Binaural Signal Integration Improves Vertical Sound Source Localization

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

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

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

Abstract

Early studies have shown that the localization of a sound source in the vertical plane can be accomplished with only a single ear and thus assumed to be based on monaural spectral cues. Such cues consists of notches and peaks in the perceived spectrum which vary systematically with the elevation of sound sources. This poses several problems to the auditory system like extracting relevant and direction-dependent cues among others. Interestingly, at the stage of elevation estimate binaural information from both ears is already available and it seems reasonable of the auditory system to take advantage of this information. Especially, since such a binaural integration can improve the localization performance dramatically as we demonstrate with a computational model of binaural signal integration for sound source localization in the vertical plane. In line with previous findings of vertical localization, modeling results show that the auditory system can perform monaural as well as binaural sound source localization given a single, learned map of binaural signals. Binaural localization is by far more accurate than monaural localization, however, when prior information about the perceived sound is integrated localization performance is restored. Thus, we propose that elevation estimation of sound sources is facilitated by an early binaural signal integration and can incorporate sound type specific prior information for higher accuracy.

Introduction

Audition is our only sensory system that let us perceive what is happening behind a wall or around the corner. To do that, extensive neural computations are in place along the auditory pathway that transform the oscillation of the eardrum, a tonotopic representation of the stimulus to, for example, comprehensible speech (phonetic representation [10, 23]) or spatial information about the location of a sound source (topographic representation [4, 8]). To transform tonotopic inputs of sounds to a topographic representation of space, the auditory system extracts three major cues from a sound signal created by the distance between the ears, their shape and the shadow of the head. The distance between the ears creates an interaural time difference (ITD) between the signal arriving time at the left and right ear [49]. Together with the interaural level difference (ILD), created by the attenuation of sounds by the head [38], these two cues provide a means for localization of sounds in the horizontal plane. However, for sounds on the median plane or on the cone-of-confusion [42] these cues provide ambiguous signals. To resolve this ambiguity and to accurately localize sounds in the vertical plane the auditory system exploits direction-dependent changes in the perceived frequency spectrum induced by the shape of the ear (pinna) [20, 29]. These so called spectral cues are characterized by direction-dependent Head Related Transfer Functions (HRTFs) [3, 5, 12, 27, 39, 47, 50].

Extracting such spectral cues from the sensory input is not straight forward. The spectra of every day sounds are very different to each other which results in very different spectra at the sensory input level after being filtered by the pinna. This poses several problems for elevation estimation: At the level of the eardrum the perceived sound spectrum has already been filtered with the elevation-dependent HRTF and, in principle, the auditory system has no indication which of the spectral cues were induced by the HRTF or were already present in the source spectrum. Thus, the estimation of sound source elevation is a ill-posed problem [21, 22]. This becomes apparent when looking at the spectra of different sounds (see Fig. 1 **B**). It is difficult to identify the fine structure imposed by the HRTFs and extracting spectral cues from such highly variable input signals for learning is challenging. One would like to have a mechanism that separates the sound type specific spectral content from the HRTF induced cues, thus leaving only the HRTF dependent frequency modulations. Some computational models tried to solve such a problem by assuming that the spectrum of the sound source is known (a prior) [31], by assuming local constancy of sound spectrum [53] or by assuming a broadband and sufficiently flat source spectrum [26].

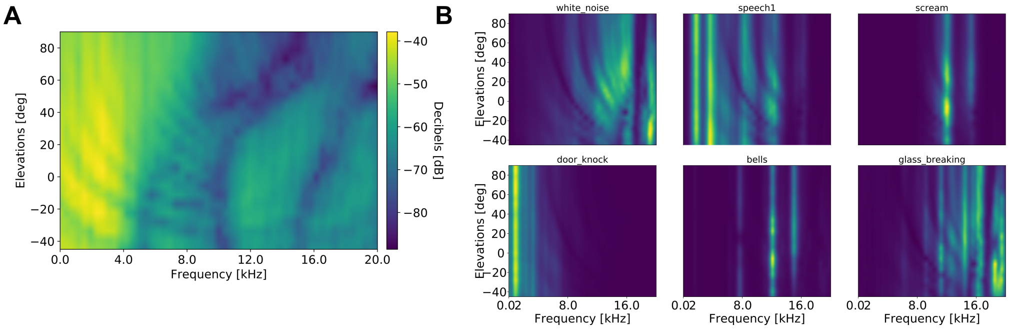


Figure 1. Head-Related Transfer Function A HRTFs for CIPIC subject no. 10 as a function over elevations. Different colors indicate the energy content. Direction dependent changes in energy content is most prominent above 4kHz. **B** Elevation spectra maps of six different natural sounds after being filtered with the HRTF of **A**.

Another issue in vertical sound source localization is the role of binaural integration. Early localization experiments demonstrated that participants are able to localize sounds with only one ear [44] and that the ability to localize sounds with one ear or two ears is similar, thus concluding that vertical localization is not binaural [18]. However, by using virtual sound sources provided over headphones Wightman et al. in a later study [50] questioned the monaural localization paradigm applied in most previous experiments including their own. Their findings demonstrate that localization is effectively degraded under monaural listing conditions. A later study confirmed that both ears contribute to the perception of elevation [28], thus supporting the hypothesis of binaural integration.

When such an integration takes place is still unclear. Hofman et al. [19] first described different schemes for elevation estimation. Based on their findings they hypothesized that there needs to be an weighted integration step of the signal from the left and right ear to get a single estimation of the elevation. Whether this integration step already takes place before the spectral-mapping (binaural) or after (monaural) was unclear. Later on, the authors tried to answer this question in another experiment [48] but the results are ambiguous and the authors were not able to derive a conclusion.

