Cross-Modal and Cognitive Processes in Sound Localization

M. Torben Pastore, Yi Zhou and William A. Yost

Abstract To perceptually situate a sound source in the context of its surrounding environment, a listener must integrate two spatial estimates, (1), the location, relative to the listener's head, of the auditory event associated with the sound-source and, (2), the location of the listener's head relative to the environment. This chapter introduces the general background of auditory localization as a multi-sensory process and reviews studies of cross-modal interactions with auditory localization for stationary/moving sound sources and listeners. Included are relevant results from recent experiments at Arizona State University's Spatial-Hearing and Auditory Computation and Neurophysiology Laboratories. Finally, a conceptual model of the integrated multisensory/multi-system processes is described.

1 Introduction

Sound-source localization is a part of the larger perceptual process wherein transduced sensation is analyzed to form an internal representation of the surrounding environment, including the listener's own position in it. The internal reference created by this process is often called a *spatial map*. For a review of spatial maps, see Stensola and Moser (2016). Localizing a sound source in relation to other perceived objects requires mapping the first-level auditory spatial estimate, which only relates sound-source position to the listener's head, into the context of the surrounding local environment.

Consider an attempt to localize a sound source without such context. Perceptually salient sound stimulation must be parsed into individual perceptual objects, perhaps in interaction with other sensory inputs such as vision. Having grouped a set of components of the sound stimulation into a specific auditory object to be localized, the listener must then extract auditory spatial cues by comparing the inputs at the two ears across frequency as well as amplitude and phase patterns across frequency. The

M. T. Pastore (✉) · Y. Zhou · W. A. Yost
College of Health Solutions, Arizona State University, Tempe, AZ 85287, USA
e-mail: m.torben.pastore@gmail.com

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_12

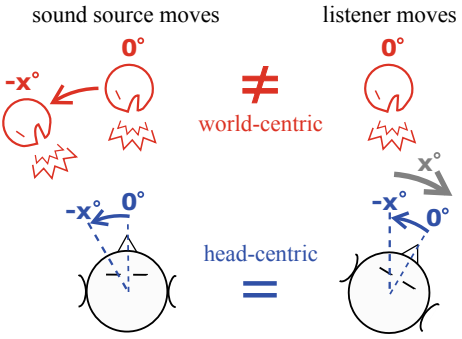

Fig. 1 A schematic illustration of the difference between head-centric and world-centric auditory localization. The actual sound source, located in the local environment, is shown in **red** with its angular displacement noted above in world-centric coordinates. In **blue**, the angular displacement of the sound source vis-à-vis the listener's head, that is, in head-centric coordinates, is shown. Positive values indicate clockwise displacement. In the **left panel**, the sound source moves from the midline to −x° in world-centric coordinates, resulting in a change in location relative to the listener's head from the midline (0°) to −x°. In the right column, the sound source is stationary at 0° in world-centric coordinates, but the listener rotates the head by +x°, as shown by the **gray arrow**, so the head-centric estimate of the sound-source location becomes −x°, that is, the same as in the **left panel**. The listener must know the position of the head in relation to the local environment to localize the sound source in world-centric coordinates

result is some estimate of the location of the sound source relative to the listener's head. Without further information, the listener cannot utilize this perceptual output for action, because there is, so far, no internal representation of the space around the listener. Figure 1 illustrates this concept. Without information about the listener's head position, the dynamic auditory spatial cues are the same for a sound source that moves while the listener is stationary versus a stationary sound source while the listener moves (−x°, printed in blue in Fig. 1). The two are therefore indistinguishable. Even with information about the listener's head position, the listener still only knows the location of the sound source relative to the head. To determine the location of the sound source relative to the surrounding environment, the listener must know the orientation of the head relative to the body and the local environment.

It is precisely because creating a perceived spatial map requires an estimate of one's location in that internally constructed context that the senses must rely on each other for reference, and that systems inputs—such as somatosensory, kinesthetic, muscular efferents, and proprioception—will necessarily interact with auditory spatial estimates at some level. Reduced to its simplest components, localizing a sound source in relation to the local environment requires mapping the estimate of the location of the sound source, relative to the listener's head, onto an internal representation of the local environment; this requires an internal representation of the listener's head position relative to the body and the surrounding environment. While this may seem obvious, the process by which this occurs is not. Many questions arise. For example, does mapping the auditory estimate into a spatial estimate of the local environment

occur at peripheral, midbrain, auditory cortex, or higher levels associated with cognitive processing—or all (some) of the above? What are the inputs to this process, and how are they combined and compared with each other? Does this combination occur according to a static rule or as a dynamic process that changes according to some set of internal and external factors, perhaps based on estimates of the reliability of the different inputs? What sort of auditory localization is possible when the internal estimate of the listener's location within the local environment is incomplete or the surrounding environment is perceptually inscrutable? Much remains to be done to address these types of questions.

This greater synthesis is likely to involve sensory, sensorimotor, and cognitive inputs. In other words, auditory localization is ultimately not merely a sensory task—it also engages non-sensory processes such as memory, attention, expectation, and motor signals. All of these questions lead inexorably to the conclusion that to fully understand spatial hearing, current inquiries must be expanded to include neural processes that occur *outside* the auditory system. Wallach (1938, 1939, 1940) was perhaps the first scientist in the modern era to enunciate and investigate these considerations. For this reason, a considerable portion of this chapter is devoted to the points he made in his seminal works on this subject. Most of the literature considered in this chapter attempts to extend findings from the laboratory toward the daily, real-world task of localizing sound sources as listeners and/or sound sources move.

The concluding section of this chapter describes the scope of this greater inquiry via a model that conceptually organizes the seemingly disparate investigations that have been reported in the literature. The model may then be used to identify future areas of study necessary to understanding auditory localization as a multisystems/multisensory process.

2 General Review

2.1 Theories of Sound-Source Localization Before the 20th Century

The early study of sound-source localization in the mid-19th century was based almost entirely on assumptions regarding the use of other sensory systems or experience in using sound to locate sound sources in the actual world—see Boring (1942). The question of whether the mind is different to the body in kind or only in degree—Cartesian Mind-Body Dualism—was a major topic in science and philosophy. Several scholars argued that the mind represents properties of the external world through sensations. These sensations had attributes, such as quality, intensity, duration, and extension, and they could be used to form percepts that the mind could use to create an internal representation of the external world. Scholars debated the exact definitions and means of measuring sensations, attributes, and perception (the mind) for nearly a half century. During this time, several scholars addressed sound-source localization.

Originally, most argued that sound has no attributes of extension (size and shape) when it impinges on the eardrum, so listeners could not use sound, on its own, to locate a sound source. This, of course, flew in the face of what most could observe, namely, that listeners can indeed localize sound sources using their hearing. Empiricists like Wundt argued that sound-source location was mediated by other senses that could sense extension, for instance, vision, touch, and the vestibular sense—see Boring (1942). Early psychologists, for example, Berkeley (1709), argued that experience helped in sound-source localization—see Pierce (1901). For at least 25 years, scholars therefore believed that sound-source localization resulted from interactions with other sensory systems and/or experience. As the 19th century ended, it became more and more accepted that sound has attributes associated with spatial extension. The question of cross-modal sound-source-localization processes became an item of increasing interest. For instance, Boring (1942) posed the question, "*Can the organism discriminate the relative positions of sounds, and, if so, how?*" This approach, exemplified in the work of Rayleigh (1876) and Thompson (1878), moved the view of sound-source localization as a multi-system process to one based on the ability of the mind (brain) to exploit differences in the inputs to the two ears to compute cues that could be used to estimate a source's location, entirely based on the sound it produces.

2.2 Auditory Input for Stationary Listeners

The auditory spatial cues are well described and documented in the literature, especially those required for azimuthal localization, since the late 19th century (Boring 1942; Mills 1972; Blauert 1997; Yost 2017a). The auditory spatial cues are commonly described in terms of the three spatial dimensions (azimuth, elevation, and distance/range)—these are discussed below.

Interaural Differences of Time and Intensity

At any given elevation, interaural differences of time (ITDs) and level (ILDs) serve as the primary cues for estimating the azimuthal location of a sound source. In normal soundfield listening conditions (i.e., excluding headphone listening) ITDs dominate the localization of low-frequency sounds ($\lesssim$1300 Hz, e.g., see Mills 1960; Macpherson and Middlebrooks 2002). Note that listeners *are* sensitive to low-frequency ILDs over headphones, and demonstrate roughly the same sensitivity to ILDs at all frequencies within the range of hearing (Yost 1981). However, in a soundfield the magnitude of low-frequency ILDs is typically small due to diffraction of long wavelengths around the head. The magnitude of high-frequency ILDs is considerably larger, and fine-structure ITDs at high-frequencies are poorly encoded, if at all, so ILDs are the dominant cue for localizing high-frequency sounds. ILDs of a decibel or more, the ILD difference threshold, are generally measured for frequencies greater than 2000 Hz—see Goupell and Stakhovskaya (2018) (but compare Hartmann et al. 2016). For further details refer to Kuhn (1977, 1987). Envelope ITDs

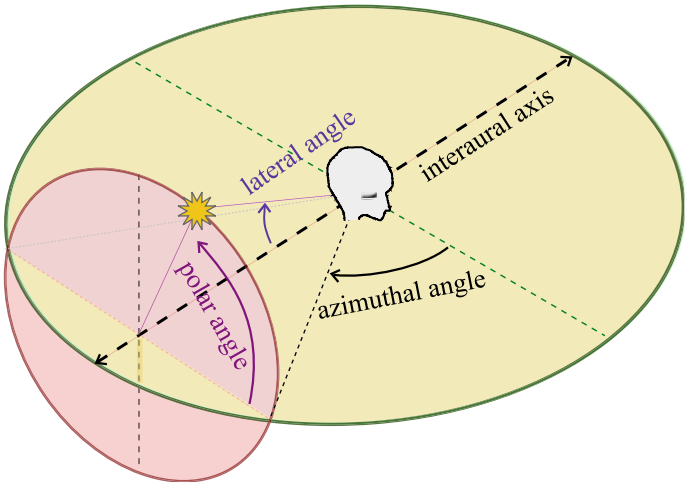

Fig. 2 The interaural double-pole coordinate system. The *interaural axis* is defined by that line which goes through both the listener's ears. The sound source is illustrated as a large **yellow asterisk**. The *lateral angle* is the angle between the sound source and the interaural axis; it is thus a combination of azimuth and elevation. The *polar angle* is the angle between the sound source and the azimuth plane, along its sagittal plane, or "cone of confusion" The *azimuthal angle*, which is complementary to the lateral angle, is the angle between the midline and the location where the sagittal plane **(in red)** meets the azimuthal plane **(in green)**. Note that the elevation angle in the single-pole coordinate system (see, for example, Fig. 6), is not interchangeable with the polar angle. Thus, it is important to specify which system is being used. Figure adapted from Morimoto (2001)

can affect perceived lateral position when sounds are presented over headphones (e.g., Blauert 1997; Bernstein and Trahiotis 2011). However, recent studies presenting similar stimuli in a sound field (Macaulay et al. 2017; Yost 2017b) failed to find a similar effect for envelope ITDs—it may be that the presence of a strong ILD cue at these frequencies renders the envelope cue redundant.

Spectral-Shape Cues

Figure 2 shows a sagittal plane (in red) intersecting with the azimuthal plane (in green) in the *interpolar coordinate system* (also called the "two-pole" system, e.g., Letowski and Letowski 2011). At any angular location on the azimuthal plane, there is a locus of possible sound-source positions that generate the same interaural disparities, especially low-frequency ITDs. Note that the iso-contours for ILDs are more complex, and the pattern of ILDs across frequency may, in itself, be useful for specifying a unique sound-source location. These loci are the so-called *cones of confusion*, (see Wallach 1938; Woodworth and Schlosberg 1938) defined by the sagittal planes in the interaural polar coordinate system (see also Baumgartner et al. 2013). Spectral-shape cues created by the filtering of sound as it passes over the torso, head, and pinna on the way to the ear canal—the head-related transfer function (HRTF)—allow listeners to determine the location of a stimulus on that locus—i.e., its polar elevation, including whether it is in front of or behind the listener. Such HRTF cues

are most useful for broadband, high-frequency (>3000 Hz) sounds. For further information, see Morimoto and Aokata (1984), Middlebrooks et al. (1989), Makous and Middlebrooks (1990), Blauert (1997). HRTF cues can aid elevation estimations if listeners have prior information about a sound's spectrum (Wightman and Kistler 1997). It would seem possible that a listener might be able to use head movements to gain familiarity with the spectrum of a stimulus by averaging across "looks" during a head-turn, or simply noting the changes in the peaks and dips of the sound spectrum as the head moves, though the authors are unaware of any such study in the literature.

