “Chirp” (https://chirp.io) and “Google Tone” (https://chrome.google.com/webstore/detail/google-tone/nnckehldicaciogcbchegobnafnjkcne) distribute URLs to nearby computers audibly (“data-over-sound”); “Ultrasonic Recognition” (http://www.lankasolution.com/ar365-usr-ultra-sonic-recognition/) embeds tags in audio tracks; and “AmpMe” (http://ampme.com) and “Tune Mob” (https://itunes.apple.com/developer/tunemob/id680664869) manage network-synchronized distributed music display. Audio steganography can embed “side-channel” information as subliminal, ultrasonic, or otherwise inaudible acoustic signals.

5.2 AI-*Empowered Conversational Agents*

Besides mobile telephony, so-called “smart speakers,” which also integrate microphone arrays and often lights or fuller displays, feature internet services for conversational interfaces backed by AI for information or control. Emergent qualities of networked sensors and the high bandwidth and low latency of wireless systems such as that promised by 5G, 5th-generation cellular networks, recall the blending of fixed-mobile convergence (FMC). As the processing is mostly on-line, intelligence cannot be attributed to the loudspeaker itself: the network makes locality of computation seamless or “cloudy.” We extend ourselves with distributed systems, and the network stretches to embrace us cyberspatially.

Such IoT devices represent an interpolation between robots and chatbots, transactional and conversational virtual assistants. Appliances, even with wireless data connections, are usually powered and fixed, but ambulatory electronic pets and consumer robots—including socially assistive models and hospitality-service bots (such as Sony Aibo (https://us.aibo.com), Honda Asimo (http://asimo.honda.com), SoftBank Pepper (https://www.softbank.jp/en/robot/), and Sharp RoBoHon (https://robohon.com/global/))—detecting and responding to human emotions, represent self-locomotive loudspeaker platforms with telepresence capability.

Acoustic devices can be wireline or wireless, spanning continua of data- and power-cordlessness: **Fixed**, as by normal loudspeakers; **Tethered**, as by many HMDs; **Bounded**, as with zones for near field communication (NFC) and area networks such wireless local area networks (WLANs) and near-me area networks (NANs), including those of Bluetooth, Wi-Fi and WiGig; and **Free-roaming**, as with cellular coverage.

Voice interfaces feature speech recognition (SR) and text-to-speech (TTS), with increasingly natural sounding synthesis, allow rendering of textual sources as auditory sources, a synæsthetic transcoding. The renaissance of machine learning and AI includes advances in big data and deep learning, for speech interpretation, machine translation, conversational intelligence, and multilingual TTS. “Vocal emotion recognition” can characterize mood from speech, using such microexpressive cues as voice dynamics, tone, timing, and metalinguals. AI can be applied to situation awareness, estimating social conditions such as user sensitivity (distractibility, attention, fatigue, multitasking, “flow”), including support functions such as face, speech and speaker recognition; optical character recognition (OCR); natural language processing (NLP); and "*kansei* (affective) engineering" sentiment analysis.

Enabled by the confluence of sensing, connectivity, computation, and machine intelligence, user recognition and characterization allow provision of personalized media and listening zones. The "quantified self" domain includes audiometric customization and individualization of ATFs. Public loudspeakers are usually around the periphery of a room—often at the walls, sometimes on the ceiling, rarely in the floor—but smart speakers among and amidst people can complement traditional loudspeakers, and along with personal displays, contribute to integrated mobile-ambient interfaces for immersive experience, taking "theatre-in-the-round" and turning it envelopeingly inside-out. Paralleling FMC, "glocal" interfaces can leverage both personal devices and shared resources. For control, smartphone-sensed orientation and GPS-like tracking can be combined with parameters such as layering and narrowcasting attributes. For display, smartphone and tablet screens can be extended by cooperative roomware lights and screens, and headphones and hearables can be augmented by speaker arrays.

