Spatial soundscape superposition, Part I: Subject motion and scene sensibility

William L. Martens$^1$,* and Michael Cohen$^2$†

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

Abstract: Spatial soundscape superposition occurs whenever multiple sound signals impinge upon a human listener’s ears from multiple sources, as in augmented reality displays that combine natural soundscapes with reproduced soundscapes. Part I of this two-part contribution on spatial soundscape superposition regards perceptual superposition of soundscapes, and therefore focuses upon human response to displayed auditory scenes, and the influence of subject (listener) motion on making sense of them in the context of information received from other sensory systems, especially the visual and vestibular systems. Consideration of listener motion and multimodal integration here is intended to lay the foundation for Part II of this contribution, which focuses upon physical stimuli, i.e., sounds and signals, and the systems used to mix, transmit, and display them. Through superposition of complex sound stimuli at the ears of a moving listener, these systems create complex auditory scenes, the nature of which cannot be predicted by simple combination of physical stimuli. As it is left to the human listener to interpret auditory scenes comprising those stimuli, this part focuses upon perceptual principles, such as grouping principles, that can aid in successfully predicting whether multiple auditory events are perceptually segregated or fused in the auditory scenes that are experienced. Furthermore, for moving listeners, four fundamental laws are identified here describing sensorimotor contingencies that enable prediction not only of what auditory images are formed, but also where in auditory space those images are likely to be perceived.

Keywords: Binaural hearing, Localization of sound sources, Auditory displays, Models and theories of auditory processes, Multimodal integration

PACS number: 43.60.Jn, 43.60.Lq, 43.66.Ba, 43.66.Pn, 43.66.Qp [doi:10.1250/ast.41.288]

1. INTRODUCTION

Spatial soundscape superposition occurs whenever multiple sound signals impinge upon the ears of human listeners from spatially distributed sources, and that superposition offers listeners the opportunity to resolve the spatial distribution of those sources within an acoustical environment, whether those sound sources occur within natural or within reproduced soundscapes. As there is no universal agreement among authors regarding the use of the term “soundscape,” some clarification is in order at the outset of this paper. First, it should be noted that there is an ISO standard [1] that provides both a definition and a conceptual framework for use of the term: “soundscape” is used to denote a perceptual construct that is related to a physical phenomenon (the acoustic environment), but distinguished from it. While some authors may use the term “soundscape” to refer to the physical acoustic environment that can be recorded or captured in a given place, review of the related nomenclature found in [2] clearly supports the framework promoted in the ISO standard, which suggests that a soundscape exists through human perception of the acoustic environment.

For the purposes of the current review, the term “spatial soundscape” is used to indicate not only a spatially distributed collection of sound stimuli perceived by a human listener in a particular place, but also the evoked apprehension of such stimuli, influenced by a spatial-perceptual context that shapes a listener’s interpretation of those sounds. Human listeners in real or virtual environments are actively engaged with the environment, including implicit “hypothesis testing,” involving overt actions such as locomotion that result in changing stimulation via multiple sensory modalities, as well as through more nuanced actions such as “idling,” postural swaying that subtly enriches audition.

Next, the focus of the current review on superposition should be clarified, as there are several types of superposition that might be addressed here. Part I of this two-
part contribution on spatial soundscape superposition focuses upon perceptual experiences (sensations) associated with such spatial soundscapes, particularly via their presentation to moving listeners (subjects). The consideration of listener motion and multimodal integration here is intended to lay the foundation for a treatment of procedural considerations (signals and systems) contained in an associated companion paper (“Part II” of this contribution [3]). A broader review of spatial soundscape superposition is presented in a forthcoming book chapter by the same authors [4], which includes a deeper exposition on acoustic superposition (sound) in the context of binaural technology and spatial sound reproduction.