An exact description of the HRTF spectral features which are responsible for elevation judgments has not been agreed upon at this time. In light of the fact that listeners do not localize elevation well with generic, KEMAR[1] HRTFs, and yet can "learn" new HRTFs (e.g., Hofman et al. 1998; Zahorik et al. 2006; Carlile and Blackman 2014) it appears likely that different listeners use different features of their own, individual HRTFs (Wenzel et al. 1993). Therefore, it seems unlikely that there is any pattern of specific spectral features, such as dips versus peaks, that is used in the same way by all listeners (see Middlebrooks 1992; Langendijk and Bronkhorst 2002). For a review on modeling of localization along sagittal planes, see Baumgartner et al. (2013).

While spectral cues are often thought of as a monaural cue, the way the spectra of the two ears are combined or weighted against each other is still not fully understood. There is evidence that the spectral cues of the ear ipsilateral to the sound source are weighted increasingly as the distance of the sound source from the midline increases. Asymmetries of the head and ears may also provide an interaural spectral difference, though it appears subservient to "monaural" spectral cues (for more information see Searle 1973; Musicant and Butler 1984; Humanski and Butler 1988; Slattery and Middlebrooks 1994; Morimoto 2001; Van Wanrooij and Van Opstal 2004; Jin et al. 2004).

When stimuli do not have sufficient high-frequency information, the acuity of auditory localization in terms of azimuth is largely unaffected, but listeners' estimation of elevation is considerably degraded and front-back reversals occur quite often. Good and Gilkey (1996) tested localization in noise, thereby disrupting high-frequency spectral cues. They found that decreased signal-to-noise ratio negatively affected listeners' ability to distinguish front from back, had less impact on elevation accuracy and affected horizontal localization the least.

Interaction of Interaural Differences and Spectral Cues

When sound stimuli do not have high-frequency information, or the pinnae are occluded with ear molds to distort HRTF cues (e.g., Morimoto 2001), listeners often tend to localize sounds in those portions of the azimuth plane that intersect with the front and the back of the cone of confusion. This may result from learning that most salient sound sources lie roughly near the azimuth plane. Spectral cues appear to specify the location on the cone of confusion that corresponds to the location of the sound source (e.g., Morimoto and Aokata 1984; Best et al. 2011; Letowski and

[1] KEMAR®is an often-used head-and-torso simulator—a so-called "dummy head".

Letowski 2011). Such a conception is subtly different to the idea that spectral cues encode elevation as it is specified in single-pole spherical coordinates, because the elevation is along the cone of confusion (i.e., on a sagittal plane), instead of being measured from the origin.

While Morimoto and Aokata (1984) and Makous and Middlebrooks (1990) have shown evidence that interaural differences and spectral cues may be estimated independently of each other, it is not clear how or at what point these two estimates are combined into a unified estimate of sound-source location. Also, the literature is somewhat mixed on whether interaural cues need to be correct for judgments of elevation to be accurate. In other words, if a listener cannot determine which cone of confusion the sound source is on, spectral cues may not be useful (for studies related to this question see Van Wanrooij and Van Opstal 2004; Morimoto 2001; Jin et al. 2004; Martin et al. 2004). It is also worth noting that the pattern of ILDs across frequency is not monotonic because of the acoustical bright spot that results from wave diffraction around the head (Macaulay et al. 2010). Therefore, the pattern of ILDs across frequency could conceivably also be used to specify where on a cone of confusion the sound source lies (e.g., Macpherson and Middlebrooks 2002). Section 4 discusses how listeners may, in the absence of spectral cues, use head movements to specify where on a cone of confusion a sound source lies.

Distance and Range

Distance cues seem to be almost completely based on listener expectations, and therefore require knowledge not only of how a sound source at a given distance relates to the head, but also of learned changes in the quality of a sound as it moves further away from, or closer to, a listener. There are several correlations between the distance of a sound source and its acoustical qualities that can be learned. If a sound source is in the near field (less than $\approx$0.3 m from the listener, depending on frequency), atypical ILDs result from the non-linear propagation of the sound around the head—this could offer a cue for judging distance (e.g., Brungart et al. 1999). For sound sources not in the near field, there are several other cues. Sounds from sources at large distances can be affected by the atmosphere, which acts as a low-pass filter, thereby providing a possible spectral cue for relative distance estimations that likely requires experience and expectation on the part of the listener (Kolarik et al. 2016). Sound intensity decreases with distance according to the inverse-square law—with expectation/memory this cue could also be exploited. In reverberant spaces, the direct-to-reverberant energy ratio decreases with increased distance, and provides a cue for judging relative distance (Zahorik 2002; Bronkhorst and Houtgast 1999). Note that this cue also relies on some expectation for the acoustics of the space. Auditory motion parallax may, in some cases, provide a cue for discerning relative sound-source distance (Genzel et al. 2018) and is discussed further below. See Kolarik et al. (2016) for a general review on auditory distance perception.

A Case for Multimodal Cues in Auditory Localization

The auditory spatial cues described above (excepting distance cues) are primarily head-centric cues. Expectation and a priori information—analyses of acoustic cues

that are based on experience—can provide indirect information to improve sound-source location. Wallach (1940) appears to have been the first to point out that the auditory spatial cues cannot, by themselves, specify the location of a sound source in the context of the local environment. Wallach (1940) demonstrated that "two sets of sensory data enter into the perceptual process of localization, (1), the changing interaural cues and, (2), the data representing the changing position of the head"—see Sect. 4 for further discussion.

While the spatial cues for sound-source localization (see above) have been well-researched for nearly 150 years, much less is known about how the cues used to estimate head position relate to sound-source localization. The literature is clear that vision is a vital cue for determining head position (Wallach 1940; Yost et al. 2015; Van Opstal 2016). The literature also suggests that additional auditory cues and/or vestibular, somatosensory, kinesthetic, proprioceptive, and neuro-motor control systems could also provide head-position information. Experience, coupled with memory as it manifests itself in spatial maps, might also provide head position information. For an exploration of some of the complexities inherent to this issue, see Buzsáki and Llinás (2017). Estimates of head (and body) position are therefore likely to be the product of a combination of cues and estimates arising from a wide range of sensory and systems inputs. The dynamic weighting of these head-position cues in determining head position, and how this weighting interacts with sound-source localization, is currently not well understood.

There is, however, a relatively rich literature on the integration of different spatial cues, related to other aspects of sound-source localization, that might also account for the integration of auditory spatial cues and head-position cues for world-centric sound-source localization. The next two sections consider evidence for sound-source localization as a multisensory/multi-systems process. Section 3 considers experiments probing audio-visual interactions under conditions where listeners and sound sources are stationary, and Sect. 4 considers investigations in which listeners and/or sound sources move.

3 Examples of Sound-Source Localization as a Multisensory Process—Localization with Stationary Listeners and Stationary Sound Sources

A great deal of study has been devoted to visual capture, in which visual stimuli affect the perceived sound-source locations. Vision is clearly an important sensory input for determining the location of the listener (body and head) with relation to the surrounding environment. Vision can perceptually situate a head-related auditory estimate of sound-source location into the spatial context of the surrounding environment. As such, interactions between audition and vision can be thought of as evidence for Wallach's 1939/1940 insight before head movement is even considered.

When visual and auditory signals are both perceptually attributed to the same source, vision improves the accuracy of sound localization. Vision often plays a dominant role in spatial judgment. Spatial visual cues can override the spatial information of a sound, causing errors in sound localization. Commonly known as the *ventriloquism effect* or visual capture, the auditory event is localized to a seen source, even though the sound source is positioned at a different location (Howard and Templeton 1966).

A bias towards vision-centered experiments has meant that most of what is known about audio-visual interactions comes from localization results in the horizontal frontal field. Limited evidence, however, reveals that vision can also enhance auditory-distance estimation (Anderson et al. 2014). The role of vision (eyes open vs. closed) is more limited in vertical localization (Shelton and Searle 1980), though vision does appear important to the calibration of vertical localization—see, for example, Zwiers et al. (2001). The horizontal and vertical difference is likely a result of the different roles of eye and head movements in gaze orientation. Recently, Solman et al. (2017) showed that eye movements preferentially exploit the horizontal span of the visual field. Head movements then shift this horizontal span up and down. Nevertheless, the vertical gaze of a listener can have a strong effect on perceived auditory elevation, as discussed in Sect. 3.2 below.

While vision is arguably involved in most everyday listening experiences, the common bias of vision over audition is not simply a result of relatively poor auditory spatial acuity. Indeed, when adequate localization cues (i.e. both ITDs and ILDs) are available across a sufficient range of frequencies, spatial hearing is remarkably accurate (Dorman et al. 2016; Yost 2016). The just-noticeable change in horizontal angular displacement, the minimum audible angle, can be as small as 1–2 ° for sound sources near midline (Mills 1972; Hartmann and Rakerd 1989). Nevertheless, most of us rely primarily on vision when localizing objects around us, whether the objects make sound or not. This is probably because the auditory spatial estimate, on its own, only specifies the position of the sound source relative to the listener's head (Yost et al. 2015) whereas the spatiotopic encoding of vision is inherently world-centric.

Over the past decades, digital technology has greatly advanced the sophistication and automation of stimulus delivery and experimental procedures, helping to uncover, (1), the structural properties of auditory and visual stimuli that are conducive to cross-modal interactions and, (2), the cognitive factors (e.g., attention, expectation and experience) that affect listeners' assumptions and awareness of the origin and cause of the multisensory inputs (Radeau and Bertelson 1977; Welch and Warren 1980). The following sections summarize empirical evidence that addresses how active vision affects auditory localization performance via frame-of-reference and perceived target position. Also, how major differences between visual and auditory spatial mechanisms may affect estimates of the center, width, and front-back location of a perceived sound source is discussed. Further, the general framework of spatial audiovisual studies is dealt with and future directions for research relevant to real-life activities are discussed.

Numerous studies have investigated how vision affects sound-source localization, mostly in the horizontal plane. The general empirical findings related to several

hypotheses of vision's role are broadly summarized below. These hypotheses are not mutually exclusive and are evolving concepts.

- *The frame-of-reference hypothesis* Sound localization is more accurate when a listener can acquire, through free or voluntary eye movements, knowledge of the spatial layout of a lighted environment (Thurlow and Kerr 1970; Warren 1970; Platt and Warren 1972; Shelton and Searle 1980).
- *The visual-dominance hypothesis* Vision is a dominant sense in spatial tasks due to its superior spatial acuity. Vision can bias the perceived direction of a source of sound towards the direction of a visual cue (Jackson 1953; Choe et al. 1975; Bertelson and Radeau 1981).
- *The cue-reliability hypothesis* The reliability of estimates for each modality determines which sense dominates perception before they are combined. Reducing the saliency of visual cues weakens visual dominance (Battaglia et al. 2003; Alais and Burr 2004; Ernst and Bülthoff 2004).

3.1 Relevance of the Frame of Reference

Boring (1926) suggested that listeners effectively map the perceived location of sound sources onto a spatial reference provided by vision. This hypothesis predicts that listeners will localize sound sources more accurately when their eyes are open, even if they cannot see the sound source—this is called visual facilitation. Warren (1970) demonstrated visual facilitation by hiding the spatial layout of a room and the loudspeakers within using a khaki cloth, so that the cloth alone constituted the "textured" environment for the task. Subjects hand-pointed to the perceived direction of a pulse-train auditory stimulus. The visual conditions were factorial combinations of eye open/closed, environment light/dark and vision free/fixated. Analyses compared the response error and response variability scores among various visual conditions.

Their results showed that active visual sensing of the physical layout of the environment, and objects in it, enhanced the acuity of listeners' auditory localization. On their own, free vision, a lit environment, or simply having the eyes open did not result in visual facilitation. The most favorable condition for visual facilitation was a combination of a lighted environment with free, target-directed eye movement. Performance under this condition was better than the lighted condition with a fixed gaze and the unlit condition with free eye movement. Warren argued that eye movement per se does not improve the accuracy of auditory localization, but that an illuminated visual environment allows better visual-motor (eye-hand) coordination by providing a spatial reference to guide action.

Shelton and Searle (1980) tested how vision affects the absolute identification of a sound-source position in a sound field. Half the subjects wore goggles painted over in black while the other half wore clear goggles. In all conditions, both sets of listeners could see the loudspeaker positions before and between testing sessions, so vision (together with memory for those listeners wearing the blacked-out goggles) could provide estimates of both the frame of reference and the target-source location. No

instructions were given to tell listeners where they could look. Listeners' auditory localization benefited most from vision with sound sources located in the frontal field along the horizontal axis. Vision also improved localization for sound sources located behind listeners and to their sides, but the improvement was far less than for the frontal horizontal span. However, there was no significant benefit to localization acuity along the vertical axis of the frontal field. These early data demonstrate that the limitation of human vision to the frontal field may have significant consequences on how auditory localization interacts with the knowledge of the frame-of-reference and target locations acquired through vision.

