Natural auditory scene statistics shapes human spatial hearing

Cesare V. Parise[a,b,1], Katharina Knorre[b], and Marc O. Ernst[a,b,1]

[a]Max Planck Institute for Biological Cybernetics and Bernstein Center for Computational Neuroscience, 72076 Tübingen, Germany; and [b]Cognitive 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.

frequency–elevation mapping | head-related transfer function | Bayesian modeling | cross-modal correspondence

The spatial connotation of auditory pitch is a universal hallmark of human cognition. High pitch is consistently mapped to high positions in space in a wide range of cognitive (1–3), perceptual (4–6), attentional (7–12), and linguistic functions (13), and the same mapping has been consistently found in infants as young as 4 mo of age (14). In spatial hearing, the perceived spatial elevation of pure tones is almost fully determined by frequency—rather than physical location—in a very systematic fashion [i.e., the Pratt effect (4, 5)]. Likewise, most natural languages use the same spatial attributes, high and low, to describe pitch (13), and throughout the history of musical notation high notes have been represented high on the staff. However, a comprehensive account for the origins of the spatial connotation of auditory pitch to date is still missing. More than a century ago, Stumpf (13) suggested that it might stem from the statistics of natural auditory scenes, but this hypothesis has never been tested. This is a major omission, as the frequency–elevation mapping often leads to remarkable inaccuracies in sound localization (4, 5) and can even trigger visual illusions (6), but it can also lead to benefits such as reduced reaction times or improved detection performance (7–12).

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 medial 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. 1*C*, *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. 1*C*, *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.

This article is a PNAS Direct Submission.

[1]To 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.

