Natural auditory scene statistics shapes human spatial hearing

Cesare V. Parisea,b,1, Katharina Knorre, and Marc O. Ernsta,b,1

*Max Planck Institute for Biological Cybernetics and Bernstein Center for Computational Neuroscience, 72076 Tübingen, Germany; and bCognitive Neuroscience Department and Cognitive Interaction Technology-Center of Excellence, Bielefeld University, 33615 Bielefeld, Germany

Edited by Dale Purves, Duke University, Durham, NC, and approved March 10, 2014 (received for review December 5, 2013)

Human perception, cognition, and action are laced with seemingly arbitrary mappings. In particular, sound has a strong spatial connotation: Sounds are high and low, melodies rise and fall, and pitch systematically biases perceived sound elevation. The origins of such mappings are unknown. Are they the result of physiological constraints, do they reflect natural environmental statistics, or are they truly arbitrary? We recorded natural sounds from the environment, analyzed the elevation-dependent filtering of the outer ear, and measured frequency-dependent biases in human sound localization. We find that auditory scene statistics reveals a clear mapping between frequency and elevation. Perhaps more interestingly, this natural statistical mapping is tightly mirrored in both ear-filtering properties and in perceived sound location. This suggests that both sound localization behavior and ear anatomy are fine-tuned to the statistics of natural auditory scenes, likely providing the basis for the spatial connotation of human hearing.

Results

To trace the origins of the mapping between auditory frequency and perceived vertical elevation, we first measured whether this mapping is already present in the statistics of natural auditory signals. When trying to characterize the statistical properties of incoming signals, it is critical to distinguish between distal stimuli, the signals as they are generated in the environment, and proximal stimuli, the signals that reach the transducers (i.e., the middle and inner ear). In the case of auditory stimuli this is especially important, because the head and the outer ear operate as frequency- and elevation-dependent filters (15), which modulates the spectra of the sounds reaching the middle ear as a function of the elevation of the sound source relative to the observer (the head-related transfer function, HRTF). Notably, the structure of the peaks and notches produced by the HRTF on the spectra of the incoming signals is known to provide reliable cues for auditory localization in the median plane (16). We therefore looked for the existence of a frequency-elevation mapping (FEM) in the statistics of natural auditory scenes and in the filtering properties of the outer ear. Hence, we effectively measured the mapping between frequency and elevation in both the distal and the proximal stimuli.

To look for the existence of an FEM in the natural acoustic environment, we recorded a large sample of environmental sounds (~50,000 recordings, 1 s each) by means of two directional microphones mounted on the head of a human freely moving indoors and outdoors in urban and rural areas (around Bielefeld, Germany). Overall, the recordings revealed a consistent mapping between the frequency of sounds and the average elevation of their sources in the external space $F(5, 57, 859) = 35.8, P < 0.0001; $ Methods, $ $, which was particularly evident in the middle range of the spectrum, between 1 and 6 kHz (Fig. 1C, Upper). That is, high-frequency sounds have a tendency to originate from elevated sources in natural auditory scenes. We can only speculate about the origins of this mapping: it could either be that at higher elevations, more energy is generated in high frequencies (e.g., leaves on the trees rustle in a higher frequency range than the footsteps on the floor), or it could also be that the absorption of the ground is frequency-dependent in a way that it filters out more of the high-frequency spectrum.

To look for the existence of an FEM in the filtering properties of the ear, we analyzed a set of 45 HRTFs [the CIPIC database (17); Methods and Fig. S1], and found again a clear mapping between frequency and elevation $F(5, 264) = 216.6, P < 0.0001; $ Fig. 1C, Lower. That is, due to the filtering properties of the outer ear, sounds coming from high (head-centered) elevations

Significance

Auditory pitch has an intrinsic spatial connotation: Sounds are high or low, melodies rise and fall, and pitch can ascend and descend. In a wide range of cognitive, perceptual, attentional, and linguistic functions, humans consistently display a positive, sometimes absolute, correspondence between sound frequency and perceived spatial elevation, whereby high frequency is mapped to high elevation. In this paper we show that pitch borrows its spatial connotation from the statistics of natural auditory scenes. This suggests that all such diverse phenomena, such as the convoluted shape of the outer ear, the universal use of spatial terms for describing pitch, or the reason why high notes are represented higher in musical notation, ultimately reflect adaptation to the statistics of natural auditory scenes.