3.2 Relevance of Visual Target Cues

Over the past decades, multisensory research has provided a broad understanding of the spatial and temporal features of sensory stimuli that are conducive to cross-modal bias. The general conclusion is that visual bias is greater when sound and light stimuli come from sources positioned close to each other and/or are presented at the same time (Jackson 1953; Pick et al. 1969; Thurlow and Jack 1973; Choe et al. 1975; Jones and Kabanoff 1975; Slutsky and Recanzone 2001). This suggests that multisensory processing follows Gestalt perceptual grouping principles—that is, spatial and temporal proximity enhance fusion between audition and vision in establishing a unitary percept. Attention appears to play a limited role in the ventriloquist effect (Bertelson et al. 2000), suggesting that audio-visual interactions may occur at early sensory stages. Studies also show that perceptual fusion between auditory and visual events is not a necessary factor for visual bias. Partial or incomplete visual capture can occur even when the auditory and visual stimuli are not perceptually fused together (Welch and Warren 1980; Bertelson and Radeau 1981; Hairston et al. 2003; Wallace et al. 2004; Kording et al. 2007). Some degree of visual capture can also occur for asynchronously presented auditory and visual stimuli (Jack and Thurlow 1973; Thurlow and Jack 1973; Radeau and Bertelson 1974; Shelton and Searle 1980; Radeau and Bertelson 1987; Recanzone 2009). However, the strength of visual bias does decrease as the spatial and temporal separation between auditory and visual spatial estimates increases. Reviews include Welch and Warren (1980), Stein and Meredith (1990) and King (2009).

While the majority of audio-visual studies have emphasized the spatial and temporal conditions underlying multisensory interactions, separate lines of work reveal that the reliability of estimates (the inverse of the variance) for each modality determines which sense dominates the fused percept. This suggests that the dominant role of visual spatial information is scalable. Indeed, results have shown that reducing the saliency of visual cues by blurring or adding corruptive noise can weaken or even reverse visual capture (Ernst and Banks 2002; Battaglia et al. 2003; Alais and Burr 2004). These empirical results have been well described in a Bayesian framework, which establishes the relationship between the stimulus, S, and response, R. See Mendonça (2020), in this volume, and further, Sivia and Skilling (2006) for a review of Bayesian analysis.

The general principle of Bayesian estimates can be expressed in terms of the relationship between two conditional probabilities of stimulus and response, that is,

$$p(S|R)p(R) = p(R|S)p(S) \tag{1}$$

where $p(S|R)$ is the posterior probability, $p(R)$ is the marginal likelihood, $p(R|S)$ is the likelihood and $p(S)$ is the prior probability. With the assumption that the distribution of neural responses is constant and stable, the equation can be expressed as the proportionality

$$p(S|R) \; \alpha \; p(R|S)p(S). \tag{2}$$

Equation (2) is the foundation of Bayesian-Inference theory. It states that the internal reconstruction of an event (the posterior probability) is the result of the likelihood estimate of whether this event leads to a neural response and an estimate of the stimulus distribution (the prior probability).

In the Bayesian model of audio-visual localization, it is assumed that auditory, A, and visual, V, cues are independently processed, $p(R_{AV}|S) = p(R_A|S) \; p(R_V|S)$. The modality-specific, neural representations, the likelihood estimates $p(R_A|S)$ and $p(R_V|S)$, typically consist of a one-to-one mapping of the auditory and visual cues associated with the position variable, in the form of a Gaussian function, $\mathcal{N}(\mu_A, \sigma_A^2)$ and $\mathcal{N}(\mu_V, \sigma_V^2)$, where $1/\sigma_V^2$ and $1/\sigma_A^2$ describe the reliability of neural estimates of the visual and auditory spatial cues, respectively. A large σ signals a greater uncertainty in the neural estimate with weak responses from many spatial channels. A small σ signals a reliable neural estimate with strong responses from selected spatial channels.

One may, for the moment, assume that the combined A and V cues lead to a fused percept (e.g., the ventriloquist effect) and that the prior distribution is flat ($p(S) = 1$). Given these assumptions, Battaglia et al. (2003) and Alais and Burr (2004) showed that the combined multisensory estimate (i.e., the mean of the posterior estimation) is equal to the weighted sum of the individual, unitary A and V estimates,

$$\mu_{AV} = \sigma_{AV}^2 \left(\frac{1}{\sigma_V^2} \mu_V + \frac{1}{\sigma_A^2} \mu_A \right). \tag{3}$$

The term σ_{AV}^2 describes the variance of the combined estimate, which is always smaller than the variances of the unisensory estimates, σ_A^2 and σ_V^2, as follows,

$$\sigma_{AV}^2 = \left(\frac{1}{\sigma_V^2} + \frac{1}{\sigma_A^2} \right)^{-1} = \frac{\sigma_A^2 \sigma_V^2}{\sigma_A^2 + \sigma_V^2} \leq \min(\sigma_A^2, \sigma_V^2). \tag{4}$$

When experimentally manipulating σ_A^2 and σ_V^2 it is important to carefully consider fundamental differences in the peripheral mechanisms of vision and audition. The visual peripheral system is spatiotopically organized—thus, it encodes space

directly. The receptive fields of ganglion cells cover different regions of space that are mapped onto the retina, and the visual system retains this mapping throughout. Therefore, manipulation of the width or quality of a visual image can directly affect the population activity of visual neurons. The auditory periphery, however, is tonotopically organized—hairs cells in the cochlea are organized according to the sound frequencies they encode and do not directly encode sound-source location. Therefore, the auditory system must estimate the location of sound sources on the basis of interaural differences of arrival time and intensity (ITDs and ILDs) as well as on the spectral characteristics imposed by the HRTFs—(see Knudsen and Brainard 1995; Middlebrooks et al. 2002, and also see Sect. 2 above). The auditory brainstem extracts these localization cues in computations that involve multiple neural structures. The resulting localization cues do not always unambiguously correspond to a single physical sound-source location but rather to a locus of possible locations—the "cone of confusion"—see Sect. 2. The computational nature of auditory space means that "blurring" an auditory image is not as straight forward as it is for a visual stimulus. Perhaps as a result, multisensory research has only seldom manipulated the reliability of auditory localization cues.

However, studies have shown that poorly localized auditory stimuli tend to facilitate visual dominance. Thurlow and Jack (1973) found that the relatively poor acuity of auditory localization in the vertical plane resulted in a stronger ventriloquist effect than in the horizontal plane. Similarly, Spence and Driver (2000) found that ventriloquism was more likely for sound stimuli that are difficult to localize (e.g., a 2-kHz tone from multiple speakers) than for sound stimuli that are readily localized (such as white noise from one loudspeaker). The reliability factor explored in these investigations is related to the quality or width of an internal, neural estimate of the auditory event, not the quality or width of the physical stimulus, as in vision. Therefore the nature of the poor localizability is not straightforward to predict. Erroneous auditory localization could be caused by reduced resolution of a wide excitation pattern across many spatial channels or interaural-cue computation in a single spatial channel, or both. To our knowledge, the neural mechanisms for the saliency of auditory spatial perception remain largely untested.

Montagne and Zhou (2016) investigated whether manipulations of the congruence between ITD and ILD affects the reliability of auditory responses and the magnitude of visual bias. Broadband noise bursts (15-ms duration) were presented from two hidden loudspeakers at $\pm 45°$ about the midline, with or without a simultaneously presented light-emitting diode (LED) flash from $-45°$, $0°$, or $+45°$.

Two auditory conditions were contrasted, (1), timing-based stereophony with incongruent ITDs and ILDs and, (2), level-based stereophony with congruent ITDs and ILDs. Figure 3 shows the relationship between the standard deviation (SD) of auditory-alone responses and the change in auditory localization when the light stimulus was present, that is, the visual bias, ΔAV. Listeners localized sound sources with greater variability and stronger visual bias for the timing stimuli than for the level stimuli. Also, the magnitude of visual bias for the timing signals correlated strongly with the variance (noise) of listeners' auditory estimate, suggesting an intrinsic link between binaural ambiguity and localization uncertainty. In turn, the putative

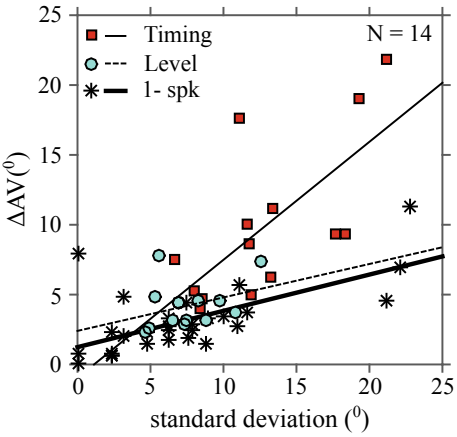

Fig. 3 Relationship of response variability and visual capture for individual subjects. Symbols indicate (average) responses of individual subjects for timing-based stereophony **(squares)**, level-based stereophony (circles) and single-speaker controls **(asterisks)**. Straight lines show linear fits for each condition. From Montagne and Zhou (2016)

uncertainty of auditory localization modulated the strength of visual bias on sound localization.

3.3 Asymmetry of Perceptual Space

When the head and body are stationary, the visual and auditory systems do not encode the same spatial range. Auditory space is broad and extends to both front and rear space, whereas human vision is restricted to the frontal region, with visual acuity declining towards peripheral locations away from the fovea (Curcio et al. 1990). The resulting asymmetry between visual and auditory space is an important factor to consider in addition to the differences between the peripheral mechanisms in vision (spatiotopic encoding) and audition (computational space based on tonotopic encoding). Despite these differences, our knowledge of cross-modal spatial bias is mostly limited to audio-visual (AV) interactions in the frontal hemifield. As mentioned earlier, the symmetry of interaural cues along sagittal planes normal to the interaural axis often leads to front-back reversals. Indeed, the question of whether frontal visual cues can interact with the auditory events that are perceived in the rear, be they real or illusory, remains an interesting and ecologically important research topic.

Montagne and Zhou (2018) investigated the influence of frontal LED flashes on the perceived front-back, left-right location of a phantom sound source generated using timing-based stereophony. Figure 4 shows that there was a considerable amount of front-back confused responses to a center-position phantom source presented either from front or back. The colored lines show that frontal visual cues increased the percent of frontal responses. Left-right response shifts can be seen to follow the direction of the light. Interestingly, the lateral visual bias is only observed for the perceived frontal sound sources at 0°. Very little lateral bias was found in the perceived sound sources at 180°. The study also revealed that increasing the stimulus duration reduced both the rate of front-back reversals and the visual bias but not

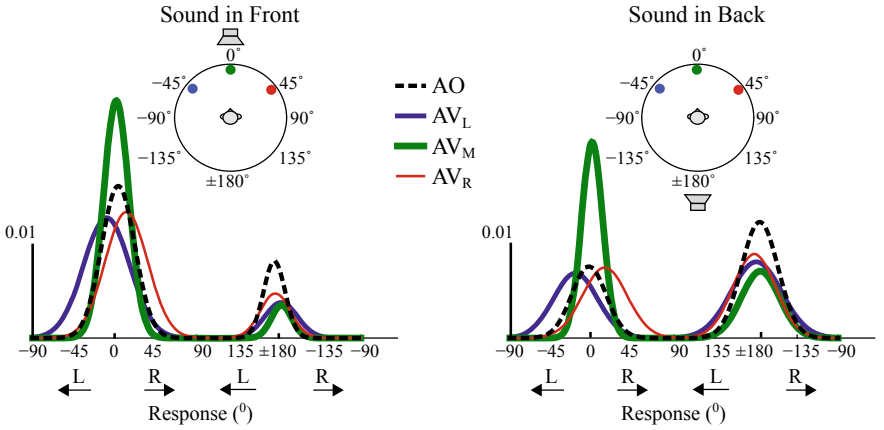

Fig. 4 Two-peak Gaussian functions for *AO*, audio only, and *AV*, audiovisual, responses for 15-ms duration Gaussian noise stimuli. The left and right figures show the results obtained using the two frontal or rear speakers, respectively. Each curve was obtained by fitting the data from all trials and all subjects using the Gaussian-mixture model. The delay between the two loudspeakers was 0 ms for both conditions shown. The speaker sign marks the expected position of the perceived "phantom sound source." The *AO* results **(black dashed line)** show that the responses were clustered on the midline at 0° and ±180°. The colored lines show changes in the left-right and front-back responses after adding visual stimulation. From Montagne and Zhou (2018)

localization errors associated with left-right judgment. These findings show that visual information separately interacts with left-right and front-back dimensions of a perceived sound source, while stimulus duration mainly modulates front-back errors in multisensory spatial processing.

The interactions between frontal vision and rear audition do not easily fit with existing Bayesian statistical models (e.g., Ernst and Banks 2002; Battaglia et al. 2003; Alais and Burr 2004) because these models are primarily based on the results of cross-modal perception of a seen target. In other words, the stimulus, S, in the prior distribution, $p(S)$, has an implicit frontal origin. Furthermore, the modality-specific, sensory representation, likelihood estimate, $p(R_A/S)$ or $p(R_V/S)$, consists of a one-to-one mapping of S in the form of a unimodal (single-peaked) likelihood function. As shown in Fig. 4, this estimate is not adequate after considering the rear sound field, where the front-back confused responses result in a bimodal likelihood function, $p(R_A/S)$. These factors complicate the variance estimate and subsequently the construction of the posterior probability using combined auditory and visual estimates as shown in Eq. (3).

