Decision making in auditory externalization perception

Robert Baumgartner[1] and Piotr Majdak

Acoustics Research Institute, Austrian Academy of Sciences

Abstract

Under natural listening conditions, humans perceptually link sounds with sound sources in their environment. This so-called auditory externalization is often distorted when sound is produced via hearing devices such as headphones or hearing aids, resulting in sources being perceptually localized inside the head. Psychoacoustic studies suggest a mixed role of various cues in the externalization process. As such, we developed a model framework able to probe the contribution of those cues and test the perceptual decision strategy of listeners performing various externalization tasks. Following previous investigations in vision, we contrasted different strategies including a static combination of cues and three dynamic selection approaches. Our results show that the decision strategy following a static combination of cues best explains externalization ratings. Further, our analysis indicates that listeners mainly base their perceptual decision on both monaural and interaural spectral shapes, whereas the contribution of other cues was not significant. The proposed model predicts auditory externalization under the most fragile condition of static listening and has the potential to significantly advance the future development of spatially-aware hearing technologies by helping to identify perceptually relevant auditory distortions.

1 Corresponding author: robert.baumgartner@oeaw.ac.at

Introduction

For a successful interaction with the world, recent and influential theories suggest that the brain's primary objective is to infer the causes of its sensory input by creating an internal model of the environment predicting the incoming cues[1,2]. These theories convince by examples coming from various areas of sensory perception and higher-order cognitive functions[3]. In the context of auditory perception, temporal predictions are often based on short-term regularities[4,5] while long-term expectations are often crucial for spatial predictions[6,7]. The underlying perceptual decision making is usually based on multiple, simultaneously accessible cues. Various decision strategies such as weighted sums or cue selection have been put forward, mainly investigated in visual search tasks[8]. The goal of this study is to transfer these concepts to a particularly puzzling perceptual phenomenon within the auditory modality, namely, the collapse of perceptual externalization, which denotes the lack to associate sensations with external objects[9]. This phenomenon constitutes a critical aspect of spatial hearing because it can be easily disrupted when listening to sounds, e.g., via headphones or other hearing devices, that do not accurately represent the spatial properties of the listener's natural acoustic exposure[10–16]. Given the high prevalence of daily hearing device usage and the rapid technological advance in augmented reality systems, this topic is timely and of high social relevance.

The spatial properties of the sound arriving at the two ears are manifold and so are the cues. Many studies aimed to identify the cues affecting the perceptual externalization of sound sources[10,11,14–20]. These cues can be grouped into two categories (Tab. 1): one category poses monaural cues, in line with studies on sound localization beyond the horizontal plane suggesting that the auditory system processes spectral cues within monaural pathways[21–24], and the other category poses interaural cues, such as an interaural time difference (ITD) or interaural intensity difference (IID), which are well known to be crucial for sound localization within the horizontal plane[25,26] but may also affect spatial hearing in other dimensions[15,27]. Besides the multitude of potential cues, a general problem is that many of those cues co-vary and that it has never been investigated which perceptual decision strategy may underlie the listeners' response behavior in those psychoacoustic tasks.

Property	Monaural (M_)	Interaural (I_)
Spectral Shape (_SS)	MSS	ISS
Spectral Standard Deviation (_SSD)	MSSD	ISSD
Coherence (_C)	-	IC
Time-Intensity Trading (_TIT)	-	ITIT
(Broadband) Intensity (_I)	MI	-

Tab. 1 | Potential auditory externalization cues considered for the evaluation.

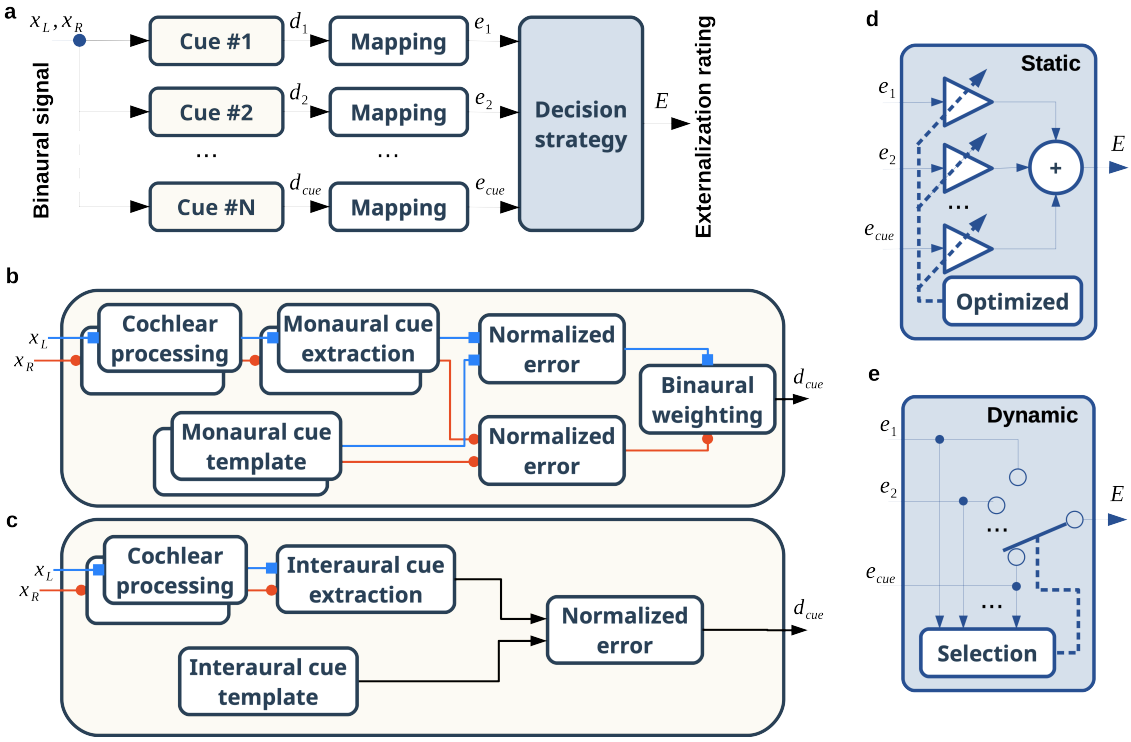

Fig. 1 | General structure of the sound externalization framework. a*, Processing of the binaural signal* x_L, x_R *resulting in a cue-based error* d_{cue}*, a stage mapping the cue-based error to a normalized externalization rating* e_{cue}*, and a decision stage combining the cue-specific ratings to a final externalization rating* E*.* ***b****, Processing of the binaural signal to calculate a single-cue error based on monaural cue templates.* ***c****, Processing based on interaural cue templates.* ***d****, Decision stage based on a static weighted-sum model (WSM).* ***e****, Dynamic decision stage based on the selection of a cue-specific rating. Minimalist (LTA), conformist (MTA), and perfectionist (WTA) approaches were considered for selection (see text).*

While the cue-based encoding can be considered as trivial, probably already happening before reaching the auditory cortex[28], more complex structures are required to combine the cues to form a decision stage yielding the final percept of externalization. Cortical representations contain an integrated code of sound source location[29–31], but also retain independent information about spatial cues[32–35], allowing the system to decide how likely they would have arisen from one or multiple sources[35], or would have arisen from a source located outside the head, i.e., manifesting as a final percept of externalization. The decision strategy used by the auditory system to form externalization is unclear yet. Here, we tested two general types of decision strategies.