Author contributions: C.V.P. and M.O.E. designed research; C.V.P. and K.K. performed research; C.V.P. analyzed data; C.V.P. and M.O.E. performed statistical modeling; and C.V.P. and M.O.E. wrote the paper.

The authors declare no conflict of interest.

1To whom correspondence may be addressed. E-mail: cesare.parise@uni-bielefeld.de or marc.ernst@uni-bielefeld.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1322705111/-/DCSupplemental.

This article is a PNAS Direct Submission.
have more energy at high frequencies. These results demonstrate that an FEM is consistently present in the statistics of both proximal and distal stimuli. This suggests that the perceptual FEM might ultimately reflect a tuning of the human auditory system to the statistics of natural sounds.

Finally, we determined the correlation between the FEM measured in proximal and distal stimuli, and found a strong similarity between the two mappings ($\rho = 0.79$, interquartile range = 0.72–0.84). That is, the filtering properties of the external ear accentuate the FEM that is present in natural auditory scenes. One possible reason for this similarity is that the elevation-dependent filtering of the outer ear is set to maximize the transfer of naturally available information. This result parallels previous findings in human vision showing a high degree of similarity between the spectra of natural images and the optical transfer function of the eye (18). This might suggest that human spatial hearing is so finely tuned to the environment that even the filtering properties of the outer ear, and hence its convoluted anatomy, evolved to mirror the statistics of natural auditory scenes.

To investigate the relation between human performance and the FEM in proximal and distal stimuli, we asked participants to localize on a 2D plane (19) a set of narrowband (∼1.8-octave) auditory noises with different central frequencies (Movie S1). Sounds were played from a set of 16 speakers hidden behind a sound-transparent projection screen, arranged on a 4 × 4 grid subtending an angle of ∼30°. Participants were asked to
point toward the sound source, while pointing direction was measured (Methods). Participants performed the sound localization experiment in three conditions in which we tilted their whole body [0°, 45°, and 90°] to dissociate head- from world-centered elevation (Fig. S2). Given that the FEM in the proximal and distal stimuli come in different reference frames (the first being head-centered, the second world-centered), tilting participants allows one to separately estimate the relationship between sound localization biases and the FEM measured in the proximal and distal stimuli. In the extreme case, when the participant lay horizontally on the side (tilt = 90°), head- and world-centered elevations were orthogonal, and as a result vertical sound localization biases on each reference frame were independent.

When participants had to localize white noise (which includes all spectral frequencies), performance was quite accurate and the orientation-dependent spatial distortions were minor (Fig. L4, Right). Conversely, sound source localization was strongly biased when the stimulus consisted of narrowband noise (Fig. L4). Such biases depended both on the spectra of the stimuli and the orientation of the observers. This bias was especially strong for those frequencies in which hearing sensitivity, as measured by equal loudness contours, was at its maximum (20). In the three frequency bands between 1.4 and 8 kHz, the localization responses were virtually independent from the actual sound source location and the reported deviation was almost entirely determined in a very consistent way by the frequency of the signals (Fig. L4, Center). Notably, such biases showed a clear mapping between frequency and elevation (Fig. 1B), which was evident in both head- and world-centered coordinates (see also refs. 5, 11). Importantly, such localization biases were significantly correlated with the FEM present in proximal and distal stimuli (ρ = 0.76 for world-centered biases with distal stimulus and ρ = 0.78 for head-centered biases with the proximal stimulus; see SI Text). Consistent with previous studies (21, 22), we also found moderate but consistent frequency-dependent biases in horizontal sound localization. These results demonstrate the existence of striking frequency- and body-orientation-dependent perceptual biases in sound localization. The results also demonstrate the dependence of such biases on the statistics of natural auditory scenes, and on the filtering properties of the outer ear.