5.3 Late Binding of Soundscape Staging: Runtime Determination of Synthesis, Filtering, Spatialization, and Multimodal Rendering

Spatial sound systems handle three different kinds of audio encodings, namely,

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

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

Audio sources for games (Collins 2008) and simulations have historically been associated with prerecorded files, but more richly parameterized applications and social media drive a shift to dynamic media streams, including physical modeling, procedural audio, algorithmic music, voice-chat, and, inevitably and imminently, "deepfake" photo- and sonorealistic multimedia. The parallel trend is away from assumed fixed loudspeaker locations and towards expectation that material will be rendered to whatever is available at the display end of the chain. As attention shifts away from prepared media towards online experiences, the process of mixing changes: instead of aggregation into "stems," raw audio tracks are pushed into dynamic rendering, configured by metadata object positions and realtime tracking. Rather than baking virtual sources into transducer channels, which is a kind of rigid compilation, sources are rendered and diffused at runtime, accommodating circumstances and exploiting opportunities. Parameterization by "late binding" display arrangement is a kind of dynamic projection mapping, configuring signal-processing to match particular loudspeaker and headphone resources and configurations.

Such freestyle improvisation lacks the broad consistency of cinematic standards such as Auro-3D (https://www.auro-3d.com), DTS:X (https://dts.com/dtsx), and Dolby Atmos (https://www.dolby.com/us/en/brands/dolby-atmos.html), but is potentially richer and is inherently future-proof. Dolby AC-4 (https://www.dolby.com/us/en/technologies/AC-4.html) combines channel- and object-based models, and DTS MDA, Multi Dimensional Audio, is a kind of interpolation between channel- and object-based encoding, with object-based channels mapped to theatrical speakers at installation time. Encoding standards for channel-, scene-, and object-based models were reviewed by Cohen and Villegas (2016). The MPEG-H (https://www.mpegh.com/en/home/) 3D Audio (https://mpeg.chiariglione.org/standards/mpeg-h/3d-audio) and the ITU ADM (Audio Definition Model; https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.2125-0-201901-I!!PDF-E.pdf) standards integrate these models. For typical instance, object-based foreground spatialization can be rendered atop both channel-based stereo (non-diegetic) BGM and scene-based Ambisonic "sweetening" atmospheric background.

Synergies among components arise even for someone alone in a room. Such mutual support includes ducking during voice chats to attenuate backgroundable media; using smartphones and smart speakers to reinforce or articulate cinematic soundtracks and conferencing channels; and using IoT addressability to integrate distributed displays (such as speakers and lights) and sensors (such as microphones and cameras).

Media device orchestration (Francombe et al. 2018) uses ad hoc arrays of appliances to augment apprehension. In the parlance of media presentation, a responsive framework serves dynamic content through an adaptive heterogeneous display. Articulation and comodulation of parameters can coordinate audio and visual displays to accommodate attention, mood, and circumstances. Synchronicity of complementary cross-modal signals—such as moving lips or flashing light, or a map or Gestalt mental model—can disambiguate otherwise indeterminate cues, or even override preliminary interpretation. Confederation of information appliances, sharing data and capabilities, can enhance awareness, expressiveness, and experience.

To recapitulate, conversation, lectures, phone calls, music, television, and announcements inundate us with sonic signals—purely acoustic, electroacoustic, These auditory stimuli comprise overlaid and attentionally oversaturated spatial sound fields, engulfing listeners cacophonously. Sound is mixed acoustically, perceptually, and cognitively—roughly and respectively associated with the air, ear, and brain—corresponding to the three kinds of spatial soundscape superposition described in this chapter, that is, physical transmission (sound), perceptual apprehension (sensation), and procedural interpretation (signal).

Together they span our anticipation for the future of auditory interfaces: heterogeneous, personal and public speakers awarely integrated into multimodal duplex interfaces leveraging idiomatic and metaphorical conventions.

