Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Natural spatial hearing relies on binaural spectral disparity to cancel source-spectrum ambiguities

Timo Oess

timo.oess@uni-ulm.de

Universität Ulm

Charlotte Roy

Universität Ulm

Heiko Neumann

Universität Ulm

Marc Ernst

Ulm University

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11441293/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Natural spatial hearing relies on binaural spectral disparity to cancel source-spectrum ambiguities

Timo Oess[1,2,*], Charlotte Roy[1], Heiko Neumann[2], Marc O. Ernst[1],

1 Applied Cognitive Psychology, Ulm University, Germany
2 Institute of Neural Information Processing, Ulm University, Germany

*** timo.oess@uni-ulm.de**

Abstract

How do we localize natural sounds? Interaural time and level differences convey little more than left versus right; elevation, front–back position and externalisation depend on spectral cues. These cues are thought to be read out monaurally, yet the spectrum at each ear confounds the unknown source spectrum with the head's directional filter. Because natural sounds typically carry their own spectral structure, monaural localization is an ill-posed inverse problem, one that laboratory studies using flat noise have largely avoided. Here we show that monaural elevation localization of unfamiliar, spectrally structured sounds is close to chance, whereas binaural listening improves it far beyond a weighted combination of two monaural estimates. A generative model explains why: the interaural spectral ratio cancels the source spectrum, analogous to binocular disparity. Spectral priors help only modestly, and head motion adds little, likely because natural spectra fluctuate rapidly. Natural spatial hearing therefore relies on binaural spectral disparity rather than monaural template matching.

Spatial hearing is commonly associated with interaural time and level differences, which provide powerful cues to the lateral position of a sound source [10,39]. Yet these cues alone are insufficient to establish a complete three-dimensional auditory percept. Elevation and front–back discrimination depend critically on direction-dependent spectral filtering by the head and outer ears [6,9,20,22,30], which also contributes to the externalization of sounds [10,19,52]. Spectral cues are therefore not merely a secondary refinement of spatial hearing but a central source of information for determining where sounds originate in the external world. Understanding how the auditory system extracts this information is fundamental to explaining spatial hearing. Sensory systems must infer stable properties of the external world from proximal signals that are often compatible with many physical causes. In vision, the same retinal image can arise from different combinations of object size and distance, and this ambiguity is resolved using additional information such as binocular disparity, motion parallax and prior knowledge [5,11,26]. Auditory elevation perception of natural, i.e. spectrally-structured sounds faces an analogous problem: the spectrum reaching the eardrum reflects both the sound-source spectrum and the direction-dependent filtering imposed by the head and outer ears [10,23,30,44]. As a result, elevation cannot in general be uniquely recovered from the sensory spectrum alone, rendering monaural elevation estimation an ill-posed inverse problem [24,29]; psychophysical studies demonstrate this ambiguity experimentally by showing that narrowband stimuli can be localized to multiple elevations depending on their spectral content [29].

Here we show that monaural spectral information [4,9,20] is generally insufficient for reliable elevation localization when the source spectrum is natural-like (i.e. spectrally structured) and unfamiliar. Binaural listening improved accuracy and precision beyond what a weighted combination of two monaural estimates predicts (two-ear summation), whereas a flat source spectrum, approximating the ideal outcome of source-spectrum priors, provided only a modest benefit. Finally, head motion did not significantly improve localization. For a demonstration of the difference in monaural and binaural elevation hearing using natural sounds see Supplementary Video V1. A generative spectral-template model explains this pattern by showing that binaural spectral comparison cancels source-related spectral structure and hence recovers direction-dependent information that is not accessible from the signals at either ear alone. Thus, auditory elevation perception relies on a binaural disparity computation that resolves source-spectrum ambiguity, revising the standard view of elevation hearing as primarily monaural spectral-template matching. Because spectral cues underlie not only elevation but also front–back discrimination and externalisation [10,19,52], this mechanism concerns the core of how sounds are located in three-dimensional space.

The elevation-localization problem is spectrally ambiguous. The sound waveform at the eardrum carries no explicit label for source position [30,39]. For lateral localization, interaural time and level differences provide strong cues, but these cues are largely ambiguous with respect to elevation and front-back position (cone of confusion) [10,39,50]. Elevation, front and back must instead be inferred from spectral filtering by the listener's head-related transfer functions (HRTFs), which are specific to the direction of the sound source relative to the head, and different for each listener and ear [6,9,14,15,17,20,23,30,42,51,52]. Spectral cues are also essential for externalisation [8,16,19]: sounds rendered with interaural time and level differences alone are heard inside the head, lateralised along a single left–right axis. Spectral cues thus do not merely refine localization; they turn a lateralised percept into a sound located in three-dimensional space. The proximal spectrum at one ear is therefore the product of two unknown quantities: the distal source spectrum and the direction-dependent HRTF [10,24,30]. Different combinations of source spectrum and HRTF can produce highly similar proximal spectra, thus matching a single-ear spectrum to an HRTF map cannot uniquely recover source elevation unless the source spectrum is known or otherwise constrained [10,30] (Fig. 1).

This ambiguity is the general case for natural sounds rather than an exceptional stimulus condition: unlike repeatedly presented flat-spectrum noise, everyday sounds have source-specific spectral structure that is not known in advance. Classical accounts of vertical localization emphasize monaural spectral cues, especially direction-dependent peaks and notches introduced by the pinnae [9,15,20,22]. These cues can support localization when the source spectrum is flat or predictable [29,55]. Notably, the evidence for the monaural account comes largely from experiments in which exactly this is the case: spectrally flat or unstructured noise, typically a single stimulus presented repeatedly, often to highly trained listeners

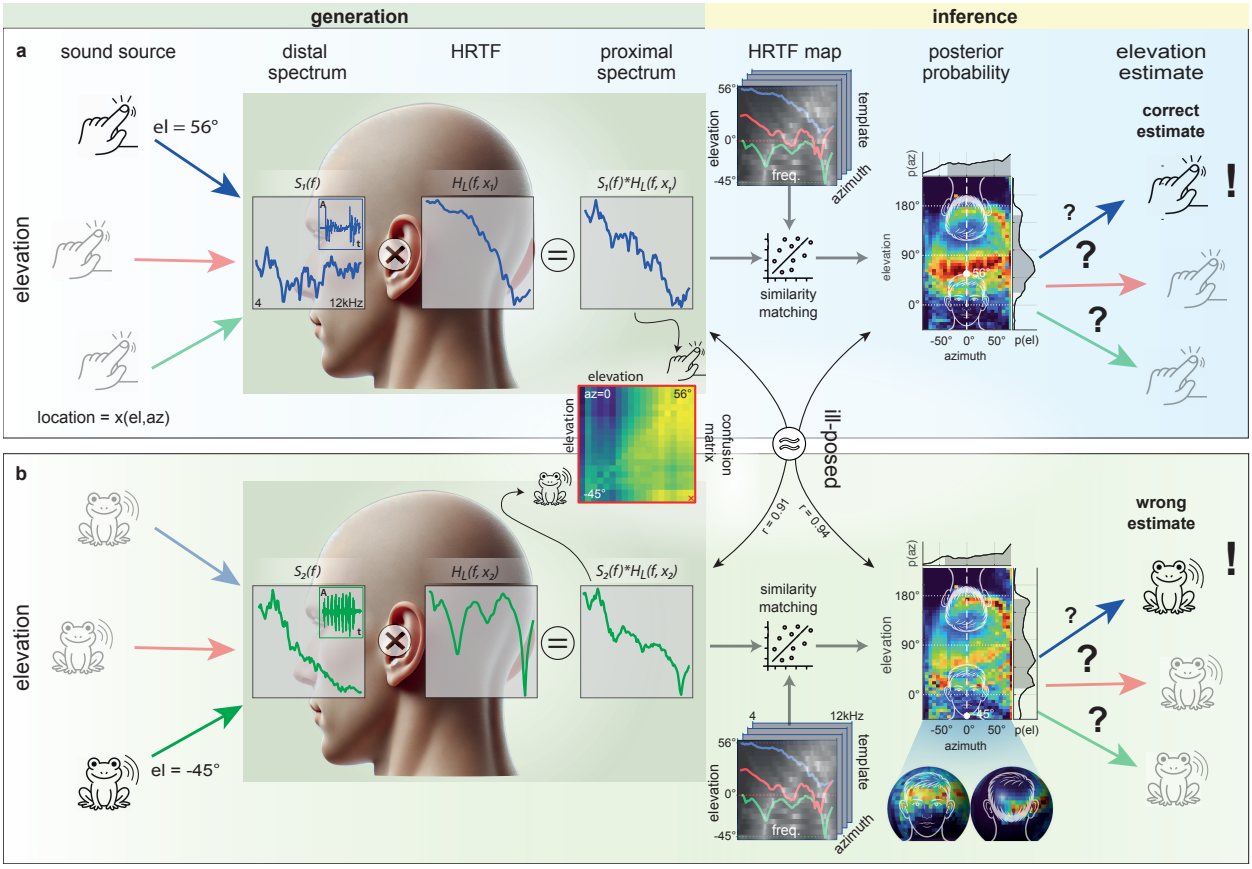


Figure 1. Localization based on monaural spectral cues: limits due to ambiguities induced by the spectral structure of natural sounds. For a sound source at location $x(\mathrm{el}, \mathrm{az})$, the distal source spectrum $S(f)$ is filtered by the left-ears HRTF $H_L(f, x)$, yielding the proximal spectrum $S(f) \cdot H_L(f, x)$ at the eardrum. This proximal spectrum is compared with all HRTFs stored in a template map spanning elevation and azimuth. Similarity matching produces a posterior distribution over source locations, which marginalized across azimuth yields an elevation estimate sampling from the posterior. **A** For one source spectrum (a knocking sound) played from above (el=56°), the filtered signal produces a localized posterior with a peak at roughly the correct elevation, however with a large uncertainty as indicated by the width of the posterior and a second smaller peak in the back indicative of frequent front/back confusions. **B** A different source spectrum (a frog sound) played from below (el=-45°) interacts with the corresponding HRTF to produce a similar proximal spectrum as the knocking sound and therefore generally a highly incorrect estimate at a high elevation, again with a big range of uncertainty and several smaller peaks. The confusion matrix shows the similarity between the two proximal signals when played from different elevations. In Supplementary Fig. S1 we show the development of the posterior for both sounds at az=0 over a wider range of elevations together with the confusion matrix for the posteriors, which highly mimics that of the proximal signals. Abbreviations: el = elevation, az = azimuth.

(e.g. [20,22,46,47]). Under these conditions the source spectrum is effectively known. Moreover, when an identical sound is repeated from different directions, HRTF filtering changes its perceived pitch and timbre with elevation (the same sound appears slightly higher in pitch from above than from below), and listeners systematically map higher-pitched sounds to higher elevations [37]. Such a pitch–elevation association can produce a monaural stimulus–response relation without any inference of the HRTF from the spectrum. Laboratory measurements may therefore overestimate monaural competence. In contrast, natural and everyday sounds often contain strong spectral structure of their own [1,2,32]. Such structure can mimic, mask or distort HRTF-induced spectral features (cf. supplementary animation: Video V2). Moreover,

isolated spectral landmarks are insufficient to explain localization performance, which depends on information distributed across the spectrum [10,27,56]. Thus, the existence of directional filtering at each ear does not by itself solve the source-spectrum ambiguity. Consequently, successful localization requires not only extracting direction-dependent spectral structure but also distinguishing it from source-related spectral variation. We considered three mechanisms that could reduce or resolve this ambiguity (Fig. 2 a, b & c).

First, because the same source spectrum reaches both ears while the left- and right-ear HRTFs generally differ characteristic ways with position, the interaural log-spectral difference can cancel the common source spectrum and leave a binaural spectral-disparity signal dominated by the HRTF ratio (Fig. 2a). Interaural spectral cues of this kind have been proposed previously [21,25], but their significance for human spatial hearing has not been demonstrated. This mechanism differs fundamentally from the prevailing view of how the two ears contribute to elevation perception, in which each ear provides its own monaural spectral estimate and the two estimates are combined by a binaural weighting that depends on perceived azimuth [20,21,47]. Weighting two estimates improves reliability, as combining any two noisy sensors would, but it cannot remove the source-spectrum ambiguity, because each monaural estimate inherits it. Binaural spectral disparity instead compares the two spectra before an estimate is formed.

Second, prior knowledge about the source spectrum could help "explain away" [5,26] source-related spectral structure, effectively whitening the incoming signal [29,55] (Fig. 2b). Spectral priors would require accurate knowledge of the source spectrum and an appropriate allocation of the prior based on context.