However, it is not immediately obvious why there is such a high degree of correspondence between the behavioral biases found in sound localization and the statistical mappings found in both the environment, and in the filtering properties of the ear. To better understand this close correspondence we would need a generative model. Recently, the Bayesian approach has been successfully used for describing such generative models and in particular for describing the effects of stimulus statistics on perceptual judgments (23–27). In Bayesian terms, the frequency dependency of sound source location can be modeled as a prior distribution p(s) representing the probability of a sound source s of a given frequency f occurring at some given 2D spatial location s = (s_x, s_y). Based on the measured statistics of natural auditory scenes, the filtering properties of the ear, and the biases in sound localization, we postulated the existence of two distinct mappings between frequency and elevation, respectively coding the expected elevation of sounds as a function of the frequency spectrum in either head- or world-centered coordinates. Therefore, we modeled two frequency-dependent priors for elevation, one being head-centered and the other world-centered. This model would involve a mechanism dedicated to the extraction and combination of relevant spectral cues from the proximal stimulus (such as the frequencies with more energy), and mapping the result to certain head- and world-centered elevations. For simplicity, such priors were modeled as Gaussian distributions, whose means represent the expected elevation given the spectrum of the incoming signal (Fig. 1E). Given that participants had to localize the auditory stimuli on a 2D plane, the Bayesian ideal observer model was also framed in 2D space (Methods and Fig. S3). In a similar fashion, we also modeled incoming sensory information in terms of Gaussian probability distributions over spatial locations: the likelihood function. According to Bayesian decision theory, prior expectations and incoming sensory information are combined to determine the final percepts. This model predicts that as soon as the sensory information from the peaks and notches of the HRTF (16) becomes unreliable, such as when sounds have a narrow spectrum as in the present experiment, the perceived elevation would be mainly determined by the prior. Given this generative model, we can use the responses from the sound localization task to estimate the expected head- and world-centered elevation of a sound given its frequency, that is, the shape of the internal FEMs.

The shapes of the estimated frequency-dependent priors on vertical sound location (Fig. 1D) reveal a strong similarity with the frequency-dependent biases measured from the responses of the participants (Fig. 1B, red lines). Given that such biases are supposedly the outcome of the estimated frequency-dependent priors, this is an expected finding that further validates the current modeling approach. Having empirically determined the shapes of the internal FEM (in both head- and world-centered coordinates), we can look for similarities (i.e., correlation) between shapes of such perceptual mappings, and the ones that we measured from both the statistics of the acoustic environment and from the HRTFs. Notably, both estimated priors significantly correlated with the statistical mappings present in proximal and distal stimuli (i.e., the maximum of the frequency spectra against spatial elevation) (Fig. 1F). However, the head-centered prior was more correlated to the FEM measured from the filtering properties of the outer ear, whereas the world-centered prior was more correlated to the FEM present in environmental sounds. These results demonstrate that the perceptual FEM in humans jointly depends on the statistics of both natural auditory scenes and the filtering properties of the outer ear.

Discussion

Previous studies have already hypothesized the grounding of cross-dimensional sensory correspondences in the statistics of incoming stimuli (13, 28). However, none of them actually measured how such mappings relate to the statistical properties of the stimuli. Our results demonstrate that an FEM is already present in the statistics of both the proximal and the distal stimuli. Moreover, we demonstrate that the perceptual FEM is in fact a twofold mapping, which separately encodes the statistics of natural auditory scenes and the filtering properties of the outer ear in different frames of reference. Interestingly, this finding provides further support for the role of vestibular and proioceptive information in sound localization (29). These results highlight the possibility of using sound spectral frequency to simulate the vertical elevation of sound sources.

The pervasiveness of the FEM in the statistics of the stimuli readily explains why previous research found this mapping to be absolute (5, 30) (i.e., each frequency is related to exactly one elevation), universal (3, 13) (cross-cultural and language independent), and already present in early infancy (14); and it argues against interpretations of cross-dimensional sensory correspondences in terms of “weak synesthesia” (9). The mapping between pitch and elevation, also reflected in musical notation and in the lexicon of most natural languages (13), has often been considered a metaphorical mapping (6, 31), and cross-sensory correspondences have been theorized to be the basis for language development (32). The present findings demonstrate that, at least in the case of the FEM, such a metaphorical mapping is indeed embodied and based on the statistics of the environment, hence raising the intriguing hypothesis that language itself might
have been influenced by a set of statistical mappings between the sensory signals. Even more, besides the FEM, human perception, cognition, and action are faced with seemingly arbitrary correspondences (33), such as for example that yellow-reddish colors are associated with a warm temperature, or that sour foods taste sharp. We may speculate here that many of these mappings are in fact the reflection of natural scene statistics.