The difficulty of separating HRTFs from the source spectrum, the contradicting results for monaural and binaural sound source localization [18, 50], and the unclear integration order of signals from the left and right ear [19] raise the question of how the auditory system processes sound signals on a neural level to learn a representative template of HRTFs in order to generate a stable and unique perception of elevation estimates.

Here, we propose a computational model of sound source localization in the median plane that extracts elevation specific cues by integrating signals from both ears and learning a sound type specific prior. Binaural integration leads to a sound type independent representation of HRTFs, thus solving the ill-posed problem of elevation estimation. Based on such signals the model reliably localizes binaural sound sources but struggles with monaural input signals. By integrating sound type specific prior information for elevation estimation, the model becomes able to localize monaural sound signals. Thereby suggesting that elevation estimation is a binaural process but can deal with monaural inputs if the sound has previously been learned.

In the following simulation experiments we demonstrate that, based on a binaurally learned map, the model is very well capable of localizing binaural sounds, but most significantly remains localization performance for monaural sounds given there is available prior information of this sound. In particular, results show that prior information constantly improves localization performance.

Results

Brief Description of the model

The model consists of two layers of neurons in the processing pathway: a normalization layer for ipsi and contralateral inputs (I), respectively, and a binaural integration layer (II). A third layer averages over all input signals thus learning a map of elevation spectra (III). Perceived signals are compared to this learned map via cross-correlation (see [21] for details) in a fourth layer (IV) to estimate the elevation of the sound source (see Fig. 2). The normalization layer of neurons receives a frequency signal of the sound signal, provided by a gammatone-filter bank [36] of the ipsi- and contralateral input signals, as an input and performs a divisive normalization with a Gaussian-filtered version of it self. This normalization already provides signals with prominent spectral cues (see Fig. 2 middle maps). Averaging these signals over elevations leads to a sound type specific prior which is in some conditions used for map learning and to improve localization of monaural and binaural sounds. In the binaural integration layer, the normalized signals from the the ipsi- and contralateral side are integrated (by a division) to provide a binaural signal. In the last layer of the model, a cross-correlation of the perceived, filtered sound signal with a previously learned map is calculated and the elevation with maximum correlation value is chosen as the elevation estimate (similar to [21]). For more details on model implementation see Methods.

We found that this simple model of elevation perception can account for typical behavioral of human auditory elevation perception and predicts that the underlying localization process is fundamentally binaural with the ability to localize monaural sounds under certain conditions.

To investigate the performance and validity of our model we used HRTFs of 45 subjects from the CIPIC database [1] and presented each with 20 different natural sound types (signal-to-noise ratio $5 : 1$), originating from 25 different elevations ($[-45°, 90°]$ in 5.625° steps according to CIPIC database recording [1]). All presented sound types are averaged for each participant to create a learned map of spectral elevations. This map is used for the cross-correlation with a perceived probing signal.

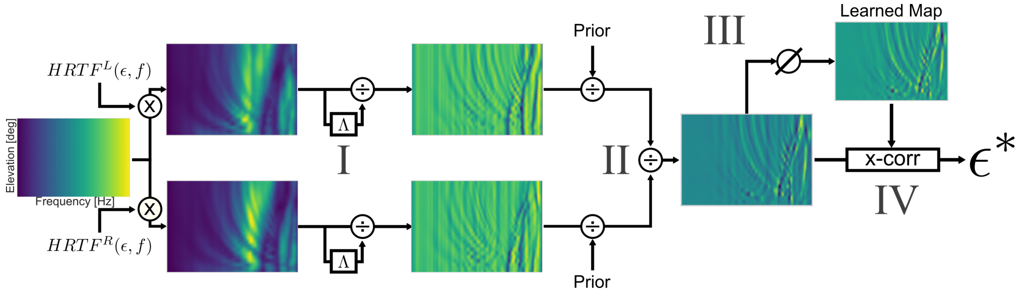


Figure 2. Model Architecture Each sound that is presented to the model is first filtered by a subject's HRTF of different elevations for the left ($HRTF^L(\epsilon, f)$) and right ear $HRTF^R(\epsilon, f)$, respectively. The resulting signals are filtered by a Gaussian normalization step (I). Then, if available, prior information is integrated, separately for the left and right ear. The binaural integration step (II) combines the signal form the left and right ear. Each perceived signal of all presented sound types contributes to build a learned map of elevation spectra (III) for later cross-correlation with a perceived sound to computed an elevation estimate ϵ^* (IV).

The resulting elevation estimate is plotted against the actual elevation of the presented sound. Consequently, for each participant a linear regression analysis is performed on this data which provides a response gain (accuracy), a response bias (spatial bias) and a response precision (coefficient of determination). Different conditions are tested to demonstrate the advantage of binaural signal integration and prior information on localization performance.

A binaurally learned map can account for binaural as well as monaural sound source localization

Experimental results demonstrate that humans can localize sounds with just a single ear [18, 44]. Based on these results, the common assumption for human vertical sound source localization is that it is fundamentally monaural, that is, localization is separately initialized for the left and right ear, respectively, and the two elevation estimates are integrated for a single estimate [21]. We question this assumption by providing simulation results that demonstrate that when a combined binaural map for the left and right ear is learned, monaural sound source localization is still possible.