Superposition of complex sound stimuli at the ears of a moving listener creates complex auditory scenes, the perception of which cannot be predicted only by simple combination of physical stimuli at a given instant. Prediction requires consideration of the changing pattern of stimulation that occurs at the ears of a binaural listener whenever that listener moves, even if the movements are only slight turns of the head made while auditioning displays of sound within an acoustic environment. A number of telling examples are presented in the following section of this paper, with treatment of cases involving the formation of binaural images and perception of spatial soundscapes given single sound sources, and also for multiple spatially distributed sound sources.

2. SUBJECT MOTION AND SCENE SENSIBILITY

This section surveys the complexity of binaural image formation under conditions in which listeners move relative to sound sources, as suggested by simultaneously received sensory input through multiple modalities, including not only auditory, but also visual and vestibular systems. Of course, the experience of self motion may also be influenced by “outflow” of information from the brain indicating initiation of voluntary motion, as well as “inflow” of sensory information to the brain. Indeed, the ability to predict the sensory consequences of self-initiated actions may be regarded as necessary for making sense of a scene. Consideration of “physical superposition” only scratches the surface of perceptual issues underlying deployment of interactive binaural technology for applications involving spatial soundscape superposition.

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 [5]). Because spatial 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 spatial hearing studies incorporating listener movement, such as listening while walking [6]. 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 (for example, listener expectations) operate during listener movement by disambiguating raw sensations through implicit hypothesis testing, such as that associated with Neisser’s Perceptual Cycle Model [7]. The important point made by Neisser is that perception and cognition are not just operations in the mind, but also transactions with the world. The model holds that the world is full of information that is potentially available to an individual through perceptual exploration driven by the individual’s constructed map of the world.

In the companion paper [3], such considerations are extended to address how auditory scenes are mentally constructed in the context of potentially abstract thoughts and concepts associated with procedural superposition. In preparation for that topic, this paper addresses key perceptual processes underlying spatial soundscape superposition, those being binaural image formation and scene sensibility through subject motion. First, the process of binaural image formation is discussed, revealing the complexity of this perceptual process.

2.1. Binaural Image Formation: Perceptual Fusion and Auditory Events

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 presented with the intention of creating an integrated unitary percept will not necessarily be successful, so principles of fusion and fission should be examined.

Under typical binaural listening conditions, when the sound of an external acoustic event impinges upon the ears of a human listener, an externalized auditory image of the sounding object typically results. Such an auditory image,
heard as occurring outside the listener’s head, 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 this discussion of binaural image formation, it is essential that this distinction between acoustic and auditory events 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, which is a mentally constructed space in which auditory events can occur, but that also exhibits auditory attributes characterizing the space itself. If the auditory space is heard as reverberant, then the auditory objects are usually heard as externalized, which means that they are mentally projected into psychological constructions of the reverberant environment in which listeners perceive themselves to be located.

In the context of this discussion on soundscape superposition, understanding principles underlying binaural image formation is key to linking physical and perceptual superposition. This is not a new idea. Plenge proposed that a sound stimulus should form a coherent auditory image if the natural processes of spatial hearing are to be engaged. His model stressed [8, p. 950] that the process of 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 a simple stimulus might be heard to be within the listener’s head (“IHL”: inside-the-head localization or locatedness), under which circumstances Plenge [8] would term the task lateralization rather than localization. Even when broadband binaural stimuli are employed, there is no guarantee that externalization and coherent auditory imagery result [9].

Consider the auditory image associated with the binaural presentation of a musical note played on a marimba. Even when a high-quality microphone captures a dry but perfectly realistic sounding marimba performance, and then that signal is transformed for headphone presentation through the 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 overtones of the note that decay more rapidly (and correspond to the brief “strike tone,” rather than the more slowly decaying resonance corresponding to the nominal pitch of the marimba note).