**Methods**

**Recordings from the Environment.** The recordings were taken by two microphones (Sennheiser ME105) mounted one above the other on the side of a baseball cap, and pointing ±25° from the horizontal midline. The distance between the microphones was 4 cm, and the experimenter kept the head in a natural upright position throughout the recording session. We did not constrain naturally occurring head movements while recording the sounds, because it was our goal to measure the natural soundscape of a listener with ordinary postures. The recordings had a sampling frequency of 44,100 Hz and a depth of 16 bits. Each recording was filtered with a pool of 71 band-pass filters (constant log-frequency width, overall range = 0.5–16 kHz), and the elevations of the resulting signals were measured from the lag that maximized the cross-correlation between the two microphones (if the cross-correlation was <0.5, elevation was not calculated). The elevation mapped to each frequency was calculated as the average elevation across recordings.

**Analysis of the HRTF.** The CIPIC HRTF (17) database includes the transfer function produced by the outer ear of 45 humans for 71 different frequency channels (linearly spaced between 0.66 and 16.1 kHz), and recorded from 50 elevations (range = 45° to 230°). The elevation mapped to each frequency channel was calculated from each individual HRTF as the elevation with the highest transfer value (dB) for that particular frequency channel (28) for each of the 45 human heads. The elevation bias (Fig. 1, left) showed a main effect of frequency [F(5,45) = 4.074, P = 0.0044], tilt [F(2,18) = 43.474, P < 0.001], and a significant interaction [F(10,90) = 8.11, P < 0.001].

To engage participants with the experiment, the whole task was presented as a shooting video game (19): A bullet-hole graphic effect (spatially aligned with the pointing response) and the sound of a gunshot accompanied each response, closely followed by the sound of a loading gun. The sound effects came from an additional speaker placed in the proximity of participants’ heads. To avoid those effects interfering with the experimental stimuli, a temporal interval randomized between 2 and 3 s separated two consecutive trials. To further motivate the participants, they were told that they could get points as a function of their performance. Every 16 trials, a fake high score list was presented, in which participants on average ranked third out of 10.

**Modeling.** In the present experiment, participants were presented with physical stimuli coming from a source s = (s_x, s_y, s_z). Using both binaural cues and the structure of the peak and notches in the frequency spectrum, the auditory system can estimate, respectively, the azimuth and the elevation of the sound source. Assuming that the sensory estimate s = (s_x, s_y, s_z) derived from the physical source of a sound with frequency f is unbiased but noisy, with some Gaussian noise ε = (σ_x,σ_y,σ_z) added independently to each spatial dimension (i.e., s = s_x + ε_x, s_y + ε_y, s_z + ε_z), the likelihood distribution p(s|s) for the spatial location of the sound source is a 2D Gaussian:

\[ p(s|s) = N(s, \Sigma) = \frac{1}{\sqrt{(2\pi)^3}|\Sigma|} \exp\left(-\frac{1}{2} (s - s)^\top \Sigma^{-1} (s - s)\right), \]

with mean \( s = (s_x, s_y, s_z) \cdot R_b \) and covariance matrix \( \Sigma = \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix} \cdot R_b \) (Fig. 33, left), where \( R_b \) is a rotation matrix that rotates the axes according to the orientation of the body with respect to gravitational vertical (\( 0° \)).

The expected elevation of a sound source of a given frequency spectrum can be modeled as a Gaussian a priori probability distribution, whose mean represents the expected location given the maximum of the frequency spectrum, and the variance the uncertainty of the mapping. Given that we empirically measured an FEM in the filtering properties of the outer ear and the statistics of the natural auditory scenes, we assumed the existence of two independent priors encoding, respectively, the FEM in head- and world-centered coordinates.