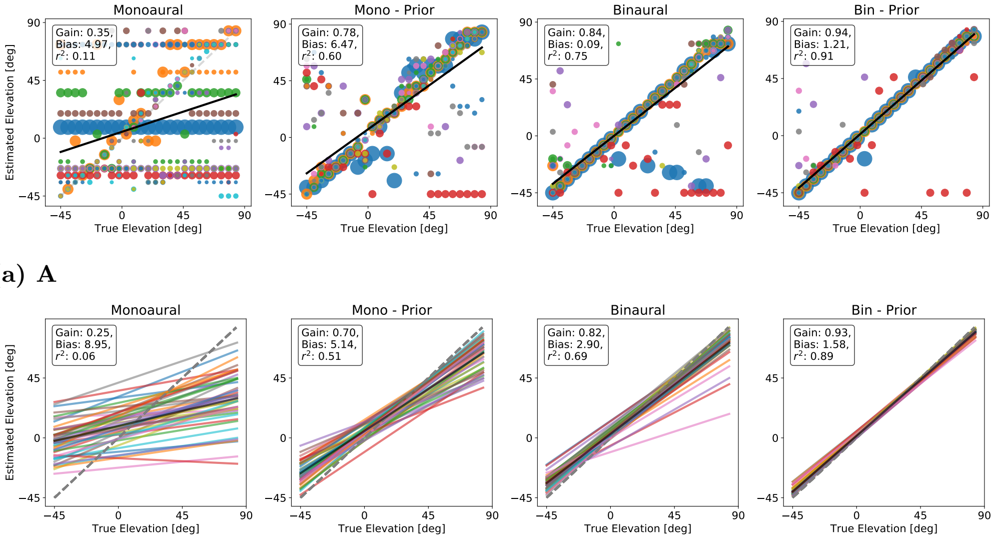
Here, we show that a binaurally learned map of elevation spectra can account for sound source localization under binaural and monaural conditions, given prior information for monaural sound signals is available. Thereby, such learned maps can account for experimental results with binaural and monaural listing conditions. Four different conditions are tested: In the monaural condition, the binaural integration layer is skipped and pure, normalized monaural signals are presented to the model for localization. These monaural signals are combined with the previously learned sound type specific prior in the monaural-prior condition. In the binaural condition the binaural integration layer remains active and the elevation estimation is calculated based on the output of this layer. The last condition (binaural-prior) combines these binaural signals with prior information before the cross-correlation with the learned map is performed.

Simulation results for a single participant (CIPIC HRIR no.8) are shown in Fig. 3**A** when a binaural map with integrated prior information is learned. Not surprisingly, pure monaural sounds (ispilateral ear, left panel) are basically not localizable (gain: 0.35, bias: 4.97, score: 0.11). However, when such monaural sounds are combined with a previously learned sound type specific prior (middle left panel), the localization quality increases dramatically (gain: 0.78, bias: 6.47, score: 0.60), thus localization ability is restored.

Localization performance for sounds that are presented binaurally with (middle right panel) and without (right panel) integrating prior information is almost perfect (gain: 0.84, bias: 0.09, score: 0.75 and gain: 0.94, bias: 1.21, score: 0.91, respectively). Such good performance in these conditions is expected since the learned map is constructed based on these binaural sounds integrating prior information.

When averaging localization quality over all participants (all 45 HRIRs from CIPCI database) the initial trend remains Fig. 3**B**. That is, localization of monaural sounds is basically non-existing (left panel, gain: 0.25, bias: 8.95, score: 0.06) but improves tremendously when prior information is integrated (middle left panel, gain: 0.70, bias: 5.14, score: 0.51). Again, localization performance for binaural sounds and binaural sounds integrating prior information remains stable (gain: 0.82, bias: 2.90, score: 0.69 and gain: 0.93, bias: 1.58, score: 0.89, respectively).

These results demonstrate that a binaural map of elevation spectra supports the localization of monaural sounds integrating prior information but is unable to localize pure monaural sounds, since their spectral information differs greatly from the learned spectra (see supplementary Fig. S1). This is a strong indication for the existence of a binaurally learned map.



(a) **A**

(b) **B**

Figure 3. Elevation estimation results. Model estimates for all presented elevations and all sound types. X-axis indicates the elevation of the presented sound. Y-axis is the model elevation estimate for a sound. Pure monaural sounds are presented in left panel. Middle-left panel shows model estimate for monaural sounds integrating prior information. Pure binaural sounds are presented in middle-right panel. Right panel depicts model estimate for binaural sounds integrating prior information. **A** Model estimates for a single participant (no. 8 CIPIC). Each dot represents one sound source elevation with color indicating sound type. . **B** Calculated regression lines for each participant are shown (colored lines). Black lines are calculated by averaging over all colored lines to achieve averaged estimation values. Regression values are shown in inset box.

Localization quality for differently learned maps

In the previous experiment a binaural map is learned to localize sounds signals of various types. Hofman and colleagues [19] described different possibilities of how a unique perception of elevation of signals from the two ears might be achieved. They hypothesize two different schemes for elevation perception: the spatial weighting scheme and the spectral weighting scheme (see [19] their Fig. 7). The spectral weighting scheme is similar to our binaural integration model with a binaurally learned map, whereas the spatial weighting scheme would correspond to our model when a monaural map is learned. To test which scheme seems more plausible and to validate our results from the previous experiment, we test the localization quality of participants when the learned map is based on different signals (i.e. monaural, monaural-prior, binaural, binaural-prior, see Fig. 4) and demonstrate that a binaural map with sound type specific prior integration produces the best localization results (Fig. 4 last row).