The first potential strategy follows the idea that based on exposure statistics our brain has established a rather fixed weighing of spatial auditory cues in order to perceive an externalized auditory event – independent of the listening context. This can be represented as a static weighted-sum model (WSM, see Fig. 1d) that scales the predictions obtained for a single cue with weights adjusted based on long-term listening experience. The WSM is often used to merge neural processing of multiple cues, with examples ranging from a lower, peripheral level such as neural binaural processing[36] over higher cortical levels integrating vis-

ual orientation cues[37] to even more abstract levels such as the representation of person identities[38]. Once the weights are settled, the WSM has the advantage of being static with no need for further adaptation during the externalization process.

The second potential decision strategy is of selective nature and has been favored in the context of visual search[39,40] or audio-visual dominance[41]. The idea is that, depending on the incoming stimulus, our brain selects the one of the externalization-related cues that fulfills (or breaks) the listener's prior expectations (see Fig. 1e). We considered three variants of such a selection strategy. First, a minimalist approach would assume that a sound can be externalized if at least one cue yields externalized perception. We implemented this by a winner-takes-all (WTA) strategy that considers the largest externalization rating (across all single-cue ratings) as the total externalization rating. Second, the contrary is the perfectionist approach, which requires all of the cues to provide an externalized percept – a single cue breaking the expectation will dominate perception. This is implemented in the loser-takes-all (LTA) strategy, which considers the smallest externalization rating (across all single-cue ratings) as the total rating. Third, the conformist approach is intermediate to these two extremes and prefers to use an average of single-cue-based externalization ratings. We implemented this as a median-takes-all (MTA) strategy, which considers the median across all single-cue-based ratings as the total externalization rating. Note that this also assumes that the brain selects the cue-specific prediction being most consistent to all the others, i.e., represents a system striving for consistency.

Here, we propose a model framework allowing us to disentangle those decision strategies potentially involved in the multi-cue-based task of sound externalization. The framework is shown in Fig. 1a) and consists of 1) cue-processing stages, each of them encoding the incoming binaural signal $[x_L, x_R]$ into a single-cue error metric, d_{cue}, 2) stages mapping that single-cue error to a normalized externalization rating, e_{cue}, and 3) a decision stage combining the single-cue based ratings into the final externalization rating, E.

We used that framework to simulate five experiments from representative auditory externalization studies[11,14–16], all focusing on the effects of spectral distortions under static listening conditions, but differing in manipulation methods, side, and bandwidth. The listener's expectations for perceptual inference are represented as internal templates and used for comparison with incoming cue-specific information, distinguishing between processing monaural and interaural cues (Fig. 1b,c). The decision strategies were not only compared for all those cues, but also for a successively reduced set of cues, in order to address the potential redundancy between them. Our results reveal that auditory externalization judgments under situations with constant reverberation underlie a fixed weighting scheme between monaural and interaural spectral cues, rather than a dynamic selection of cues for sound externalization.

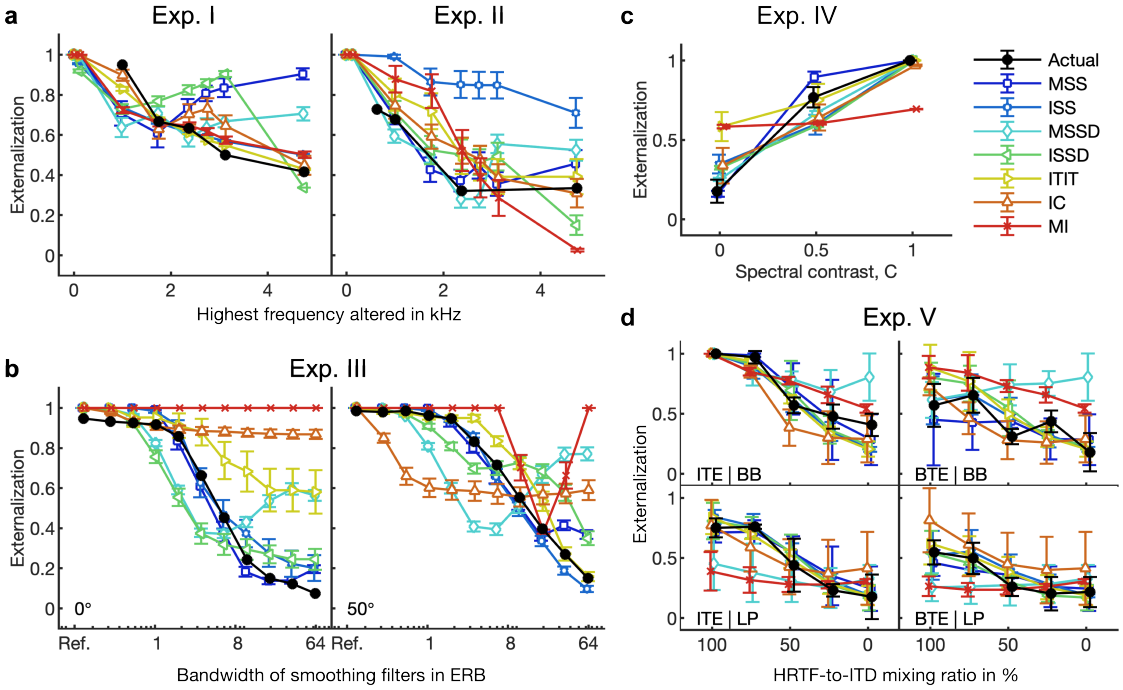

Fig. 2 | Externalization ratings: actual data from psychoacoustic experiments *(closed circles)* ***and predictions from the single-cue models*** *(open symbols).* ***a****, Effects of low-frequency modifications tested by Hartmann and Wittenberg [14]. Exp. I: IID set to zero (bilateral average magnitude); actual data from their Fig. 7. Exp. II: ipsilateral magnitudes flattened (IID compensated by contralateral magnitudes); actual data from their Fig. 8.* ***b****, Exp. III: effect of spectral smoothing of low-frequency sounds presented from various azimuths (left: 0°; right: 50°); actual data represents direct-sound condition from Hassager et al. [15].* ***c****, Exp. IV: effect of spectral smoothing in high frequencies as a function of spectral contrast (C=1: natural listener-specific spectral coloration; C=0: flat spectra); actual data calculated from the paired-comparison data from Baumgartner et al. [16].* ***d****, Exp. V: effects of stimulus bandwidth and microphone casing for various mixes between simulations based on listener-specific binaural room impulse responses (100%) and time-delay stereophony (0%); actual data from Boyd et al. [11]. ITE: in-the-ear casing; BTE: behind-the-ear casing; BB: broadband stimulus; LP: low-pass filtered at 6.5 kHz; HRTF: head-related transfer function (spectral coloration due to directional acoustic filtering).*

Results

Individual cues: large differences across experimental conditions

We first investigated the ability of each individual cue to predict the effects of signal manipulations on auditory externalization as tested in five previous headphone experiments[11,14–16]. Figure 2 shows the predicted externalization ratings together with the actual ratings replotted from the corresponding experiments with normalized rating scales. The model's normalization stage (Fig. 1b,c) requires a listener-specific sensitivity[42,43], which was fitted to optimally predict the externalization ratings separately for each experiment and cue (Tab. 2).