Now consider what is possible if a series of marimba notes is performed (as in a roll), and the presentation involves headphone display incorporating active head-tracking. With low-latency coupling, coordinated lateral shifts in all the tone’s partials that accompany voluntary head-turning increase the likelihood of perceptual fusion of those partials. Then, if the presentation includes an effective (i.e., spatially realistic) binaural simulation of indirect sound, binaural imagery of the marimba notes is likely heard as both unified and externalized. It is tempting to propose that a Gestalt principle could be operating in this case, where the fundamental frequency that normally segregates perceptually from the strike tone of each note might be integrated based upon the ‘common fate’ of all partials as they shift together laterally as a consequence of head-turning. Whereas in free-field conditions it seems reasonable to elicit a listener’s report of the direction and distance of the marimba as a sounding object in physical space, if a spatially static and dry marimba note is presented via headphones, it is not so easy for listeners to make the same sort of subjective estimate when partials are segregated perceptually.

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. In the early 1980s, Shaw 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 [10]. 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. With an emphasis on spatially static sources and listeners, many reported research results have contributed less to practical applications of binaural technology than desired. The philosophical underpinnings of the above issues are well addressed in a paper by Blauert [11] that introduces into this discussion the concept of “Perceptionism.” The perceptionist’s approach to psychoacoustics is also a perspective on methods used in evaluating the effectiveness of binaural technology, emphasizing methods that should benefit those engaged in optimizing spatial auditory display in real-world applications (rather than in research laboratories). It is fortunate that there has been renewed interest recently in interactive binaural displays, associated with emphasis on multimodal sensory integration in spatial hearing research, including focus upon the influence of visual, proprioceptive, and vestibular sensations.
2.2. 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 listener position. There are ample sources of information for active localization available via sensorimotor contingencies, which for moving listeners can vary along six dimensions, three dimensions describing changing location (a.k.a. translation), and three describing changing orientation (rotation, as tabulated in Table 1). The focus of this section is on the relative importance of the dynamic interaural cues to source location, in comparison to pinna-based spectral cues. A particularly telling result in this regard was found in experiments in which listeners were fitted with pseudophonic displays that swap left-ear and right-ear signals. When those listeners, with effectively interchanged ear signals, walked through an environment attempting to localize naturalistic sound sources such as speech sounds, head-motion-coupled variation in interaural directional cues appeared to dominate other cues [6]. If, however, sources containing more high-frequency content than speech were presented, there was less dominance of head-motion-coupled changes in low-frequency interaural cues over spectral cues associated with the pinna [12]. In fact, directional ambiguities can result from the cue conflict that results from such pseudophonic displays when broadband noise bursts comprise the source stimuli to be localized [13], and segregation of the frequency content of the noise stimuli into separate auditory images also can occur [14].

However, when speech is the stimulus, spectral fusion is much more likely, and continuous changes in orientation of the head during walking (head-turning, tipping, and rolling) contribute to strong auditory localization illusions that are hard to negate, even when the mouth of the talker is clearly visible. Despite having in clear view the stationary talker who produces the speech sound, as listeners walked toward the talker (in which case 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. Indeed, when listeners with interchanged ear signals were asked to walk past a continuously viewed speech source emanating from a fixed spatial position, the source was heard to be located in a spatial region that was reversed with respect to all three spatial axes: left for right, front for back, and above for below (as explained in [6]). For naturalistic sound sources that form unified auditory images, these head-motion-coupled interaural cues are so strong that they defeat the contradictory “pinna-based” directional cues, and also the visual cues (associated with 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 to create illusions of frontward incidence [15], through a reversal of interaural cues accompanying head-turning. While these results replicate those of the classic study by Wallach [16], a related, but potentially surprising result was found when walking listeners rolled their heads while listening to speech sounds arriving from elevated loudspeakers in an analogous reversal of interaural cues accompanying rotations of the head [6]. The increased variation in head-roll angle that occurs during walking is illustrated in Fig. 1.

Just as front–back reversals are associated with pseudophonic treatment during head-turning [12], above–below reversals were shown to be associated with pseudophonic treatment during head-rolling. This result argues for the dominance of dynamic interaural cues over spectral directional cues [17], at least for speech sounds containing power mostly below 5 kHz. In contrast, when sources containing more high-frequency energy are presented, presumably allowing pinna-based spectral cues greater influence on binaural image formation, the rate of such illusory reversals is greatly reduced [12,13].