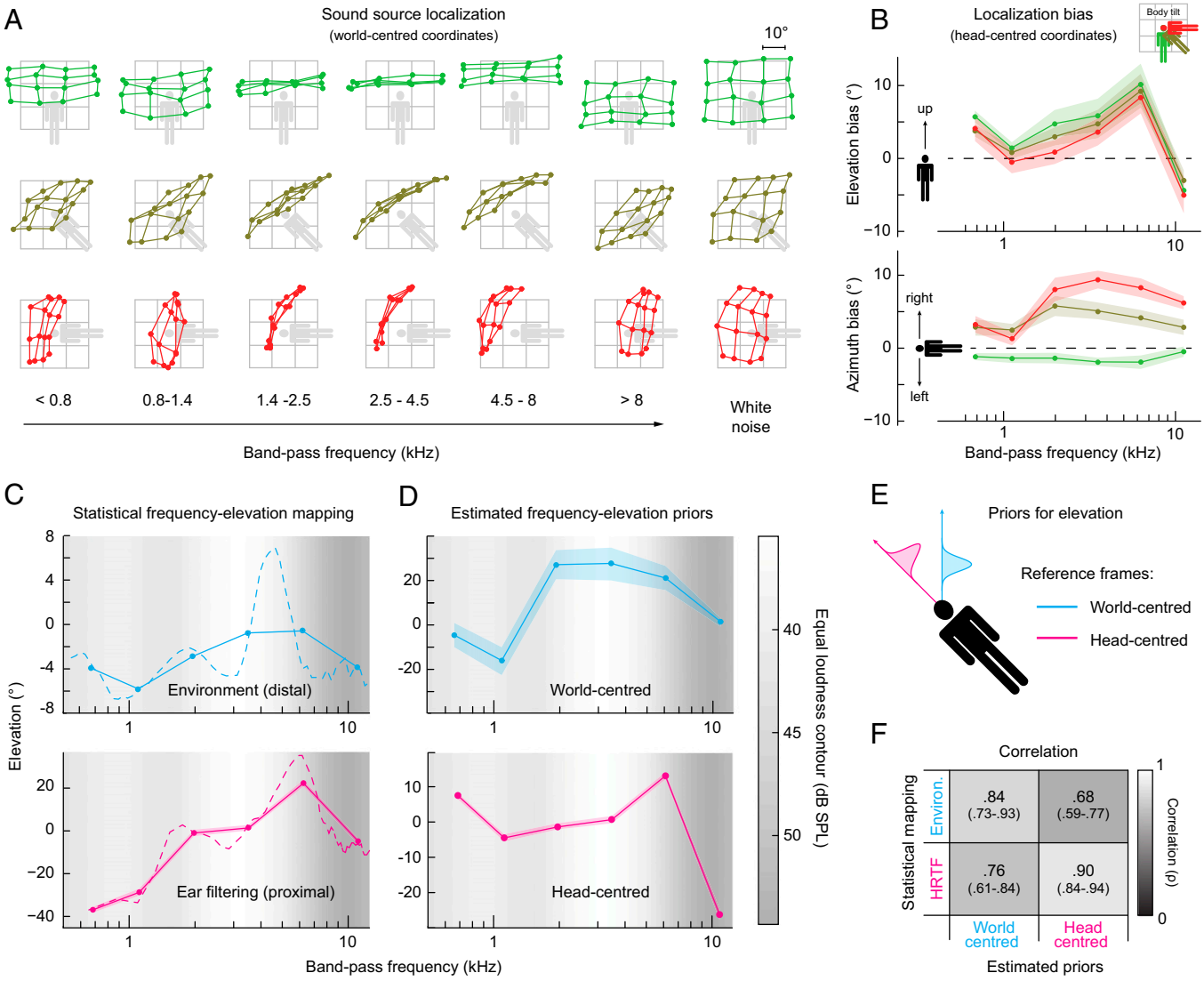

Fig. 1. (*A*) Average endpoint of pointing responses for the various frequency bands (column) and body tilts (row). The filled points correspond to the average responses; the thin gray grid represents the actual position of the stimuli. Colors represent tilt (green = 0°, brown = 45°, red = 90°). (*B*) Frequency-dependent bias (±SEM) in sound localization in head-center elevation (*Upper*) and azimuth (*Lower*). The magnitude of the frequency-dependent elevation biases was only mildly affected by body tilt, reflecting the contribution of a frequency–elevation mapping encoded in head-centered coordinates. The frequency-dependent azimuth biases increase in magnitude with increasing body-tilt angle reflecting the contribution of a frequency–elevation mapping encoded in world-centered coordinates. (*C*) Statistical mapping (±SEM) between frequency and elevation recorded in the environment (*Upper*) and measured from the HRTFs (*Lower*). The dashed lines represent the frequency–elevation mapping using nonbinned data. (*D*) Shapes of the estimated priors coding for the frequency–elevation mapping in world-centered (*Upper*) and head-centered coordinates (*Lower*). Lightness within the panels represents the equal loudness contour (International Organization for Standardization 226:2003): lighter gray represents higher sensitivity. (*E*) Schematic 1D representation of the model illustrating the head- (magenta) and world-centered priors (cyan). (*F*) Correlation (and 95% confidence intervals) between the estimated priors and the frequency–elevation mapping measured from the environment and the HRTFs. (*C–F*) Colors indicate the reference frame (magenta = head-centered; cyan = world-centered).

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 × 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. 1*A*, *Right*). Conversely, sound source localization was strongly biased when the stimulus consisted of narrowband noise (Fig. 1*A*). 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 elevation was almost entirely determined in a very consistent way by the frequency of the signals (Fig. 1*A*, *Center*). Notably, such biases showed a clear mapping between frequency and elevation (Fig. 1*B*), 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 ($\rho = 0.76$ for world-centered biases with distal stimulus and $\rho = 0.78$ 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 developing 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 $p_f(s)$ 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. 1*E*). 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. 1*D*) reveal a strong similarity with the frequency-dependent biases measured from the responses of the participants (Fig. 1*B*, 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 the 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. 1*F*). 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). None of them, however, directly 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 proprioceptive 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 laced 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 sounds coming from the midsagittal plane (Fig. S1).