Montagne and Zhou (2018) suggested an alternative mode of AV interaction for when the stimulus space extends outside the field of vision. They proposed that visual processing might affect the left-right and front-back auditory judgment independently in two different stages, (1), an initial coarse and broad auditory detection to decide the relative front vs. back direction of an event and, (2), if the perceived target location is in front, visual analysis to refine the estimate using integrated auditory and visual information. According to the causal-inference theory, the brain should limit the

extent of integration between sensory events perceived to rise from different sources (Kording et al. 2007). Montagne and Zhou (2018) argued that the causality test likely occurs during the initial auditory detection stage, which includes front-back discrimination.

4 Sound-Source Localization with Moving Listeners and/or Moving Sound Sources

Section 3 showed evidence for the integration of head-centric auditory spatial estimates with world-centric visual estimates under conditions where listeners and sound sources were stationary. This section considers evidence from scenarios where listeners and/or sound sources move, especially with sound stimuli that offer no spectral cues to specify where a target sound source is on a given cone of confusion. Wallach (1939, 1940) has been continuously cited in the literature with regard to the role head motion plays in avoiding front-back reversals. However, Wallach's foundational insight that multisensory, multi-systems information about head position must be integrated with interaural-difference cues in order to localize sound sources to their position in the surrounding environment, has received little attention until very recently.

This section, therefore, begins by reviewing some of Wallach's experiments and the logic that inspired them. To begin with, the simplest case for Wallach's hypothesis, that listeners could resolve spatial ambiguities in the azimuth plane by using head movements to compare the change in head-related auditory cues to the change in head position, is examined. The section then considers how Wallach extended this insight to propose a possible mechanism for estimating the elevation of sound sources without using spectral cues. Current knowledge about the head-position cues that might be integrated with the interaural cues in determining world-centric sound-source location is then reviewed. Finally, there is a brief review of some current investigations of the integration of interaural and head-motion cues.

4.1 The Wallach Azimuth Illusion

Wallach (1938, 1939, 1940) noted that interaural difference cues alone (especially ITDs) specify not just a single location, but an entire locus of positions, a "cone of confusion," all with the same angular relation to the head—see Sect. 2.2 for further details. As Wallach (1939) showed, head movements can be used to determine the front/back location of a stationary sound source—see Fig. 5. Wallach hypothesized that the relation between the change in interaural difference cues, relative to a given change in head position, would allow listeners to reduce the cone of confusion to a single point, thereby avoiding front-back reversals. An essential component of this hypothesis is that the listener makes some assumption about the movement, or

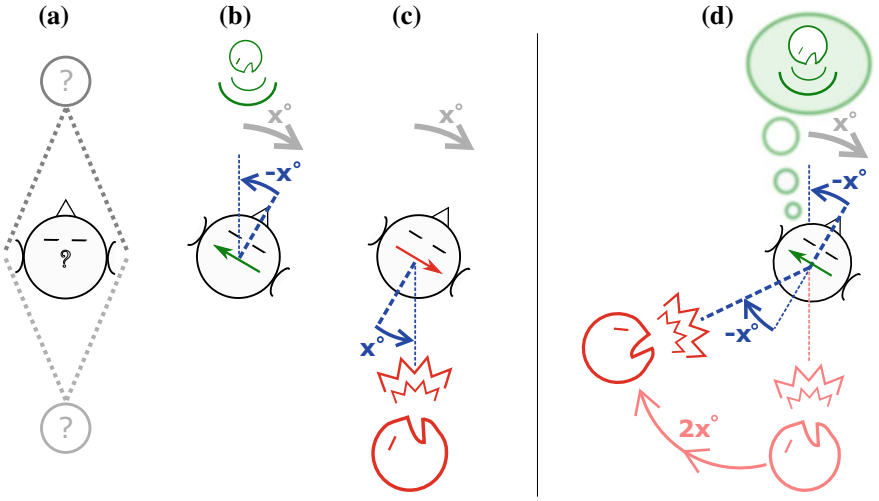

Fig. 5 **Left (a, b, c)** The basic idea of how head movements can be used to disambiguate front-back confusions. For world-centric changes of sound-source position (**red arrows**) and head position (**gray arrows**), clockwise rotation is notated as positive. For head-centric interaural differences, changes (**blue arrows**) that favor the right ear are notated as positive. Column (**a**) shows that low-frequency interaural differences, especially ITDs, are the same regardless of whether the sound source is in front or in back of the listener. (**b**) For a sound source in front of the listener on the azimuth plane at 0° elevation, a head turn of x° (**grey arrow**) results in a change in interaural differences equivalent to a −x° (**blue arrow**) change in sound-source position. (**c**) The same head turn results in a change in interaural cues equivalent to x° **(blue arrow)** for a sound source behind the listener. (**d**) A visual explanation of Wallach's Azimuth Illusion. By rotating a sound source at twice the rate of the listener's head turn of 2x° **(red)**, the same change in interaural-difference cues that would occur for a stationary sound source in the opposite front/back hemifield (front, in this case: −x°) is produced. Provided there are no spectral cues to disambiguate front from back, the listener hears a stationary sound source in front, at the location of the green sound source, even though the actual **(red)** sound source is moving behind the listener at twice the listener's rate of rotation. Note that the difference in sign between world-centric and head-centric angular rotation highlights the disconnection between the two coordinate systems that must somehow be bridged

lack thereof, of the sound source during the head movement. Specifically, Wallach assumed that "of all the directions which realize the given sequence of lateral angles, that one is perceived which is covariant with the general content of the surrounding space." That is, assuming the sound source is stationary, there will only be one point in space, at or above the height of the pinnae, that is common to all cones of confusion that exist along the trajectory of the listener's head rotation—the *Selective Principle of Rest*.

To test this notion, Wallach (1939, 1940) created an experimental apparatus that was coupled to the listener's head. The device had electrical switches that activated, as a function of the listener's head movements, one of 20 equidistantly spaced loudspeakers on a 120° circular arc. Wallach calculated the rate at which the head-centric auditory spatial estimate, derived from interaural-difference cues, would change dur-

ing a head turn for a sound source in front of the listener. He then produced the same changes in sound-source location (relative to the listener's head) that would occur for a frontal sound source. However, he presented the sound from *behind* the listener, rotating at twice the rate of the listener's head rotation. Figure 5D offers a graphical explanation of this basic concept—see Sect. 4.2 and Yost et al. (2019) for more detailed, mathematical explanations. Given a stimulus conducive to front-back reversals, the listener hears a stationary sound source in the front-back hemifield opposite to the one from which the stimulus was initially presented, despite the fact that the sound source is actually rotating around the listener in the same direction but at twice the rate of the listener's rotation. This suggests that the listener determines the front/back location of the sound source using the concomitant changes in interaural-difference cues for a given head turn. Since the change in interaural cues is commensurate with the magnitude of the head turn, the listener assumes the sound source is static. This basic result was reported by Wallach (1940) for all five tested listeners. For a review of perceived auditory motion, see Carlile and Leung (2016).

There are at least two possibilities for how the Wallach Illusion, and dynamic world-centric localization in general, could occur. It is possible that the world-centric location of auditory objects is updated at relatively sparse intervals, and that localization is head-centric between these intervals. For example, localization could be world-centric before and after a head turn, but head-centric during the turn due to the increased complexity and reduced resolution of dynamic sound-source localization. Following this notion, the change in interaural cues would be compared with the change in head position. The result of this comparison would then be mapped to world-centric coordinates for a "spatial update." Under such conditions, one might expect vestibular cues to provide useful information regarding the change in head position in between spatial updates. Another, perhaps more computationally intensive possibility, is that the auditory system continuously updates world-centric coordinates of a perceived sound source. In this case, changes in the world-centric estimate(s), which could be bimodal if the possibility for front-back reversals exists, or even a locus of possible source positions in the form of a cone of confusion, would be compared with the head position. The comparative trajectory of the sound source and head position estimates would then determine the singular estimate of the sound-source position in the local environment. Targeted experiments will be required to reveal which of the two hypothesized processes is more appropriate—(see also Brimijoin and Akeroyd 2014, reviewed below).

4.2 The Wallach Vertical Illusion

The direction-dependent filtering provided by the pinnae, head, and torso—the so-called head-related tranfer function (HRTF)—may not be the only elevation/front-back cue. Wallach (1938, 1939, 1940) extended his Azimuth Illusion—see Sect. 4.1—to include the judgment of elevation, pointing out that the rate at which interaural cues change relative to head motion could be used, assuming a stationary

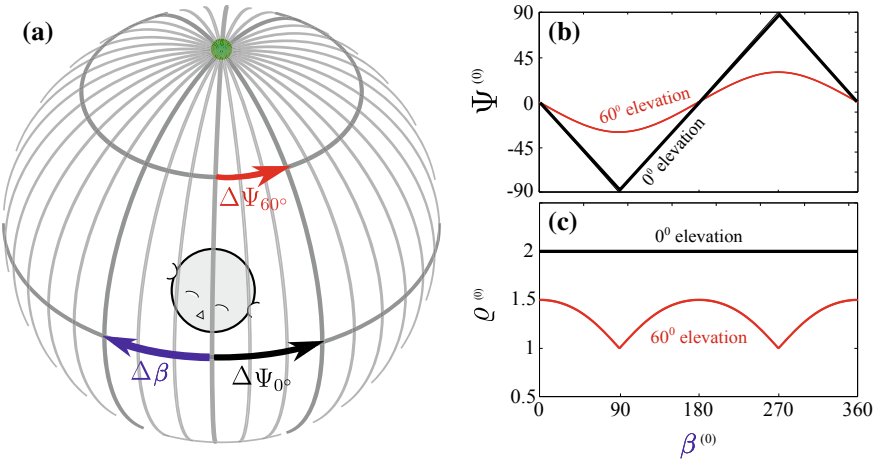

Fig. 6 **a** Visual description of Wallach's basic concept. For a stationary stimulus on the azimuth plane, that is, with 0° elevation, a listener's head rotation, $\Delta\beta$, results in the same but opposite change in angular displacement of the sound source relative to the head as would occur if the sound source had traveled the same angular distance in the opposite direction, $\Delta\Psi_{0^\circ}$. If the sound source is at an elevation ν that is off the azimuth plane, for example, Ψ_{60°, then the change in interaural-difference cues will be less for the same head movement, $\Delta\beta$. For a sound source above the head **(green dot)**, there are no changes in interaural differences for a given head turn. **b** The lateral angle, Ψ, as a function of head position, β, where the frontal midline is 0°. **c** The ratio, ϱ, of sound-source rotation relative to listener head rotation required to induce the Wallach Illusion, is shown as a function of the head position, β

sound source, to determine elevation of a sound source in the absence of spectral HRTF cues.

Figure 6a illustrates Wallach's basic insight. The position of the head, β, is measured relative to the midline. Interaural differences are instead considered relative to the *interaural axis*, which can be imagined as a line passing through both ears. This is also the axis about which the *cones of confusion* are centered. Wallach called the angle between the sound source and the interaural axis the *lateral angle*, Ψ. This value corresponds to an interaural difference. Note that *both* azimuth and elevation contribute to the lateral angle, Ψ, in the following way:

$$\Psi = 90^\circ - \left(\cos^{-1}(\sin\beta\cos\nu)\right). \tag{5}$$

That is, despite the common conception that interaural disparities are used only for encoding azimuth, a given interaural difference actually corresponds to a range of positions at many elevations—see Sect. 2.2 for further details. If a sound source is located anywhere on the sagittal plane corresponding to the midline, the interaural differences are approximately zero. Note that, in Wallach (1940), this condition would be notated as $\Psi = 90^\circ$. For this chapter, Ψ is reduced by 90°, so that the midline corresponds to $\Psi = 0^\circ$. Therefore, the "lateral angle" in this case is really

the displacement of the sound source from the median plane instead of the interaural axis.

Turning again to Fig. 6a, our listener makes a head turn of β. If the sound source lies on the 0° elevation azimuth plane, the corresponding change in Ψ (indicated by the **black arrow**) relative to β is $\frac{\Delta\Psi}{\Delta\beta} = 1$, the same as β. This is the maximal change in Ψ for a given head turn. At the other extreme, a sound source directly above the listener will elicit the smallest possible change, $\Delta\Psi = 0$.

Wallach realized that for a sound source at some intermediary elevation, say $\nu = 60°$, the change in Ψ will also be intermediary, as indicated by the red arrow in Fig. 6a. Figure 6b shows how Ψ, the angular relation of the sound source to the median plane, changes as a function of head rotation, β. Note that, at 0° elevation, there is a unity gain between β and Ψ, whereas for 60° sound-source elevation a change in β results in far less of a change in Ψ. The maximum value, $|\Psi|$ can take at any elevation is the complement of ν, for example, 30° for a sound source at 60° elevation—see Mills 1972 for a similar derivation.