Exp. I: Effect of IID modifications at low frequencies

The first experiment presented listeners with harmonic tone complexes from a mid-left position[14] (Fig. 2a, left panel). The magnitudes of all harmonics up to a certain harmonic were set to the interaural average, effectively removing IIDs up to that harmonic's frequency (see Methods for further details).

Exp.		MSS	ISS	MSSD	ISSD	ITIT	IC	MI
I	S_cue	0.10	0.88	0.02	0.62	0.80	0.01	0.03
	Error	0.29	0.11	0.21	0.23	**0.06**	0.07	0.11
II	S_cue	0.15	0.01	0.02	0.01	0.02	0.00	0.00
	Error	**0.05**	0.21	0.10	0.17	0.15	0.12	0.20
III	S_cue	0.24	0.33	0.13	0.17	0.17	0.05	0.01
	Error	**0.18**	0.25	0.42	0.29	0.49	0.54	0.61
IV	S_cue	0.09	0.42	0.39	0.50	0.03	0.02	0.12
	Error	**0.19**	0.28	0.19	0.23	0.41	0.25	0.36
V	S_cue	0.18	0.45	0.05	0.24	0.53	0.37	0.11
	Error	0.20	**0.16**	0.31	0.18	0.19	0.22	0.25
W. Avg.	S_cue	0.16	0.60	0.09	0.27	0.67	0.03	0.04
Mean	Error	**0.18**	0.20	0.25	0.22	0.26	0.24	0.31

Tab. 2 | Cue-specific sensitivities and RMS prediction errors *revealed as best explaining the actual externalization ratings. Per experiment, the smallest error indicates the most informative cue and is emphasized in bold.*

The monaural spectral cues (MSS and MSSD) showed a non-monotonic relationship between the modified frequency range and predicted ratings, which largely diverged from the actual ratings especially for the largest frequency ranges and resulted in large prediction errors (Tab. 2, Exp. I). Further inspection showed that the modification of interaural averaging only marginally affected those monaural cues because at low frequencies the complex acoustic filtering of the pinnae is negligible and thus monaural spectral shapes are quite similar at both ears. In contrast, the broadband monaural cue (MI) was able to predict the actual ratings surprisingly well because the IID removal caused a small but systematic increase in sound intensity, being in line with the systematic decrease in externalization ratings.

Most interaural cues (ISS, ITIT, and IC) were able to predict the actual ratings very well. Differences in one interaural cue (IC) were, however, very small and thus required a very steep mapping function in order to become informative (indicated by a low sensitivity in Tab. 2, Exp. I). Predictions based on interaural spectral contrasts (ISSD) failed presumably because the evaluation of the standard deviation metric became unstable for spectral IID distributions strongly deviating from a normal distribution. Overall, the broadband interaural cue (ITIT) performed best with the minimum RMS prediction error of 0.06.

Exp. II: Effect of monaural modifications at low frequencies

The second experiment tested the same stimulus as in Exp. II, but instead of removing interaural differences, the ipsilateral magnitude spectrum was flattened up to a certain harmonic while the contralateral magnitudes were shifted in parallel [14]. This procedure effectively maintained the original IIDs but changed the monaural spectral profiles and by doing so also gradually distorted perceived externalization (Fig. 2a, right panel).

In contrast to Exp. I, the predictions based on monaural spectral cues (MSS and MSSD) reflected the decrease in actual ratings very well, as indicated by RMS prediction error as low as 0.05 (for MSS). As expected, predictions based on interaural cues failed because the degradation induced by flattening the ipsilateral spectrum was designed to maintain the original interaural features.

Exp. III: Effect of spectral smoothing at low frequencies

The third experiment investigated the effect of spectral smoothing on the auditory externalization of noise stimuli band-limited up to 6 kHz and presented sounds either from the front (0° azimuth) or from the left (50° azimuth) [15]. Spectral smoothing was implemented by applying filters of various equivalent rectangular bandwidths (ERBs), with broader filters removing more spectral details. This procedure caused a degradation in perceived externalization, which was more gradual for the lateral as compared to the frontal direction (Fig. 2b).

Again, only some of the cues were able to predict the systematic trend of externalization degradation. The stimulus manipulation affected some cues only hardly (ITIT, IC, and MI) or in a non-systematic manner (MSSD) whereas both monaural and interaural spectral shape cues (MSS and ISS) yielded much better predictions (errors of 0.18 and 0.25, respectively). Both cues yielded predictions consistent with the actual results in that they were insensitive to spectral smoothing below one ERB. It appears noteworthy that the monaural (MSS) outperformed the interaural (ISS) cue in this particular experiment, which has been used to promote an earlier modeling approach only based on the interaural cue (ISS) [15]. Moreover, neither cue was able to explain why actual ratings were slightly below the maximum for the frontal reference sound (and bandwidths ≤ 1 ERB). Additional factors not represented in the model seem to be necessary to explain this lateral dependence of reference externalization.

Exp. IV: Effect of spectral smoothing at high frequencies

The fourth experiment also tested the effect of spectral smoothing [16], however, focusing on the high-frequency range from 1 kHz to 16 kHz, where the pinnae induce the most significant directional spectral variations. There, listeners were presented to noise stimuli being spectrally flat (C = 0), spectrally shaped according to their individual acoustics (C = 1), or something in between (C = 0.5). Flattening the spectral contrast concordantly degraded externalization perception (Figure 2c).

Broadband cues (ITIT and MI) were hardly affected by this method and were not able to explain the actual results. Best predictions were obtained with monaural spectral cues (MSS and MSSD), followed by the interaural cues (ISS, ISSD, and IC). Hence, the template comparison based on monaural spectral shapes (MSS) allowed accurate predictions of the effects of spectral smoothing independent of both the particular smoothing methods and frequency ranges considered in Exps. III and IV.