Third, head motion could provide auditory motion parallax by sampling the same source from different directions over time [18,48] (Fig. 2c). However, such motion-based cues require source position and source spectrum to remain stable over the relevant temporal window, which may often be violated for natural sounds [28,31,45]. We therefore predicted that binaural spectral disparity would provide the strongest disambiguating cue, with weaker contributions from sounds with either flat or flattened spectra due to the application of a spectral prior and possibly also from head motion.

Monaural localization is highly ambiguous. We first tested whether elevation localization can be achieved from monaural spectral information alone using natural-like sounds, that is, sounds with a distinct spectral structure. To minimize the contribution of source-spectrum priors, semantic information, or motion cues, we created spectrally structured broadband noise stimuli (structured noise),), drawn on each trial from a set of 1,000 distinct spectra, so that, unlike in most previous studies, no source spectrum was presented repeatedly, while head movements were constrained and tracked (Methods; Fig. 5). Localization accuracy was quantified as the gain of indicated elevation against true elevation at the median plane; a gain of 1 indicates veridical localization, whereas a gain of 0 indicates no systematic performance for elevation hearing.

Monaural localization was severely impaired (Fig. 3a, blue traces, structured spectrum, no-motion). Across participants, the mean gain was 0.19. Although this was significantly above zero (two-sided t-test, $t(23) = 4.30$, $p < .001$, 95% CI [.09, .27], $d_z = .88$), it indicates highly inaccurate localization, and many participants showed gains not significantly different from zero (Fig. 3b, blue distribution, structured spectrum, no-motion). Individual response variability was correspondingly high, with an average standard deviation of $\approx 30°$ (Fig. 3c, blue distribution, structured spectrum, no-motion); even the best-performing participant showed substantial uncertainty (standard-deviation (SD) $\approx 22°$). These results are consistent with the ill-posed nature of the monaural inference problem: the proximal spectrum alone provides insufficient information to uniquely determine elevation because source spectrum and HRTF filtering are confounded [29,52].

A spectral-template model captures monaural ambiguity We next asked whether the behavioral limitations of monaural localization arise directly from ambiguity in the sensory signal itself using a generative spectral-template model. The model generates the proximal spectrum $X(f, x)$ by filtering a source spectrum $S(f)$ through a recorded, direction-dependent HRTF $H(f, x)$ (CIPIC HRTF database [3]). Localization is then simulated by matching this observed spectrum against a learned HRTF-template map to compute a posterior distribution over directions, from which responses are sampled (Figs. 1, 2).

If ambiguity arises from the sensory representation itself, the model should reproduce the characteristic failures of human monaural localization. Using the same structured-noise stimuli, the model reproduced the

main features of the human monaural behavior (Fig. 3d, blue traces, structured spectrum, no-motion). Monaural posteriors were broad, often multi-modal and frequently peaked at incorrect elevations, even in the back (Supplementary Fig. S1: example of highly structured, natural stimuli played from different locations). Some source spectra still supported limited elevation tracking, whereas others produced posterior distributions that remained largely invariant across elevation, rendering localization effectively impossible (Fig. 3g, correlation $r = 0.51, p < .01$, structured spectrum, no-motion). The model also reproduced typical biases of monaural listening, including front-back confusions, a tendency to localize sounds too high, and especially poor performance for sources above the head.

We validated the model using 215 natural and man-made sounds from the MIT [35] and ESC-50 [38] databases together with 123 HRTF sets from CIPIC [3] and SoundSphere 2.0 [49] (Supplementary Fig. S3). Across sounds and virtual listeners, monaural gains remained shallow and variable, demonstrating that the ambiguity is a general consequence of the ill-posed inverse problem rather than a peculiarity of a specific stimulus set. The model further revealed structured idiosyncrasies across both sounds and listeners: some sounds were easier to localize than others, and some HRTFs provided richer monaural elevation information than others (Fig. 3g for experimental sounds and Supplementary Fig. S3d for natural sounds and Supplementary Fig. S5 for different heads). Across simulated listeners, monaural and binaural gains were positively correlated (Pearson $r = .39$, $p < .01$, $n = 123$, natural sounds, no motion), indicating that listeners with more informative HRTFs tended to localize better under both conditions.

Binaural disparity resolves source-spectrum ambiguity In contrast to monaural listening, binaural listening produced a strong improvement in localization (Fig. 3a, green data, structured noise, no motion). For unfamiliar, spectrally structured sounds without head motion, the mean behavioral gain increased from 0.19 to 0.60 (paired two-sided t-test, $t(23) = 7.42$, $p < .0001$, 95% CI of the difference [.31, .55], $d_z = 1.51$, individual gains Fig. 3b, structured spectrum, no-motion) and average response variability decreased from SD $\approx 30°$ to 18.6° (Fig. 3c, structured spectrum, no-motion). To test whether this binaural advantage requires a dedicated binaural computation, we compared human performance with a Bayesian two-ear-summation observer that used each participant's monaural gain and variability to predict the binaural performance expected from combining two independent ear-specific monaural estimates, without exploiting interaural disparities. This observer formalises the binaural-weighting account of elevation perception [21,33,47]: each ear yields its own spectral elevation estimate, and the two estimates are combined according to their reliability. It therefore sets a conservative benchmark that a dedicated binaural comparison must exceed.

The empirically observed improvement substantially exceeded the prediction of the simple two-ear-summation observer, for which the predicted average gain increase was modest (to 0.38) and the predicted average reduction in variability limited to SD = 26.6°. The observed binaural gain, however, was significantly higher than this two-ear-summation prediction (among participants with valid summation estimates: n = 19; 0.58 vs. 0.38; paired $t(18) = 4.07$, $p < .001$, 95% CI [0.10, 0.30], $d_z = 0.93$ - i.e., for participants with negative gains this prediction could not be calculated), and the observed binaural SD significantly lower (18.6° vs. 26.6°; $t(18) = 5.77$, $p < .0001$, 95% CI $[-10.9, -5.1]$, $d_z = 1.32$). Since behavioral measurements were obtained only from each participant's dominant ear, the behavioral two-ear summation prediction assumes equal uncertainty in both ears. As localization with the non-dominant ear is likely less reliable, this assumption provides a conservative estimate of the benefit expected from simple probability summation. To verify that this conclusion did not depend on assumptions about human ear dominance, we also derived an assumption-free two-ear summation prediction directly from the spectral-template model, where both monaural posterior maps are available. This prediction closely matched the behaviorally derived benchmark (Fig. 3e,f), providing converging evidence and indicating a substantial computational benefit of binaural localization performance beyond simply combining two independent monaural estimates. Together, these findings indicate the existence of an additional dedicated binaural mechanism: the two ears are not merely weighted but compared to cancel the source spectrum. Binaural weighting of monaural estimates, the prevailing account of how the two ears contribute to elevation perception, is therefore insufficient to explain the localization of natural, i.e. spectrally structured sounds.

The *spectral-template model* accounts for this benefit through binaural spectral disparity. Because the source spectrum is common to both ears, whereas the left- and right-ear HRTFs differ with location, forming

the interaural log-spectral difference cancels the common source term and leaves a representation determined primarily by the HRTF ratio:

$$\frac{S(f) \cdot H_L(f,x)}{S(f) \cdot H_R(f,x)} = \frac{H_L(f,x)}{H_R(f,x)}.$$

Localization is then achieved by matching this binaural signature to a learned spectral-disparity template map. This binaural model reproduced the behavioral improvement, with steep stimulus-response gains and reduced variability across a wide range of sounds and HRTFs (monaural vs. binaural model gain, paired two-sided t-test across simulated listeners, $t(44) = 33.2$, $p < .0001$, $d_z = 4.9$; binaural also exceeded the model's own two-ear-summation, $t(44) = 20.9$, $p < .0001$; Fig. 3e and f, structured spectrum, no-motion). Thus, binaural disparity transforms a severely ambiguous monaural problem into a well-constrained inference problem.

Perfectly symmetric heads and ears would leave residual ambiguity along the median plane (Fig. 2a, upper row). However, measured HRTFs, including those from standard head simulators (KEMAR [13], KU100, HATS) and human databases [3, 49], contained sufficient left-right asymmetry to make the binaural spectral ratio informative for elevation also in the median plane (Supplementary Fig. S5). Because the behavioral experiment was purposely conducted along the median plane, where conventional interaural timing and broadband level cues provide little information about elevation, the observed binaural advantage is particularly diagnostic of elevation-dependent spectral differences between the two ears.

Spectral priors can provide only limited disambiguation. A second potential mechanism for resolving ambiguity is the use of prior knowledge about the sound-source spectrum. Listeners are known to use prior expectations about sound-source locations [37], and prior assumptions about source spectra have also been proposed to improve spatial hearing [29, 55]. Such a spectral prior could effectively "explain away" [26] the source spectrum by whitening the incoming signal, thereby making the remaining spectral pattern more informative about the HRTF (Fig. 2b). To test this mechanism independently of how accurately such priors are learned or applied, we compared localization performance for unfamiliar, structured-noise sounds with that for flat-spectrum noise (white noise), which corresponds to a perfectly whitened signal with all source-related spectral structure removed (i.e. $S(f) = \text{const}$). Under these conditions, any remaining spectral structure in the filtered signal $X_i(f,x) = S(f) \cdot H_i(f,x)$ arises solely from the HRTF (i.e. $X(f,x) = \text{const} \cdot H_i(f,x)$), thereby disambiguating the filtered signal as far as the HRTF by itself is unique for any location.

Flat-spectrum noise improved localization in both monaural and binaural listening (Fig. 3a&b, flat spectrum, no-motion, Fig. 3h&i, main effect sizes). Monaural gains increased from 0.19 to 0.36 (paired two-sided t-test of flat-spectrum vs. structured gains, $t(23) = 5.03$, $p < .0001$, 95% CI of the difference [.11, .26], $d_z = 1.03$), and binaural gains increased from 0.60 to 0.69 (paired $t(23) = 5.69$, $p < .0001$, 95% CI [.06, .13], $d_z = 1.16$), with modest reductions in response variability (monaural SD $\approx 24.4°$, binaural SD $\approx 13.6°$; Fig. 3c, flat spectrum, no-motion). Thus, reducing source-spectrum uncertainty by using flat-spectrum sounds improved localization, consistent with a role for spectral priors in explaining away source-related structure. However, the benefit was much smaller than the binaural advantage, and monaural localization remained substantially impaired even for a flat source spectrum. Monaural performance for flat noise was also lower than sometimes reported in studies using a single, repeatedly presented noise stimulus (e.g. Van Wanrooij and colleagues [47]). In our experiment, the flat noise was high-pass filtered at a low cut-off that varied randomly from trial to trial (1–200 Hz; Methods). This left the elevation-relevant spectrum unchanged but made successive sounds differ slightly in pitch, preventing listeners from learning a fixed source spectrum or exploiting the elevation-dependent pitch shift of a repeated sound [37]. Higher monaural performance in earlier studies may thus partly reflect such knowledge of the source rather than inference of the HRTF.

The model captured this pattern. Replacing structured-spectrum noise with flat-spectrum (white) noise reduced posterior ambiguity and improved localization, while preserving the strong advantage of binaural over monaural processing (Fig. 3e, flat spectrum, no-motion). For the natural example sounds this can also be seen from Figure 2b where the marginal width of the posterior narrows with the application of prior knowledge to explain away the spectrum, but much less compared with the binaural mechanism Fig. 3h&i,

main effect sizes. These results indicate that spectral priors can contribute to elevation judgments, but they do not by themselves provide the robust disambiguation achieved by binaural spectral disparity.

Motion-based cues are limited by source-spectrum instability. Head motion could, in principle, provide auditory motion parallax (Fig. 2c). As the listener moves, the source is sampled from different directions, and the resulting spectral gradient could be matched to a learned HRTF-gradient map. Unlike binaural disparity, however, motion information is acquired sequentially. It therefore requires the source spectrum to remain stable throughout the temporal integration window and, for self-generated head motion, the source position to remain fixed. This requirement is restrictive because the spectra of many natural sounds fluctuate on the scale of tens to hundreds of milliseconds (spectral entropy) [28,31,45] .

In the main experiment, residual head movement (mean motion 10.1 mm/s) did not significantly improve localization. For structured noise stimuli, localization gains with motion were 0.24 in the monaural condition and 0.61 in the binaural condition (Fig. 3b, structured spectrum, motion). For flat-spectrum noise, gains were 0.34 and 0.70, respectively (Fig. 3b, flat spectrum, motion). In none of these conditions did residual motion produce a significant improvement (paired two-sided t-tests of motion vs. no-motion gains: structured noise monaural, $t(23) = 1.71$, $p = .10$, $d_z = .35$; structured binaural, $t(23) = .34$, $p = .74$; flat-spectrum monaural, $t(23) = -.62$, $p = .54$; flat-spectrum binaural, $t(23) = .63$, $p = .53$, Fig. 3h&i, main effect sizes). Bayesian paired t-tests (two-sided JZS prior, $r = 0.707$) provided moderate evidence for the null in three of these comparisons ($\mathrm{BF}_{01} = 4.4$, structured binaural; 3.9, flat monaural; 3.9, flat binaural), with the structured-monaural trend inconclusive ($\mathrm{BF}_{01} = 1.3$). Thus, residual head motion did not measurably improve localization, whereas the large advantage of binaural over monaural listening remained unchanged.