Using this information, the ratio of sound-source rotation to head rotation which is necessary to induce the Wallach Illusion, ϱ, can be calculated for a sound source, presented from the 0°-azimuth plane, as

$$\varrho = \frac{\Delta\Psi}{\Delta\beta} + 1. \tag{6}$$

For a signal without sufficient high-frequency information to allow a listener to exploit pinna-based cues, a purely rotational head movement will not allow the listener to determine if a sound source is above or below the 0° azimuth plane. If even a small head tilt is included in the head movement, however, this ambiguity could also be avoided.

To test this, Wallach could have asked rotating listeners to judge the elevation of stationary sound sources. Instead, Wallach (1940) employed the same argument that leads to the Wallach Azimuth Illusion to show how azimuthal head rotation, coupled to azimuthally-rotating sound sources, could lead to the illusory perception of a stationary sound source at an elevation specified by the speed of sound-source rotation relative to the listener's rotation, ϱ. In his main experiment, Wallach simulated a sound source at an elevation of 60°, above the horizontal plane. He did so by rotating a listener passively sitting in a chair (either blindfolded or not) with the sound source rotated at 1.5 times the rate of head rotation from behind the listener. This rate of rotation, an approximation to (6), was expected to induce a perceived elevation angle of $\nu = 60°$, given that the listener only rotated within a relatively narrow angular range. Fifteen listeners indicated that the musical sounds were perceived above them in elevation, more so when their eyes were open than when they were closed. However in many cases, the listeners' judgments of elevation underestimated the predicted elevation of 60°.

Since Wallach's (1940) calculations only indicate a change in elevation relative to the horizontal plane and not whether the vertical angle is positive (above the pinnae) or negative (below the pinnae), a response below the horizontal plane would be consistent with his calculations. In this regard, Wallach (1940) made two some-

what inconsistent assumptions. First, he assumed that listeners' experience naturally biased them to perceive sounds above them rather than below them. Second, Wallach argued that perceived sound-source locations below the listener may have influenced listeners to underestimate elevation. It is worth noting that, at elevations other than directly above/below the listener or on the 0°-elevation azimuthal plane, the rate of change in interaural cues for a given head rotation is not constant but rather essentially a rectified sinusoidal function—see Fig. 6c. Thus, another possibility is that the linear estimation of the rate of sound-source rotation was too coarse an approximation to elicit the full illusion.

4.3 What Are the Cues for Head Position?

Both the horizontal and vertical illusions reported by Wallach (1940) suggest that head motion is a crucial variable in sound-source localization. Wallach (1940) assumed that "three types of sensory data represent a displacement of the head, that is, proprioceptive stimulation from the muscles engaged in active motion, stimulation of the eyes, and stimulation of the vestibular apparatus." In this section, some of the current knowledge regarding these and other possible head motion cues is reviewed.

Clearly, vision provides an important estimate of head position—except when our eyes are closed. Previous visual experience is nevertheless likely to be useful even with eyes shut (Zwiers et al. 2001)—see Sect. 3. Head and eyes often move independently, and nearly constant eye movements could make the formation of a stabilized image of the outside world impossible. To cope with this, the visual system employs an eye-centric reference system in addition to a head-centric reference system. To stabilize perception of visual objects, the vestibulo-ocular reflex (VOR) and the optokinetic reflex (OKR) work together to provide a means to correct the retinal output for retinal movement. There is some evidence that, in addition to a head-centric reference system, an eye-centric reference system that involves eye motion and sound-source localization may also play a role in sound-source localization—see Van Opstal (2016).

The vision literature shows that head-position signals can be used to "correct" spatial visual cues by use of efferent (efference) copies or corollary discharge signals—see Van Opstal (2016). The general idea is that when a neural signal is generated to control head position, a copy (*efferent copy or corollary discharge*) is also made. This copy is then integrated with the retinal spatial signal to yield a stable perception of the world. For instance, if there is a stationary light source and the head moves, the retinal output would change. The efferent copy/corollary discharge would indicate that it is the head that moved and not the light source. This efferent-copy signal could be used to effectively cancel the retinal change signal, yielding a veridical estimate of the location of the stationary visual source. In the visual literature, there are several well-established examples of such a "cancellation" based on both eye movements and head movements (Bridgeman and Stark 1991).

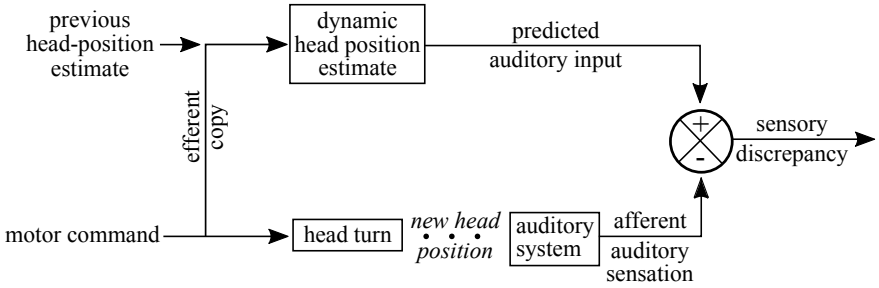

Fig. 7 A simplified schematic model for a possible role of an efferent-copy process in auditory localization

Figure 7 offers a simplified schematic of how efferent copy might work with auditory localization, based on an *efferent-copy* process in the visual system. Before a listener rotates the head, an internal estimate of the head position already exists. A motor command sends the signal to neck muscles and other involved systems to turn the head. Another "copy" of this signal is sent elsewhere in the brain so that a series of new, dynamic head position estimates can be made. Based on these estimates, the change in auditory spatial cues and/or auditory spatial estimates can be calculated. At approximately the same time, new afferent auditory activity offers a new spatial estimate of sound-source position that can be compared to the predicted auditory output, allowing the listener to determine whether the sound source has remained stationary or, if not, the position to which it has moved. While there is no direct physiological evidence for such efferent copy/corollary discharge processes in the mammalian auditory system, several authors (e.g., Wallach 1940; Brimijoin and Akeroyd 2012; Genzel et al. 2018; Freeman et al. 2017) have suggested such processes for sound-source localization.

The vestibular system also offers cues for determining a change in the head position. Vestibular cues result from the head's angular acceleration, which triggers hair-cell responses in the semi-circular canals that in turn elicit neural impulses to inform an estimate of head-position (Lackner and DiZio 2005). Because the otoliths in the vestibular system act as accelerometers, there is no vestibular output when the head is kept still, nor is there any output when the head rotates at constant velocity. In most experiments and most everyday experience, both passive and active listener rotation include changes in velocity—self-rotation necessarily includes acceleration and deceleration. Yost et al. (2015) appears to be the only study in which sound-source localization judgments were partitioned according to whether listeners rotated in an accelerating, decelerating, or constant manner—compare see Sect. 4.4 for details.

The arguments presented in this chapter imply that having access to the head-position angle is important to establishing a head-position cue. It is worth noting that the vestibular system provides information that the head is rotating, the direction of rotation, and the relative velocity of rotation, but the vestibular system cannot by itself indicate the world-centric position of the head since it directly encodes only

the change of the position of the head. The absolute head angle could be computed if the time over which the rotation had occurred and the starting head location were known—this would, however, require memory and other sensory inputs. This idea, in terms of establishing head-position cues for world-centric sound-source localization, appears to be unexplored.

Several sound-source-localization studies either indirectly infer or directly implicate proprioception and/or neural motor control of head rotation as ways to gain information about the current head position. In most cases, ideas from the vision literature are used to infer how proprioceptive outputs could inform head-position cues. When listeners move, neural-motor control signals are required to initiate and control the movement. These signals could indicate the angle of the head. In addition, it is possible that when listeners are rotated by some external means and must keep their heads still, that resistance to the rotational motion would stimulate muscles (e.g., neck muscles) which would trigger neural signals as a means of indicating head rotation. However, it isn't clear how such resistance would inform the estimate of head-position angle, nor is there much physiological evidence for how such proprioceptive/neural control signals interact with physiological sound-source localization processes.

The other three possible processes that might provide head-position cues, namely, auditory cues from sound sources other than the "target" stimulus, somatosensory cues, and cognitive processes (spatial maps) have not been studied as far as the authors can tell. It seems logically possible that these cues could inform the spatial system about head position and thereby contribute to sound-source localization—they should thus be investigated.

4.4 Recent Studies of Sound-Source Localization as a Multisensory Process

While Wallach's research is seminal in establishing a multisensory approach to understanding sound-source localization, there are several aspects of his work that need to be considered in light of current relevant knowledge. First, Wallach (1940) presented music played by a Victrola record player. Due to the constraints of the technology of the time, this likely means that the sound stimuli were essentially low-pass filtered, removing any useful HRTF/pinna cues (note that noise from scratches and dust on the record would also be filtered in the same way). This resulted in listener performance that led Wallach to believe the "pinna factor" was likely subservient to the integration of changing interaural cues with changing head position. Later experiments, reviewed below, would show that HRTF cues can remediate front-back reversals so that the listener hears the rotating sound source circling around the azimuth plane, and the Wallach Illusion fails.

Second, Wallach manually rotated listeners in a swivel chair back and forth over an arc of approximately 60°, with the eyes closed and the head fixed in a head holder.

Wallach also ran experiments with a rotating visual screen that induced the sensation of listener motion in the direction opposite to the screen's rotation to show that the Wallach Illusion could also be induced without listener movement, provided the listener received the same visual stimulation as would accompany a head movement (c.f., McAnally and Martin 2008). Unfortunately, the relative weightings of the different sensory and systems inputs were not measured in Wallach's experiments.

Perrett and Noble (1997a, b) attempted to replicate Wallach's elevation experiments. However, they were only able to replicate Wallach's findings for low-frequency sounds, suggesting that when reliable HRTF cues are present in the stimulus they override elevation cues derived from listener motion. For low-frequency stimuli, the correspondence between the predicted elevation and the actually judged elevations was only approximately 2/3 of the target elevation. Given the limited acuity of dynamic sound-source localization together with "binaural sluggishness," this result is not altogether surprising. Indeed, auditory resolution of elevation along the midline is also considerably poorer compared to localization on the azimuth plane.

Thus, until there are additional data, the current literature suggests that elevation cues provided by head motion are subservient and considerably less useful than HRTF spectral information for judging elevation. However, this may not be the case for machine listening. Zhong et al. (2016) used the Wallach concept to show that machine-learning algorithms (e.g., Kalman filters) could learn to use simulated head motion to determine the location of up to three different simultaneously presented sound sources located in different azimuthal and vertical locations.

Early work relating to head movement for the avoidance of front-back reversals and judging elevation can also be found in the 1938 thesis of Alva Wilska—see Kohlrausch and Altosaar (2011) and de Boer and van Urk (1941), also referenced in Blauert (1997).

Macpherson (2011) was interested in the relative weighting of spectral cues versus dynamic interaural differences in resolving front-back reversals. He designed an analogous version of the Wallach-Azimuth-Illusion experiment in a virtual auditory space, whereby he presented stimuli with various center frequencies and bandwidths. Data from only one listener have been reported. They indicate that when the stimulus was a low-pass noise (0.5–1 kHz), so that spectral cues were not available, listeners perceived a static sound source, front-back reversed to where it had originally been presented—as in Wallach (1940). Macpherson (2011) also tested narrow-band, high-frequency-noise stimuli and found that the Wallach Illusion failed. Macpherson thus suggested that this result could indicate that ILDs may not provide a sufficient basis for the dynamic auditory processing required for the Wallach Illusion. It should be noted, however, that Macpherson (2011) presented stimuli from in front of the listener, so that listeners would have to confuse a frontally-presented stimulus for one presented from behind. However, it has been repeatedly demonstrated that listeners tend to localize narrow-band, high-frequency stimuli to the frontal hemifield, independent of the actual location of presentation—e.g., Blauert 1969, 1997; Morimoto and Aokata 1984; Middlebrooks et al. 1989; Middlebrooks 1992. It may therefore be the case that the so-called "directional bands" are implicated in this result.

Brimijoin and Akeroyd (2012, 2017) also investigated the Wallach Azimuth Illusion. In their experiments, normal-hearing and hearing-impaired listeners moved their heads back and forth between ±15° of the midline. A camera system recorded the head motion and the system's output controlled amplitude panning of the sound such that the location of the *phantom sound source* was at twice the angle of the listener's head angle, thereby generating the "2–1" rotation necessary for the Wallach Azimuth Illusion. A low-pass filtered speech signal was presented and the filter cutoff was raised from 500 Hz to 16 kHz between conditions in octave steps. As listeners started to rotate their heads, a moving speech sound was either presented from a loudspeaker directly in front, or from a loudspeaker directly behind the listener. Listener responses indicated that they perceived a stationary sound source in the hemifield opposite to where the rotating sound was first presented. However, listeners responses were less robust in terms of replicating the Wallach Azimuth Illusion as the speech sounds included more and more high-frequency information. The authors state that "signals with the most high-frequency energy were often associated with an unstable location percept that flickered from front to back as self-motion cues and spectral cues for location came into conflict," perhaps suggesting that the brief duration of the presentations did not allow for the listeners to fully experience rotation.