Of course, head-motion-coupled directional cues do not 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 readily measurable amounts [18], and they seem to move their heads just as much when engaged in natural listening activities, such as watching movies [19]. These recent

| 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” |

Table 1  Angular motions of the head (“cocking”).
studies of motion-based influences on binaural perception of auditory direction are preceded by important earlier studies. In introducing the topic of such influences on binaural perception, Lackner noted [20] in the early 1980s 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.

Also, when sound sources are presented via headphones without adjusting the displayed sound signal via active head-tracking data, the natural contribution of head movement to sound localization is precluded [17]. It is also interesting to note that when sound sources are presented via multiple loudspeakers, signals that reach the ears are modified in ways that generally do not match the ways in which those signals vary with respect to single sounding objects. Nonetheless, proper head-motion-coupled directional cues (which are naturally superimposed upon incoming sound by a listener’s own ATFs) can be experienced when a large number of loudspeakers are deployed around a listening position, and both sources and simulated indirect sound components are each delivered from a unique and singular loudspeaker (i.e., with no panning or other multi-loudspeaker synthesis). Systems using this approach, which have been termed “discrete multichannel simulation” systems, are useful when auditory imagery should be robust in the presence of listener head movements [21]. For example, when accurate auralization is required for a yet-to-be-constructed acoustical enclosure, it is important to have virtual acoustic components modified by listeners’ own acoustical characteristics (their ATFs), this being the most natural variation with head motion, resulting in robust imagery given rotation with three degrees of freedom, as well as translation across multiple listening positions.

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. Most notable was early work by Wallach [16], 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 the listener. He noted that when the head was turned to the left, a frontal stationary sound source was shifted towards the right, and a dorsal source was shifted towards the left. Thus forward directions could be distinguished from rearward directions on the basis of head-motion-coupled changes in lateral angle.

In order to establish the dominance over pinna cues of these dynamic interaural cues that depend upon head-turning, he constructed a mechanism that presented sound stimuli such that pinna cues and movement cues indicated conflicting angles of incidence. The lateral angle of a frontal sound stimulus was made to shift with head movement as it would for a rearward incidence angle as follows: Loudspeakers in a horizontal row spanning 120° in front of a listener were switched on and off in response to movements of the listener’s head such that a 3° turn of the head to the left caused a shift to a loudspeaker 6° to the left of the previously sounding loudspeaker. Listeners were blindfolded and their heads firmly positioned on a chin rest, while they were rotated passively by the experimenter (affording primarily vestibular cues to self-rotation). All listeners reported that the sound was spatially stationary and located to the rear, despite the fact that it arrived from frontal angles while shifting to the left at twice the angular velocity of the head rotation.

In a subsequent experiment, Wallach presented such dynamic sound stimuli under conditions in which an illusion of self-rotation was induced by placing a stationary subject 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 (i.e., they experiencedvection, the illusion that they moved and that the visual world was
stationary). Again, a front–back reversal was experienced because of an imposed mismatch between apparent head rotation and the lateral shift of the sound source. As subjects experienced this illusory self-rotation, the direction of the sound remained fixed in relation to the apparently stationary visual environment, so they felt themselves passing the sound as they turned. The “principle of rest” [16] describes this and related phenomena: “Of all the directions which realize the given sequence of lateral angles, that one is perceived which is covariant with the general content of surrounding space.”