When the learned map of elevation spectra is based of pure monaural input signals, localization performance is best for monaural signals integrated with the sound type specific prior (gain: 0.91, bias: 0.72, score: 0.87). Surprisingly, these sound signals are even better to localize than pure monaural signals (the basis for the map, gain: 0.52, bias: 12.76, score: 0.27). This demonstrates the benefit of the integration of a sound type specific prior. This advantage of prior integration can be also seen in the binaural sound conditions. For pure binaural sounds the localization performance is worse compared to binaural signals integrating prior information (gain: 0.42, bias: 3.76, score: 0.24 and gain: 0.57, bias: 2.71, score: 0.42, respectively). Here, the binaural - prior condition even outperforms the pure monaural condition, which is surprising since binaural signals differ substantially from monaural signals (see supplementary Fig. S1).

Taken together, these simulation results demonstrate that a pure monaural map is not sufficient to localize pure monaural sounds (first row). Prior information is required to localize sounds in monaural and binaural conditions. Even if this prior information is integrated in the learned map (second row), localization of sounds is difficult and again prior information of the input signals is crucial. However, if the learned map is based on binaural signals with or without the integration of prior information (third and fourth row, respectively) localization performance for each condition except the pure monaural condition is close to optimal. Thus, we hypothesize that elevation estimation is essentially based on binaural signals but can deal with monaural signals when prior information of these signals is available. Furthermore, these results demonstrate the prior information of sounds consistently improves localization performance of sound sources.

Neural network model

To demonstrate the biological plausibility of the model we present localization results of a neural network model. This model uses first-order differential equations to describe membrane potentials of neurons in different populations (see Methods for details). The neuron populations are implemented and connected with each other according to the different layers presented in the computational model. In the following experiments the signal-to-noise ratio is set to 0.

The localization performance for all participants from the CIPIC database is presented in Fig. 5. Even though different in the linear regression values the overall trend of the localization quality in the different conditions is similar to the computational model. Pure monaural sounds (ispilateral ear, left panel) are basically not localizable (gain: 0.28, bias: 2.89, score: 0.09). When combined with prior information such monaural sounds (middle left panel) the localization quality increases (gain: 0.40, bias: 1.29, score: 0.19). For binaurally presented sounds the localization

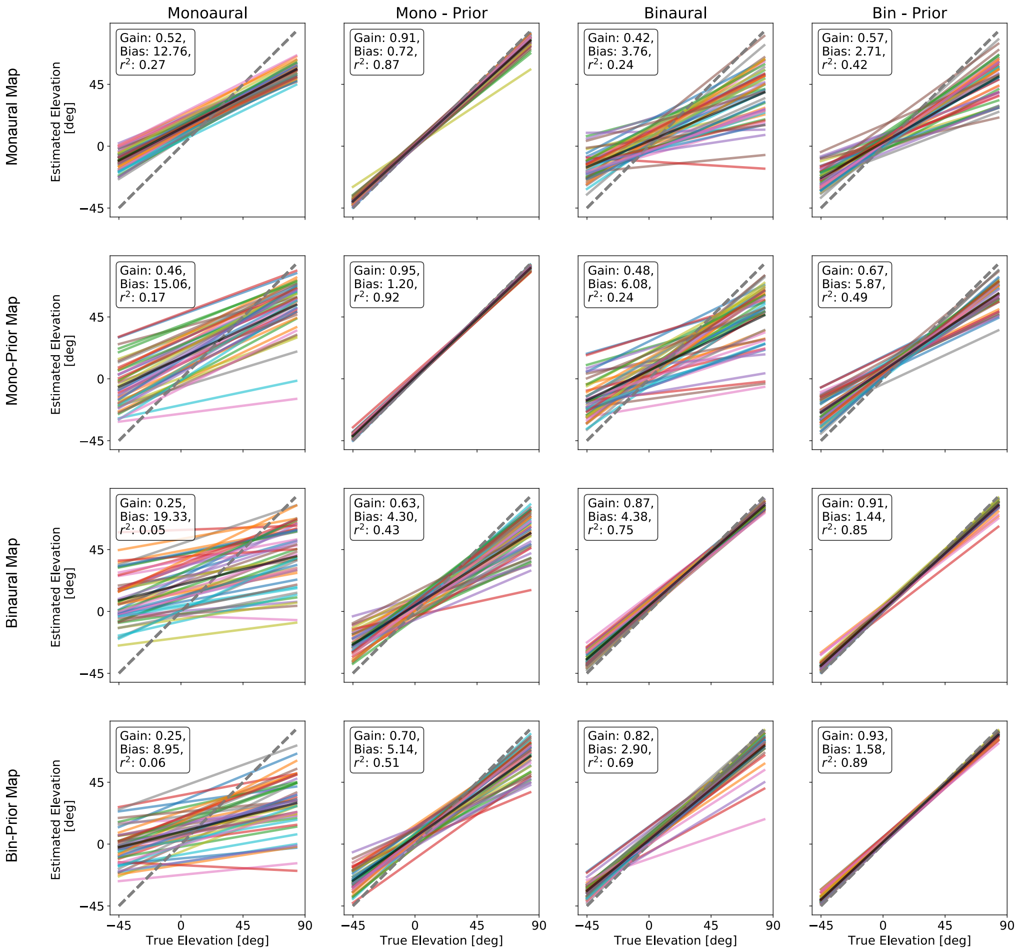


Figure 4. Estimation results over differently learned maps. Each column depicts localization results of the model for the different conditions, similar to Fig. 3. In each row the learned map which is used to compare perceived sounds to, is learned based on different signals. In the first row, the map is based on pure monaural sounds. Monaural sounds integrating prior information are used to build the learned map for the second row. In the third row, the map is based on pure binaural sounds. Binaural sounds integrating prior information are used to build the learned map for the fourth row.

performance is again improved (gain: 0.67, bias: −1.82, score: 0.51) and for binaural sounds integrating prior information the localization performance is close to the computational model (gain: 0.87, bias: −0.76, score: 0.81).