In head-centered coordinates the prior distribution \( p_{\text{head}}(s) \) for the location \( s_{\text{head}} \) of a sound with frequency \( f \) is defined as a 2D Gaussian:

\[ p_{\text{head}}(s) = N(s, \Sigma_{\text{head}}) = \frac{1}{\sqrt{(2\pi)^2|\Sigma_{\text{head}}|}} \exp\left(-\frac{1}{2} (s - s_{\text{head}})^\top \Sigma_{\text{head}}^{-1} (s - s_{\text{head}})\right), \]

with mean \( s_{\text{head}} = (s_{x_{\text{head}}}, s_{y_{\text{head}}}) \cdot R_b \) and covariance matrix \( \Sigma_{\text{head}} = \begin{pmatrix} \sigma_{x_{\text{head}}}^2 & 0 \\ 0 & \sigma_{y_{\text{head}}}^2 \end{pmatrix} \cdot R_b \) (Fig. 33, second column), where \( \sigma_{x_{\text{head}}} \) represents the expected spatial elevation and the variance \( \sigma_{y_{\text{head}}}^2 \) the mapping uncertainty. For simplicity, we assumed no mapping between frequency and the head-centered left–right location of a sound source; therefore, the prior had a mean azimuth of zero and \( \sigma_{y_{\text{head}}} \) variance. In a similar fashion, the world-centered prior distribution \( p_{\text{world}}(s) \) for the location \( s_{\text{world}} \) of a sound with frequency \( f \) is defined as a 2D Gaussian:

\[ p_{\text{world}}(s) = N(s, \Sigma_{\text{world}}) = \frac{1}{\sqrt{(2\pi)^2|\Sigma_{\text{world}}|}} \exp\left(-\frac{1}{2} (s - s_{\text{world}})^\top \Sigma_{\text{world}}^{-1} (s - s_{\text{world}})\right), \]

with mean \( s_{\text{world}} = (s_{x_{\text{world}}}, s_{y_{\text{world}}}, s_{z_{\text{world}}}) \cdot R_b \) and covariance matrix \( \Sigma_{\text{world}} = \begin{pmatrix} \sigma_{x_{\text{world}}}^2 & 0 & 0 \\ 0 & \sigma_{y_{\text{world}}}^2 & 0 \\ 0 & 0 & \sigma_{z_{\text{world}}}^2 \end{pmatrix} \cdot R_b \) (Fig. 33, third column), where \( \sigma_{x_{\text{world}}} \) represents the expected spatial elevation and the variance \( \sigma_{y_{\text{world}}}^2, \sigma_{z_{\text{world}}}^2 \) the mapping uncertainty. Again, the prior was made uninformative as to the world-centered azimuth location of the sound source.

The statistically optimal way to combine noisy sensory information with prior knowledge is described by the Bayesian theorem, according to which the posterior \( p(s|f) \) (Fig. 33, right), on which the percept is based, is proportional to the product of the likelihood (i.e., the sensory information) and the prior (here, the FEM):

\[ p(s|f) \propto p_{\text{head}}(s) \cdot p_{\text{world}}(s) \cdot p_f(s). \]

Assuming all of the noise in the data to be due to sensory (as opposed to response-motor) noise (19), participants’ responses would represent random samples of the posterior distribution \( p_f(s) \). Therefore, given the psychophysical data it is possible to estimate the parameters of the model and eventually estimate the shape of the internal FEMs. Using a maximum-likelihood
procedure, we fitted the mean of the priors $\sigma_{f|\theta}$ and $\sigma_{\omega|\theta}$ for each frequency band that we tested and, assuming for simplicity that the strength of the FEM is independent of frequency, we fitted the two mapping uncertainties $\sigma_{f|\omega}$ and $\sigma_{\omega|f}$. We also fitted the covariance matrix $\Sigma_f$ of the likelihood function (given that sound frequency is known to impact the sensitivity to the elevation of a sound source, we fitted a different variance $\sigma_f$ for each frequency band tested). Overall, the model had 21 free parameters fitted over 13,440 trials, that is, 640 trials per parameter.