Pastore and Yost (2017) and Yost et al. (2019, 2020) conducted an experiment that was an approximate replication of Wallach's (1940) study, but with a different means of rotating the listener and sound sources. The rate of front-back reversals (FBRs) was measured for noise stimuli under static listener/sound source conditions. Listeners were then rotated via a computer-controlled chair at a constant velocity of 45°/s. The sound-source rotated at twice the rate of listener rotation by way of saltatory motion from loudspeaker to loudspeaker around a circular array consisting of 24 equally spaced (15° apart) loudspeakers. Five differently filtered 200-ms noise bursts were tested, namely, three that generated more than 35% FBRs (FBR likely) and two that generated fewer than 6% FBRs (FBR unlikely)—for further details, see Fig. 8.

After eight seconds of stimulus presentation, the listener indicated the direction of rotation (clockwise or counterclockwise) for stimuli perceived as rotating, or the loudspeaker (separated by 60°, the same as in the first experiment) that most closely corresponded with the perceived static sound-source location.

Figure 8 depicts the effects of the stimulus spectrum and whether the listeners' eyes were open or shut. The left two panels show results when seven listeners' eyes were open, giving them information about the head position. The right two panels are the results from six of the same seven listeners when the listeners' eyes were closed—in a dark room and wearing a blindfold. One might expect that, in the eyes-closed condition, listeners have little or no access to information about the position of their head, thereby restricting their localization to the angular relation of the sound source to the head. In this case one would expect listeners to perceive a rotating sound source with their eyes closed, regardless of the stimulus frequency—see Yost et al. (2015) for more details about the assumptions regarding head-centric versus world-centric sound-source localization when attempts are made to eliminate head-position cues.

For the filtered noises that were prone to FBRs (FBR likely), listeners perceived the sound as being stationary when the eyes were open (consistent with the Wallach

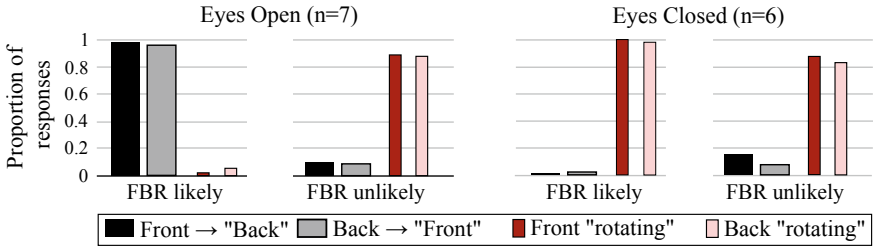

Fig. 8 Data from Pastore and Yost (2017). Results are pooled across noise stimuli that were likely to elicit front-back reversals ("FBR likely")—250-Hz center frequency, 2-octave and 1/10-octave bandwidths—and unlikely to elicit front-back-reversals ("FBR unlikely")—4-kHz center frequency, 2-octave and 1/10-octave bandwidths. At the beginning of either listener or sound-source rotation, the sound stimulus was presented from directly in front or from directly behind the listener. The listeners indicated whether the sound was perceived at a fixed location or whether it was rotating. For example, the proportion of those responses is shown by the black bars, for which stimuli that were presented either in front or from behind were indicated to be stationary

Azimuth Illusion), but rotating when the eyes were closed. For the filtered noises that are not prone to FBRs (FBR unlikely), listeners nearly always perceived the sound as rotating in both the eyes-open and eyes-closed cases, indicating that the Wallach Illusion fails for stimuli that are not prone to FBRs. In the eyes-open condition with the listener facing forward, the perception of a stationary noise source was nearly 100% of the time at the rear loudspeaker when the sound was presented from the front, and at the 'frontal loudspeaker when the sound was presented from behind. When listeners' eyes were closed and the sounds were not likely to produce FBRs, listeners always indicated that the sound rotated clockwise—as the actual sound did. When the eyes were closed and the noise was likely to elicit FBRs, listeners indicated that the sound was rotating in a clockwise direction most of the time, but occasionally counterclockwise rotation was also indicated. However, one listener's responses in the eyes-closed condition when FBRs existed was not consistent with the other five listeners' responses. Thus, listeners' perception with their eyes closed, needs further investigation.

Brimijoin and Akeroyd (2014) studied the moving minimum-audible angle (MMAA), that is, the minimum-perceivable angle between two sound sources when the angular displacements of two sound sources change relative to a listener's head. They reported that, when listeners rotated their heads and the sound sources were stationary, the MMAA was 1–2° smaller than when listeners kept their heads still and the sound sources rotated around the listener with the same angular velocity and displacement as the listeners' previous head turns. Brimijoin and Akeroyd (2014) concluded that "spatial processing involves an ongoing and highly accurate comparison of spatial acoustic cues with self-motion cues."

Brimijoin (2018) showed that the perceived motion of a moving sound source differs depending on its angular displacement. Sounds to the sides of listeners needed to be moved more than twice as far as sounds near midline for both sounds to

appear to have moved the same amount. How this relative compression/expansion of auditory space interacts with head position cues is unknown. One possibility is that the comparison of the two rates of motion is not very precise. Another possibility is that other inputs are employed, such as vision, to compensate for these distortions of auditory space, as considered in Sect. 3.

Yost et al. (2015) investigated several aspects of sound-source location when listeners were rotated in a chair and their eyes were either open or closed. They argued that listeners had little or no information about the position of their head when they rotated in the chair at constant velocity, and, with their eyes closed, there were no visual cues. Under these conditions, listeners perceived stationary sound sources as rotating. When sound sources and the listener rotated at the same rate, listeners perceived a stationary sound source—entirely consistent with localization based on a head-centric reference system. When these listeners' eyes were closed and the rotation was accelerating or decelerating, the results were somewhat mixed. Yost et al. (2015) point out that there were possible confounds in their procedures, making it difficult to unambiguously determine the role of vestibular acceleration/deceleration cues in judging head position. How these cues might thereby influence sound-source localization is therefore also unclear.

Genzel et al. (2016) investigated "spatial updating," the process of mapping the head-centric auditory estimate of sound-source position to the listener's spatial map of the surrounding environment using successive estimates of head position. In three different experimental conditions, blindfolded listeners were either, (1), asked to move their head according to a trained rotational trajectory, (2), passively moved along the same trajectory, or (3), counter-rotated as a function of head rotation, such that a given head rotation resulted in no change in head position relative to the surrounding environment. In a two-alternative forced-choice experiment, listeners reported whether they heard a test sound to the right or left of a previously presented reference sound. Listeners were most accurate when passively rotated and least accurate when they moved their own heads. Genzel et al. (2016) modeled the integration of head-centric auditory spatial inputs and world-centric head position information as a linear addition, dividing head-motion cues into vestibular cues and proprioceptive/efference copy cues, with visual inputs zeroed out due to the listener being blindfolded. They determined that both proprioceptive/efference copy and vestibular cues play a role in determining head position, but that vestibular cues are weighted more heavily. While there are several untested assumptions underlying their interpretations, their data clearly indicate support for the notion that sound-source localization depends on the integration of head-motion and auditory spatial cues, and that vestibular function and proprioception/efferent copy are possibly used as indicators of head position.

Wightman and Kistler (1999) investigated whether head movements could be used to disambiguate front/back sound-source localization along cones of confusion. The authors tested this under four scenarios: (1), no head movement allowed, (2), the listeners moved their heads, (3), the listeners did not move their heads, but the sound source was moved by the experimenter, and (4), the listener did not move their heads, but they themselves moved the sound source via key presses on a computer keyboard.

Wightman and Kistler found that head movements reduced the front/back errors to almost zero (see Braasch et al. 2013, for related modeling). Listeners also reduced front/back reversals to a minimum when they themselves moved the sound source via keyboard with no visual or vestibular feedback. No such benefit was found when the experimenter moved the sound source. This important finding suggests that mapping head-related sound-source localization to the local environment involves cognitive processing that uses whatever information and spatial estimates that appear to be useful.

Motion parallax is a powerful cue used in vision to judge relative distance (Steinman and Garzia 2000). Genzel et al. (2018) demonstrated the possibility that motion parallax might play a similar role in judging the relative distance of sound sources. Their main experiment used a virtual panning process to present two sounds (a low-pitched and a high-pitched sound) at two different panned (virtual) distances. With no motion of the sounds or the listeners, stationary listeners could not determine whether one sound was further away from the other, since all known distance cues were eliminated. When listeners moved, they were much better at discriminating the differences in panned distances than when the sound source moved and the listener was stationary. In other words, listeners could infer that one sound was closer than the other one by exploiting the perceptual effect that the near-panned source appeared to move faster than the far-panned one while the head was moving. This is consistent with the visual analogy. There was a small decrement in performance when the listeners were moved on a platform rather than moving themselves. This suggests that proprioceptive cues for self-motion are involved when judging sound-source distance.

5 A Concept for a Model

This section offers a descriptive model of sound-source localization, based on Wallach's (1940) insight that auditory spatial information must be integrated with an estimate of the location of the listener's head relative to the surrounding environment to provide an estimate of the location of the sound source in that environment. Because the range of possible inputs is large, and their respective temporal-processing speeds and parameter spaces are potentially very different, the model offered here is not yet actually implemented but rather of a conceptual kind. In particular, it does not yet specify details of how the various inputs to the model are combined and compared.

Two crucial points should be mentioned at the outset. First, full development of the model requires further studies of the multi-system/multisensory interactions that are involved in auditory localization and in the generation of dynamic, multisensory spatial maps. Second, the model is not yet available as a flowchart, because this would be too complex. In fact, the overall process is not simply feed-forward, but rather includes feedback and other interactions between system elements and sensory input/output—compare Chap. 1, this volume.

Also, there is, as of yet, no controlled experiment available to test such a model even if it were precisely specified. As the literature reviewed in this chapter shows, only small parts of the overall process can be investigated at a time and subsequently modeled. As such, the individual model structures may differ in kind. For example, audio-visual interaction, considered in Sect. 3, may be adequately modeled within a Bayesian framework. Whether interactions between memory, attention, motor processes, etc., can also be modeled in this way is unclear. Also, the putative *spatial map* may be a dynamic system of various spatial maps with different references along many dimensions. In other words, the proposed model concept of auditory localization as a multisensory/multi-system process is primarily intended to be a tool for orientation in a yet largely unknown territory.

This section uses a notation where Θ'_{a_b} denotes an estimate (indicated by the prime) of angular displacement in polar space, Θ, relative to the frame of reference, a, and in terms of the type of input, b. It is worth noting that one cannot be entirely sure what all the *frames of reference* might actually be. There appear to be *head-centric* and *world-centric* frames of reference as illustrated in Fig. 1, but there may be *body-centric* and other frames of reference as well. For example, a listener with closed eyes may be able to point to the location of a sound source while not being able to place that location into the context of other objects in the room, for instance, for reaching out to grab a buzzing mosquito in the dark. Even this example still requires some internal map of, at least, the body. Nevertheless, the basic argument of the model concept is that the auditory estimate of the location of a perceived sound source within the context of the local environment (world-centric localization, Θ'_{w_A}), is determined by the integration of an auditory estimate of the sound source's location relative to the head based on auditory spatial cues, Θ'_{A_h}, with a multisensory/multi-systems estimate of the location of the head relative to the local environment, Θ'_{w_h}.

This descriptive model does not specify how Θ'_{A_h} and Θ'_{w_h} would be combined, but rather suggests that sound-source localization requires the integration of information—including, but not limited to, perceptual cues from several (perhaps many) neural systems. This includes cognitive processes such as experience and memory—compare Buzsáki and Llinás (2017). The model assumes that any such integration also involves an assessment of the reliability of the cues employed for each estimate. Furthermore, the model assumes that each cue estimate and each estimate resulting from the integration of those estimates introduces error, (ξ_i). For example, Θ'_{h_A} is determined by weighted integration of the auditory spatial cues, and Θ'_{w_h} is determined by a weighted sum of the multi-system head-position cues mentioned above. The weight, w_i, of any particular auditory spatial or head-position estimate would be proportional to the external noise of the cue, due to variability in the stimulus along the relevant dimension, together with an internal noise term, ξ_i, that arises from the variability inherent to neural processes in general. Combining these estimates to arrive at Θ'_{w_A} introduces further error, again due to internal noise. The initial auditory estimate Θ'_{h_A} relates sound-source position to the head as follows.

$$\Theta'_{h_A} \propto [w_{\Psi'}\Psi', w_{\nu'}\nu'], \tag{7}$$

where Ψ' is an estimate of the angle of the sound source relative to the interaural axis, the *lateral angle* in Wallach's terminology—see Sect. 4.2. Ψ' is therefore a component of Θ'_{h_A} that is based on interaural differences of time, Ψ'_{ITD}, and level, Ψ'_{ILD}. Note that Ψ' is not simply an estimate of azimuth—the same interaural differences exist along a range of locations on the *cones of confusion* as discussed throughout this chapter. ν' is a polar elevation estimate on a sagittal plane normal to the interaural axis of Θ'_{h_A}, based on spectral HRTF cues. Therefore, *both* the component estimates, Ψ' and ν', are required for a single-valued estimate of the sound-source location, Θ'_{h_A}. Furthermore, it is unclear whether elevation, ν', can be estimated without an initial interaural-difference estimate, Ψ', to specify which sagittal plane will be the basis for the elevation estimate,

$$\Theta'_{w_h} \propto [w_V V, w_\Gamma \Gamma, w_B B, w_A A, w_C C], \tag{8}$$

where V ... vision, Γ ... vestibular cues, B ... body awareness (e.g., proprioceptive, somatosensory, kinesthetic, neuro-motor control), A ... auditory, and C ... cognitive processes, which include expectation, memory, and attention. To determine an estimate of the sound-source position in the surrounding environment, the head-centric estimate, Θ'_{h_A}, must be combined with the estimate of head position, Θ'_{w_h}. This process could be analogous to simply adding the two estimates, or perhaps one is mapped onto the other—the actual mechanism is not understood at this time and, consequently, not specified in the following expression.

$$\Theta'_{w_A} \propto [\Theta'_{h_A}, \Theta'_{h_w}, C, \chi], \tag{9}$$

where χ denotes interactions between various estimates of Θ_w, such as auditory, Θ'_{w_A}, and visual, Θ'_{w_V}.