Discussion

We propose that binaural signal integration can solve the ill-posed problem of vertical sound source localization. Model results demonstrate that integrating signals form the left and right ear improves localization of sounds without the need for prior information about the spectrum. If only monaural signals are available, as tested in several behavioural experiments [51], sound source localization remains difficult. However, when sound type specific prior information is integrated, localization performance of monaural sounds is restored. In addition to these experimental results, the structure of

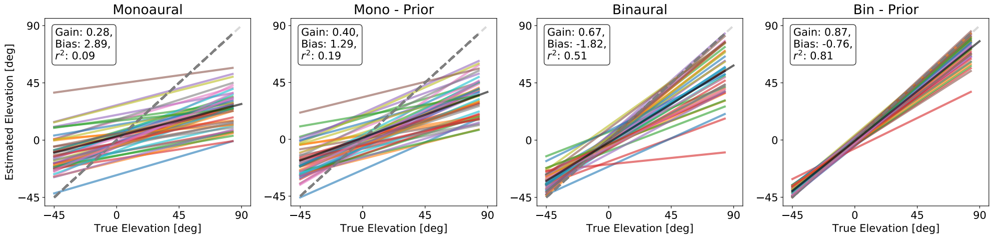


Figure 5. Neural Network Elevation Estimates. Model estimates for a single participant (no. 8 CIPIC). Each dot represents one sound source elevation with color indicating sound type X-axis indicates the elevation of the presented sound. Y-axis is the model elevation estimate for a sound. Pure monaural sounds are presented in left panel. Middle-left panel shows model estimate for monaural sounds integrating prior information. Pure binaural sounds are presented in middle-right panel. Right panel depicts model estimate for binaural sounds integrating prior information. Regression values are shown in inset box.

the model also provides a hint on when the integration of the signals from the left and right ear are integrated. Therefore, the proposed model for binaural integration provides an excellent computational basis for understanding vertical sound source localization and guides future behavioral and physiological experiments.

Implications on the current view of vertical sound source localization

Findings of Wightman and colleagues [51] questioned the results of several previous experiments on monaural localization performance and demonstrated that when participants are presented with a pure monaural signal over headphones, localization is basically not possible. The authors suggested that in previous experiments with contradicting results, the occlusion of one ear was not sufficient to block all informative signals or that small head movements have been used to localize a sound, therefore This is inline with our results from the first experiment which demonstrates that localization is essentially eliminated for monaural sounds without integrating any further information.

However, these findings are in contrast to the results of [18], in which they presented participants with white noise and presumably unfamiliar filtered noise to test the localization performance under monaural and binaural conditions for known and unknown sounds, respectively. Their results indicate that, the additional binaural information in the binaural condition does not improve localization performance and that known and unknown sounds are localized equally well. Our results clearly indicate that unknown sounds are basically impossible to localize under monaural conditions. Thus, we believe that these behavioral results are misleading because of two major factors: The method to occlude one ear might not be sufficient, as already pointed out by [50], to ensure pure monaural information. The second factor is the choice of the unknown sound, which is a filtered white noise stimuli, with random peaks and notches similar the ones provided by the HRTF. However, this is not necessarily an unknown sound but might merely lead to a confusion between sounds from different elevations. Therefore, we are planning on implementing a new behavioral experiment that avoids these two factors by providing virtual sounds over headphones and applying a stimuli which is indeed unknown. If our model predictions hold true, unknown monaural sounds will be very difficult to localize. Though, unknown binaural sounds should be localized accurately and quickly.

Hofman and van Opstal [19] already suggested that the elevation estimation is facilitated by a binaural interaction of the left and right ear. They introduced two conceptual schemes for this interaction. However, it is still unclear which of the scheme is applied [48]. Our model architecture and results from the second experiment demonstrate that binaural integration is most likely taking place before the computation of an elevation estimate (spatial mapping stage), since it enables the system to extract unique elevation dependent cues and remove unnecessary source spectrum induced spectral information.

The process of binaural signal integration is an integral part of horizontal sound source localization and provides major cues like interaural level or time difference. The fundamental principals used for the computation of these cues are similar and are based on the integration of excitatory and inhibitory input signal [6, 15, 16, 52]. It is therefore plausible that the process of binaural integration, as shown in our model, is adapted to provide distinct cues for vertical sound sources.

Prior Information Another major finding of our model is that the integration of sound type specific prior information facilitates monaural sound source localization as well as it improves binaural localization performance. By learning sound type specific prior information, which consists of the mean frequency components over elevations, localization performances for all conditions are improved (see Fig. 4). In our model, we assume that this prior information is learned in higher layers of auditory processing, which are able to identify a sound or at least categorize it (as has been observed in [2, 17]) and provided by feedback connections [25, 37, 41]. If this is the case one could measure a difference in localization speed between monaural and binaural sounds, since monaural sounds can be localized only after they have been categorized and a feedback signal has been sent back. However, binaural signals can be localized immediately without the use of prior information, the prior information just increases accuracy.