Alternatively, we used the responses in the white noise condition to estimate further frequency-independent distortions of perceived space. This was modeled by shifting the mean of the posterior, for each position and orientation, by the bias calculated from the white noise (i.e., the discrepancy between physical and perceived position in the white noise condition).

The parameters were fitted over the mean pointing response for each condition (i.e., frequency, tilt, and spatial location) across participants (i.e., orientation, by the bias calculated from the white noise (i.e., the discrepancy parameters fitted over 13,440 trials, that is, 640 trials per parameter. The results in Fig. 1D represent the mean of the 10 iterations. To minimize the effect of the starting parameter values we iteratively repeated each fitting procedure 30 times using random starting values, and selected the set of parameter values that provided the best fit.

Given that in this study we were especially interested in the effects of frequency on perceived elevation, we only included frequency-dependent priors for elevation in our model. However, previous studies also demonstrated the existence of frequency-dependent biases for azimuth (22), and such biases have also been related to the filtering properties of the outer ear (21). That said, biases on azimuth had a much smaller magnitude in the present study (–21°; Fig. 1B, green line) and they were almost frequency independent. The reason why these azimuth biases here were so small compared with Butler (22)—and thus could be safely neglected in the modeling—might be because our task involved binaural hearing, thus having time difference and loudness difference between the ears as a main cue to azimuth, whereas Butler (22) determined azimuth biases for monaural hearing only.

**Comparison Between the Estimated Priors and the Statistics of the Natural Sounds and Filtering Properties of the Outer Ear.** To calculate the relation between the priors and the statistics of the proximal and distal stimuli, we first divided the spectra of the HRTF and the recordings into the same six frequency bands that we used for the experiment. The elevation mapped to each frequency band corresponded to the mean of the elevations within the frequency range. This procedure was carried out individually for each recording and HRTF, and the results were used for statistical inference on the existence of a FEM in the proximal and distal stimuli (see Results) and for the correlation between the statistics of the stimulus and human performance (estimated priors and biases). The similarity between the shapes of the FEM measured from the psychophysical task and from the statistics of the stimulus was measured in terms of Pearson correlation (Fig. 1F). A correlation of 1 means that the mappings are identical in shape, irrespective of potential shifts and scaling factors, whereas a correlation of 0 means that the two mappings are statistically independent. The correlation was only calculated for the frequency bands between 0.8 and 8 kHz, as above and below such frequencies the estimated priors and the measurements from statistics of the signals were estimated over different ranges of frequencies. To estimate the mean and the confidence interval of the correlation, we used a resampling procedure, whereby the correlation was iteratively calculated from the mean of a subset of one-fifth of the whole recordings ($n = 9, 962$), one-fifth of the HRTFs ($n = 9$), and one-fifth of the 10 estimated parameter sets ($n = 2$). This procedure was repeated 1,000 times.

The results of these analyses are reported in Fig. 1F. Note that despite the strong similarities between the shapes of the FEM in the statistics of the natural stimuli and in the estimated priors, the scale of the FEM in the statistics of the distal stimulus is much smaller than all of the other mappings (Fig. 1C and D). Something similar has been found in human vision, where the filtering properties of the retina seem to exaggerate the statistics of natural visual scenes (18). It would be a matter of future research to understand why the brain and the filtering of the outer ear encode the same FEM present in the environment on a different scale.

**ACKNOWLEDGMENTS.** The authors would like to thank the Cognitive Neuroscience research team in Bielefeld for precious support throughout this study, and J. Burge and J.M. Aché for insightful comments on a previous version of the manuscript. C.V.P. and M.O.E. were supported by the 7th Framework Programme European Projects “The Hand Embodied” (248587) and “Wearhp” (601165). This study is part of the research program of the Bernstein Center for Computational Neuroscience, Tübingen, funded by the German Federal Ministry of Education and Research (German Federal Ministry of Education and Research; Förderkennzeichen: 01GQ1002).