Exp. V: Bandwidth limitation and sensor displacement

The last experiment presented broadband speech samples from a mid-left position and compared externalization ratings for different sensor positions, with sensors placed in the ear (ITE) distorting natural spectral cues less than behind the ear (BTE), stimulus bandwidths, and mixing ratios with stereophonic recordings providing only an ITD but no IID. The predictions showed much variability (Fig. 2d) across cues and conditions. The broadband BTE con-

dition caused the most distinct spectral modifications and was particularly informative about the predictive power of considered cues. Most cues were able to follow the actual ratings quite well (except MSSD and MI). On average across conditions, expectations based on interaural spectral templates (ISS, prediction error: 0.16) seem to have been most relevant to the listeners in this experiment.

Single-cue summary

The overall picture of prediction errors suggests that there is no single cue able to predict externalization for all experiments but that there is a small set of cues which listeners base their externalization ratings on. These cues correspond to the evaluation of monaural and interaural spectral features (MSS and ISS, respectively) as well as the broadband interaural disparities between ITD and IID (ITIT), respectively. This is consistent with previous observations showing that both interaural and monaural changes affect externalization.

The monaural cue (MSS) yielded best predictions in three out of five experiments as well as on average across the five experiments (prediction error of 0.18). The evaluation of this cue focuses mainly on positive gradients in the spectral profile as motivated by physiological findings in the cat dorsal cochlear nucleus [44]. To assure that our results were not confounded by a suboptimal choice of gradient sign, we repeated all simulations also with two different model configurations either favoring negative or considering both positive and negative gradients [technically, we implemented this by either switching the signs or removing the $\pm\pi/2$ shifts in Eq. (6), respectively]. We found that the average prediction error increased to 0.24 (for negative gradients only) and to 0.23 (for both negative and positive gradients), consolidating our choice of focussing the evaluation on positive spectral gradients.

The sensitivity parameters used to scale the mapping function from cue-specific error metrics to externalization scores were optimized separately for every cue and experiment. The separate optimization was reasoned by the limited quantitative comparability of subjective externalization ratings because of various factors such as differently trained subjects, different contexts, different experimental procedures (especially with respect to presenting reference stimuli), and many other methodological differences. Nevertheless, the optimization procedure yielded somewhat similar sensitivity parameters for the same cue across experiments whenever that cue was informative as indicated by small errors.

In summary, we found considerable variance in cue-specific predictions. Predictions based on a single monaural cue (MSS) turned out to be most potent in explaining the set of the considered experiments. However, the best cue clearly depended on the experiment and future predictions based on that particular would fail in situations similar to Exp. I. This indicates that the auditory system does not simply rely on a single cue and rises the question of which perceptual decision strategy is used to combine cues in different listening situations.

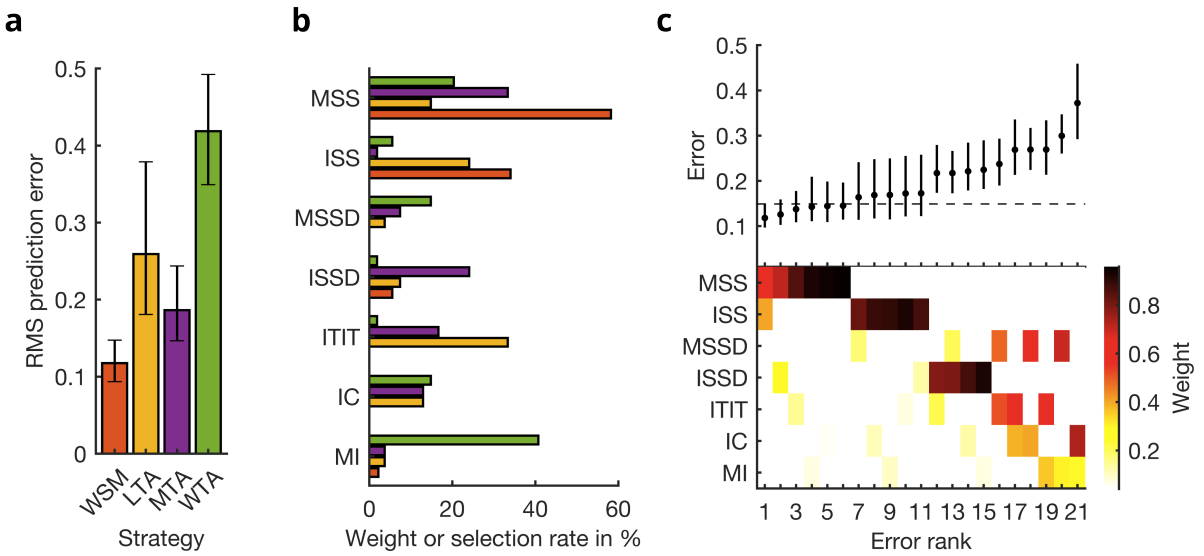

Fig. 3 | Prediction errors for different decision strategies and cue combinations show that static combination (WSM) based on monaural and interaural spectral shape cues (MSS, ISS) performs best. a, *Prediction errors for different strategies and pooled experimental data. Error bars denote 95% confidence intervals estimated via boot strapping. WSM: weighted-sum model; L/M/WTA: looser/median/winner takes all.* ***b,*** *Individual cue contributions. Cue abbreviations defined in Tab. 1.* ***c, Top:*** *Prediction errors for pairwise combinations of considered cues. Dashed line shows the border of significant difference to the best pair (MSS and ISS).* ***Bottom****: Considered cue pairs with their respective weights.*

Decision strategy: static weighting outperforms dynamic selection

We aimed to determine which decision strategy best explains externalization ratings across all experiments without an a-priori knowledge about the context. In the static decision strategy (WSM), the simulated listener derives the final externalization rating from a linearly weighted combination of single-cue predictions (Fig. 1b). The weights were obtained from an optimization procedure minimizing the prediction errors. The dynamic decision strategies, that is, WTA, MTA, and LTA, selected the largest, median, and smallest predicted externalization rating, respectively, obtained across the considered cues. By doing so, the cue weights become binary: the cue providing the largest, median, or smallest externalization rating is considered as the total rating, all other cues are ignored.

Overall, the static cue combination (WSM) outperformed the dynamic strategies, as indicated by significantly lower prediction errors (Fig. 3a). The static strategy (WSM) predicted the externalization ratings based on the weighted sum of mainly two spectral cues: a monaural cue (MSS) weighted by about two thirds and an interaural cue (ISS) weighted by the remaining third, while the other cues contributed marginally only (Fig. 3b). The contribution of the interaural cue is in line with previous findings in the context of Exp. III, for which only interaural cues have previously been used to explain externalization ratings[15]. Across other experiments, as shown here, the monaural cue becomes essential.