Acknowledgements We thank Yôiti Suzuki for his valuable comments and suggestions. This chapter has been reviewed by two anonymous experts.

References

Alam, S., M. Cohen, J. Villegas, and A. Ahmed. 2009. Narrowcasting in SIP: Articulated privacy control. In *SIP Handbook: Services, Technologies, and Security of Session Initiation Protocol*, ed. S.A. Ahson, and M. Ilyas, 323–345. Boca Raton: CRC Press, Taylor & Francis. Chap. 14. https://doi.org/10.1201/9781315218939.

Bailey, R. 2007. Spatial emphasis of game audio: How to create theatrically enhanced audio. In *Audio Anecdotes III*, ed. K. Greenebaum, and R. Barzel, 399–406. Wellesley: A K Peters/CRC Press. https://doi.org/10.1201/9781439864869.

Bauck, J.L., and D.H. Cooper. 1996. Generalized transaural stereo and applications. *Journal of the Audio Engineering Society* 44 (9): 683–705. http://www.aes.org/e-lib/browse.cfm?elib=7888.

Begault, D.R. 1994. *3-D Sound for Virtual Reality and Multimedia*. Boston: Academic Press. ISBN 978-0120847358

Blauert, J. 2012. A perceptionist's view on psychoacoustics. *Arch. Acoust*. 37 (3): 365–371. https://doi.org/10.2478/v10168-012-0046-z.

Blauert, J. 2017. "Reading the World with Two Ears" Keynote at Int. Congress on Sound and Vibration, London, https://www.youtube.com/watch?v=p1kDtgqmTdw.

Blauert, J., D. Kolossa, K. Obermayer, and K. Adiloğlu. 2013. Further challenges and the road ahead. *Modern Acoustics and Signal Processing*, 477–501. Berlin: Springer. https://doi.org/10.1007/978-3-642-37762-4_18. Chap. 18.

Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge: MIT Press.

Brimijoin, W.O., and M.A. Akeroyd. 2012. The role of head movements and signal spectrum in an auditory front/back illusion. *i-Perception* 3 (3): 179–182. https://doi.org/10.1068/i7173sas.

Broll, W., I. Lindt, I. Herbst, J. Ohlenburg, A.-K. Braun, and R. Wetzel. 2008. Towards next-gen mobile AR games. *Computer Graphics and Animation* 28 (4): 40–48. https://doi.org/10.1109/MCG.2008.85.

Cabrera, D., H. Sato, W. Martens, and D. Lee. 2009. Binaural measurement and simulation of the room acoustical response from a person's mouth to their ears. *Acoustics Australia* 37: 98–103.

Choi, J.-W., B.J. Cho, and I. Shin. 2016. Toward the holographic reconstruction of sound fields using smart sound devices. *IEEE MultiMedia* 23 (3): 64–74. https://doi.org/10.1109/MMUL.2016.46.

Choueiri, E. 2018. Binaural audio through loudspeakers, in Roginska and Geluso. https://doi.org/10.4324/9781315707525. Chap. 5.

Cohen, M. 1993. Throwing, pitching, and catching sound: Audio windowing models and modes. *IJMMS: Journal of Person-Computer Interaction* 39 (2): 269–304. https://doi.org/10.1006/imms.1993.1062.

Cohen, M. 1998. Quantity of presence: Beyond person, number, and pronouns. In *Cyberworlds*, ed. T.L. Kunii, and A. Luciani, 289–308. Tokyo: Springer. https://doi.org/10.1007/978-4-431-67941-7_19. Chap. 19.

Cohen, M. 2000. Exclude and include for audio sources and sinks: Analogs of mute & solo are deafen & attend. *Presence: Teleoperators and Virtual Environments* 9 (1): 84–96. https://doi.org/10.1162/105474600566637.