Supporting Information

Parise et al. 10.1073/pnas.1322705111

SI Text

Comparison Between the Localization Biases and the Statistics of Natural Sounds and Filtering Properties of the Outer Ear

Strictly speaking, the perceptual biases measured in the psychophysical task do not represent the internal mappings between frequency and elevation but the outcome of such mappings which we estimated with a Bayesian model. Nevertheless, to further prove the link between the measured statistics and observed behavior without relying on a model (which is necessarily based on a set of assumptions, which might as such be wrong), we also directly measured the correlation between the perceptual biases and the frequency–elevation mapping (FEM) in the environment and in the filtering properties of the outer ear. To do so, we used the frequency-dependent bias observed when participants were tilted by 90° (Fig. 1B, red lines), that is, when the head- and world-centered FEMs were made orthogonal, so that the elevation bias in head-centered coordinates should reflect the contribution of the head-centered FEM, whereas the azimuth bias (again in head-centered coordinates) should reflect the contribution of the world-centered FEM.

As in the previous section, the similarity between the shapes of the FEM measured from the psychophysical task (Fig. 1B, red lines) and from the statistics of the stimulus (Fig. 1C) was measured in term of Pearson correlation. The mean and confidence interval of the correlation were calculated by an iterative resampling procedure whereby the estimated priors were correlated to the FEM measured from the mean of a subset of one-fifth of the all recordings (n = 9,962), one-fifth of the head-related transfer functions (HRTFs) (n = 9), and one-fifth of the observers (n = 2). The procedure was repeated 1,000 times.

The correlation between the azimuth bias and the statistics of the environment was 0.76 [95% confidence interval (c.i.) = 0.56–0.87], the correlation between the azimuth bias and the FEM in the proximal stimulus was 0.89 (95% c.i. = 0.72–0.95). The correlation between the elevation bias and the statistics of the environment was 0.90 (95% c.i. = 0.83–0.96), and the correlation between the elevation bias and the FEM in the proximal stimulus was 0.78 (95% c.i. = 0.65–0.86). Notably, this pattern of correlation by and large confirms the findings based on the priors estimated using the Bayesian model and provides further converging evidence supporting the conclusion that the perceptual FEMs reflect the statistics of the proximal and distal stimuli.

Fig. S1. Average HRTF, obtained by averaging all 45 HRTFs of the CIPIC database. The black dots, representing the elevation with maximum transfer for each frequency, show a clear mapping between frequency and elevation. The FEM reported in Fig. 1C (Lower) was obtained by calculating for each individual HRTF the elevation with maximum transfer for each frequency (i.e., the black dots here), and then averaging the results across the 45 HRTFs.
**Fig. S2.** (A) Schematic representation of the experimental setup. (B) Representation of the three different body orientations. The vertical arrows represent the world-centered elevation; the tilted arrows represent the head-centered elevation. When the body of the participant is not tilted (Left), head- and world-centered elevation overlap, whereas when the body is tilted by 90° (Right), the head- and world-centered elevations are orthogonal. The gray grids represent the physical position of the speakers.

**Fig. S3.** Schematic illustration of the Bayesian model. The icons on the left represent the different orientations of the observers (in rows). The left column represents the likelihood function (the sensory information). The red dots represent the physical position of the stimulus \( s = (s_x, s_y) \). The second and the third columns represent the frequency-dependent priors on elevation in head- and world-centered coordinates, respectively. The last column on the right represents the posterior distribution; the red dot represents the physical position of the stimuli, whereas the green dot represents the maximum a posteriori, that is, the perceived position of the stimuli. Note how the perceived position is shifted away from the actual position as a function of both the frequency-dependent priors and body orientation. Colors indicate the reference frame of the priors (magenta = head-centered; cyan = world-centered). This figure represents the case of the localization of a 4.5–8-kHz band-pass auditory stimulus coming from the bottom-right speaker \( s_{wc} = [15, -15]^\circ \). Frequency-independent distortions of perceived auditory space (Modeling) are not represented.
Movie S1. Auditory stimuli used in the sound localization experiment. To better appreciate how perceived spatial elevation changes as a function of the spectra of the stimuli, we recommend playing the sounds using loudspeakers (not headphones), and listening with the eyes closed.

Movie S1