Another illuminating result was also reported when the revolving screen induced an illusion of self-rotation and the angle of the source was not varied over time. The sound emitted by a stationary loudspeaker located in front of the subject (but hidden behind the screen) was heard to be located directly above the listener’s head, presumably because of its fixed and laterally centered angle in the presence of apparent head rotation. These conditions mimic the natural phenomenon in which head-turning provides information regarding elevation of a sound source. Wallach observed that the further from the horizontal plane a sound source is located, the smaller the shift in lateralization of the auditory image with a constant angular rotation of the head. Indeed, when a sound source is located directly above or below the head (at the zenith or nadir), no lateral shift of auditory imagery accompanies head-turning. He demonstrated that when the lateral shift typically associated with head-turning was reduced artificially using the apparatus described above, listeners typically reported that the sound source was elevated toward the zenith, rather than shifted downwards towards the nadir. What was missing from that narrative was a recognition of the four fundamental sensorimotor contingencies that come into play here, two of which can serve to disambiguate source direction in terms of the hemifields of origin (not just the front versus the rear hemifield, but also the above versus the below hemifield), and two of which can serve to determine how far removed a source is from a horizontal or vertical plane (as illustrated in Fig. 2).

In another relatively early study, Thurlow and Runge [22] 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 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 early paper [22] was called ‘pivoted,’ as tabulated in Table 1), 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 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.

In order to determine how frequently human listeners use different types of head movements in sound source localization, Thurlow et al. [23] observed over 50 subjects during a free-field localization task. First, they noted that most subjects always included some amount of head rotation in exploration of sound stimuli. Subjects rotated their heads 42° on the average when a sound source was located at high elevation, but only an average of 29° when a source was at lower elevation. The type of head movement observed most often was a combination of rotation (yaw) and tipping (pitch). The second most common type of head movement observed was rotation without tipping or pivoting. (A movement was placed in this category when the extent of both tipping and pivoting was less than 3°.) Third most common was a combination of rotation, tipping, and pivoting, and fourth was rotation with pivoting. The relatively high frequency of the fourth type of movement is somewhat surprising because rotation and pivoting produce confounding lateral shifts of the auditory image that should not disambiguate direction, due to the uncertain interpretation of binaural changes accompanying combined rotation and pivot movements.

In a more recent study of immersive spatial impression [15], the relative influence of tipping and pivoting was reported for a spatial soundscape created using many simultaneous sources, rather than the single source considered in [23]. Furthermore, the spatial soundscape had separable components associated with two sets of spatially distributed sources, one of which could be superposed upon the other. Presentation of multichannel program material that included an elevated soundscape (which was presented via a 10-channel array of loudspeakers that included ‘height channels’) produced a sense of auditory spatial diffuseness comparable to that of a more truly diffuse stimulus presented using twice as many loudspeakers, distributed about a hemispherical array. 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., reproduction employing loudspeakers located only on a single plane near the listener’s ear level). However, this 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. In effect, the elevated portion of the soundscape could only be “heard out” under these conditions.

Performance on a two-alternative forced choice (2AFC) discrimination task was significantly improved when listeners rolled their heads from side to side. Discrimination between the diffuse reference and either of the two 10-channel loudspeaker configurations was impossible when listeners held their heads stationary. Furthermore, if listeners pitched their heads up and down while listening to the pairs of stimuli to be compared, no clear evidence for diffuseness discrimination could be found. This result is consistent with the general hypothesis that it is the head-rotation-coupled lateral shifts of auditory images along the interaural axis that allow potentially ambiguous source directions to be disambiguated. This of course has been well established and replicated for front–back disambiguation ever since the influence of head-turning was reported by Wallach [16]. It is not so well known, however, that the analogous above-below disambiguation is made possible by head-rolling-coupled lateral shifts of auditory images along the interaural axis, as demonstrated in [6]. Head pitching cannot produce analogous changes in lateralization for static sources that are removed from the horizontal plane.

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. Under these circumstances, HRTF-based binaural technology does not produce a satisfying perceptual result, since the greatest benefits of head-tracked processing are frustrated when no head-motion-coupled variation in lateral angle is presented. Conversely, when static HRTF-processed sounds are presented with head-motion-coupled dynamic changes in interaural time difference (ITD), listeners typically report “natural” sound localization [17], and front-rear confusion is greatly reduced by head-turning of the subject, especially for sources near the median plane. In fact, the head-motion-coupled changes in ITD were included in that study of static HRTF-processed sound [17] primarily to ensure that directional percepts remained in the front hemifield for frontward HRTFs, and in the rear hemifield for rearward HRTFs.