Cohen, M. 2015. Hierarchical narrowcasting. In *Proceedings of HCII: International Conference on Human-Computer Interaction– DAPI: International Conference on Distributed, Ambient and Pervasive Interactions*, ed. N. Streitz and P. Markopoulos, 274–286. Los Angeles: LNCS 9189. https://doi.org/10.1007/978-3-319-20804-6_25.

Cohen, M. 2016. Dimensions of spatial sound and interface styles of audio augmented reality: Whereware, wearware, & everyware. In *Fundamentals of Wearable Computers and Augmented Reality*, ed. W. Barfield, 277–308. Mahwah: CRC Press. https://doi.org/10.1201/b18703. Chap. 12.

Cohen, M., and O.N.N. Fernando. 2009. Awareware: Narrowcasting attributes for selective attention, privacy, and multipresence. In *Awareness Systems: Advances in Theory, Methodology and Design*, ed. P. Markopoulos and W. Mackay, 259–289. London: Springer. https://doi.org/10.1007/978-1-84882-477-5. Chap. 11.

Cohen, M., O.N.N. Fernando, U.C. Dumindawardana, and M. Kawaguchi. 2009. Duplex narrowcasting operations for multipresent groupware avatars on mobile devices. *IJWMC: International Journal of Wireless and Mobile Computing* 3 (4): 280–287. https://doi.org/10.1504/IJWMC.2009.029348.

Cohen, M., and H. Kojima. 2018. Multipresence and autofocus for interpreted narrowcasting. In *AES: Audio Engineering Society International Conference on Spatial Reproduction—Aesthetics and Science*, Tokyo. http://www.aes.org/e-lib/browse.cfm?elib=19653.

Cohen, M., W.L. Martens. 2020. Spatial soundscape superposition, Part II: Signals and systems. *Acoustical Science and Technology* 41.1 (Jan. 2020). ed. by Masato Akagi, Masashi Unoki, and Yoshifumi Chisaki. JASJ 76 (1): 297–307. ISSN: 1347-5177, 1346-3969, 0369-4232. https://doi.org/10.1250/ast.41.297.

Cohen, M., and J. Villegas. 2016. Applications of audio augmented reality: Wearware, everyware, anyware, & awareware. In *Fundamentals of Wearable Computers and Augmented Reality*, 2nd ed, ed. W. Barfield, 309–330. Mahwah: CRC Press. https://www.taylorfrancis.com/books/9780429192395. Chap. 13.

Collins, K. (ed.). 2008. *Game Sound*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/7909.001.0001. ISBN 978-0-262-03378-7.

Crum, P. 2019. Here come the hearables: Technology tucked inside your ears will augment your daily life. *IEEE Spectrum* 56 (5): 38–43. https://doi.org/10.1109/MSPEC.2019.8701198.

Francombe, J., J. Woodcock, R.J. Hughes, R. Mason, A. Franck, C. Pike, T. Brookes, W.J. Davies, P.J.B. Jackson, T.J. Cox, F.M. Fazi, and A. Hilton. 2018. Qualitative evaluation of media device orchestration for immersive spatial audio reproduction. *Journal of the Audio Engineering Society* 66 (6): 414–429. http://www.aes.org/e-lib/browse.cfm?elib=19581.

Hartmann, W.M. 1999. *Signals, Sound, and Sensation*. New York: AIP Press.

Herder, J., and M. Cohen. 2002. The helical keyboard: Perspectives for spatial auditory displays and visual music. *JNMR: Journal of New Music Research* 31 (3): 269–281. https://doi.org/10.1076/jnmr.31.3.269.14180.

Hugonnet, C., and P. Walder. 1998. *Stereophonic Sound Recording: Theory and Practice*. Chichester: Wiley. ISBN 978-0471974871.

Inoue, A., Y. Ikeda, K. Yatabe, and Y. Oikawa. 2017. Three-dimensional sound-field visualization system using head mounted display and stereo camera. In *Proceedings of ASA Meetings on Acoustics*, vol. 29. https://doi.org/10.1121/2.0000381.