Within the dynamic strategies, the conformist approach (MTA) performed significantly better than all the other selection approaches (Fig. 3a). The perfectionist approach (WTA) performed particularly poorly, being close to the chance rate of 0.5. The cues selected by the

dynamic strategies contributed more diversely than the weights of the static strategy (WSM, Fig. 3b). In the MTA strategy, the monaural cue (MSS), also favored by the static strategy, played an important role, suggesting that it provides the most consistent ratings across all considered cues. In the LTA strategy, broadband interaural cues (ITIT) were most frequently selected, which is in accordance with our single-cue prediction results where this particular cue also provided good results for some conditions. Concordantly, the large variance in prediction errors (large error bars for LTA in Fig. 3a) indicates here that such a break down of broadband interaural disparities is only meaningful in some listening conditions.

Cue relevance: both monaural and interaural spectral shapes matter

Our analysis suggests that sound externalization is mainly based on two cues. However, most of the cues co-vary to some degree, and thus the relevance of a particular cue as a contributor to the cue combination may be strongly affected by the joint consideration of dependent cues. To investigate the effect of such interdependencies, we simulated listeners using the static decision strategy (WSM) with only two cues but all possible combinations.

The results, sorted by increasing prediction error, are shown in Fig. 3c. As expected there is a considerable variance in the prediction errors obtained by different pairs, with an order of magnitude between the best and worst pair. The condition considering the previously favored cues (MSS and ISS) yielded the smallest prediction error, confirming our findings from simulations considering more than two cues.

The results underline the importance of both cues by two means. First, all pairs including the monaural cue (MSS) were ranked highest and the prediction error remained small regardless of the other considered cue. As long the that cue was involved, the prediction errors did not increase significantly (remained within the confidence intervals of the best condition). Thus, the monaural spectral shape cue seems to be the most important cue for auditory externalization under static listening conditions with fixed reverberation. Second, by not considering that monaural cue, the error increased (with the mean above the confidence intervals of the best condition), but was still not significantly different to the best condition, as long as the spectral interaural cue (ISS) was considered in the modeling. In fact, the prediction errors significantly increased ($p < 0.05$, compared to the best condition) as soon as neither the monaural (MSS) nor the interaural spectral shape cue (ISS) were considered. Other spectral cues that only evaluate the amount of spectral contrast (MSSD and ISSD) instead of its shape fail to explain perceived externalization.

Discussion

To uncover the decision strategy underlying perceptual auditory externalization in various static listening contexts, we developed a template-based modeling framework and applied it to a set of psychoacoustic experiments investigating how spectral distortions degrade the au-

ditory externalization percept. Contrary to our expectations derived from previous investigations in vision[39–41], our results suggest that a static, weighted combination (WSM) rather than a dynamic selection (LTA, MTA, WTA) of multiple cues drives the perceptual decision of auditory externalization. Hence, although listeners are sensitive to individual cues they seem to apply fixed weights to integrate cues in order to judge perceptual externalization.

The highly predictable world we live in allows us to build internal models preparing the sensory systems for processing. By doing so, we can react faster and more accurately to incoming stimuli. Potentially, the sensory system could use such predictions not only to establish internal templates but also to build up an adaptive strategy optimizing the perception based on ambiguous sensory information. The perceptual inference of auditory externalization is known to be adaptive[45,46]. However, our findings suggest that this is not due to adaptive strategies. Instead, listeners may be quite flexible in the generation of expectations, i.e., their cue templates.

Our model selection shows that the number of spectral cues contributing to auditory externalization can be limited to two and comprise both the evaluation of monaural and interaural spectral shapes. Those two cues are well-known to be important for the perceptual inference of sound direction. Monaural spectral shapes comprise localization cues within sagittal planes, i.e., including both the vertical and front/back dimension[22,42]. The actual implementation of this cue monaurally processes positive spectral gradient profiles (MSS). In line with previous assumptions[22,42] and electrophysiological measurements in the dorsal cochlear nucleus[44], limiting the contribution of negative spectral gradients turned out to benefit predictive performance here. Future psychoacoustic experiments specifically targeting the spectral processing asymmetry proposed by our model investigations may still be important because conclusions drawn from existing data are limited by the fact that positive and negative spectral gradients were affected quite equally and simultaneously in most conditions.

Interaural intensity differences are very well established as cues for auditory lateralization[25,26]. Given the inherent frequency selectivity of the auditory periphery it is reasonable to believe that relative differences to an internal reference are also evaluated on a frequency-specific level[15], in line with our results.

The strong overlap of cues for externalization and directional localization of sounds suggests their joint evaluation in the future. The externalization model proposed here does not explicitly consider interactions with directional sound localization[47–50]. In our simulations, template comparisons were only performed for the reference direction, ignoring that there could be a strong match to a template from another direction yielding strong externalization, e.g., within sagittal planes because of spectral cue changes. We did that because, except for one study[16], listeners in the experimental paradigms were asked to rate externalization against a fixed reference stimulus and, thus, we assumed that listeners were able to ignore directional localization changes. Future experiments should combine assessments of externalization

and directional localization allowing future modeling approaches to consider these directional changes jointly with the externalization ratings.

Reverberation smears the spectral profile, decorrelates the binaural signal, and thus, increases the variance of interaural cues. Externalization usually benefits from reverberation but the interaction of underlying cues is still debated and ranges from the temporal fluctuation of short-term IIDs[19,51] to the spectral variance of long-term ICs[52]. This raises the need for model investigations being extended to naturally reverberant stimuli. Moreover, investigating this dissociation in combination with a systematic variation of the amount of reverberation would allow to evaluate the transition in perceptual relevance from MSS and ISS under anechoic conditions to a potentially broader set of cues under reverberant conditions.

From a technical perspective, our model predicts sound-externalization ratings based on arbitrary binaural signals within static listening conditions without prior knowledge about the experiment – a feature often requested by scientists from the area of psychophysics and neuroscience, as well as engineers working on audio and human interfaces such as augmented and virtual reality. In particular, it can be applied to predict auditory externalization in a variety of acoustic applications relying on spectral cue modifications such as headphones and hearing aids.

Methods

Structure of the model mapping signals to externalization ratings

The model flexibly comprises a variety of cues as summarized in Tab. 1. Depending on the considered cue, slightly different model architectures were required for processing either monaural (Fig. 1b) or interaural cues (Fig. 1c). Both architectures used a binaural signal as an input and predicted externalization ratings as an output after performing comparisons with an internal cue templates. Those templates were derived from listener-specific head-related transfer functions (HRTFs) or binaural room impulse responses (BRIRs), depending on the modeled experiment.

For both monaural and interaural cues, the template comparison may either be based on whole spectral shapes or scalar spectral metrics. Hence, in addition to monaural spectral shape (MSS)[22,42] and interaural spectral shape (ISS)[15], we considered various cues that could be associated with externalization or distance perception of anechoic sounds: the difference in overall monaural intensity (MI), the spectral standard deviation of IIDs (ISSD)[53], the broadband interaural coherence (IC)[52], and the inconsistency between ITD and IID (ITIT). In analogy to the ISSD as a direct measure of ISS, we also introduced a direct measure assessing the spectral standard deviation of spectral gradients (MSSD).