Several points are worth noting. First, although this model concept is expressed as a series of mathematical expressions, this form has only been chosen for convenience. The inputs and interactions between them for each of the spatial estimates are still largely unspecified. For example, the model could further include head-position cues that future research may suggest. The model is not expressed as addition of individual estimates since they may interact in non-linear ways. The model concept is meant to, hopefully, provide a structure to motivate experiments, the results of which could alter this putative model considerably. Second, the relative weighting of different sensory/systems input can be such that one or several are completely disregarded in a given estimate. For example, when listeners' eyes are closed, their visual input is probably not considered in any internal head-position estimate. Third, it is worth noting that Θ'_{A_w} is only one spatial estimate of a perceived sound source in the context of the surrounding environment among several other estimates from other sensory modalities, as well as from cognition. For example, if an estimate, Θ'_{V_w}, of the location of a visual object is perceptually grouped with the sound object associated with Θ'_{A_w}, these estimates will likely interact, either reinforcing each other or leading to cross-modal capture. Memory or expectation could play a similar role in this regard. This possibility is denoted by χ.

In summary, it is hoped that the contribution of the model concept is to point out that the roles in auditory localization as being played by several components of the assumed input, such as proprioceptive, somatosensory, kinesthetic, neuromotor control, cognitive processing, and spatial auditory input used to determine head position, are still not clearly understood. Thus, the model primarily points to what remains unknown rather than at what is known.

Acknowledgements The work reported here is supported by the National Science Foundation (No. NSF BCS-1539376), the National Institute for Deafness and Communication Disorders (Nos. R0101DC015214 and F32DC017676), and Facebook Reality Labs. The authors are indebted to two anonymous reviewers for constructive comments and suggestions.

References

Alais, D., and D. Burr. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* 14: 257–62. https://doi.org/10.1016/j.cub.2004.01.029.

Anderson, P.W., P. Zahorik, J.A. Schirillo, and W. Forest. 2014. Auditory/visual distance estimation: accuracy and variability. *Frontiers in Psychology* 5 (October): 1–11. https://doi.org/10.3389/fpsyg.2014.01097.

Battaglia, P.W., R.A. Jacobs, and R.N. Aslin. 2003. Bayesian integration of visual and auditory signals for spatial localization. *The Journal of the Optical Society of America* 20 (7): 1391–1397.

Baumgartner, R., P. Majdak, and B. Laback. 2013. Assessment of sagittal-plane localization performance. In *The Technology of Binaural Listening*, ed. J. Blauert, 93–119. Berlin- Heidelberg-New York: Springer and ASA Press.

Berkeley, G. 1709. An Essay towards a New Theory of Vision. https://www.maths.tcd.ie/~dwilkins/Berkeley/Vision/1709A/Vision.pdf (last accessed Dec. 20, 2020).

Bernstein, L.R., and C. Trahiotis. 2011. Lateralization produced by envelope-based interaural temporal disparities of high-frequency, raised-sine stimuli: empirical data and modeling. *The Journal of the Acoustical Society of America* 129 (3): 1501–8. https://doi.org/10.1121/1.3552875.

Bertelson, P., and M. Radeau. 1981. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception and Psychophysics* 29 (6): 578–584.

Bertelson, P., J. Vroomen, B. de Gelder, and J. Driver. 2000. The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception and Psychophysics* 62 (2): 321–332.

Best, V., D.S. Brungart, S. Carlile, N. Jin, E.A. Macpherson, R.L. Martin, K.I. McAnally, A.T. Sabin, and B.D. Simpson. 2011. A meta-analysis of localization errors made in free field. In *Principles and Applications of Spatial Hearing*, vol. 1, ed. Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato, 14–23. Singapore: World Scientific Publishing. https://doi.org/10.1142/7674.

Blauert, J. 1969. Sound localization in the median plane. *Acustica* 22, 205–213.

Blauert, J. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*, 222–237. Cambridge: MIT Press.

Boring, E.G. 1926. Auditory theory with special reference to intensity, volume, and localization. *The American Journal of Psychology* 37 (2): 157–188.

Boring, E.G. 1942. *Sensation and Pareception in the History of Experimental Psychology*. New York: Appleton-Century-Crofts.

Braasch, J., S. Clapp, A. Parks, M.T. Pastore, and N. Xiang. 2013. A binaural model that analyses aural spaces and stereophonic reproduction systems by utilizing head movements. In *The Technology of Binaural Listening*, vol. 8, ed. J. Blauert, 201–224. Springer and ASA Press.

Bridgeman, B., and L. Stark. 1991. Ocular proprioception and efference copy in registering visual direction. *Vision Research* 31 (11): 1903–1913. https://doi.org/10.1016/0042-6989(91)90185-8.

Brimijoin, W.O. 2018. Angle-dependent distortions in the perceptual topology of acoustic space. *Trends in Hearing* 22: 1–11. https://doi.org/10.1177/2331216518775568.

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): 179–181. https://doi.org/10.1068/i7173sas.

Brimijoi, W.O., and M.A. Akeroyd. 2014. The moving minimum audible angle is smaller during self motion than during source motion. *Frontiers in Neuroscience* 8: 1–8. https://doi.org/10.3389/fnins.2014.00273.

Brimijoin, W.O., and M.A. Akeroyd. 2017. The effects of hearing impairment, age, and hearing aids on the use of self motion for determining front/back location. *Journal of the American Academy of Audiology* 27 (7): 588–600. https://doi.org/10.3766/jaaa.15101.

Bronkhorst, A.W., and T. Houtgast. 1999. Auditory distance perception in different rooms. *Nature* 397: 517–520.

Brungart, D.S., N.I. Durlach, and W.M. Rabinowitz. 1999. Auditory localization of nearby sources. II. Localization of a broadband source. *Journal of the Acoustical Society of America* 106 (4): 1956–1968. https://doi.org/10.1121/1.427943.

Buzsáki, G., and R. Llinás. 2017. Space and time in the brain. *Science* 358 (October): 482–485.

Carlile, S., and T. Blackman. 2014. Relearning auditory spectral cues for locations inside and outside the visual field. *Journal of the Association for Research in Otolaryngology* 15 (2): 249–263. https://doi.org/10.1007/s10162-013-0429-5.

Carlile, S., and J. Leung. 2016. The perception of auditory motion. *Trends in Hearing* 20: 1–19. https://doi.org/10.1177/2331216516644254.

Choe, C.S., R.B. Welch, R.M. Gilford, and J.F. Juola. 1975. The "ventriloquist effect": Visual dominance or response bias? *Perception and Psychophysics* 18 (1): 55–60.

Curcio, C.A., K.R. Sloan, R.E. Kalina, and A.E. Hendrickson. 1990. Human photoreceptor topography. *Journal of Comparative Neurology* 292 (4): 497–523. https://doi.org/10.1002/cne.902920402.

de Boer, K., and A.T. van Urk. 1941. Some particulars of directional hearing. *Philips Technical Review* 6: 359–364.

Dorman, M.F., L.H. Loiselle, S.J. Cook, W.A. Yost, and R.H. Gifford. 2016. Sound source localization by normal hearing listeners, hearing-impaired listeners and cochlear implant listeners. *Audiology and Neurotology* 21: 127–131.

Ernst, M.O., and M.S. Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415 (6870): 429–433.

Ernst, M.O., and H.H. Bülthoff. 2004. Merging the senses into a robust percept. *Trends in Cognitive Sciences* 8 (4): 162–169.

Freeman, T.C.A., J.F. Culling, M.A. Akeroyd, and W.O. Brimijoin. 2017. Auditory compensation for head rotation is incomplete. *Journal of Experimental Psychology* 43 (2): 371–380. https://doi.org/10.1037/xhp0000321.

Genzel, D., U. Firzlaff, L. Wiegrebe, and P.R. MacNeilage. 2016. Dependence of auditory spatial updating on vestibular, proprioceptive, and efference copy signals. *Journal of Neurophysiology* 116 (2): 765–775. https://doi.org/10.1152/jn.00052.2016.

Genzel, D., M. Schutte, W.O. Brimijoin, and P.R. MacNeilage. 2018. Psychophysical evidence for auditory motion parallax. *Proceedings of the National Academy of Sciences* 115 (6): 4264–4269. https://doi.org/10.1073/pnas.1712058115.

Good, M.D., and R.H. Gilkey. 1996. Sound localization in noise: the effect of signal-to-noise ratio. *The Journal of the Acoustical Society of America* 99 (2): 1108–17. https://doi.org/10.1121/1.415233.

Goupell, M.J., and O.A. Stakhovskaya. 2018. Across-channel interaural-level-difference processing demonstrates frequency dependence. *The Journal of the Acoustical Society of America* 143 (2): 645–658. https://doi.org/10.1121/1.5021552.

Hairston, W.D., M.T. Wallace, J.W. Vaughan, B.E. Stein, J.L. Norris, and J.A. Schirillo. 2003. Visual localization ability influences cross-modal bias. *Journal of Cognitive Neuroscience* 15 (1): 20–29.

Hartmann, W.M., and B. Rakerd. 1989. On the minimum audible angle-a decision theory approach. *The Journal of the Acoustical Society of America* 85 (5): 2031–2041.

Hartmann, W.M., B. Rakerd, Z.D. Crawford, and P.X. Zhang. 2016. Transaural experiments and a revised duplex theory for the localization of low-frequency tones. *The Journal of the Acoustical Society of America* 139 (2): 968. https://doi.org/10.1121/1.4941915.

Hofman, P.M., J.G.A. van Riswick, and A.J. van Opstal. 1998. Relearning sound localization with new ears. *Nature Neuroscience* 1 (5): 417–421.

Howard, I.P., and W.B. Templeton. 1996. *Human Spatial Orientation*, 359–362. New York: Wiley.

Humanski, R.A., and R.A. Butler. 1988. The contribution of the near and far ear toward localization of sound in the sagittal plane. *The Journal of the Acoustical Society of America* 83 (6): 2300–2310. https://doi.org/10.1121/1.396361.

Jack, C.E., and W.R. Thurlow. 1973. Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and Motor Skills* 37 (3): 967–979.

Jackson, C.V. 1953. Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology* 5: 52–65.

Jin, C., A. Corderoy, S. Carlile, and A. van Schaik. 2004. Contrasting monaural and interaural spectral cues for human sound localization. *The Journal of the Acoustical Society of America* 115 (6): 3124–3141. https://doi.org/10.1121/1.1736649.

Jones, B., and B. Kabanoff. 1975. Eye movements in auditory space perception. *Perception and Psychophysics* 17 (3): 241–245. https://doi.org/10.3758/BF03203206.

King, A.J. 2009. Visual influences on auditory spatial learning. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1515): 331–339. https://doi.org/10.1098/rstb.2008.0230.

Knudsen, E.I., and M.S. Brainard. 1995. Creating a unified representation of visual and auditory space in the brain. *Annual Review of Neuroscience* 18: 19–43. https://doi.org/10.1146/annurev.neuro.18.1.19.

Kohlrausch, A., and T. Altosaar. 2011. Early research on spatial hearing by Alvar Wilska (1911–1987). *Forum Acusticum*, 1103–1108. Aalborg: European Acoustics Association.

Kolarik, A.J., B.C. Moore, P. Zahorik, S. Cirstea, and S. Pardhan. 2016. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, and Psychophysics* 78 (2): 373–395. https://doi.org/10.3758/s13414-015-1015-1.

Kording, K.P., U. Beierholm, W.J. Ma, S. Quartz, J.B. Tenenbaum, and L. Shams. 2007. Causal inference in multisensory perception. *PLoS One* 2 (9): e943. https://doi.org/10.1371/journal.pone.0000943.