To test whether this absence of a motion benefit reflected the quite small residual head movements or the short stimulus durations, we conducted a follow-up control experiment with much faster instructed nodding and unrestricted free head motion listening to either 400-ms or 1000-ms structured-noise stimuli (Fig. 4a, columns). To maximize sensitivity to a motion benefit, the control experiment was performed under monaural listening. However, neither instructed nor free motion improved localization (Fig. 4b &c). Localization gains remained low and essentially unchanged across motion conditions (400-ms stimuli: 0.14 no motion, 0.14 instructed, 0.14 free; 1000-ms stimuli: 0.14, 0.16, 0.14, respectively), as did response variability (400-ms: 31.9°, 32.9°, 33.0°; 1000-ms: 32.3°, 33.8°, 34.0°). Bayesian paired t-tests (two-sided JZS prior, $r = 0.707$) yielded moderate evidence for the null, with instructed and free motion versus no motion both giving $\mathrm{BF}_{01} \approx 3.3$–$3.5$ across the 400- and 1000-ms stimuli. Instead, trajectory analyses suggested that participants tended to slow their movements once the sound began. Although this observation is descriptive, it is consistent with the limited contribution of head motion under these stimulus conditions.

The model reproduced the limited behavioral effect of motion. We simulated motion-based localization using spectral snapshots separated by $\Delta t = 100\,\mathrm{ms}$ and HRTF-gradient templates constructed for a 1° spatial displacement. Under these assumptions, motion cues provided only a marginal improvement beyond the information already available from static spectral cues (Fig. 3e, motion rows). Additional simulations with natural sounds showed that increasing the temporal integration window improved performance only when the source spectrum was stationary.In contrast, for natural sounds with time-varying spectra, the benefit of motion declined rapidly once the temporal integration window exceeded approximately 20 ms, and increasing the integration window up to the full 2-s sound duration did not further improve localization (Supplementary Fig. S3f).

The binaural advantage is robust to noise. The preceding simulations contained no additional neural noise; uncertainty arose from matching the filtered signal to the template map. To assess robustness, we added noise either early, before binaural-ratio formation, or late, during template comparison. In the absence of noise, the interaural ratio exactly removes the source spectrum (Supplementary Video V2). With early independent noise in the two ears, this invariance becomes approximate:

$$\frac{X_L(f,x)}{X_R(f,x)} = \frac{S(f)\cdot H_L(f,x) + \sigma_L(f)}{S(f)\cdot H_R(f,x) + \sigma_R(f)} \approx \frac{H_L(f,x)}{H_R(f,x)} + \varepsilon,$$

provided that the noise terms are small relative to the source spectrum.

As expected, increasing noise degraded localization performance (Supplementary Fig.S4). However, early noise did not impair performance substantially more than late noise. A likely reason is that localization integrates information across many frequency channels, so independent corruption of a subset of channels has limited influence on the overall similarity match. Importantly, the binaural advantage over monaural localization persisted across the tested noise range.

Together, the behavioral results show that monaural spectral cues alone are highly ambiguous for unfamiliar, spectrally structured sounds as they occur naturally, whereas binaural listening improves both accuracy and precision beyond the prediction of binaural weighting (two-ear summation). A flat source spectrum provides a smaller benefit, whereas head motion contributed little under the tested conditions. The spectral-template model reproduces this full pattern (Fig. 3h and i, Pearson's correlation $r = 0.98,\ p < .0001$) by showing that monaural ambiguity arises from the ill-posed source-spectrum/HRTF inverse problem, whereas binaural spectral disparity cancels the common source spectrum and provides a more invariant cue to elevation. These principles remained robust across sounds, HRTFs, and noise levels, supporting binaural spectral disparity as a computational solution to elevation localization.

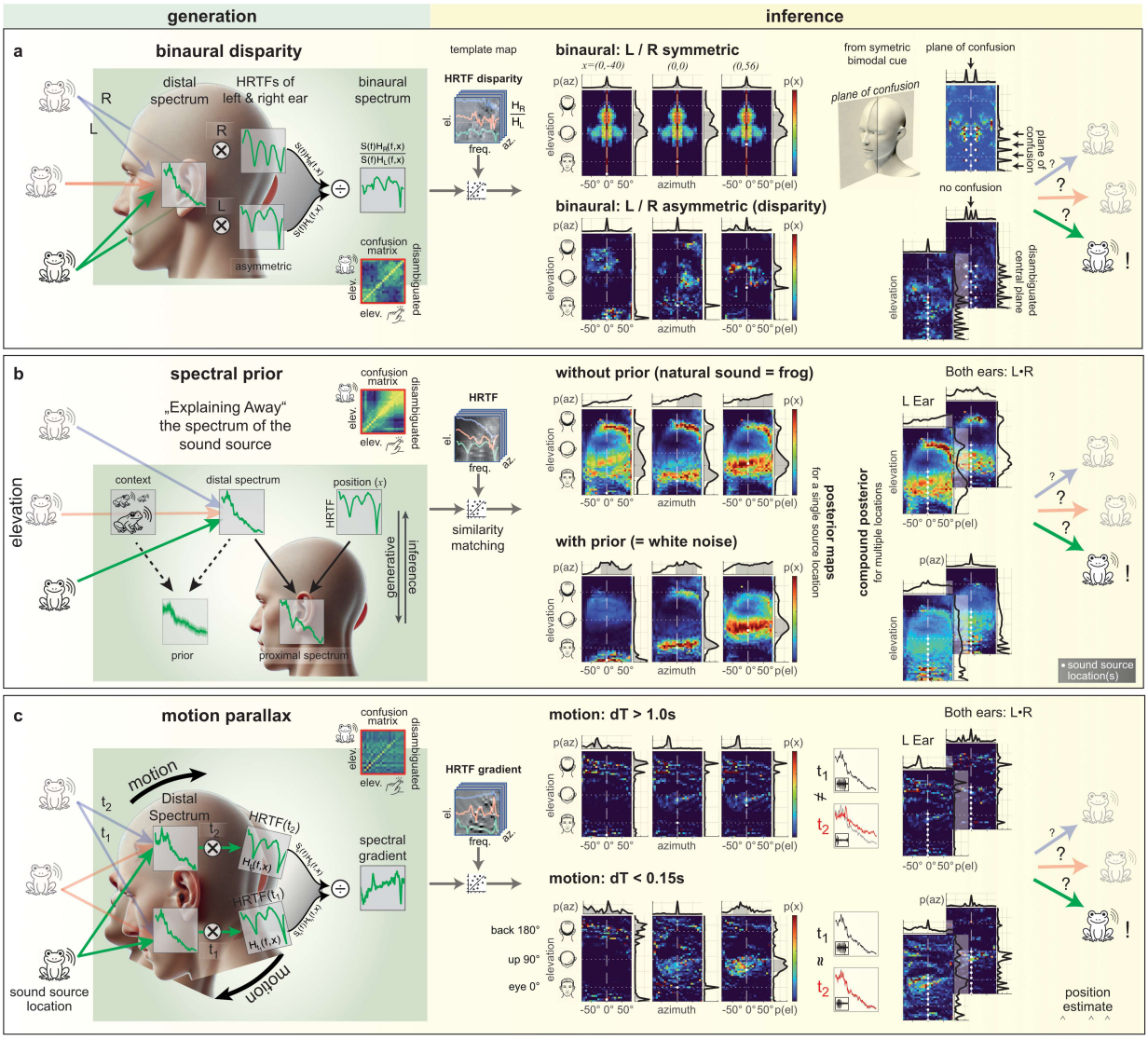


Figure 2. Disambiguation in vertical localization using binaural disparity, spectral priors and motion parallax. Illustrated are three candidate mechanisms for reducing source-spectrum ambiguity exemplary shown with the frog sound. Across mechanisms, inference is formulated as matching the observed cue to a learned template map, producing a posterior distribution over candidate source positions. **a. Binaural disparity:** A common source spectrum is filtered by left- and right-ear HRTFs. Their log ratio canceling the incoming sound spectrum defines a binaural spectral-disparity cue that can be matched to an interaural HRTF-disparity map. Natural left-right asymmetries can break median-plane ambiguities (upper row) and support unique localization (lower row). **b. Spectral prior:** Prior knowledge about the source spectrum can "explain away" source-related spectral variability in the proximal signal reducing ambiguity, i.e., sharpening the posterior. **c. Motion parallax:** Head or source motion generates time-varying HRTFs. The resulting spectral gradient can, in principle, be matched to a motion-gradient template, but this cue depends on temporal stability of the source spectrum. With the exemplary frog sound there is almost no benefit using a temporal window above 1 sec, while a theoretical benefit emerges below 150 ms with sufficiently similar spectra at T_1 and T_2 so they cancel.