Psychophysical Task. Ten healthy observers with normal audition and normal or corrected-to-normal vision took part in the experiment (six females, mean age 25 y, range 21–33). All of them were students or employees at the University of Bielefeld and provided written informed consent before participating. The study was conducted in accordance with the Declaration of Helsinki and had ethical approval from the ethics committee of the University of Tübingen.

Observer's head was fixed 130 cm away from a sound-transparent projection screen (220 × 164 cm) mounted in front of a set of 16 speakers (Fig. S2). On each trial, one of the speakers played a 300-ms band-pass noise (band-pass kHz: <0.8; 0.8–1.4; 1.4–2.5; 2.5–4.5; 4.5–8; >8; or white noise; Movie S1). Participants were instructed to indicate where they have heard the stimulus come from, using a cursor projected on the screen in front of the speakers. Localization was visually guided (closed loop) and temporally unconstrained. When participants were happy with the position of the cursor, they had to press on the touchpad to submit their response. In different blocks (in a counterbalanced order), we tilted participants' bodies (0°, 45°, or 90° counterclockwise; Fig. S2) with respect to the gravitational vertical using custom-built chairs that maintained the line of sight aligned with the center of the screen, without covering the ears. Each combination of stimulus frequency, tilt, and spatial location was repeated 4 times (1,344 trials per participants).

Before running the localization task, the auditory stimuli were perceptually equalized in loudness using the method of adjustment to prevent any perceived loudness differences to affect our results. That is, we used the white noise stimulus as the standard, and participants adjusted the intensity of each band-pass stimulus until the loudness of the band-pass stimuli perceptually matched the standard. Each band-pass stimulus was adjusted six times, and we repeated the procedure with four participants. The gain factor used to equalize each stimulus was determined as the median value of all adjustments.

The experiment was conducted in a dark anechoic chamber, and controlled by a custom-built software based on the Psychtoolbox (34). Participants were tested in three sessions taking place on three consecutive days. Different body-tilt conditions were tested in separate blocks (4 blocks/d), with the order of the blocks counterbalanced within and across participants. Within each block, sounds with different frequencies and positions were presented in a pseudorandom fashion.

For each orientation and frequency, the localization bias was calculated separately for head-centered elevation and azimuth as the grand mean of the responses for each participant. The elevation bias (Fig. 1*B*, *Upper*) showed a main effect of frequency [$F(5,45) = 11.564$; $P < 0.001$], without significant effects of tilt [$F(2,18) = 1.157$, $P = 0.337$] or interactions [$F(10,90) = 1.313$; $P = 0.235$]. The azimuth bias (Fig. 1*B*, *Lower*) showed a main effect of frequency [$F(5,45) = 4.074$; $P = 0.004$], 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 $s = (s_x, s_y)$. 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 $\hat{s} = (\hat{s}_x, \hat{s}_y)$ derived from the physical source of a sound with frequency f is unbiased but noisy, with some Gaussian noise $\sigma = (\sigma_{f,x}, \sigma_{f,y})$ added independently to each spatial dimension i ($\hat{s}_i = s_i + \sigma_{f,i}$), the likelihood distribution $p_f(\hat{s}|s)$ for the spatial location of the sound source is a 2D Gaussian:

$$p_f(\hat{s}|s) = N\left(s_\theta, \Sigma_{f,\theta}\right),$$

with mean $s_\theta = (s_x, s_y) \cdot R_\theta$ and covariance matrix $\Sigma_{f,\theta} = \begin{pmatrix} \sigma_{f,x}^2 & 0 \\ 0 & \sigma_{f,y}^2 \end{pmatrix} \cdot R_\theta$ (Fig. S3, *Left*). Assuming the likelihood to be encoded in head-centered coordinates, R_θ is a rotation matrix that rotates the axes according to the orientation of the body with respect to gravitational vertical (θ).