3. HEMIFIELD DISTINCTIONS AND SENSORIMOTOR CONTINGENCIES

Auditory spatial perception can be understood as the activity of exploring an environment in ways mediated by tacit knowledge of the four fundamental sensorimotor contingencies outlined below. Prior to delineation of those sensorimotor contingencies, it is helpful to review the sets of iso-angular contours presented in Fig. 2. This apparently tangential presentation of iso-angular contours and spherical coordinates is important for the exposition of the fundamental sensorimotor contingencies in that there are confusions in the literature that stem from a failure to distinguish between two different forms of azimuth angle measurements, those being vertical-polar (VP) azimuth and double-polar (DP) azimuth angles (which in this latter case should more properly be termed lateral angles).

A spherical double-polar (DP) coordinate system is shown in Fig. 2 by which sound source directions are specified using a lateral angle and an elevation angle, both of which range from $-90^\circ$ to $+90^\circ$. According to the predominant convention in the literature, the elevation angle is specified relative to the horizontal plane (also called the visuoaural plane, which contains the interaural axis). Against the predominant vertical-polar (VP) convention, this DP elevation angle here is coupled with a DP lateral angle, rather than a VP azimuth angle. The reason...
for use of a DP coordinate system here is to support more adequately the below explanation of sensorimotor contingencies underlying active localization. However, before launching into that, the coordinate system itself is briefly introduced here.

Figure 2(a) plots circular lines that depict the intersections between a series of sagittal planes and the surface of a sphere, with the listener situated at its center. These lines define iso-lateral-angle contours for lateral angles that are offset in 15° steps from the 0° lateral angle of the median sagittal plane. Unlike the azimuth angle used in the conventional VP coordinate system, the use of a DP lateral angle denotes a directional component that makes no distinction between source incidence from within the frontward hemifield versus the rearward hemifield. This feature is consistent with the potential ambiguity of static binaural cues to frontward versus rearward incidence angles, in that interaural time and level differences are very nearly constant for sound sources that arrive from directions sharing the same lateral angle.

Figure 2(b) plots circular lines that depict iso-elevation-angle contours for elevation angles that are offset in 15° steps from the 0° elevation angle of the visuoaural (horizontal) plane. In both the VP and DP coordinate systems, elevation angles make no distinction between source incidence from left versus right, or from front versus rear, matching the longitudinal angles of the globe (which define full circles of constant latitude, termed the parallels). However, lateral angles do not match the latitudinal angles of the globe as do azimuth angles. Whereas VP azimuth varies over a 360° range regardless of source elevation (which define semicircles of constant longitude, termed the meridians), lateral angles vary over a decreasing range as source elevation increases.

To illustrate this elevation dependence, Fig. 2(c) plots lateral angle as a function of VP azimuth angle, with elevation angle as the parameter of the graph. For sources arriving from directions on the visuoaural plane (i.e., at elevation 0°), lateral angle and VP azimuth angle are of identical magnitude. However, as source elevation increases, measured lateral angle reaches a degenerate point for sources incident from directly above (at elevation 90°). Of course, an inverted graph would show the same form of dependence as source elevation decreases from 0° to −90° (i.e., the extent of variation in lateral angle reduces from a maximum at ear level to a minimum at the point directly below the listener’s head).

Contemplation of the fore-mentioned elevation dependence should reveal two sorts of sensorimotor contingencies that might aid listeners in actively localizing sound sources in the world around them: those that make unsigned distinctions in source incidence angles, and those that make hemifield distinctions, such as the above–below distinctions enabled by head-rolling. What is made clear in the following delineation is that there are two sets of analogous contingencies useful in determining source direction, and these are proposed here as governing laws underlying head-movement-based spatial hearing. The first two sensorimotor contingencies enable hemifield distinctions:

1) Variation in the yaw angle of the head (turning) produces variation in the lateral angle of incidence of a static sound source in a manner that depends on, and therefore can determine, whether the source originates in the frontward or rearward hemifield.