Kuhn, G.F. 1977. Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America* 62 (1): 157–167. https://doi.org/10.1121/1.381498.

Kuhn, G.F. 1987. Physical acoustics and measurements pertaining to directional hearing. In *Directional Hearing*, eds. W.A. Yost and G. Gourevitch, Chap 1, 3–25. Springer Nature. https://doi.org/10.1007/978-1-4612-4738-8_1.

Lackner, J.R., and P. DiZio. 2005. Vestibular, proprioceptive, and haptic contributions to spatial orientation. *Annual Review of Psychology* 56 (1): 115–147. https://doi.org/10.1146/annurev.psych.55.090902.142023.

Langendijk, E.H.A., and A.W. Bronkhorst. 2002. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America* 112 (4): 1583. https://doi.org/10.1121/1.1501901.

Letowski, T., and S. Letowski. 2011. Localization error: accuracy and precision in auditory localization. In *Advances in Sound Localization, Chap. 4*, ed. P. Strumillo, 55–78. London: Intech Open. https://doi.org/10.5772/597.

Macaulay, E.J., W.M. Hartmann, and B. Rakerd. 2010. The acoustical bright spot and mislocalization of tones by human listeners. *Journal of the Acoustical Society of America* 127 (3): 1440–1449. https://doi.org/10.1121/1.3294654.

Macaulay, E.J., B. Rakerd, T.J. Andrews, and W.M. Hartmann. 2017. On the localization of high-frequency, sinusoidally amplitude-modulated tones in free field. *Journal of the Acoustical Society of America* 141 (2): 847–863. https://doi.org/10.1121/1.4976047.

Macpherson, E.A. 2011. Head motion, spectral cues, and Wallach's 'principle of least displacement' in sound localization. In *Principles and Applications of Spatial Hearing, Chap. 9*, ed. Y. Suzuki, D. Brungart, and H. Kato, 103–120. Singapore: World Scientific.

Macpherson, E.A., and J.C. Middlebrooks. 2002. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America* 111 (5): 2219. https://doi.org/10.1121/1.1471898.

Makous, J.C., and J.C. Middlebrooks. 1990. Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America* 87 (5): 2188–2200. https://doi.org/10.1121/1.399186.

Martin, R.L., M. Paterson, and K.I. McAnally. 2004. Utility of monaural spectral cues is enhanced in the presence of cues to sound-source lateral angle. *Journal of the Association for Research in Otolaryngology* 5 (1): 80–89. https://doi.org/10.1007/s10162-003-3003-8.

McAnally, K.I., and R.L. Martin. 2008. Sound localisation during illusory self-rotation. *Experimental Brain Research* 185 (2): 337–40. https://doi.org/10.1007/s00221-007-1157-z.

Mendonça, C. 2020. Psychophysical models of sound localisation with audiovisual interactions. In *The Technology of Binaural Understanding*. Springer, ed. J. Blauert, and J. Braasch, 289–314. Cham, Switzerland: Springer and ASA Press.

Middlebrooks, J.C. 1992. Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America* 92 (5): 2607–24.

Middlebrooks, J.C., J.C. Makous, and D.M. Green. 1989. Directional sensitivity of sound—pressure levels in the human ear canal. *The Journal of the Acoustical Society of America* 86 (1): 89–107. https://doi.org/10.1121/1.398224.

Middlebrooks, J.C., L. Xu, S. Furukawa, and E.A. Macpherson. 2002. Cortical neurons that localize sounds. *Neuroscientist* 8 (1): 73–83.

Mills, A.W. 1960. Lateralization of high-frequency tones. *The Journal of the Acoustical Society of America* 32 (1): 132–134.

Mills, A.W. 1972. Auditory localization. In *Foundations of Modern Auditory Theory*, ed. J.V. Tobias, 303–348. New York: Academic Press.

Montagne, C., and Y. Zhou. 2016. Visual capture of a stereo sound : Interactions between cue reliability, sound localization variability, and cross-modal bias. *The Journal of the Acoustical Society of America* 140 (July): 471–485. https://doi.org/10.1121/1.4955314.

Montagne, C., and Y. Zhou. 2018. Audiovisual interactions in front and rear space. *Frontiers in Psychology* 9 (MAY): 1–15. https://doi.org/10.3389/fpsyg.2018.00713.

Morimoto, M. 2001. The contribution of two ears to the perception of vertical angle in sagittal planes. *The Journal of the Acoustical Society of America* 109 (4): 1596–1603. https://doi.org/10.1121/1.1352084.

Morimoto, M., and H. Aokata. 1984. Localization cues of sound sources in the upper hemisphere. *Journal of the Acoustical Society of Japan* 5 (3): 165–173. https://doi.org/10.1250/ast.5.165.

Musicant, A.D., and R.A. Butler. 1984. The influence of pinnae-based spectral cues on sound localization. *Journal of the Acoustical Society of America* 75 (4): 1195–1200. https://doi.org/10.1121/1.390770.

Pastore, M.T., and W.A. Yost. 2017. Sound source localization as a multisensory process: The Wallach azimuth illusion. *The Journal of the Acoustical Society of America* 141 (5): 3635–3635.

Perrett, S., and W. Noble. 1997a. The contribution of head motion cues to localization of low-pass noise. *Perception and Psychophysics* 59 (7): 1018–1026. https://doi.org/10.3758/BF03205517.

Perrett, S., and W. Noble. 1997b. The contribution of head motion cues to localization of low-pass noise. *Perception and Psychophysics* 59 (7): 1018–1026.

Pick, H.L., D.H. Warren, and J.C. Hay. 1969. Sensory conflict in judgments of spatial direction. *Perception and Psychophysics* 6 (4): 203–205.

Pierce, A. 1901. *Studies in Auditory and Visual Space Perception*. New York: Longmans, Green, and Co.
Platt, B.B., and D.H. Warren. 1972. Auditory localization: The importance of eye movements and a textured visual environment. *Perception and Psychophysics* 12 (2B): 245–248.
Radeau, M., and P. Bertelson. 1974. The after-effects of ventriloquism. *The Quarterly Journal of Experimental Psychology* 26 (1): 63–71. https://doi.org/10.1080/14640747408400388.
Radeau, M., and P. Bertelson. 1977. Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception and Psychophysics* 22 (2): 137–146. https://doi.org/10.3758/BF03198746.
Radeau, M., and P. Bertelson. 1987. Auditory-visual interaction and the timing of inputs. Thomas (1941) revisited. *Psychological Research* 49 (1): 17–22.
Rayleigh, L. 1876. On our perception of the direction of a source of sound. In *Proceedings of the Musical Association*, vol. 2, 75–84.
Recanzone, G.H. 2009. Interactions of auditory and visual stimuli in space and time. *Hearing Research* 258 (1–2): 89–99. https://doi.org/10.1016/j.heares.2009.04.009.
Searle, C.L. 1973. Cues required for externalization and vertical localization. *The Journal of the Acoustical Society of America* 54: 308. https://doi.org/10.1121/1.1978213.
Shelton, B.R., and C.L. Searle. 1980. The influence of vision on the absolute identification of sound-source position. *Perception and Psychophysics* 28 (6): 589–96.
Sivia, D.S., and J.S. Skilling. 2006. *Data Analysis: A Bayesian Tutorial*, 2nd ed. New York: Oxford University Press.
Slattery, W.H., and J.C. Middlebrooks. 1994. Monaural sound localization: acute versus chronic unilateral impairment. *Hearing Research* 75 (1): 38–46.
Slutsky, D.A., and G.H. Recanzone. 2001. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* 12 (1): 7–10.
Solman, G.J., T. Foulsham, and A. Kingstone. 2017. Eye and head movements are complementary in visual selection. *Royal Society Open Science* 4 (1): 160569. https://doi.org/10.1098/rsos.160569.
Spence, C., and J. Driver. 2000. Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. *Neuroreport* 11 (9): 2057–2061.
Stein, B.E., and M.A. Meredith. 1990. Multisensory integration. Neural and behavioral solutions for dealing with stimuli from different sensory modalities. *Annals of the New York Academy of Sciences* 608: 51–70.
Steinman, S.B., and R.P. Garzia. 2000. *Foundations of Binocular Vision: A Clinical perspective*, 2–5. McGraw-Hill Professional
Stensola, T., and E.I. Moser. 2016. Grid cells and spatial maps in entorhinal cortex and hippocampus. In *Micro-, Meso- and Macro-Dynamics of the Brain. Research and Perspectives in Neurosciences*, ed. G. Buzsáki, and Y. Christen, 59–80. Cham: Springer. https://doi.org/10.1007/978-3-319-28802-4.
Thompson, S.P. 1878. On binaural audition. *Philosophical Magazine* 2 (6): 383–391.
Thurlow, W.R., and C.E. Jack. 1973. Certain determinants of the 'ventriloquism effect'. *Perceptual and Motor Skills* 36 (3): 1171–1184. https://doi.org/10.2466/pms.1973.36.3c.1171.
Thurlow, W.R., and T.P. Kerr. 1970. Effect of a moving visual environment on localization of sound. *The American Journal of Psychology* 83 (1): 112–118.
Van Opstal, A.J. 2016. *The Auditory System and Human Sound-Localization Behavior*, 1st ed, 436. Amsterdam: Academic Press.
Van Wanrooij, M.M., and A.J. Van Opstal. 2004. Contribution of head shadow and pinna cues to chronic monaural sound localization. *Journal of Neuroscience* 24 (17): 4163–4171. https://doi.org/10.1523/JNEUROSCI.0048-04.2004.
Wallace, M.T., R. Ramachandran, and B.E. Stein. 2004. A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Sciences* 101 (7): 2167–2172. https://doi.org/10.1073/pnas.0305697101.
Wallach, H. 1938. Über die Wahrnehmung der Schallrichtung (On the perception of sound direction). *Psychologische Forschung* 22 (3–4): 238–266.

Wallach, H. 1939. On sound localization. *The Journal of the Acoustical Society of America* 10 (4): 270–274.
Wallach, H. 1940. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology* 27 (4): 339–368.
Warren, D.H. 1970. Intermodality interactions in spatial localization. *Cognitive Psychology* 1 (2): 114–133. https://doi.org/10.1016/0010-0285(70)90008-3.
Welch, R.B., and D.H. Warren. 1980. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin* 88 (3): 638.
Wenzel, E.M., M. Arruda, D.J. Kistler, and F.L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94 (1): 111–123.
Wightman, F.L., and D.J. Kistler. 1997. Monaural sound localization revisited. *The Journal of the Acoustical Society of America* 101 (2): 1050–63. https://doi.org/10.1121/1.418029.
Wightman, F.L., and D.J. Kistler. 1999. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America* 105 (5): 2841–53. https://doi.org/10.1121/1.426899.
Woodworth, R., and H. Schlosberg. 1938. *Experimental Psychology*. New York: Henry Holt and Company.
Yost, W.A. 1981. Lateral position of sinusoids presented with interaural intensive and temporal differences. *The Journal of the Acoustical Society of America* 70 (2): 397–409. https://doi.org/10.1121/1.386775.
Yost, W.A. 2016. Sound source localization identification accuracy: Level and duration dependencies. *The Journal of the Acoustical Society of America* 140 (1): EL14–EL19. https://doi.org/10.1121/1.4898045.
Yost, W.A. 2017a. History of sound source localization: 1850–1950. *The Journal of the Acoustical Society of America* 30: 1–15. https://doi.org/10.1121/2.0000529.
Yost, W.A. 2017b. Sound source localization identification accuracy: Envelope dependencies. *The Journal of the Acoustical Society of America* 142 (1): 173–185. https://doi.org/10.1121/1.4990656.
Yost, W.A., M.T. Pastore, and K.R. Pulling. 2019. Sound source localization as a multisystem process: the Wallach azimuth illusion. *The Journal of the Acoustical Society of America* 146 (1): 382–398.
Yost, W.A., M.T. Pastore, and M.F. Dorman. 2020. Sound source localization is a multisystem process. *Acoustical Science and Technology* 41 (1).
Yost, W.A., X. Zhong, and A. Najam. 2015. Judging sound rotation when listeners and sounds rotate: Sound source localization is a multisystem process. *The Journal of the Acoustical Society of America* 138 (5): 3293–3310. https://doi.org/10.1121/1.4920001.
Zahorik, P. 2002. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America* 112 (5): 2110. https://doi.org/10.1121/1.1506692.
Zahorik, P., P. Bangayan, V. Sundareswaran, K. Wang, and C. Tam. 2006. Perceptual recalibration in human sound localization: Learning to remediate front-back reversals. *The Journal of the Acoustical Society of America* 120 (1): 343–359. https://doi.org/10.1121/1.2208429.
Zhong, X., L. Sun, and W.A. Yost. 2016. Active binaural localization of multiple sound sources. *Robotics and Autonomous Systems* 85: 83–92. https://doi.org/10.1016/j.robot.2016.07.008.
Zwiers, M.P., and A.J. van Opstal. 2001. A spatial hearing deficit in early-blind humans. *Journal of Neuroscience* 21 (9): RC142.