Neural implementation In a last experiment we introduced a neural implementation of our computational model, that implements different neuron populations and interactions of excitatory and inhibitory signals among them to replicate computations of the computational model in a biologically plausible fashion. In [30] the authors investigated typical neural responses of neurons in the dorsal cochlear nucleus to stimuli with spectral notches and discovered they these neurons already show a sensitivity to spectral notches. Our proposed model is similar to their type II and type IV neurons in a sense that it receives excitatory inputs from the best frequency of a neuron and inhibitory inputs from neighbouring frequency bands (wide-band inhibition, see Fig. 6). Similar investigations of the inferior colliculus have found neurons that specialized in processing directional dependent features of the HRTF [9]. Our neural model follows these findings and additionally, assumes inhibitory input to neurons in the inferior colliculus from the contralateral side to enable binaural integration. Such connections have been shown to exist [41]. The fact that in the neural model the same neuron parameters are used for all participants demonstrates on the one hand the robustness of the model and on the other hand provides a possibility to improve the performance for each participant by tuning the neuron parameters specifically for each participant.

The presented experiments for monaural and binaural sounds challenges the view on the fundamentals of vertical sound source localization. We propose that vertical sound source localization takes advantage of binaural signal integration, which can be found throughout the early auditory pathway, in every day situations but is capable of localizing monaural signals providing they have been heard (learned) beforehand.

Methods

Input data creation

Inputs $S^s_{m,i}$ to the model are generated by, first, convolving a mono sound signal $x_i(t)$ of sound type i with recorded head-related impulse responses (HRIR), separately for the ipsi- $s = Ipsi$ and contralateral ear $s = contra$ of listener m provided by the CIPIC database [1] to model simulated sound signals arriving at the cochlea

$$I^s_{m,i}(\epsilon, t) = HRIR^s_m(\epsilon, t) * x_i(t) \cdot (1 - \eta) + \eta \cdot (x_i(t) + \mathcal{U}(0, 1) \cdot \eta)), \tag{1}$$

where $*$ is the convolution of two signals, $U(0, 1)$ the uniform distribution and η describes the signal-to-noise ratio and is commonly set to 0.2. The input noise of the data is modeled so that a part of the original, unfiltered signal ($x_i(t)$ in second term) is perceived together with random noise ($\mathcal{U}(0, 1)$). For the influence of the signal-to noise ratio parameter on the localization ability see supplementary Fig. 2.

The cochlea response over frequencies for a perceived sound signal can be simulated using gammatone-filter banks [36]. This transformation from time into frequency domain is implemented by using a python implementation (https://github.com/detly/gammatone) of the auditory toolbox [43] with 128 frequency bands, window length of $twin = 0.1s$ and step time $thop = \frac{twin}{2}$. Thus, each signal $I^s_{m,i}$ at the eardrum is transformed to its frequency domain by

$$\bar{S}^s_{m,i}(\epsilon, f, t) = GBF(I^s_{m,i}(\epsilon, t)), \tag{2}$$

where GBF is the gammatone-filter bank as described in [43]. The resulting spectrum is set to be in range $[20, 20000]Hz$ to resemble the perceivable range of humans [35]. After this filtering step, the log power of the signal is calculated by $S^s_{m,i}(\epsilon, f, t) = 20 \cdot log_{10}(\bar{S}^s_{m,i,t}(\epsilon, f) + 1)$. This power spectrogram is averaged over time to for the final spectrum of the perceived sound

$$S^s_{m,i}(\epsilon, f) = \frac{1}{k} \sum_{t=0}^{k} \bar{S}^s_{m,i}(\epsilon, f, t) \tag{3}$$

with k the number of time steps calculated by the gammatone filter bank.

To provide signals with similar energy levels each spectrum is normalized over frequencies:

$$S^s_{m,i}(\epsilon, f) = \frac{S^s_{m,i}(\epsilon, f)}{\sum_{i=0}^{128} S^s_{m,i}(\epsilon, f_i)}, \tag{4}$$

These transformation steps are separately initiated for each listener (45 HRIR from the CIPIC database), each sound type (in total 20 different sounds) and each elevation (25 in total) ranging from $[-45°, +90°]$ in 5.625° steps on the median plane.

Sounds

Sounds that are used for the presented experiments can be found under `TODO`

Model Description

Two different version of the binaural integration model were simulated: a computational model that uses a sequence of mathematical operations and a neural model that is based on different neuron populations implementing similar operations as the first model. Response of each neuron in such populations is described by a first-order differential equation of its membrane potential. This model is provided to demonstrate the biologically plausibility of our model. If not stated otherwise all presented results are based on the computational model.

Computational Model The basic model consists of three consecutive processing layers with an optional layer for the integration of prior information, which is used only in "prior" conditions or when a prior integrating map is learned.

The first layer in the model is a normalization layer that receives the frequency signal $S^s_{m,i}(\epsilon, f)$ as an input and normalizes it with a Gaussian-filtered version of it self

$$\hat{S}^s_{m,i}(\epsilon, f) = \frac{S^s_{m,i}(\epsilon, f)}{S^s_{m,i}(\epsilon, f) * \Lambda(f)} \tag{5}$$

where $\Lambda(f)$ is a Gaussian kernel with $\sigma = 1$ (see supplementary Fig. 3). This normalization already provides signals with prominent spectral cues.

The optional prior integration step calculates a sound type specific prior by averaging these filtered signals over elevations:

$$p_i(\hat{S}^s) = \frac{1}{n}\sum_{j=0}^{n} \hat{S}^s_{m,i}(\epsilon_j, f), \tag{6}$$

Such prior information is used in the "prior" conditions to effectively remove sound type specific peculiarities in the perceived frequency spectrum. Thus, enabling monaural localization. It is combined with the filtered sound signal by a simple division $(\hat{S}^s_{m,i}(\epsilon, f)/p_i(\hat{S}^s))$. Note, that this step is omitted for conditions in which no prior information is considered.