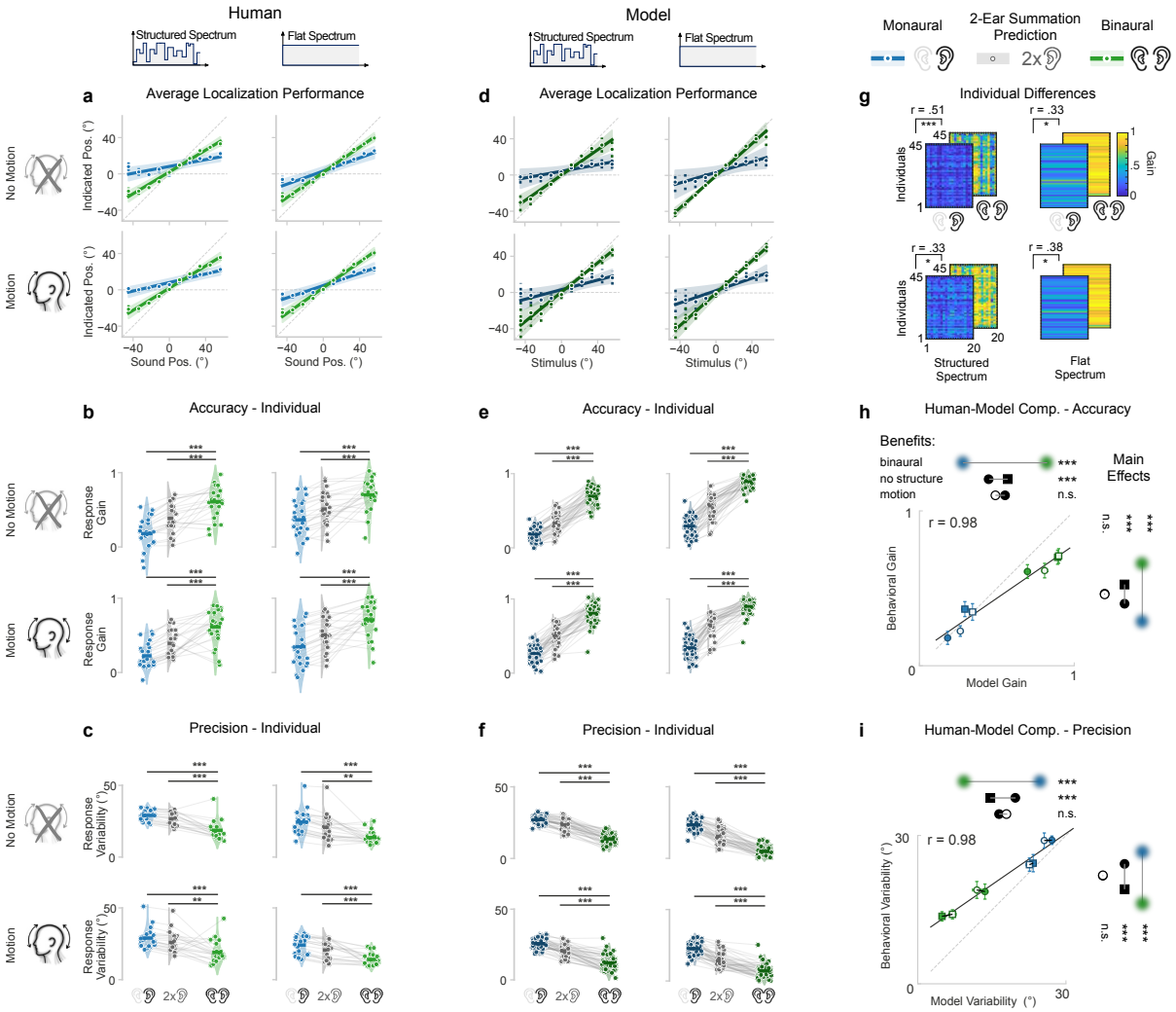


Figure 3. Localization performance: Human vs. Model a–c human performance; **d–f** performance of spectral-template model. Each of the panels **a-f** shows binaural (green) and monaural (blue) performance for both factors, Sound Type and Head Motion (horizontal: structured vs. flat spectrum vertical: no-motion vs. motion). **a**, **d** Sound source elevation against indicated position: average localization performance for human and simulated participants (dots) with linear-mixed-model (LMM) regression fits (line) and 95% CI (shaded area). Dashed lines indicate chance and veridical performance. **b**, **e** Accuracy of individuals measured as regression gain. **c, f** Precision of individuals, quantified as the standard deviation (SD) of localization errors within each condition. Lower SD indicates higher precision. Accuracy and precision are determined for monaural (blue), binaural (green) and the prediction derived from the monaural condition for simple two-ear-summation (grey). Dots show individual human and simulated participants together with their distribution (violin) with the mean group indicated by a horizontal bar. Individuals are connected by grey lines. **g** Individual differences: Color coded is the accuracy (regression gain) for all model-simulated listeners and experimental sounds from blue (gain=0) to yellow (gain=1). For both factors, the Sound Type (horizontal) and Head Motion (vertical) conditions monaural performance is shown in the front panel, binaural performance in the back panel. Correlation coefficients quantify the consistency across the monaural and binaural listening conditions. **h, i** Direct comparison of human and model performance (Pearson correlation r) together with main effect sizes for **h** accuracy (regression gains) and **i** precision (mean SD). Each data point represents the average performance per condition with the three factors: Ear (Monaural: blue / Binaural: green), Sound Type (structured: circles / flat: squares) and Head Motion (no motion: unfilled / motion: filled). Marginals show the benefits for binaural listening (main effect: monaural vs. binaural), no structure (main effect: structured spectrum vs. flat spectrum) and motion (main effect: no motion vs. motion) for both the spectral-template model (top) and for the human observers (right). Significance levels: $p < .05(*)$; $p < .01(**)$; $p < .001(***)$; not significant (n.s.).

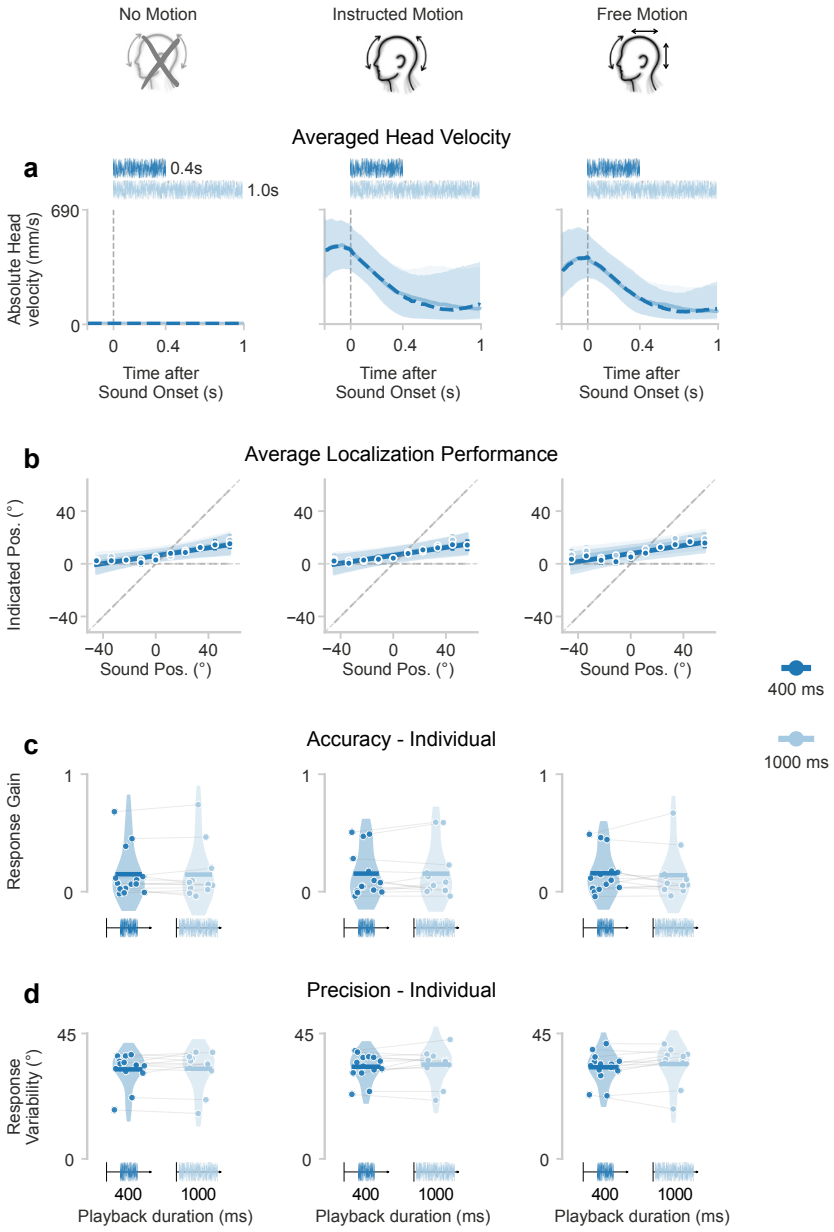


Figure 4. Localization performance under controlled head motion. Participants localized structured-spectrum sounds under three movement conditions (different columns): *No Motion*, *Instructed Motion*, and *Free Motion*. Each condition was tested with 400,ms (dark blue) and 1000,ms (light blue) playback durations. **a** Head motion over time, aligned to sound onset, quantified as Manhattan translational absolute velocity across the three position axes. Lines show median trajectories across trials; shaded regions indicate the 20th–80th percentile range. **b** Average localization performance. Dots show mean indicated position for each sound-source location; lines show LMM regression fits; shaded regions indicate 95% cluster-bootstrap confidence intervals. Dashed lines indicate chance and veridical performance. **c** Accuracy of individual participants, measured as response gain. **d** Precision of individual participants, measured as the standard deviation of signed localization error. For **c** and **d**, dots show individual participants together with their distribution (violin); horizontal bars indicate group means; grey lines connect the same participant across playback durations. Overall, neither head motion nor longer playback duration produced a clear improvement in accuracy or precision.

Discussion

A central challenge for sensory systems is that the proximal signal available to the brain is often compatible with many different states of the world. Here we show that this ambiguity is particularly severe for auditory elevation localization. When listeners had access only to monaural spectral information from unfamiliar sounds, localization was highly inaccurate and imprecise, with many listeners performing close to chance. These findings support the idea that the proximal sound spectrum alone is insufficient to uniquely infer elevation, because it reflects an inseparable combination of the unknown source spectrum and the direction-dependent filtering imposed by the listener's HRTF.

Traditionally, elevation localization has been described as relying primarily on monaural spectral cues, because direction-dependent filtering is physically present at each ear [4,9,20,44]. Our results suggest, however, that the existence of directional spectral structure at each ear does not imply that the auditory system uses these spectra independently. The binaural improvement substantially exceeded what would be expected from simply combining two independent monaural estimates, indicating that the two ears do not merely provide redundant samples of the same uncertain cue. Instead, comparing the spectra at the two ears can effectively cancel the unknown source spectrum and expose a more invariant binaural spectral-disparity signal that depends primarily on source direction. This mechanism is closely analogous to binocular disparity in vision, where comparing two individually ambiguous retinal images provides information that is unavailable from either image alone.

Our results also clarify how the two ears contribute to elevation perception. Rather than merely weighting two independent monaural estimates, as proposed by binaural-weighting accounts [24,33,47] and formalised here as two-ear summation, the auditory system appears to employ a dedicated binaural mechanism that cancels the source spectrum and thereby resolves the ambiguity. Interaural spectral cues have been proposed before [25,43]; here we provide behavioural evidence for their use together with a computational account of why they are needed. To reveal this mechanism, we approximated natural listening conditions while excluding prior knowledge of the source. Instead of the spectrally flat noise used in most previous studies [20,22,46,47], we used spectrally structured noise; we tested naïve participants with little training; and no stimulus spectrum was repeated, so successive sounds differed slightly in pitch. This also prevented listeners from relying on the elevation-dependent pitch shift that HRTF filtering imposes on a repeatedly presented sound [37]. Under these conditions, the source-spectrum ambiguity becomes apparent and monaural cues no longer suffice. Earlier findings are therefore not in conflict with ours: they describe a special case, in which the source spectrum is known and monaural cues suffice, that our account explains alongside the natural case, in which they do not.

Rather than have listeners first learn a source-specific spectral prior — which would depend on uncertain assumptions about individual learning, and recent evidence suggests yields only limited improvement [7] — we used white noise, a signal that is flat from the outset and thus leaves nothing to "explain away." White noise therefore provides an upper-bound benchmark for a perfectly learned prior, without depending on individual listeners' learning ability (see supplementary text). Using this approach, we show that prior knowledge of the sound spectrum can improve localization performance, but only modestly: monaural localization remained substantially impaired even under this favorable condition. Thus, spectral priors can alleviate the inverse problem, but do not provide the robust disambiguation achieved by binaural spectral disparity.

In contrast, head motion contributed little to localization performance under the tested conditions. Although head motion could in principle provide auditory motion parallax by sampling the same sound source from different directions, this information is acquired sequentially rather than simultaneously. Thus to be useful, unlike binaural spectral disparity which is available instantaneously, motion-based localization requires the source spectrum to remain sufficiently stable over the movement period.

Model simulations provide a computational explanation for why this requirement is difficult to satisfy under natural listening conditions. The approximately 20-ms temporal integration limit closely matches the timescale over which natural sound spectra fluctuate (spectral entropy), implying that successive spectral samples rapidly become inconsistent with the demand for a single underlying source spectrum. Consequently, extending the temporal integration window provides little additional information once the source spectrum is no longer stationary as in most natural sounds. Furthermore, given a minimum audible angle in elevation of

approximately 10°, useful motion-based localization would require head movements substantially faster than those typically observed. Together, these constraints might explain why auditory motion parallax contributes at best modestly to resolving elevation ambiguity, despite its prominent role in visual depth perception.

Our conclusions are specific to motion-parallax cues arising from changes in HRTF during head movements in the median plane, where ITDs and broadband ILDs provide little or no elevation information. They do not exclude other mechanisms by which motion may benefit auditory perception, including dynamic changes in ITD or ILD during lateral head movements or motion-based tracking of auditory objects based on identity and other higher-order scene properties [12,40]. Rather, our results indicate that, for the single-source elevation-localization problem studied here, sequential spectral information obtained through head motion provides little additional disambiguation beyond the simultaneous binaural spectral comparison.

The spectral-template model reproduced the main behavioral findings across monaural, binaural, prior-based and motion-based conditions. It captured the broad and often multimodal monaural posteriors, the reduction of front–back confusions under binaural listening, the modest benefit of spectral whitening, and the limited contribution of head motion. The model remained robust across a wide range of sounds, HRTFs, and noise levels, suggesting that the proposed mechanism is not restricted to a particular stimulus set or listener anatomy.

An important implication of these findings is that individual differences in spatial hearing may partly reflect differences in HRTF structure. That is, the model revealed that some sounds are inherently easier to localize than others, but also that some listeners' HRTFs provide more informative spatial structure for disambiguating elevation. This suggests a direct link between anatomical filtering, the richness of the learned HRTF templates, and behavioral localization performance.

More generally, our results suggest that binaural spectral disparity provides a computationally powerful solution to the auditory elevation-localization problem. Rather than relying primarily on isolated spectral landmarks, spectral priors, or head motion, the auditory system may exploit comparisons between the two ears to remove uncertainty associated with the source spectrum and recover direction-dependent information. In this sense, auditory elevation localization may belong to a broader class of sensory computations in which ambiguity in one signal is resolved through comparison with another.

The proposed binaural spectral-disparity mechanism is also neurobiologically plausible. In the auditory pathway, binaural convergence occurs remarkably early, already at the level of the superior olivary complex [53], where neurons compute interaural timing and level differences with submillisecond precision. Such early binaural integration may be particularly advantageous for spectral-disparity computations. Forming ratios or differences between the two ears is expected to be most robust before substantial independent neural noise accumulates along the processing hierarchy. This is analogous to classical binaural computations of ITDs and ILDs, which also require precise and low-noise comparisons between the two ears. Spectral information relevant for elevation localization is known to emerge in parallel pathways involving the dorsal cochlear nucleus, inferior colliculus, and auditory cortex [34,54], suggesting that binaural spectral-disparity signals could potentially be computed and refined across multiple stages of the auditory system.