2) Variation in the roll angle of the head (rolling) produces variation in the lateral angle of incidence of a static sound source in a manner that depends on, and therefore can determine, whether the source originates in the upward or downward hemifield.

The second two sensorimotor contingencies enable assessment of displacement from the planes dividing those hemifields:

1) Variation in the yaw angle of the head produces variation in the lateral angle of incidence of a static sound source in a manner that indicates how far from the horizontal plane the source is displaced (the unsigned magnitude of displacement, which is ambiguous regarding whether the displacement is upward or downward).

2) Variation in the roll angle of the head produces variation in the lateral angle of incidence of a static sound source in a manner that indicates how far from the frontal plane the source is displaced (the unsigned magnitude of displacement, which is ambiguous regarding whether the displacement is frontward or rearward).

4. CONCLUSIONS

Results of reviewed recent and classic experiments indicate bidirectional interaction between perception of head orientation and auditory spatial perception. The results do not suggest that HRTFs are not important for successful application of binaural technology. Rather, they imply that deployed HRTFs should be tested using methods that recognize the importance of listener interaction with the environment (as in [17]). Application of perceptual principles, such as grouping principles, can aid in successfully predicting whether multiple auditory events are perceptually segregated or fused in the auditory scenes that are experienced. Furthermore, principles based upon multimodal integration, particularly for moving listeners, aid in predicting not only what auditory images are formed, but also where in space those auditory images are likely to be localized. For moving listeners actively engaged in sound localization, these principles have been presented as governing laws that underly the use of head-movement-based dynamic interaural cues in spatial hearing based upon sensorimotor contingencies.
These conclusions have implications for the understanding of spatial soundscape superposition. Most importantly, practitioners should be aware of the contrast between content designed according to physical principles versus content designed according to focus upon perceptual constructs that characterize soundscapes. These perceptual constructs are sensitive to listener motion in ways described above. Therefore, when spatial soundscape are superposed for moving listeners, the four identified governing laws describing sensorimotor contingencies must be heeded in the design of user interaction, since the tracking of some dimensions, such as head-turning, while ignoring other dimensions, such as head-rolling, may disable desired hemifield distinctions, resulting in less effective display of superposed spatial soundscapes. This point was exemplified in the results of a study of immersive spatial impression [24] in which a spatial soundscape was created using many simultaneous sources. Based upon that study’s results, it was concluded that listeners rolling their heads could discriminate an increase in spatial diffuseness when elevated components were added to a reproduced soundscape, while no clear evidence for diffuseness discrimination could be found for listeners who pitched their heads up and down, or for listeners who held their heads still.