These signals from the ipsi- and contralateral side are integrated (by a division) in the integration layer to provide a binaural signal $S^b(\epsilon, f)$:

$$S^b_{m,i}(\epsilon, f) = \frac{\hat{S}^{Ipsi}_{m,i}(\epsilon, f)}{\hat{S}^{Contra}_{m,i}(\epsilon, f)}. \tag{7}$$

This step effectively removes sound type specific information in the signal so that only HRTF induced frequency modulations remain, making it simple for the model to localize such signals. The resulting signal is normalized over frequencies to ensure values in a feasible range $S^b_{m,i}(\epsilon, f) = \frac{S^b_{m,i}(\epsilon, f)}{\sum_{j=1}^{128} S^b_{m,i}(\epsilon, f_j)}$.

Finally, the output layer performs a cross correlation of either $\hat{S}^s_{m,i}$ for the monaural (omitting the prior integration step) and monaural-prior conditions or $S^b_{m,i}$ for the binaural (omitting the prior integration step) and binaural-prior conditions with a previously learned map $M_m(\epsilon, f)$ to estimate the elevation $\epsilon*$ of the perceived sound source

$$\epsilon^* = \underset{\epsilon}{\operatorname{argmin}} \left[xcorr\big(S_{m,i}(\epsilon, f), M_m(\epsilon, f)\big) \right] \tag{8}$$

The learned map $M_m(\epsilon, f)$ for a participant is previously constructed by averaging over all presented sound types

$$M_m(\epsilon, f) = \frac{1}{n}\sum_{j=0}^{n} S_{j,m}(\epsilon, f), \tag{9}$$

here, S again depends on which condition is tested. For the monaural and monaural-prior conditions $S = \hat{S}^s$ and for the binaural and binaural-prio conditions $S = S^b$.

Neural Model The following neural model for elevation estimation is based on the computational layers introduced with the computational model (see Fig.6). Each layer is realized with one or two populations of N neurons selective to frequency band f which are modeled by a first-order differential equation of the neuron's membrane potential. This membrane potential is transformed to a firing rate by an activation function $g(\bullet)$ which is a simple linear rectified function

$$g(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x > 1, \\ x, & \text{else,} \end{cases} \tag{10}$$

with saturation level of 1.

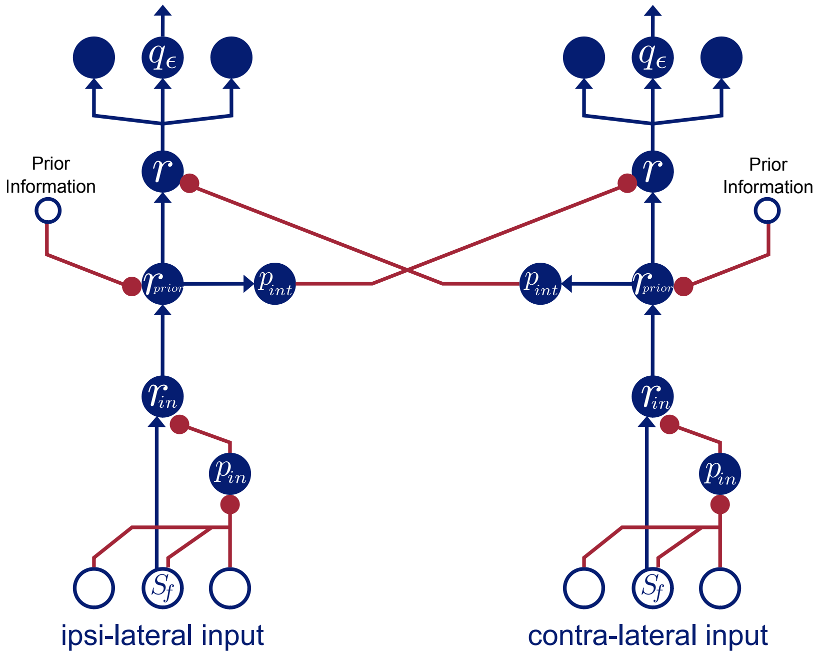


Figure 6. Neural Network Architecture. Blue filled circles indicate model neurons. Blue empty circles represent model inputs. Blue arrow-headed connections are excitatory and red bullet-headed connections are inhibitory connections from inputs to neurons and from neurons to other neurons, respectively.

The core of the model is an integration population that receives excitatory input from neurons of the ipsilateral side and inhibitory inputs from neurons of the contralateral side, thus performs a binaural signal integration

$$\tau \dot{r}_f^{Bin} = -\alpha_d \cdot r_f^{Bin} \cdot g(r_{in,f}^{s}) \tag{11}$$

$$+ (\beta_{r^{Bin}} - r_f^{Bin}) \cdot g(r_{in,f}^{s}) - \kappa_{r^{Bin}} \cdot r_f^{Bin} \cdot g(p_{sum,f}^{Contra}) \tag{12}$$

here, parameter τ defines the membrane capacitance, α_d is a default passive membrane leak conductance, $\beta_{r^{Bin}}$ describes a saturation level of excitatory inputs and $\kappa_{r^{Bin}}$ define the divisive influence of the inhibitory input. A special feature of the neuron is that the the input modulates the decay rate of the neuron so that higher inputs lead to a faster decay which leads an alignment of signals with very different intensities. Such a generic neuron model has been previously demonstrated to resemble typical neuronal response and to successfully solve a variety of tasks [24, 32, 33, 45]. Since such a model approach does not lead to specific voltage traces of neurons the nomenclature differs from typical electrophysiological descriptions but is in line with previous computational models [7, 14, 34, 40].