Several generalizations merit further consideration. First, our behavioral experiments were purposely restricted to the median plane, where interaural timing and broadband level differences carry little or no elevation information and where the ratio of perfectly symmetric left and right HRTFs would be ambiguous (plane of confusion; Fig. 2a). This design makes elevation-dependent spectral differences between the ears particularly diagnostic. How these results generalize to the full sphere of directions, and how this cue combines with other binaural cues, remains to be determined. The spectral-template model suggests one possibility (Supplementary Fig.S5): the ratio often becomes more structured for off-center directions as the left- and right-ear HRTFs diverge, predicting better elevation discrimination for off-center sounds.

Second, the binaural advantage depends on listener-specific differences between the left- and right-ear HRTFs. While it is known how individual HRTFs are learned in general [22,36], it remains unclear how precisely the auditory system learns or calibrates its own interaural-disparity map. The spectral-template model presented here assumes knowledge of this map.

Third, our stimuli were deliberately presented under controlled conditions: a single, stationary source with natural-like spectral structure, carrying no semantic information and unfamiliar to the listener. The

spectral-template model shows that these stimuli elicit results comparable to natural stimuli (Supplementary Fig. S3a & e). The Supplementary Video V1 further illustrates that the HRTF ratio remains the determining factor for natural sounds under real-world listening conditions. Experimentally testing the robustness of binaural spectral disparity in everyday listening — with reverberation, competing sources, and moving talkers — would be a valuable direction for future work.

Beyond elevation, interaural spectral disparity may also provide information about source distance and lateral position. The model predicts systematic changes in the ratio of the two HRTFs with both azimuth and, particularly at near distances, source distance, reflected in the sharply peaked azimuth marginals of the posterior distributions (Fig. 2a). Testing whether listeners exploit this information will require future experiments.

These findings may have implications beyond basic auditory neuroscience. In the clinical domain, current hearing-assistive technologies are primarily optimized for speech amplification and intelligibility, while spatial cues — particularly elevation-dependent spectral cues — are often degraded or lost. Beyond microphone placement, which affects spectral cues at each ear, bilateral devices typically process the two ears independently, for example through separate dynamic-range compression and noise reduction. Our results suggest that this may matter more than previously assumed: even if the spectral cues at each ear were preserved, independent processing could distort the interaural spectral ratio on which elevation perception for natural sounds relies. Preserving or reconstructing individualized binaural spectral-disparity cues may therefore improve spatial hearing, source segregation, and environmental awareness in these devices.

More broadly, our results suggest that auditory spatial perception, like vision, may resolve ambiguity by comparing partially ambiguous sensory signals to infer stable properties of the external world. This principle could inform applications well beyond hearing research: more realistic audio rendering in VR and gaming, spatialized content for music, film, and entertainment, and spatial-audio-based navigation in robotics and engineering. Because spatial-audio rendering and HRTF personalisation are typically evaluated with noise bursts, our results suggest that, for natural content, the decisive quantity is the interaural HRTF ratio rather than each ear's HRTF alone.

Methods

In the main experiment, we measured and computationally modeled (*spectral-template model*) vertical sound localization using a three-factor design: 1) Factor "Ear" (Monaural / Binaural), 2) Factor "Sound Type" (structured spectrum / flat spectrum), 3) Factor "Head Motion" (no motion / motion) resulting in eight conditions (Fig. 3). The spectral-template model simulates monaural, binaural-disparity, spectral-prior and motion cues from recorded head-related transfer functions (HRTFs). As a baseline for binaural performance, we applied a Bayesian summation observer (*two-ear-summation observer*) that predicts the improvement expected if the two ears provide only independent monaural samples (the two-ear summation prediction). In the control experiment we further tested human localization performance under different head motion conditions varying the duration of a spectrally structured noise sound and we modeled the usefulness of the motion cue using natural, time-varying sounds (Supplementary Fig. S3f).

Behavioral experiment

Participants. In the main experiment, we tested n=24 participants (f=16/m=8, mean age 24.04 ± 2.92 y), and n=12 (f=8 /m=4, mean age, 25.01 ± 2.21 y) in the control experiment. All participants reported no known hearing deficit, and their hearing thresholds were confirmed before testing using pure tones ranging from 1 to 10 kHz in a staircase procedure. The dominant ear of each participant was determined before testing [41]. To verify that sounds were inaudible with plugged ears during monaural testing, prior to testing both ears were occluded with earplugs and noise-blocking headphones (3M™ PELTOR™ Optime™ III Over-the-Head Earmuffs, 33–34 dB sound reduction). A staircase procedure then verified that no sound was audible through the blocked ears: a test sequence was played from all ten speakers at an initial level of 54 dBA and reduced in steps (to 47.5, 43, 41, and 40 dBA) until the participant reported no detection. The resulting sound level was individually adjusted and subsequently used for all conditions. The starting

condition was randomly chosen between binaural and monaural for each participant. If the binaural condition was performed first, the earplugs and earmuffs were removed during that condition and the staircase procedure was repeated before the subsequent monaural condition to verify that the selected sound level remained inaudible through the occluded ear. All experiments were approved by the local ethics board (approval no. EG-IIP-2025065) of Ulm University and informed consent was provided by all participants.

Setup. Experiments were conducted in a sound-attenuated, dark room. Thirteen loudspeakers (Genelec 8020c, linear frequency response) were mounted on a median-plane arc at a constant distance of 120 cm from the center of the participant's head (Fig. 5); the ten speakers spanning the frontal range were used for stimulus presentation. These ten speaker elevations ranged from $-45°$ to $+56.25°$ in $11.25°$ steps, yielding positions at $-45°$, $-33.75°$, $-22.50°$, $-11.25°$, $0°$, $+11.25°$, $+22.50°$, $+33.75°$, $+45°$, and $+56.25°$. Speakers were driven by a digital audio interface (RME Fireface 802) and individually calibrated to equal output level with a sound-level meter placed at head position. The flat spectrum output of each speaker was confirmed with a measurement microphone (Beyerdynamic MM 1). Before experimentation started, the speakers were covered with a sound-transparent cloth to avoid any response biases.

Participants were seated at the center of the arc with the head supported by a chin rest. Participants indicated the perceived source location with a red laser pointer mounted on a rotary potentiometer. The laser dot was the only visible feedback in an otherwise dark room. Participants confirmed each response with a button press on the device. Potentiometer readings were converted to perceived elevation using a geometric calibration that compensated for the offset between the head rotation axis and the potentiometer axis, as well as the fact that $0°$ elevation corresponded to eye height.

Motion tracking, classification and head-motion control experiment. Although participants' heads were supported by a chin rest in the main experiment, it was desired that small residual head movements could occur, to be exploited for the motion condition. To this end, head position was recorded at 120 Hz with an OptiTrack motion-capture system (Trio 3, Motive v2.2.0) using retro-reflective markers attached to the participant's head. For each trial, head motion was quantified from the three-dimensional position trace as the mean absolute frame-to-frame displacement across the x, y, and z axes,

$$d = \frac{1}{N-1} \sum_{k=1}^{N-1} \big(|\Delta x_k| + |\Delta y_k| + |\Delta z_k|\big), \tag{1}$$

where N is the number of tracked samples in the classification window. The window spanned from 100 ms before the logged sound timestamp to the end of the sound interval. The distribution of d was summarized by its mean, median, standard deviation, and range. Trials in the main experiment were split post hoc into *motion* and *no-motion* trials using a median split of this trial-wise motion metric. Across participants, the median threshold corresponded to approximately 4.4 mm/s (sample median 3.67×10^{-5} m/frame), yielding average head speeds of 10.1 mm/s in the *motion* trials and 3.3 mm/s in the *no motion* trials (very still fixation, Supplementary Fig. S2a). Because head motion was not experimentally imposed, the number of trials in each *Sound Type* × *Ear* × *Motion* cell varied across participants. We therefore required at least 3 trials per participant in each analyzed cell; no cells were excluded (the minimum cell count was four trials).

Because residual head movements under chin-rest support were small and the main-experiment stimulus duration was limited to 400 ms, we conducted a dedicated head-motion control experiment investigating whether larger or longer-lasting self-generated movements could improve localization. Procedure was the same as in the main experiment but restricted to structured-noise sounds listened monaurally, as this constitutes the condition with the lowest baseline performance such that any motion benefit should be most detectable. The experiment was two factorial with one factor being *Stimulus Duration* (400 ms, 1000 ms) and the other *Head Motion* (no motion, instructed motion, free motion), yielding six conditions. Each condition comprised 200 trials. In the *no-motion* condition of the control experiment, the participant's head was stabilized with a chin rest. Motion was analyzed online, and if the mean per-frame summed translational and rotational movement during stimulus playback exceeded a threshold of 0.05, the trial was marked as failed and re-queued, and an auditory prompt instructed the participant to reduce movement. In the two head

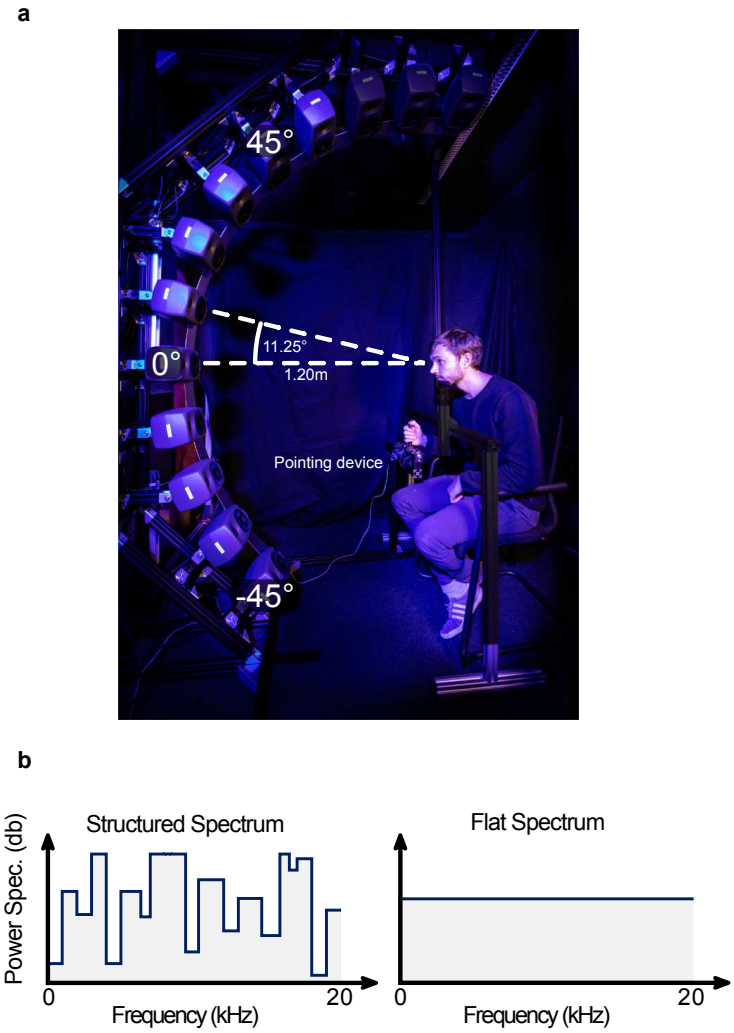


Figure 5. Experimental setup. **a** shows the experiment setup in a sound-attenuated chamber. The setup consists of 13 speakers (spaced 11.25°) covered with a sound-transparent blanket during the experiment. For the actual experiment, sounds were played from the ten frontal speakers in the range of −45° to +56.25°. These ten speakers were chosen so they were in the visual field of participants such that they did not have to move their head for the response. Participants were seated 1.2 m in front of the setup with their head supported by a chin-rest of fixed height. Using a red-dot laser pointer attached to a potentiometer participants indicated the perceived elevation after stimulus presentation. The room was completely dark during the experiment so that the only feedback provided was from the red laser. **b** shows exemplary sound stimuli for structured-spectrum (left panel) and flat-spectrum noise (right panel).

motion conditions, the chin rest was removed and a reference head posture was recorded before the first trial. Participants fixated the speaker at 0° elevation and remained still for 2 s while white noise played from the highest speaker. The reference recording was accepted when the mean per-frame summed translational and rotational movement during the 2 s calibration playback fell below 0.03. Subsequently, the reference posture was defined as the mean position across all recorded translational and rotational axes. In the *instructed-motion* condition, participants executed either a forward–backward translation along the y-axis or a pitch rotation before returning to the reference posture. Sound playback was triggered when y-axis translational motion exceeded 144 mm/s or y-axis rotation exceeded 90°/s, provided the participant had returned within 50 mm (translation) and 5.0° (rotation) of their calibrated resting position. In the *free-motion* condition, participants moved without specific instruction; the playback-trigger mechanism was identical, but participants independently discovered how to initiate trials. In both motion conditions, participants actively triggered each sound. They could terminate playback at any time by confirming their response with a button press.