REFERENCES

[1] ISO 12913-1: 2014, Acoustics — soundscape — part 1: Definition and conceptual framework. Technical Report, International Organization for Standardization (2014).
[2] A. L. Brown, T. Gjestland and D. Dubois, “Acoustic environments and soundscapes,” in Soundscape and the Built Environment, J. Kang and B. Schulte-Fortkamp, Eds. (CRC Press, Boca Raton, 2016), pp. 1–17.
[3] M. Cohen and W. L. Martens, “Spatial soundscape superposition, Part II: Signals and systems,” Acoust. Sci. & Tech., 41, 297–307 (2020).
[4] M. Cohen and W. L. Martens, “Spatial soundscape superposition and multimodal interaction,” in The Technology of Binaural Understanding, J. Blauert and J. Braasch, Eds. (Springer, Berlin. Heidelberg, 2019), Chapt. 13, in press.
[5] S. Shamma, “On the emergence and awareness of auditory objects,” PLoS Biology, 6(6), June (2008).
[6] W. L. Martens, D. Cabrera and S. Kim, “The ‘phantom walker’ illusion: Evidence for the dominance of dynamic interaural over spectral directional cues during walking,” in Principles and Applications of Spatial Hearing, Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera and H. Kato, Eds. (World Scientific, Singapore, 2011), pp. 81–102.
[7] U. Niesser, Cognition and Reality: Principles and Implications of Cognitive Psychology (Freeman, San Francisco, 1976).
[8] G. Plenge, “On the difference between localization and lateralization,” J. Acoust. Soc. Am., 56, 944–951 (1974).
[9] F. E. Toole, “In-head localization of acoustic images,” J. Acoust. Soc. Am., 48, 943–949 (1969).
[10] E. A. G. Shaw, “1979 Rayleigh Medal Lecture: The elusive connection,” in Localization of Sound: Theory and Applications, R. Gatehouse, Ed. (Amphora Press, Groton, 1982), pp. 13–29.
[11] J. Blauert, “A perceptionist’s view on psychoacoustics,” Arch. Acoust., 37, 365–371 (2012).
[12] W. O. Brimijoin and M. A. Akeroyd, “The role of head movements and signal spectrum in an auditory front/back illusion,” i-Perception, 3, 179–182 (2012).
[13] W. Martens, S. Sakamoto, L. Miranda and D. Cabrera, “Dominance of head-motion-coupled directional cues over other cues during walking depends upon source spectrum,” Proc. Meet. Acoustics, Vol. 19, page 050129 (2013).
[14] H. Pontynen, O. Santala and V. Pullki, “Conflicting dynamic and spectral directional cues form separate auditory images,” Audio Eng. Soc. Convention 140, Audio Engineering Society (2016).
[15] E. A. Macpherson, “Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation,” Proc. Meet. Acoustics, Vol. 19, page 050131 (2013).
[16] H. Wallach, “The role of head movements and vestibular and visual cues in sound localization,” J. Exp. Psychol., 27, 339–368 (1940).
[17] J. Kawaura, Y. Suzuki, F. Asano and T. Sone, “Sound localization in headphone reproduction by simulating transfer functions from the sound source to the external ear,” J. Acoust. Soc. Jpn. (E), 12, 203–216 (1991).
[18] G. Wersényi and J. Wilson, “Evaluation of head movements in short-term measurements and recordings with human subjects using head-tracking sensors,” Acta Techn. Iaurinensis, 8, 218–229 (2015).
[19] C. Kim, R. Mason and T. Brookes, “Head movements made by listeners in experimental and real-life listening activities,” J. Audio Eng. Soc., 61, 425–438 (2013).
[20] J. R. Lackner, “Influence of posture on the spatial localization of sound,” J. Audio Eng. Soc., 31, 650–661 (1983).
[21] W. L. Martens, “Display of three-dimensional virtual acoustic sound fields for human listeners: Value of discrete multichannel simulation,” ICSV: Proc. 21st Int. Congress on Sound and Vibration, pp. 13–17, Beijing (2014).
[22] W. R. Thurlow and P. S. Runge, “Effect of induced head movements on localization of direction of sounds,” J. Acoust. Soc. Am., 42, 480–488 (1967).
[23] W. R. Thurlow, J. W. Mangles and P. S. Runge, “Head movements during sound localization,” J. Acoust. Soc. Am., 42, 489–493 (1967).
[24] W. L. Martens and Y. Han, “Discrimination of auditory spatial diffuseness facilitated by head rolling while listening to ‘with-height’ versus ‘without-height’ multichannel loudspeaker reproduction,” Proc. Audio Eng. Soc. Int. Conf. Spatial Reproduction, Tokyo, August (2018).

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

Michael Cohen is a Prof. at the University of Aizu in Japan, where he heads the Spatial Media Group in the Computer Arts Lab. His research primarily concerns interactive multimedia, including virtual & mixed reality, spatial audio & stereotelephony, ubicomp, mobile computing, hypermedia and digital typography. He is a Senior Member of the IEEE.