Jekosch, U. 2005. Assigning meaning to sounds—semiotics in the context of product-sound design. In *Communication Acoustics*, ed. J. Blauert. Berlin: Springer. https://doi.org/10.1007/3-540-27437-5_8. Chap. 8.

Jo, H., W.L. Martens, Y. Park, and S. Kim. 2010. Confirming the perception of virtual source elevation effects created using 5.1 channel surround sound playback. In *VRCAI: Proceedings of International Conference on Virtual-Reality Continuum and Its Applications in Industry*, 103–110. Seoul: ACM. https://doi.org/10.1145/1900179.1900200.

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

Kawaura, J., Y. Suzuki, F. Asano, and T. Sone. 1991. Sound localization in headphone reproduction by simulating transfer functions from the sound source to the external ear. *Journal of the Acoustical Society of Japan (E)* 12 (5): 203–216. https://doi.org/10.1250/ast.12.203.

Kendall, G. 2010. Spatial perception and cognition in multichannel audio for electroacoustic music. *Organised Sound* 15 (3): 228–238. https://doi.org/10.1017/S1355771810000336.

Kim, C., R. Mason, and T. Brookes. 2013. Head movements made by listeners in experimental and real-life listening activities. *Journal of Audio Engineering Society* 61 (6): 425–438. http://www.aes.org/e-lib/browse.cfm?elib=16833.

Kleiner, M., B.-I. Dalenbäck, and P. Svensson. 1993. Auralization— an overview. *Journal of Audio Engineering Society* 41 (11): 861–875. http://www.aes.org/e-lib/browse.cfm?elib=6976.

Kondo, H.M., D. Pressnitzer, I. Toshima, and M. Kashino. 2012. Effects of self-motion on auditory scene analysis. *Proceedings of the National Academy of Sciences* 109 (17): 6775–6780.

Lackner, J.R. 1977. Induction of nystagmus in stationary subjects with a rotating sound field. *Aviation, Space and Environmental Medicine* 48 (2): 129–131.

Lackner, J.R. 1983. Influence of posture on the spatial localization of sound. *Journal of Audio Engineering Society* 31 (9): 650–661. http://www.aes.org/e-lib/browse.cfm?elib=18987.

Leung, J., D. Alais, and S. Carlile. 2008. Compression of auditory space during rapid head turns. *Proceedings of the National Academy of Sciences* 105 (17): 6492–6497. https://doi.org/10.1073/pnas.0710837105.

Lossius, T., P. Baltazar, and T. de la Hogue. 2009. DBAP—Distance-based amplitude panning. In *Proceedings of the International Computer Music Conference, ICMC*, (Aug. 16–21, 2009) Montréal, Quebec, Canada. https://hdl.handle.net/2027/spo.bbp2372.2009.111.

Lyon, E., ed. 2016. *Computer Music J.: High-Density Loudspeaker Arrays, Part 1: Institutions*, 40, https://doi.org/10.1162/COMJ_e_00388.

Lyon, E., ed. 2017. *Computer Music J.: High-Density Loudspeaker Arrays, Part 2: Spatial Perception and Creative Practice*, 41, https://doi.org/10.1162/COMJ_a_00403.

Macpherson, E.A. 2013. Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation. In *Proceedings of Meetings on Acoustics*, Vol. 19, p. 050131. https://doi.org/10.1121/1.4799913.

Martens, W., S. Sakamoto, L. Miranda, and D. Cabrera. 2013. Dominance of head-motion-coupled directional cues over other cues during walking depends upon source spectrum. In *Proceedings of Meetings on Acoustics*, Vol. 19, p. 050129. https://doi.org/10.1121/1.4800124.

Martens, W.L., D. Cabrera, and S. Kim. 2011. The ‘phantom walker’ illusion: Evidence for the dominance of dynamic interaural over spectral directional cues during walking. In *Principles and Applications of Spatial Hearing*, ed. Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato, 81–102. Singapore: World Scientific. https://doi.org/10.1142/7674.