Stimuli. Two stimulus classes, *structured-spectrum* and *flat-spectrum noise*, were used in the main experiment, each with a duration of 400 ms and 5 ms linear on- and offset ramps. In the control experiment, only the structured-spectrum noise was used with either 400 ms or 1000 ms sound duration.

Structured-spectrum noise stimuli ("natural-like spectrum" stimulus) were generated by imposing randomly positioned spectral peaks and notches over the entire spectrum: for the notches relative to the peaks the energy within a frequency window (width drawn uniformly from 0.5–2 kHz) was attenuated by a randomly chosen factor of 50–99 %; the window was then shifted by a random offset (0.5–2 kHz) and the procedure repeated up to 20 kHz. The average power of the spectrum was the same as for the white noise stimulus (Fig. 5**b**). The resulting deeply structured spectra were designed to mimic natural sound-like spectral structure, but to be unlike any naturally occurring sound, thereby minimizing the availability of learned spectral–elevation priors, while carrying no semantic information. On every trial one of a set of 1,000 distinct structured-noise sounds was randomly chosen.

Flat-spectrum (*white noise*) stimulus was generated digitally with a flat power spectrum. To prevent listeners from learning the fixed spectrum of a repeatedly presented flat-noise stimulus and from exploiting the elevation-dependent change in perceived pitch that HRTF filtering imposes on such sounds [37], we high-pass filtered the sounds slightly with a cutoff frequency randomly selected on each trial in the range between 1 and 200 Hz, thereby leaving the frequency range that carries spectral localization cues unaffected. Because the cut-off varied from trial to trial, successive stimuli differed slightly in pitch; the high-frequency elevation-cue band remained unchanged.

Procedure. The design comprised three factors—*Ear* (monaural, binaural), *Sound Type* (structured-spectrum noise, flat-spectrum noise), and *Head Motion* (motion, no motion)—yielding eight analysis conditions. Head Motion was defined *post hoc* by a median split of naturally occurring head movement (see *Motion tracking, classification and head-motion control experiment*). The two ear conditions were tested in separate blocks because the non-dominant ear had to be occluded in the monaural condition. Block order was randomly assigned and counterbalanced across participants, with a 5,min break between blocks.

Each ear block comprised 400 trials: 200 flat-spectrum noise trials and 200 structured-spectrum noise trials. Within each block, sound type and speaker location were presented in randomized order, with each of the 10 speaker locations presented 20 times per sound type. Thus, each participant completed 800 trials in total. During testing, participants localized flat-spectrum or structured-spectrum noise while a chin rest supported their head. In the binaural condition, both ears remained unoccluded; in the monaural condition, the non-dominant ear was occluded as described above. On each trial, participants indicated the perceived sound elevation using a laser pointer mounted on a potentiometer in the dark. Successive trials were separated by a 700,ms inter-trial interval.

Computational modeling

Spectral-template model

This model simulates typical localization behavior by comparing incoming filtered spectra with stored HRTF-template maps using cosine similarity, producing a probability distribution (posterior distribution) over possible positions in space. Sampling from this distribution provides an estimate of sound source location. The subplots in Fig. 1 & Fig. 2 are generated by this model. Individual head-related transfer functions (CIPIC database [3]; or SoundSphere 2.0 [49]) are used to generate filtered input signals replicating spectral cues of human listeners. The processing steps of the model can be divided into three consecutive stages: generation of the proximal stimulus (generative), followed by template matching & posterior computation, and sampling & response prediction (inference). For modeling the main results we used 45 CIPIC virtual listeners and the same sounds as for the main experiment: 20 randomly selected structured-spectrum noise stimuli from the 1000 generated, together with a single flat-spectrum noise stimulus. We verified that the high-pass cut-off at the very low frequency end (1–200 Hz) used for the human experiment did not affect the modeling which was conducted for frequencies between 4 and 12 kHz. The modeling was repeated with 215 natural and man-made sounds from the MIT [35] and ESC-50 [38] databases (400 ms after onset threshold of 25% max db) to verify its general applicability under natural listening conditions and to confirm that the structured-noise sounds used with human participants behave identical. To further demonstrate the models' generality we added another HRTF database with a set of 123 simulated listeners (CIPIC [3] and SoundSphere 2.0 HRTFs [49] merged).

Stimulus preprocessing. For each sound source position $x(az, el)$, with azimuth az and elevation el, on the CIPIC spatial grid (25 azimuths $\times$ 50 elevations, spanning $[-80°, 80°]$ azimuth and $[-45°, 230.625°]$ elevation in 5.625° steps, SoundSphere 2.0 grid was interpolated onto the CIPIC spatial grid for the supplementary analyses), the stimulus $s(t)$ was filtered with the left- and right-ear head-related impulse responses (HRIRs) $h_L(x,t)$ and $h_R(x,t)$ of the simulated listener using frequency-domain convolution. This yields the proximal signals $\hat{s}_L(x,t) = s(t) * h_L(x,t)$ and $\hat{s}_R(x,t) = s(t) * h_R(x,t)$; equivalently, in the log-spectral domain used below, the proximal spectrum at each ear is the sum of the source spectrum and that ear's HRTF.

The power spectral density (PSD) of the source stimulus $s(t)$ was estimated using Welch's method. For each 400 ms stimulus segment, spectra were computed using a 512-sample Hann window (≈10.7 ms at 48 kHz) with 50% overlap and a 512-point FFT. This yielded a one-sided PSD estimate, denoted $PSD_s(f)$, describing the distribution of signal power across frequencies. To obtain a representation proportional to spectral amplitude, the square root of the PSD was taken and subsequently transformed into logarithmic units,

$$S_{\log}(f) = \log\left(\sqrt{PSD_s(f)} + \varepsilon\right),$$

where $\varepsilon = 10^{-12}$ prevents numerical underflow. Thus, $S_{\log}(f)$ is the distal (source) log-amplitude spectrum of the stimulus, independent of source location and ear. The proximal log-amplitude spectrum received at ear $E \in \{L, R\}$ for a source at location x is formed in the log-spectral domain by adding the corresponding head-related transfer function, $S_{\log}(f) + \mathcal{H}_E(f, x)$ (equivalent to convolution in the linear domain). Only frequencies between 4 and 12 kHz were retained for subsequent analyses, as this range contains the dominant elevation-dependent spectral peaks and notches introduced by the HRTF [30, 56].

Both stimulus spectra and HRTF templates (see below) underwent identical pre-processing: the mean across frequency bins was subtracted from each spatial location independently ($S_{\log} = S_{\log} - \overline{S_{\log}}$), removing any overall level offset and making the representation sensitive only to spectral shape.

For motion trials, two source spectra were computed at times t_1 (stimulus onset) and $t_2 = t_1 + \Delta T$ ($\Delta T = 100$ ms), yielding $S_{\log}(t_1, \cdot)$ and $S_{\log}(t_2, \cdot)$. The inter-snapshot source gradient $\Delta S = S_{\log}(t_2, \cdot) - S_{\log}(t_1, \cdot)$ (common to both ears) is combined with the HRTF change to form the motion cue that the model matches against spatiotemporal HRTF templates (see next paragraph).

Template matching and posterior computation. Incoming signals were matched against previously stored templates, based on the HRTFs (HRTF template maps). For each simulated listener from the HRTF

database, the complex HRTFs $H_L(f,x)$ and $H_R(f,x)$ were obtained by taking the 512-point FFT of the measured HRIRs (head-related impulse responses). Log-magnitude HRTF templates were formed as $\mathcal{H}_L(f,x) = \log(|H_L(f,x)| + \varepsilon)$ and likewise for $\mathcal{H}_R$ (with ε to prevent underflow, see above). After extracting the high-frequency band (4–12 kHz) and applying per-spectrum mean removal (identical to the stimulus preprocessing), the template matrix $\mathbf{T} \in \mathbb{R}^{N_{\text{loc}} \times N_f}$ contains one spectral shape vector per candidate location ($N_{\text{loc}} = 1{,}250$, where N_f denotes the number of frequency bins in the 4–12 kHz band).

For a given condition and reference (source) position x, a similarity score $k(x,\ell)$ is computed between the observation $\mathbf{X}(x)$ (the feature vector at the reference position, written $X(f,x)$ in the main text, i.e. a vector across frequency f) and the template vector $\mathbf{T}(\ell)$ at every candidate location ℓ (azimuth and elevation), using exponentiated cosine dissimilarity:

$$k(x,\ell) = \exp\left(-\frac{1 - \cos(\mathbf{X}(x),\, \mathbf{T}(\ell))}{\tau}\right), \qquad \tau = 0.08$$

where $\cos(\mathbf{a},\mathbf{b}) = \mathbf{a}^\top\mathbf{b} / (\|\mathbf{a}\|\,\|\mathbf{b}\|)$ is the cosine similarity, clipped to $[-1,1]$. The parameter τ controls the sharpness of the posterior. The unnormalized similarities are converted to a proper probability distribution over all N_{loc} candidate locations ℓ:

$$p(\ell \mid x) = \frac{k(x,\ell)}{\sum_{\ell'} k(x,\ell') + \varepsilon}$$

This posterior represents the model's inferred probability distribution over candidate source locations ℓ given the observed spectral evidence $\mathbf{X}(x)$ at reference position x.

Depending on the condition, a different observation $\mathbf{X}(x)$ and template $\mathbf{T}(\ell)$ is used to compute the posterior. For the monaural condition, the observation (input signal to the template-matching stage) at reference position x is the left-ear feature vector

$$\mathbf{X}_L^{\text{mono}}(x) = S_{\log}(\cdot) + \mathcal{H}_L^{\text{true}}(x,\cdot),$$

where addition is element-wise across frequency (equivalent to the log-spectrum of the HRTF-filtered stimulus). The likelihood of each candidate location ℓ is evaluated by comparing $\mathbf{X}_L^{\text{mono}}(x)$ to the corresponding template vector $\mathbf{T}_L(\ell)$. The right-ear condition ($\mathcal{H}_R$) is defined symmetrically.

For the binaural summation posterior, posteriors from the left and right ear are computed independently and then combined by element-wise multiplication followed by renormalization:

$$p_{\text{LR}}(\ell \mid x) \propto p_L(\ell \mid x) \cdot p_R(\ell \mid x)$$

This is equivalent to assuming that the two ears provide conditionally independent spectral observations given the source location. This posterior, p_{LR}, constitutes the spectral-template model's two-ear summation prediction (the gray L+R prediction in the model panels of Fig. 3), against which the binaural-disparity posterior is compared.

The binaural disparity condition uses the interaural log-magnitude difference, equivalent to the logarithm of the left-to-right magnitude ratio, as the observation:

$$\mathbf{X}^{\text{ratio}}(x) = \big(\mathcal{H}_L(x,\cdot) + S_{\log}(\cdot)\big) - \big(\mathcal{H}_R(x,\cdot) + S_{\log}(\cdot)\big) = \mathcal{H}_L(x,\cdot) - \mathcal{H}_R(x,\cdot)$$

Because the same source spectrum is present at both ears, the source term cancels in the interaural log-spectral difference. The resulting interaural log-magnitude difference depends only on $\log(|H_L|/|H_R|)$. The template matrix for binaural disparity is $\mathbf{T}_{\text{ratio}}(\ell) = \mathcal{H}_L(\ell,\cdot) - \mathcal{H}_R(\ell,\cdot)$.

For the motion condition, the model used spectral-gradient templates rather than static spectral templates. The upward-elevation HRTF-motion template at candidate location ℓ was defined using a one-sided finite difference normalized by the template step:

$$\Delta\mathcal{H}_L^{+\text{El}}(\ell,f) = \frac{\mathcal{H}_L(\ell + \Delta_{\text{templ}}, f) - \mathcal{H}_L(\ell,f)}{\Delta_{\text{templ}}}.$$

The template step was $\Delta_{\text{templ}} = 1°$. Downward, rightward-azimuth, and leftward-azimuth templates were defined analogously using one-sided differences in the corresponding directions, yielding four directional templates.