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 $p_{hc,f}(s)$ for the location $s_{hc,f}$ of a sound with frequency f is defined as a 2D Gaussian:

$$p_{hc,f}(s) = N\left(s_{hc,f,\theta}, \Sigma_{hc,f,\theta}\right),$$

with mean $s_{hc,f,\theta} = (0, s_{hc,f,y}) \cdot R_\theta$ and covariance matrix $\Sigma_{hc,f,\theta} = \begin{pmatrix} \infty & 0 \\ 0 & \sigma_{hc,f,y}^2 \end{pmatrix} \cdot R_\theta$ (Fig. S3, second column). The mean $s_{hc,f,y}$ represents the expected spatial elevation and the variance $\sigma_{hc,f,y}^2$ 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 ∞ variance (i.e., the prior is uninformative with respect to the head-centered azimuth).

In a similar fashion, the world-centered prior distribution $p_{wc,f}(s)$ for the location $s_{wc,f}$ of a sound with frequency f is defined as a 2D Gaussian:

$$p_{wc,f}(s) = N\left(s_{wc,f}, \Sigma_{wc,f}\right),$$

with mean $s_{wc,f} = (0, s_{wc,f,y})$ and covariance matrix $\Sigma_{wc,f} = \begin{pmatrix} \infty & 0 \\ 0 & \sigma_{wc,f,y}^2 \end{pmatrix}$ (Fig. S3, third column). The mean $s_{wc,f,y}$ represents the expected spatial elevation and the variance $\sigma_{wc,f,y}^2$ 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 Bayes theorem, according to which the posterior $p_f(s|\hat{s})$ (Fig. S3, *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):

$$p_f(s|\hat{s}) \propto p_{hc,f}(s) \cdot p_{wc,f}(s) \cdot p_f(\hat{s}|s).$$

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 $p_f(s|\hat{s})$. 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 $s_{hc,f,y}$ and $s_{wc,f,y}$ 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 $\sigma^2_{hc,f,y}$, and $\sigma^2_{wc,f,y}$. We also fitted the covariance matrix $\Sigma_{f,\theta}$ 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 $\sigma^2_{f,y}$ for each frequency band tested). Overall, the model had 21 free parameters fitted over 13,440 trials, that is, 640 trials per parameter.

Additionally, 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., the dots in Fig. 1*A*). The fitting was based on an unconstrained nonlinear optimization procedure (fminsearch, Matlab). Parameters were fitted using a leave-one-out Jackknife procedure, consisting of iteratively estimating the parameters of the pointing responses excluding one participant at a time. The results in Fig. 1*D* 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 parameters 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 (~2°; Fig. 1*B Lower*, green line) compared with elevation biases (~15°, Fig. 1*B Upper*, 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. 1*F*). 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. 1*F*. 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. 1 *C* and *D*). Something similar has been found in human vision, where the filtering properties of the eye 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. Ache 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 "Wearhap" (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).