Martens, W.L., and M. Cohen. 2020. Spatial soundscape superposition, Part I: Subject motion and scene sensibility. In *Acoustical Science and Technology* 41.1 (Jan. 2020). ed. by Masato Akagi, Masashi Unoki, and Yoshifumi Chisaki. JASJ 76 (1): 288–296. ISSN: 1347-5177, 1346-3969, 0369-4232. https://doi.org/10.1250/ast.41.288.

Martens, W.L., Y. Han. 2018. Discrimination of auditory spatial diffuseness facilitated by head rolling while listening to ‘with-height’ versus ‘without-height’ multichannel loudspeaker reproduction. In *Proceedings of Audio Engineering Society International Conference on Spatial Reproduction*, Tokyo. http://www.aes.org/e-lib/browse.cfm?elib=19608.

Marui, A., and W.L. Martens. 2006. Spatial character and quality assessment of selected stereophonic image enhancements for headphone playback of popular music. In *AES: Audio Engineering Society Conv. (*120th *Conv.)*, Paris. http://www.aes.org/e-lib/browse.cfm?elib=13622.

Middlebrooks, J.C., J.Z. Simon, A.N. Popper, and R.R. Fay (eds.). 2017. *The Auditory System at the Cocktail Party*. Cham: Springer. https://doi.org/10.1007/978-3-319-51662-2.

Milgram, P., and H. Colquhoun Jr. 1999. A taxonomy of real and virtual world display integration. In *Mixed Reality: Merging Real and Virtual Worlds*, ed. Y. Ohta and H. Tamura, 5–30. Omsha: Springer. Chap. 1. ISBN 978-3-642-87514-4

Ochiai, Y., T. Hoshi, and I. Suzuki. 2017. Holographic whisper: Rendering audible sound spots in three-dimensional space by focusing ultrasonic waves. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, 4314–4325. New York. https://doi.org/10.1145/3025453.3025989.

Pastore, M.T., Y. Zhou, and W.A. Yost. 2020. Cross-modal and cognitive processes in sound localization. In *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 315–350. Cham, Switzerland: Springer. Chap. 12. https://doi.org/10.1007/978-3-030-00386-9_12.

Pereira, F., and W.L. Martens. 2018. Psychophysical validation of binaurally processed sound superimposed upon environmental sound via an unobstructed pinna and an open-ear-canal earspeaker. In *Proceedings of Audio Engineering Society International Conference on Spatial Reproduction*, Tokyo. http://www.aes.org/e-lib/browse.cfm?elib=19626.

Perrott, D.R., H. Ambarsoom, and J. Tucker. 1987. Changes in head position as a measure of auditory localization performance: Auditory psychomotor coordination under monaural and binaural listening conditions. *Journal of the Acoustical Society of America* 82 (5): 1637–1645.

Plenge, G. 1974. On the difference between localization and lateralization. *Journal of the Acoustical Society of America* 56: 944–951. https://doi.org/10.1121/1.1903353.

Pulkki, V. 1997. Virtual source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society* 45 (6): 456–466.

Pulkki, V., T. Lokki, and D. Rocchesso. 2011. Spatial effects. In *DAFX: Digital Audio Effects*, 2nd ed, ed. U. Zölzer, 139–184. West Sussex: Wiley. https://doi.org/10.1002/9781119991298.ch5. Chap. 5.

Ranaweera, R., M. Cohen, and M. Frishkopf. 2015. Narrowcasting and multipresence for music auditioning and conferencing in social cyberworlds. Presence: Teleoperators and Virtual Environments 24 (3): 220–242, https://doi.org/10.1162/PRES_a_00232.

Roginska, A., and P. Geluso (eds.). 2018. *Immersive Sound: The Art and Science of Binaural and Multi-channel Audio*. Routledge: Taylor & Francis. https://doi.org/10.4324/9781315707525.