The observed left-ear motion feature was

$$\Delta\mathbf{X}_L^{\text{mot}}(x) = \Delta S + \Delta\mathcal{H}_L(x, \cdot),$$

where

$$\Delta S = S_{\log}(t_2, \cdot) - S_{\log}(t_1, \cdot)$$

is the temporal source-spectrum gradient between two snapshots separated by $\Delta T = 100\,\text{ms}$, and

$$\Delta\mathcal{H}_L(x, \cdot) = \frac{\mathcal{H}_L(x + \Delta_{\text{true}}, \cdot) - \mathcal{H}_L(x, \cdot)}{\Delta_{\text{true}}}$$

is the true HRTF gradient for a displacement of $\Delta_{\text{true}} = 1°$.

The binaural-disparity motion feature was

$$\Delta\mathbf{X}_{\text{ratio}}^{\text{mot}}(x) = \Delta\mathbf{X}_L^{\text{mot}}(x) - \Delta\mathbf{X}_R^{\text{mot}}(x),$$

and was matched to the template

$$\Delta\mathbf{T}_{\text{ratio}}(\ell) = \Delta\mathcal{H}_L(\ell, \cdot) - \Delta\mathcal{H}_R(\ell, \cdot).$$

Additionally, we assume a Gaussian prior over elevation defined as:

$$p_{\text{prior}}(\ell) \propto \exp\left(-\frac{(\hat{e}(\ell) - \mu_{\text{prior}})^2}{2\sigma_{\text{prior}}^2}\right) + p_{\text{floor}}$$

with $\mu_{\text{prior}} = 5.5°$, $\sigma_{\text{prior}} = 15°$, $p_{\text{floor}} = 10^{-12}$, replicated uniformly across azimuth. When enabled, the posterior is multiplied element-wise by $p_{\text{prior}}(\ell)$ and renormalized.

Final monaural and binaural posteriors. The elementary posteriors described above were combined to define the two model outputs used for all regression analyses. We refer to these outputs as the *monaural* and *binaural* model posteriors.

For the static monaural model, the final posterior was the left-ear monaural posterior,

$$p_{\text{mono}}^{\text{static}}(\ell \mid x) = p_L(\ell \mid x).$$

The right-ear monaural posterior was computed analogously and used for diagnostic analyses, but the main monaural condition was based on the left ear.

For the static binaural model, the final posterior combined monaural evidence from both ears with the binaural interaural-disparity posterior. Thus,

$$p_{\text{bin}}^{\text{static}}(\ell \mid x) \propto p_L(\ell \mid x)\, p_R(\ell \mid x)\, p_{\text{ratio}}(\ell \mid x),$$

followed by normalization over all candidate locations ℓ. Here, p_{ratio} is obtained by matching the observed interaural log-spectral difference,

$$D_{\text{obs}}(f, x) = \left[S_{\log}(f) + \mathcal{H}_L^{\text{true}}(f, x)\right] - \left[S_{\log}(f) + \mathcal{H}_R^{\text{true}}(f, x)\right],$$

to the binaural-disparity template,

$$D_{\text{templ}}(\ell, f) = \mathcal{H}_L^{\text{templ}}(\ell, f) - \mathcal{H}_R^{\text{templ}}(\ell, f).$$

For identical source spectra at the two ears, the source term cancels in the log-ratio, leaving a cue that depends only on the interaural HRTF difference.

For motion analyses, the same logic was applied after adding motion-gradient evidence. The monaural motion posterior combined the static left-ear monaural posterior with the left-ear elevation-motion posterior,

$$p_{\text{mono}}^{\text{motion}}(\ell \mid x) \propto p_L(\ell \mid x)\, p_{\Delta L}(\ell \mid x).$$

The binaural motion posterior combined static monaural evidence, static binaural-disparity evidence, and binaural motion-gradient evidence,

$$p_{\text{bin}}^{\text{motion}}(\ell \mid x) \propto p_L(\ell \mid x)\, p_R(\ell \mid x)\, p_{\text{ratio}}(\ell \mid x)\, p_{\Delta L}(\ell \mid x)\, p_{\Delta R}(\ell \mid x)\, p_{\Delta\text{ratio}}(\ell \mid x).$$

All posterior products were renormalized after multiplication. When the elevation prior was enabled, the final posterior was multiplied by $p_{\text{prior}}(\ell)$ and renormalized before response sampling.

Generation, inference and fixed parameters. By default the model is an ideal observer in which the same individual HRTF set is used both to generate the proximal observation (the "true" HRTF $\mathcal{H}^{\text{true}}$ at the source location) and to construct the template map (the "template" HRTF $\mathcal{H}^{\text{templ}}$ across all candidate locations); generation and inference are therefore matched ($\mathcal{H}^{\text{true}} = \mathcal{H}^{\text{templ}}$). HRTFs are individualized rather than population-averaged: each simulated listener is processed with its own measured HRTF set (CIPIC or SoundSphere 2.0). The same HRTF set was used for generation and template matching in all main analyses. Importantly, the spectral-template model has no participant-level free parameters and is not fit to individual behavioral responses. Simulated listeners are taken from independent HRTF databases rather than from our behavioral cohort (whose individual HRTFs were not measured), so the model yields a population-level prediction rather than a per-participant fit. All parameters were held constant across conditions, sounds and listeners: the analysis band (4–12 kHz), the similarity temperature ($\tau = 0.08$), the elevation prior ($\mu_{\text{prior}} = 5.5°$, $\sigma_{\text{prior}} = 15°$), the motion window ($\Delta T = 100\,ms$), the template step ($\Delta_{\text{templ}} = 1°$), and the number of posterior samples ($M = 250$). The analysis band, motion window, template step and number of posterior samples were set *a priori*. The similarity temperature and the mean and standard deviation of the elevation prior were calibrated once at the population level to reproduce the overall range and distribution of behavioral localization performance. These three quantities were global model parameters: they were not fit separately for individual participants, sounds or experimental conditions, and the same values were used throughout all analyses.

In particular, no parameters were fit per participant or per condition, and differences among the monaural, binaural-disparity, prior and motion conditions emerged from the HRTF-derived information available to each model variant rather than from condition-specific tuning. The only participant-specific behaviorally estimated parameter was the prior width of the separate Bayesian summation observer. This parameter was fit for each participant using monaural data alone and was then held fixed when generating that participant's binaural summation prediction (see Bayesian summation observer). The spectral-template model therefore tests whether the information carried by the correctly learned HRTF cues, together with a single globally calibrated read-out and elevation prior, is sufficient to reproduce the across-condition behavioral pattern, rather than whether participant- or condition-specific parameters can be adjusted to fit the data.

Sampling and Response Prediction. The CIPIC grid extends from $-45°$ to $230.625°$ in elevation (covering both the front and rear hemispheres). Because behavioral targets and responses were restricted to the frontal field, candidate elevations were projected onto a folded frontal coordinate:

$$\hat{e}(\ell) = 90° - |e(\ell) - 90°|$$

This folds the full elevation range onto $[-45°, 90°]$, collapsing front–back ambiguity.

For each reference position x, $M = 250$ responses were randomly drawn from the posterior $p(\ell \mid x)$ by inverse-transform sampling. The sampled indices were converted to folded elevations $\hat{e}(\ell^{(k)})$.

For each reference position x, the mean perceived elevation was taken as $\bar{y}(x) = \frac{1}{M}\sum_{k=1}^{M} \hat{e}(\ell^{(k)})$, which defined the model's mean predicted response at location x. The trial-to-trial variability was then calculated as $\sigma_y(x) = \text{std}\left\{\hat{e}(\ell^{(k)})\right\}$.

To obtain a single estimate of localization precision for comparison with the behavioral data, responses sampled from all reference elevations were pooled, and the standard deviation of the pooled response distribution was calculated.

Predicted responses $\bar{y}(x)$ were regressed against target elevation $\theta(x) \equiv \hat{e}(x)$ (the folded reference elevation) across all reference positions sampled along the frontal median plane (azimuth 0°, elevation −45° to +56.25°):

$$\bar{y}(x) = \alpha + \beta \cdot \theta(x) + \epsilon$$

The slope β (localization gain) is the primary model output: $\beta = 1$ indicates veridical localization, $\beta = 0$ no systematic dependence of predicted elevation on target elevation.

Natural-stimuli set. For Supplementary Fig. S3, we tested how well the model can predict the location of real-world sounds. Natural sounds were taken from the ESC-50 corpus (one token randomly chosen from each of the 50 categories) [38] and from the MIT natural sounds stimulus set [35]. For each sound we took a 400 ms window with an onset determined by a sliding RMS envelope crossing 25% of the maximum.

Bayesian summation observer

To test whether the binaural advantage in localization gain could be explained by reliability summation of the two ears, rather than by an additional binaural spectral-disparity computation, we developed a Bayesian summation observer. Whereas the binaural summation posterior p_{LR} described above provides the two-ear summation prediction for the spectral-template model, this observer provides the corresponding two-ear summation prediction for the behavioral data (the gray L+R prediction in the behavioral panels of Fig. 3).

The observer models responses as Bayesian estimates of target elevation θ. The prior is Gaussian,

$$p(\theta) = \mathcal{N}\big(0, \sigma^2_{\mathrm{prior}}\big)\,, \tag{2}$$

centered at the default straight-ahead response in the absence of reliable sensory information. The likelihood is an unbiased Gaussian centered on the true target elevation,

$$p(y \mid \theta) = \mathcal{N}\big(\theta, \sigma^2_{\mathrm{lik}}\big)\,. \tag{3}$$

Under precision-weighted cue combination, the expected posterior estimate is

$$\hat{\theta} = w_{\mathrm{lik}}\,\theta, \tag{4}$$

where

$$w_{\mathrm{lik}} = \frac{1/\sigma^2_{\mathrm{lik}}}{1/\sigma^2_{\mathrm{prior}} + 1/\sigma^2_{\mathrm{lik}}} \tag{5}$$

is the weight assigned to the likelihood ($w_{\mathrm{lik}} = 1 - w_{\mathrm{prior}}$, with w_{prior} the complementary prior weight). Because the prior mean is zero and the likelihood mean is θ, this likelihood weight equals the predicted localization gain, defined as the regression slope of indicated elevation against target elevation.

The posterior variance is

$$\sigma^2_{\mathrm{post}} = \left(\frac{1}{\sigma^2_{\mathrm{prior}}} + \frac{1}{\sigma^2_{\mathrm{lik}}}\right)^{-1}. \tag{6}$$

Using the expression for w_{lik}, this can equivalently be written as

$$\sigma^2_{\mathrm{post}} = w_{\mathrm{lik}}\,\sigma^2_{\mathrm{lik}}. \tag{7}$$

We estimated the empirical posterior variance as

$$\sigma^2_{\mathrm{post}} = \mathrm{Var}[y_i - \theta_i]\,, \tag{8}$$

where y_i is the response and θ_i the target elevation on trial i. Rearranging gives the likelihood variance without introducing an additional free parameter:

$$\sigma_{\text{lik}}^2 = \frac{\sigma_{\text{post}}^2}{w_{\text{lik}}}. \tag{9}$$

The prior standard deviation σ_{prior} was estimated separately for each participant from the monaural conditions by minimizing the mismatch between the empirically observed posterior variance and the posterior variance predicted by the observer. For a candidate value of σ_{prior}, the predicted posterior variance was

$$\hat{\sigma}_{\text{post}}^2 = \left(\frac{1}{\sigma_{\text{prior}}^2} + \frac{1}{\sigma_{\text{lik}}^2} \right)^{-1}. \tag{10}$$

The optimization objective was

$$\mathcal{L} = \sum \left[\log \hat{\sigma}_{\text{post}}^2 + \frac{\left(\sigma_{\text{post}}^2 - \hat{\sigma}_{\text{post}}^2\right)^2}{2\hat{\sigma}_{\text{post}}^2} \right], \tag{11}$$

which was minimized using SciPy `minimize`.

Under the two-ear summation hypothesis, the binaural likelihood standard deviation is reduced by a factor of $\sqrt{2}$ relative to the monaural likelihood standard deviation,

$$\sigma_{\text{lik}}^{\text{bin}} = \frac{\sigma_{\text{lik}}^{\text{mon}}}{\sqrt{2}}. \tag{12}$$

The predicted binaural posterior variance is therefore

$$\left(\sigma_{\text{post}}^{\text{bin}}\right)^2 = \left(\frac{1}{\sigma_{\text{prior}}^2} + \frac{1}{\left(\sigma_{\text{lik}}^{\text{bin}}\right)^2} \right)^{-1}. \tag{13}$$

The corresponding prior weight is

$$w_{\text{prior}}^{\text{bin}} = \frac{1/\sigma_{\text{prior}}^2}{1/\sigma_{\text{prior}}^2 + 1/\left(\sigma_{\text{lik}}^{\text{bin}}\right)^2}. \tag{14}$$

The predicted binaural posterior mean is

$$\mu_{\text{post}}^{\text{bin}} = \mu_{\text{prior}}\, w_{\text{prior}}^{\text{bin}} + \mu_{\text{lik}} \left(1 - w_{\text{prior}}^{\text{bin}}\right). \tag{15}$$

Because $\mu_{\text{prior}} = 0$ and $\mu_{\text{lik}} = 1$ in gain units, this posterior mean is equivalent to the predicted binaural localization gain,

$$w_{\text{lik}}^{\text{bin}} = 1 - w_{\text{prior}}^{\text{bin}}. \tag{16}$$

Comparing this predicted binaural gain and variance with the empirically observed binaural gain and response variability tests whether the binaural advantage exceeds what would be expected from reliability summation of two independent ear-specific monaural estimates. All other contrasts, including monaural versus binaural, structured versus flat-spectrum, and motion versus no-motion, used the full cohort of 24 participants. Contrasts against the summation-predicted gain and SD included only participants for whom a valid per-participant summation prediction could be estimated ($n = 19$–22 of 24), which is reflected in the corresponding degrees of freedom.