For each cue, we first calculated an internal error metric relating an acoustic signal feature with externalization. The error metric was then mapped to an externalization rating, which

has been scaled individually for each experiment. Thus, both the mapping and calculation of error metrics depended on the considered cue and experiment, unless otherwise noted.

Error metrics

For the MI cue, the overall level differences were considered as:

$$d_{MI} = \Delta MI / MI_{template} \,, \tag{1}$$

with $\Delta MI = |MI_{target} - MI_{template}|$ denoting the difference in root mean square (RMS) levels (in dB) of the incoming broadband signals. Differences smaller than ±1 dB were considered to be below the just-noticeable difference[54] and thus set to zero. The error based on MI, d_{MI} was calculated for each ear separately and then averaged across both ears.

The IC cue was calculated as $IC = \lim_{\tau} \inf \int x_L(t-\tau) x_R(t) dt$ for the binaural signals x_L and $\pm\pi/2$ within the range of $\tau \in [-1,1]$ ms. The error based on IC was then calculated by comparing the target IC with the template IC:

$$d_{IC} = \frac{|IC_{target} - IC_{template}|}{IC_{template}} \,. \tag{2}$$

The errors for the other cues were calculated after filtering the target and template signals through a bank of fourth-order Gammatone filters with a regular spacing of one equivalent rectangular bandwidth[55]. The spectral excitation profiles were computed from the logarithm of the RMS energy within every frequency band[22]. Audibility thresholds for band-limited signals were roughly approximated by generally considering stimulus SPLs to be 70 dB and a within-band threshold of 20 dB. Assuming stationary input signals, the spectral profiles were averaged over time, yielding spectral profiles as a function of frequency band, $p(f)$. Further, the interaural difference of the spectral profiles yielded IIDs as a function of frequency band, $ILD(f)$.

For the ISSD cue, the model evaluated standard deviations (SD) of $ILD(f)$ and computed the negative difference of these deviations between the target and template, relative to the template deviation, yielding:

$$d_{ISSD} = \frac{|SD_f(ILD_{target}(f)) - SD_f(ILD_{template}(f))|}{SD_f(ILD_{template}(f))} \,. \tag{3}$$

For the ISS cue, the absolute values of frequency-specific differences between the target and template IIDs were evaluated. Then differences smaller than 1 dB were set to zero and larger differences, $\Delta|IID(f)|$, were normalized by the template IIDs and averaged across frequency bands, yielding:

$$d_{ISS} = \frac{1}{N_f} \sum_f \frac{\Delta|ILD(f)|}{|ILD_{template}(f)|} \,, \tag{4}$$

with N_f being the number of frequency bands.

The ITIT error was calculated as the broadband deviation between target-to-template ratios of ITD and IID:

$$d_{ITIT}=\left|\frac{\Delta ITD}{ITD_{template}}-\frac{\Delta ILD}{ILD_{template}}\right| , \tag{5}$$

with $\Delta ITD=ITD_{target}-ITD_{template}$ and $\Delta IID=IID_{target}-IID_{template}$. ΔITD and ΔIID smaller than ±20 µs and ±1 dB, respectively, were set to zero. The ITDs were derived from binaural signals following a procedure, in which the signals were low-pass filtered at 3 kHz and the ITD was the time lag that yielded maximum IC of the temporal energy envelope[56].

For the MSS and MSSD cues, positive spectral gradient profiles were derived exactly as in Baumgartner et al.[42]. Briefly, first monaural spectral gradients were obtained by differentiating the excitation profiles ($p(f)\rightarrow p'(f)$) and softly restricting the value range by an elevated arctangent, yielding

$$MSG(f)=\arctan\left(p'(f)-\pi/2\right)+\pi/2 . \tag{6}$$

For the MSS cue, these gradients were then compared between the target and template separately for each ear by applying the same procedure as for the ISS metric, that is, calculating absolute target-to-template differences, normalizing differences larger than 1 dB $(\Delta|MSG(f)|)$ by the template gradients, and averaging those differences across frequencies:

$$d_{MSS,L/R}=\frac{1}{N_f}\sum_f \frac{\Delta|MSG_{L/R}(f)|}{|MSG_{template,L/R}(f)|} , \tag{7}$$

separately for the left and right ear as indexed by L and R, respectively. The MSS error metric was defined in analogy to ISSD:

$$d_{MSSD,L/R}=\frac{\left|SD_f\left(MSG_{target,L/R}(f)\right)-SD_f\left(MSG_{template,L/R}(f)\right)\right|}{SD_f\left(MSG_{template,L/R}(f)\right)} . \tag{8}$$

These unilateral error metrics were then combined according to a binaural weighting function[22,23], effectively increasing the perceptual weight of the ipsilateral ear with increasing lateral eccentricity:

$$d=d_R+\frac{d_L-d_R}{1+e^{-\phi/\Phi}} , \tag{9}$$

with $\phi\in[-90°,90°]$ denoting the lateral angle (left is positive) and $\Phi=13°$.

Mapping to externalization ratings

A sigmoidal mapping function scaled by $2e_{range}$, shifted by e_{offset}, and slope-controlled by a sensitivity parameter S_{cue} was used to map the deviation metrics d_{cue} to externalization ratings e_{cue}:

$$e_{cue} = \frac{2e_{range}}{1+\exp\left(d_{cue}/S_{cue}\right)} + e_{offset} \,. \tag{10}$$

The nominator was doubled because the rating scale used in previous experiments was usually one-sided with respect to the reference sound, i.e., listeners were asked to rate only decreases and not increases of perceived externalization.

The mapping function in Eq. (10) contains one free model parameter, S_{cue}, inversely related to the slope of the function mapping changes in the deviation metrics to changes in the externalization ratings. Mapping sensitivity is denoted by $1/S_{cue}$ because a larger $1/S_{cue}$ yields a steeper mapping function that projects smaller errors to smaller externalization ratings. This cue- and experiment-specific sensitivity was obtained by minimizing the squared prediction error, which was defined as the RMS of differences between actual and predicted externalization ratings (normalized scales). For the minimization, we applied the Nelder-Mead simplex (direct search) method (fminsearch, Matlab Optimization Toolbox, The Mathworks Inc.).

Decision stage

For the WSM, the optimal weights were obtained by optimizing weights scaling the contribution of a cue to the ratings, i.e., minimizing the prediction error. We used the same optimization technique as used for S_{cue}. Weights smaller than 0.005 were considered as negligible and were set to zero.

For a fair comparison across our simulations of the decision strategies, the same number of model parameters was considered in the optimization. For the dynamic strategies, the mapping parameters were optimized to best fit the data of all experiments. For the weighted-sum strategy, the mapping parameters were fixed and corresponded to those from single-cue predictions, and the individual summation weights were optimized to best fit the data of all experiments.