Satongar, D., C. Pike, Y.W. Lam, and A.I. Tew. 2015. The influence of headphones on the localization of external loudspeaker sources. *Journal of the Audio Engineering Society* 63 (10): 3–19. https://doi.org/10.17743/jaes.2015.0072.

Seldess, Z. 2014. "MIAP: Manifold-interface Amplitude Panning in Max/MSP and Pure Data" in *Audio Engineering Society Convention 137*, Los Angeles, http://www.aes.org/e-lib/browse.cfm?elib=17435.

Shaw, E.A.G. 1982. 1979 Rayleigh medal lecture: The elusive connection. In *Localization of Sound: Theory and Applications*, ed. R. Gatehouse, 13–29. Groton: Amphora Press.

Sodnik, J., and S. Tomažič. 2015. *Spatial Auditory Human-Computer Interfaces*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-22111-3.

Streicher, R., and F.A. Everest. 2006. *The New Stereo Soundbook*, 3rd ed. Pasadena: Audio Engineering Associates. ISBN 978-0-9665162-1-0.

Sullenberger, R.M., S. Kaushik, and C.M. Wynn. 2019. Photoacoustic communications: Delivering audible signals via absorption of light by atmospheric H_2O. *Optics Letters* 44 (3): 622–625. https://doi.org/10.1364/OL.44.000622.

Suzuki, Y., A. Honda, Y. Iwaya, M. Ohuchi, and S. Sakamoto. 2020. Binaural display supporting active listening: Perceptual bases and welfare applications initial proposal: Training of spatial perception with binaural displays supporting active listening. In *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 665–695. Cham, Switzerland: Springer and ASA Press. Chap. 22. https://doi.org/10.1007/978-3-030-00386-9_22.

Tachi, S. 2015. *Telexistence*, 2nd ed. Singapore: World Scientific Publishing Company.

Tanno, K., A. Saji, and J. Huang. 2014. A 3D sound generation system with horizontally arranged five-channel loudspeakers. *IEICE Transactions on Information and Systems* 2 (J97-D(5)): 1044–1052.

Thurlow, W.R., J.W. Mangles, and P.S. Runge. 1967. Head movements during sound localization. *Journal of the Acoustical Society of America* 42 (2): 489–493. https://doi.org/10.1121/1.1910605.

Thurlow, W.R., and P.S. Runge. 1967. Effect of induced head movements on localization of direction of sounds. *Journal of the Acoustical Society of America* 42 (2): 480–488. https://doi.org/10.1121/1.1910604.

Toole, F.E. 1969. In-head localization of acoustic images. *Journal of the Acoustical Society of America* 48: 943–949. https://doi.org/10.1121/1.1912233.

Villegas, J., and M. Cohen. 2010. `Hrir`: Modulating range in headphone-reproduced spatial audio. In *VRCAI: Proceedings of International Conference on Virtual-Reality Continuum and Its Applications in Industry*, Seoul. https://doi.org/10.1145/1900179.1900198.

Wallach, H. 1940. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology* 27: 339–368. https://doi.org/10.1037/h0054629.

Wang, D. 2017. Deep learning reinvents the hearing aid. *IEEE Spectrum* 54 (3): 32–37. https://doi.org/10.1109/MSPEC.2017.7864754.

Wenzel, E.M., D.R. Begault, and M. Godfroy-Cooper. 2018. Perception of spatial sound. In Roginska and Geluso (2018), 5–39. https://doi.org/10.4324/9781315707525. Chap. 1.

Wersényi, G., and J. Wilson. 2015. Evaluation of head movements in short-term measurements and recordings with human subjects using head-tracking sensors. *Acta Technica Jaurinensis* 8 (3): 218–229.

Wolfe, J. 2018. From idea to acoustics and back again: the creation and analysis of information in music. Substantia 1: 77–91. https://doi.org/10.13128/Substantia-42.