Data analysis

Localization response functions. Localization responses were analyzed separately for each of the eight conditions constituting the three factors: Ear (monaural or binaural), Sound Type (structured-spectrum noise or flat-spectrum noise), and Head Motion (motion or no motion). For each condition, repeated trial

responses were first averaged within each participant and target elevation. Population response functions were then estimated with linear mixed-effects models relating indicated elevation to target elevation:

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\theta_{ij} + \varepsilon_{ij}, \tag{17}$$

where y_{ij} is the mean indicated elevation of participant i at target elevation j, θ_{ij} is the target elevation, β_0 and β_1 are the population intercept and slope, and b_{0i} and b_{1i} are participant-specific deviations from these values. Thus, each model included participant-specific random intercepts and random slopes for target elevation.

Target elevation was mean-centered during fitting, and coefficients were back-transformed to degrees for reporting. The fixed-effect slope was interpreted as localization gain, and the fixed-effect intercept as localization bias. Behavioral response functions are shown with 95% confidence intervals for the fixed-effect fit and per-location standard errors of the mean (Fig. 3). Participant-wise gains were additionally estimated with separate regressions for each participant and compared between monaural and binaural conditions using paired t-tests.

The same condition-wise regression framework was applied to the spectral-template model outputs. For model panels, uncertainty around the population fit was shown as a 95% prediction interval for a new participant, and per-location variability was shown as standard deviation. Predicted binaural response functions derived from the summation observer were fit for visualization with pooled ordinary least squares.

For the model, the L+R prediction (the binaural summation posterior p_{LR}) is fit using the previously described population regression setting.

Acknowledgements

We thank Maximilian Rupp for assistance with data collection.

Funding

This work was supported by the Baden-Württemberg Stiftung through the VA-MORPH project within the Neurorobotik programme (Project No. NEU012) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Code & Data availability

The model code will be made available on GitHub upon publication.

The behavioral datasets generated and analyzed during the current study, together with the analysis code required to reproduce the reported results, will be made publicly available upon publication in a permanent online repository (https://osf.io) and assigned a DOI. The repository link and DOI will be included in the published article.

Author contributions

T.O., C.R. H.N. and M.O.E. conceived the study. T.O. conducted the experiments, curated the data, and implemented the computational models. T.O. and M.O.E. developed the methodology, performed the formal analyses, validated the results, interpreted the findings, and prepared the figures. M.O.E. and H.N. secured funding and provided resources and supervision. T.O., C.R. and M.O.E. wrote the manuscript. M.O.E., C.R. and H.N. reviewed and revised the manuscript. All authors approved the final version of the manuscript.

Competing interests

The authors declare that no competing interests exist.

Correspondence

Timo Oess, timo.oess@uni-ulm.de

References

1. L. Abdoune and M. Fezari. Feature extraction for everyday life sounds. In *5th International Conference on Control & Signal Processing (CSP-2017) Proceeding of Engineering and Technology-PET*, volume 26, pages 186–191, 2017.

2. K. Adiloglu, A. Annies, E. Wahlen, H. Purwins, and K. Obermayer. A graphical representation and dissimilarity measure for basic everyday sound events. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1542–1552, 2012.

3. V. Algazi, R. Duda, D. Thompson, and C. Avendano. The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pages 99–102, 2001.

4. J. R. Angell. Monaural localization of sound. *Science*, 13(333):775–777, 1901.

5. P. W. Battaglia, M. Di Luca, M. O. Ernst, P. R. Schrater, T. Machulla, and D. Kersten. Within-and cross-modal distance information disambiguate visual size-change perception. *PLoS computational biology*, 6(3):e1000697, 2010.

6. D. W. Batteau. The Role of the Pinna in Human Localization. *Proceedings of the Royal Society B: Biological Sciences*, 168(1011):158–180, 1967.

7. R. Baumgartner, R. Barumerli, B. Brands, and P. Majdak. Short-term statistical learning mitigates the ill-posed problem of sound localization. *Trends in Hearing*, 30:23312165261465030, 2026.

8. R. Baumgartner, D. K. Reed, B. Tóth, V. Best, P. Majdak, H. S. Colburn, and B. Shinn-Cunningham. Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias. *Proceedings of the National Academy of Sciences*, 114(36):9743–9748, 2017.

9. J. Blauert. Sound localization in the median plane. *Acta Acustica united with Acustica*, 22(4):205–213, 1969.

10. J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

11. E. G. Boring. Size constancy and emmert's law. *The American Journal of Psychology*, 53(2):293–295, 1940.

12. A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1990.

13. M. Burkhard and R. Sachs. Anthropometric manikin for acoustic research. *The Journal of the Acoustical Society of America*, 58(1):214–222, 1975.

14. R. A. Butler. An analysis of the monaural displacement of sound in space. *Perception & Psychophysics*, 41(1):1–7, 1987.

15. R. A. Butler and R. A. Humanski. Localization of sound in the vertical plane with and without high-frequency spectral cues. *Perception & psychophysics*, 51(2):182–186, 1992.

16. Y. Deng, I. Choi, B. Shinn-Cunningham, and R. Baumgartner. Impoverished auditory cues limit engagement of brain networks controlling spatial selective attention. *NeuroImage*, 202:116151, 2019.

17. M. B. Gardner. Some monaural and binaural facets of median plane localization. *The Journal of the Acoustical Society of America*, 54(6):1489–1495, Dec. 1973.

18. D. Genzel, M. Schutte, W. O. Brimijoin, P. R. MacNeilage, and L. Wiegrebe. Psychophysical evidence for auditory motion parallax. *Proceedings of the National Academy of Sciences*, 115(16):4264–4269, 2018.

19. W. M. Hartmann and A. Wittenberg. On the externalization of sound images. *The Journal of the Acoustical Society of America*, 99(6):3678–3688, 1996.

20. J. Hebrank and D. Wright. Spectral cues used in the localization of sound sources on the median plane. *The Journal of the Acoustical Society of America*, 56(6):1829–1834, 1974.

21. P. Hofman and A. Van Opstal. Binaural weighting of pinna cues in human sound localization. *Experimental Brain Research*, 148(4):458–470, 2003.

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

23. P. M. Hofman and A. J. Van Opstal. Spectro-temporal factors in two-dimensional human sound localization. *The Journal of the Acoustical Society of America*, 103(5):2634–2648, May 1998.

24. P. M. Hofman and A. J. Van Opstal. Bayesian reconstruction of sound localization cues from responses to random spectra. *Biological Cybernetics*, 86(4):305–316, Apr. 2002.

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

26. D. Kersten, P. Mamassian, and A. Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55(1):271–304, 2004.

27. A. Kulkarni and H. S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–749, Dec. 1998.

28. J. H. McDermott and E. P. Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011.

29. J. C. Middlebrooks. Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America*, 92(5):2607–2624, 1992.

30. J. C. Middlebrooks and D. M. Green. Sound localization by human listeners. *Annu Rev Psychol*, 42:135–159, 1991.

31. H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky. Spectral entropy based feature for robust asr. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–193. IEEE, 2004.

32. W. Młynarski and J. Jost. Statistics of natural binaural sounds. *PloS one*, 9(10):e108968, 2014.

33. M. Morimoto. 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, Mar. 2001.

34. I. Nelken and E. Young. Two separate inhibitory mechanisms shape the responses of dorsal cochlear nucleus type IV units to narrowband and wideband stimuli. *Journal of neurophysiology*, 71(6):2446–2462, 1994.

35. S. Norman-Haignere, N. G. Kanwisher, and J. H. McDermott. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88(6):1281–1296, 2015.

36. C. Parise, M. Gori, S. Finocchietti, M. Ernst, D. Esposito, and A. Tonelli. Happy new ears: Rapid adaptation to novel spectral cues in vertical sound localization. *iScience*, 27(12), 2024.

37. C. V. Parise, K. Knorre, and M. O. Ernst. Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, 111(16):6104–6108, Apr. 2014.

38. K. J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015.

39. L. Rayleigh. Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

40. D. K. Reed, M. Chait, B. Tóth, I. Winkler, and B. Shinn-Cunningham. Spatial cues can support auditory figure-ground segregation. *The Journal of the Acoustical Society of America*, 147(6):3814–3818, 2020.

41. M. Reiß and G. Reiß. Ear preference: Association with other functional asymmetries of the ears. *Perceptual and motor skills*, 86(2):399–402, 1998.

42. S. K. Roffler and R. A. Butler. Factors That Influence the Localization of Sound in the Vertical Plane. *The Journal of the Acoustical Society of America*, 43(6):1255–1259, June 1968.

43. C. Searle, L. Braida, D. Cuddy, and M. Davis. Binaural pinna disparity: another auditory localization cue. *The Journal of the Acoustical Society of America*, 57(2):448–455, 1975.

44. E. A. G. Shaw. The External Ear. In H. W. Ades, A. Axelsson, I. L. Baird, G. v. Békésy, R. L. Boord, C. B. G. Campbell, O. Densert, D. H. Eldredge, H. Engström, J. Fex, J. M. Harrison, O. W. Henson, M. E. Howe, S. Iurato, A. Michelsen, A. R. Møller, R. R. Pfeiffer, S. Rauch, I. Rauch, E. A. G. Shaw, J. Wersäll, E. G. Wever, W. D. Keidel, and W. D. Neff, editors, *Auditory System: Anatomy Physiology (Ear)*, Handbook of Sensory Physiology, pages 455–490. Springer, Berlin, Heidelberg, 1974.

45. N. C. Singh and F. E. Theunissen. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6):3394–3411, 2003.

46. M. M. Van Wanrooij and A. J. Van Opstal. Contribution of head shadow and pinna cues to chronic monaural sound localization. *The Journal of neuroscience*, 24(17):4163–4171, 2004.

47. M. M. Van Wanrooij and A. J. Van Opstal. Sound localization under perturbed binaural hearing. *Journal of neurophysiology*, 97(1):715–726, 2007.

48. H. Wallach. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339, 1940.

49. M. Warnecke, S. Clapp, Z. Ben-Hur, D. Alon, S. Amengual Garí, and P. Calamia. Sound sphere 2: A high-resolution hrtf database. In *Proceedings of the 5th AES International Conference on Audio for Virtual and Augmented Reality (AVAR 2024)*. Audio Engineering Society, Aug. 2024.

50. F. L. Wightman and D. J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J Acoust Soc Am*, 91(3):1648–1661, 1992.

51. F. L. Wightman and D. J. Kistler. Sound localization. In *Human psychophysics*, pages 155–192. Springer, 1993.

52. F. L. Wightman and D. J. Kistler. Monaural sound localization revisited. *The Journal of the Acoustical Society of America*, 101(2):1050–1063, Feb. 1997.

53. T. C. T. Yin. Neural Mechanisms of Encoding Binaural Localization Cues in the Auditory Brainstem. In *Integrative Functions in the Mammalian Auditory Pathway*, volume 15 of *Springer Handbook of Auditory Research*, pages 99–159. Springer-Verlag, Berlin, 2002.

54. E. D. Young, G. A. Spirou, J. J. Rice, and H. F. Voigt. Neural Organization and Responses to Complex Stimuli in the Dorsal Cochlear Nucleus. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 336(1278):407–413, 1992.

55. P. Zakarauskas and M. S. Cynader. A computational theory of spectral cue localization. *The Journal of the Acoustical Society of America*, 94(3):1323, 1993.

56. B. Zonooz, E. Arani, K. P. Körding, P. A. T. R. Aalbers, T. Celikel, and A. J. V. Opstal. Spectral Weighting Underlies Perceived Sound Elevation. *Scientific Reports*, 9(1):1642, Feb. 2019.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NaturalSoundsBinauralRatio.mp4
- Supplementary.pdf
- DemoElevationHearing.mp4
- VSSLcode.zip
- VSSLdata.zip