1. Douglas KM, Bilkey DK (2007) Amusia is associated with deficits in spatial processing. *Nat Neurosci* 10(7):915–921.
2. Rusconi E, Kwan B, Giordano BL, Umiltà C, Butterworth B (2006) Spatial representation of pitch height: The SMARC effect. *Cognition* 99(2):113–129.
3. Dolscheid S, Shayan S, Majid A, Casasanto D (2013) The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychol Sci* 24(5):613–621.
4. Pratt CC (1930) The spatial character of high and low tones. *J Exp Psychol* 13(3): 278–285.
5. Roffler SK, Butler RA (1968) Factors that influence the localization of sound in the vertical plane. *J Acoust Soc Am* 43(6):1255–1259.
6. Maeda F, Kanai R, Shimojo S (2004) Changing pitch induced visual motion illusion. *Curr Biol* 14(23):R990–R991.
7. Chiou R, Rich AN (2012) Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception* 41(3):339–353.
8. Melara RD, O'Brien TP (1990) Effects of cuing on cross-modal congruity. *J Mem Lang* 29(6):655–686.
9. Melara RD, O'Brien TP (1987) Interaction between synesthetically corresponding dimensions. *J Exp Psychol Gen* 116(4):323–336.
10. Bernstein IH, Edelstein BA (1971) Effects of some variations in auditory input upon visual choice reaction time. *J Exp Psychol* 87(2):241–247.
11. Mossbridge JA, Grabowecky M, Suzuki S (2011) Changes in auditory frequency guide visual-spatial attention. *Cognition* 121(1):133–139.
12. Evans KK, Treisman A (2010) Natural cross-modal mappings between visual and auditory features. *J Vis* 10(1):1–12.
13. Stumpf K (1883) *Tonpsychologie* (Hirzel, Leipzig, Germany).
14. Walker P, et al. (2010) Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychol Sci* 21(1):21–25.
15. Batteau DW (1967) The role of the pinna in human localization. *Proc R Soc Lond B Biol Sci* 168(11):158–180.
16. Iida K, Itoh M, Itagaki A, Morimoto M (2007) Median plane localization using a parametric model of the head-related transfer function based on spectral cues. *Appl Acoust* 68(8):835–850.
17. Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The CIPIC HRTF database. *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (IEEE, New Paltz, NY), pp 99–102.
18. Burge J, Geisler WS (2011) Optimal defocus estimation in individual natural images. *Proc Natl Acad Sci USA* 108(40):16849–16854.
19. Parise CV, Spence C, Ernst MO (2012) When correlation implies causation in multisensory integration. *Curr Biol* 22(1):46–49.
20. Suzuki Y, Takeshima H (2004) Equal-loudness-level contours for pure tones. *J Acoust Soc Am* 116(2):918–933.
21. Carlille S, Pralong D (1994) The location-dependent nature of perceptually salient features of the human head-related transfer functions. *J Acoust Soc Am* 95(6): 3445–3459.
22. Butler RA (1987) An analysis of the monaural displacement of sound in space. *Percept Psychophys* 41(1):1–7.
23. Adams WJ, Graf EW, Ernst MO (2004) Experience can change the 'light-from-above' prior. *Nat Neurosci* 7(10):1057–1058.
24. Weiss Y, Simoncelli EP, Adelson EH (2002) Motion illusions as optimal percepts. *Nat Neurosci* 5(6):598–604.
25. Tassinari H, Hudson TE, Landy MS (2006) Combining priors and noisy visual cues in a rapid pointing task. *J Neurosci* 26(40):10154–10163.
26. Zhang R, Kwon O-S, Tadin D (2013) Illusory movement of stationary stimuli in the visual periphery: Evidence for a strong centrifugal prior in motion processing. *J Neurosci* 33(10):4415–4423.
27. Girshick AR, Landy MS, Simoncelli EP (2011) Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nat Neurosci* 14(7):926–932.
28. Rogers ME, Butler RA (1992) The linkage between stimulus frequency and covert peak areas as it relates to monaural localization. *Percept Psychophys* 52(5):536–546.
29. Goossens HH, van Opstal AJ (1999) Influence of head position on the spatial representation of acoustic targets. *J Neurophysiol* 81(6):2720–2736.
30. Cabrera D, Ferguson S, Tilley S, Morimoto M (2005) Recent studies on the effect of signal frequency on auditory vertical localization. *Proceedings of International Conference on Auditory Display* (ICAD, Limerick, Ireland).
31. Sadaghiani S, Maier JX, Noppeney U (2009) Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. *J Neurosci* 29(20): 6490–6499.
32. Ramachandran V, Hubbard E (2001) Synaesthesia: A window into perception, thought and language. *J Conscious Stud* 8(12):3–34.
33. Parise C, Spence C Audiovisual cross-modal correspondences in the general population. *Oxford Handbook of Synaesthesia*, eds Simner J, Hubbard EM (Oxford Univ Press, Oxford, UK).
34. Kleiner M, Brainard D, Pelli D (2007) What's new in Psychtoolbox-3. *Perception* 36(14): 1–16.