Considered studies

For the modeling, we considered results of five experiments[11,14–16]. The pool of experiments was a compromise of data availability and the degree to which the tests isolate specific cues. The experiments differed in whether they provided visual information about the reference source and tested in anechoic or reverberant environments. While the constant reverberation and visual information may or may not have stabilized auditory externalization, they certainly

did not prevent the tested signal modifications to be effective within the tested condition. In our study, we thus assumed that such differences in experimental procedures do not modulate our effects of interest.

Exp. I and II: Hartmann and Wittenberg[14] synthesized the vowel /a/ with a tone complex consisting of 38 harmonics of a fundamental frequency of 125 Hz, yielding a sound limited up to 4750 Hz. This sound was presented via headphones and filtered with individualized HRTFs corresponding to 37° azimuth (0° elevation). For modeling, we used HRTFs from 21 exemplary listeners contained in the ARI database.

Exp. III: Hassager et al.[15] investigated the effect of spectral smoothing on auditory externalization of Gaussian white low-frequency noises, i.e., band-limited from 50 Hz to 6000 Hz. These sounds were filtered with individualized BRIRs (reverberation time of ~0.2 s) in order to simulate sound sources positioned at azimuths of 0° and 50°. As independent experimental variable, Gammatone filters with various equivalent rectangular bandwidths (ERBs) were used to spectrally smooth either the direct path portion (until 3.8 ms) or the reverberation of the BRIRs. Filters with larger ERBs more strongly smoothed the shape of the magnitude spectrum and reduced the degree of externalization only when applied to the direct path. Smoothing applied to the reverberation was perceptually ineffective. Because the original BRIRs were not accessible, our model predictions were again based on the same 21 (anechoic) HRTFs from the ARI database and only addressed the effect of spectrally smoothing the direct path.

Exp. IV: Absolute externalization ratings from Baumgartner et al.[16] were estimated from the listeners' paired binary judgments by calculating mean relative frequencies of "farther" judgments. We evaluated only data from their "discontinuous trial" condition of their Exp. II, which did not elicit an auditory looming bias, and from the ten out of twelve listeners who perceived the spectrally flat sound (C = 0) as being closer than the individualized reference sound (C = 1).

Exp. V: Boyd et al.[11] used individualized BRIRs to simulate a talker positioned at 30° azimuth. They compared externalization ratings for in-the-ear (ITE) and behind-the-ear (BTE) microphone casings as well as broadband (BB) and 6.5-kHz-low-pass (LP) filtered stimuli at various mixing ratios with stereophonic recordings providing only an ITD but no IID. The reverberation time of 0.35 s remained constant across mixing ratios. For our model simulations, original BRIRs were only available for three out of seven (normal-hearing) listeners.

References

1. Hawkins, J. & Blakeslee, S. *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*. (Times Books, 2004).
2. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 815–836 (2005).

3. Den Ouden, H. E., Kok, P. & De Lange, F. P. How Prediction Errors Shape Perception, Attention, and Motivation. *Front. Psychol.* **3**, (2012).
4. Ulanovsky, N., Las, L. & Nelken, I. Processing of low-probability sounds by cortical neurons. *Nat. Neurosci.* **6**, 391–398 (2003).
5. Baldeweg, T. Repetition effects to sounds: evidence for predictive coding in the auditory system. *Trends Cogn. Sci.* **10**, 93–94 (2006).
6. van der Heijden, K., Rauschecker, J. P., de Gelder, B. & Formisano, E. Cortical mechanisms of spatial hearing. *Nat. Rev. Neurosci.* **20**, 609–623 (2019).
7. Majdak, P., Baumgartner, R. & Jenny, C. Formation of three-dimensional auditory space. in *The Technology of Binaural Understanding* (Springer, Berlin, Heidelberg, 2020).
8. Gardner, J. L. Optimality and heuristics in perceptual neuroscience. *Nat. Neurosci.* **22**, 514–523 (2019).
9. Weber, E. H. On the circumstances under which one is led to refer sensations to external objects. in *Proceedings of the royal saxon society for science in Leipzig* vol. 2 226–237 (1848).
10. Kulkarni, A. & Colburn, H. S. Role of spectral detail in sound-source localization. *Nature* **396**, 747–9 (1998).
11. Boyd, A. W., Whitmer, W. M., Soraghan, J. J. & Akeroyd, M. A. Auditory externalization in hearing-impaired listeners: The effect of pinna cues and number of talkers. *J. Acoust. Soc. Am.* **131**, EL268–EL274 (2012).
12. Durlach, N. I. *et al.* On the Externalization of Auditory Images. *Presence Teleoper Virtual Env.* **1**, 251–257 (1992).
13. Blauert, J. *Spatial hearing. The Psychophysics of Human Sound Localization*. (MIT-Press, 1997).
14. Hartmann, W. M. & Wittenberg, A. On the externalization of sound images. *J Acoust Soc Am* **99**, 3678–88 (1996).
15. Hassager, H. G., Gran, F. & Dau, T. The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment. *J. Acoust. Soc. Am.* **139**, 2992–3000 (2016).
16. Baumgartner, R. *et al.* Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias. *Proc. Natl. Acad. Sci.* **114**, 9743–9748 (2017).
17. Zhang, P. X. & Hartmann, W. M. On the ability of human listeners to distinguish between front and back. *Hear Res* **260**, 30–46 (2010).
18. Catic, J., Santurette, S., Buchholz, J. M., Gran, F. & Dau, T. The effect of interaural-level-difference fluctuations on the externalization of sound. *J. Acoust. Soc. Am.* **134**, 1232–1241 (2013).
19. Catic, J., Santurette, S. & Dau, T. The role of reverberation-related binaural cues in the externalization of speech. *J. Acoust. Soc. Am.* **138**, 1154–1167 (2015).
20. Romigh, G. D., Simpson, B. D. & Iyer, N. IN EAR TO OUT THERE: A MAGNITUDE BASED PARAMETERIZATION SCHEME FOR SOUND SOURCE EXTERNALIZATION. in (2016).
21. Slattery III, W. H. & Middlebrooks, J. C. Monaural sound localization: Acute versus chronic unilateral impairment. *Hear. Res.* **75**, 38–46 (1994).
22. Baumgartner, R., Majdak, P. & Laback, B. Modeling sound-source localization in sagittal planes for human listeners. *J. Acoust. Soc. Am.* **136**, 791–802 (2014).
23. Macpherson, E. A. & Sabin, A. T. Binaural weighting of monaural spectral cues for sound localization. *J Acoust Soc Am* **121**, 3677–3688 (2007).
24. May, B. J. Role of the dorsal cochlear nucleus in the sound localization behavior of cats. *Hear. Res.* **148**, 74–87 (2000).

25. Lord Rayleigh, J. W., F. R. S. ,. or, Strutt. On our perception of sound direction. *Philos Mag* **13**, 214–232 (1907).
26. Macpherson, E. A. & Middlebrooks, J. C. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *J. Acoust. Soc. Am.* **111**, 2219–2236 (2002).
27. Shinn-Cunningham, B. G., Santarelli, S. & Kopco, N. Tori of confusion: binaural localization cues for sources within reach of a listener. *J Acoust Soc Am* **107**, 1627–36 (2000).
28. Devore, S., Ihlefeld, A., Hancock, K., Shinn-Cunningham, B. & Delgutte, B. Accurate Sound Localization in Reverberant Environments Is Mediated by Robust Encoding of Spatial Cues in the Auditory Midbrain. *Neuron* **62**, 123–134 (2009).
29. Wood, K. C., Town, S. M. & Bizley, J. K. Neurons in primary auditory cortex represent sound source location in a cue-invariant manner. *Nat. Commun.* **10**, 1–15 (2019).
30. Higgins, N. C., McLaughlin, S. A., Rinne, T. & Stecker, G. C. Evidence for cue-independent spatial representation in the human auditory cortex during active listening. *Proc. Natl. Acad. Sci.* **114**, E7602–E7611 (2017).
31. Salminen, N. H. *et al.* Integrated processing of spatial cues in human auditory cortex. *Hear. Res.* **327**, 143–152 (2015).
32. Altmann, C. F. *et al.* Independent or integrated processing of interaural time and level differences in human auditory cortex? *Hear. Res.* **312**, 121–127 (2014).
33. Schröger, E. Interaural time and level differences: Integrated or separated processing? *Hear. Res.* **96**, 191–198 (1996).
34. Tardif, E., Murray, M. M., Meylan, R., Spierer, L. & Clarke, S. The spatio-temporal brain dynamics of processing and integrating sound localization cues in humans. *Brain Res.* **1092**, 161–176 (2006).
35. Edmonds, B. A. & Krumbholz, K. Are Interaural Time and Level Differences Represented by Independent or Integrated Codes in the Human Auditory Cortex? *J. Assoc. Res. Otolaryngol.* **15**, 103–114 (2014).
36. Colburn, H. S. & Isabelle, S. K. Models of Binaural Processing Based on Neural Patterns in the Medial Superior Olive. in *Auditory Physiology and Perception* (eds. Cazals, Y., Horner, K. & Demany, L.) 539–545 (Pergamon, 1992). doi:10.1016/B978-0-08-041847-6.50066-6.
37. Baldassi, S. & Burr, D. C. Feature-based integration of orientation signals in visual search. *Vision Res.* **40**, 1293–1300 (2000).
38. Thornton, M. A., Weaverdyck, M. E. & Tamir, D. I. The brain represents people as the mental states they habitually experience. *Nat. Commun.* **10**, 1–10 (2019).
39. Palmer, J., Verghese, P. & Pavel, M. The psychophysics of visual search. *Vision Res.* **40**, 1227–1268 (2000).
40. Baldassi, S. & Verghese, P. Comparing integration rules in visual search. *J. Vis.* **2**, 3–3 (2002).
41. Song, Y.-H. *et al.* A Neural Circuit for Auditory Dominance over Visual Perception. *Neuron* **93**, 940-954.e6 (2017).
42. Baumgartner, R., Majdak, P. & Laback, B. Modeling the Effects of Sensorineural Hearing Loss on Sound Localization in the Median Plane. *Trends Hear.* **20**, 1–11 (2016).
43. Majdak, P., Baumgartner, R. & Laback, B. Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Front. Psychol.* **5**, 1–10 (2014).
44. Reiss, L. A. J. & Young, E. D. Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus. *J. Neurosci.* **25**, 3680–3691 (2005).

45. Gil-Carvajal, J. C., Cubick, J., Santurette, S. & Dau, T. Spatial Hearing with Incongruent Visual or Auditory Room Cues. *Sci. Rep.* **6**, 37342 (2016).
46. Mendonça, C., Campos, G., Dias, P. & Santos, J. A. Learning Auditory Space: Generalization and Long-Term Effects. *PLOS ONE* **8**, e77900 (2013).
47. Brimijoin, W. O., Boyd, A. W. & Akeroyd, M. A. The Contribution of Head Movement to the Externalization and Internalization of Sounds. *PLoS ONE* **8**, e83068 (2013).
48. Li, S., Schlieper, R. & Peissig, J. The Role of Reverberation and Magnitude Spectra of Direct Parts in Contralateral and Ipsilateral Ear Signals on Perceived Externalization. *Appl. Sci.* **9**, 460 (2019).
49. Hendrickx, E. *et al.* Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. *J. Acoust. Soc. Am.* **141**, 2011–2023 (2017).
50. Kates, J. M., Arehart, K. H., Muralimanohar, R. K. & Sommerfeldt, K. Externalization of remote microphone signals using a structural binaural model of the head and pinna. *J. Acoust. Soc. Am.* **143**, 2666–2677 (2018).
51. Li, S., Schlieper, R. & Peissig, J. The effect of variation of reverberation parameters in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source in a listening room. *J. Acoust. Soc. Am.* **144**, 966–980 (2018).
52. Leclère, T., Lavandier, M. & Perrin, F. On the externalization of sound sources with headphones without reference to a real source. *J. Acoust. Soc. Am.* **146**, 2309–2320 (2019).
53. Georganti, E., May, T., Par, S. van de & Mourjopoulos, J. Sound Source Distance Estimation in Rooms based on Statistical Properties of Binaural Signals. *IEEE Trans. Audio Speech Lang. Process.* **21**, 1727–1741 (2013).
54. Mills, A. W. On the Minimum Audible Angle. *J. Acoust. Soc. Am.* **30**, 237–246 (1958).
55. Lyon, R. F. All pole models of auditory filtering. in *Diversity in auditory mechanics* (eds. E. R. Lewis et al.) 205–211 (World Scientific Publishing, Singapore, 1997).
56. Katz, B. F. G. & Noisternig, M. A comparative study of interaural time delay estimation methods. *J. Acoust. Soc. Am.* **135**, 3530–3540 (2014).
57. Majdak, P. & Co. The Auditory Modeling Toolbox. http://amtoolbox.sourceforge.net (2019).
58. Søndergaard, P. & Majdak, P. The Auditory Modeling Toolbox. in *The Technology of Binaural Listening* (ed. Blauert, J.) 33–56 (Springer, 2013).

Acknowledgements

We want to thank Bill Withmer for kindly providing parts of the original data from Boyd et al. (2012) as well as Henrik Gert Hassager, Barbara Shinn-Cunningham and H. Steven Colburn for fruitful discussions. This work was supported by the Austrian Science Fund (FWF, grant J 3803-N30 to R.B.) and Facebook Reality Labs.

Author Contributions

R.B. designed the research and performed the simulations; R.B. and P.M. interpreted the results and wrote the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests. All model simulations are publicly available as part of the Auditory Modelling Toolbox[57,58] (model name: baumgartner2020).