The inhibitory input $p^{Contra}_{sum,f}$ is modeled by an intermediate inhibitory population of the contralateral side

$$\tau \dot{p}^s_{sum,f} = -\alpha_d \cdot p^s_{sum,f} + (\beta_d - p^s_{sum,f}) * g(r^s_{prior,f}) \tag{13}$$

The input $r^s_{prior,f}$ to such neurons is provided by a population of so called prior integration neurons and is modeled by

$$\tau \dot{r}^s_{prior,f} = -\alpha_d \cdot r^s_{prior,f} + (\beta_{r_{prior}} - r^s_{prior,f}) \cdot g(r^s_{in,f}) - \gamma_{r_{prior}} \cdot g(\bar{w}^{s,prior}_{f,\epsilon}) \tag{14}$$

These neurons receive, presumably, cortical inhibitory input which is the mean over elevations based on a previously learned, sound type specific signal $w^{s,prior}_{f,\epsilon}$ for the ipsi- and contralateral side, respectively.

Similarly, the excitatory input $g(r^s_{in,f})$ is modeled by neurons at side s

$$\tau \dot{r}^s_{in,f} = -\alpha_d \cdot r^s_{in,f} \cdot In^s_f + (\beta_{r_{in}} - r^s_{in,f}) \cdot In^s_f - \kappa_{r_{in}} \cdot r^s_{in,f} \cdot g(p^s_{in,f}) \tag{15}$$

This population of neurons in the neural model realizes the Gaussian normalization layer of the computational model by integrating inhibitory inputs from an inhibitory input population

$$\tau \dot{p}^s_{in,f} = -\alpha_d \cdot p^s_{in,f} + (\beta_d - p^s_{in,f}) \cdot \sum_{f'=1}^{128} In^s_f \cdot \Lambda_{f'f} \tag{16}$$

The input kernel $\Lambda_{f'f}$ enables an integration of inputs over several frequency bands f and is defined as $\Lambda_{f'f} = \exp(-\frac{(f-f')^2}{2 \cdot \sigma^2})$ with $\sigma = 3$.

To ensure valid input values to the neural model, the input $S^s_{m,i}(\epsilon, f)$ over frequency band f for a single participant m, elevation ϵ and a sound type i is normalized by

$$In^s_f = \frac{S^s_{m,i}(\epsilon, f)}{\sum_{j=0}^{128} S^s_{m,i}(\epsilon, f_j)}, \tag{17}$$

again $s \in \{Ipsi, Contra\}$ depending on the input side.

To estimate the elevation of a perceived sound source the final readout layer of the network is defined as a set of 25 neurons q^{Bin}_ϵ, each tuned to a certain elevation ϵ

$$\tau \dot{q}^{Bin}_\epsilon = -\alpha_d \cdot q^{Bin}_\epsilon + (\beta_d - q^{Bin}_\epsilon) \cdot \sum_{f=1}^{128} r^{Bin}_f \cdot w_{f\epsilon} \tag{18}$$

It receives excitatory inputs form the binaural integration layer and integrates them according to a previously learned weight kernel $w_{f\epsilon}$. For an elevation estimate the index ϵ^* of the neuron with maximal activity is determined

$$\epsilon^* = \underset{\epsilon}{\operatorname{argmax}}(q^{Bin}_\epsilon) \tag{19}$$

All presented results of the neural network model are calculated from the network responses readout at a single neuron level after keeping the input stimuli constant for at least 3000 time steps. This duration is sufficient for the neuron to dynamically converge to its equilibrium membrane potential of numerical integration of the state equations. For the numerical integration of the state equations we chose Euler's method with a step size of $\Delta t = 0.0001$ (for details see [46]).

Learning The weight kernel $w_{f\epsilon}$ is learned using a supervised learning approach similar to instar learning [13].

$$\Delta w_{f,\epsilon} = \eta \cdot (r_f^{Bin} - w_{f\epsilon}) \cdot v \tag{20}$$

where $\eta = 0.00005$ is the learning rate and v is a vector of 25 entries, one for each elevation and is assumed to provide a visual guidance signal. That is, for a sound signal arriving from elevation ϵ entry v_ϵ of the vector is set to 1 while all other entries remain 0.

The sound type specific prior is learned separately for the ipsi- and contralateral side and is based on the activation of the prior integration neurons:

$$\delta w_{f,\epsilon}^{s,prior} = \eta \cdot (r_{prior,f}^{s} - w_{f,\epsilon}^{s,prior}) \cdot v \tag{21}$$

here, the values of η and v are set as described above.

The learning phase consists of 15000 trials. In each trial a sound signal from a randomly chosen elevations and sound type is presented to the model. After this learning phase the weights are normalized over frequencies to ensure similar energy content ($w_{f,\epsilon} = w_{f,\epsilon} / \sum_{i=0}^{128} w_{f_i,\epsilon}$). Subsequently, localization performance is tested by presenting all sound signals to the model and calculating the elevation response ϵ^*. For this localization phase η is set to 0 to disable learning.

Table 1. Model parameters

General Parameters			
N (# Neurons)	128		
σ_{Kernel}	3		
τ_d	0.005	α_d	1.0
β_d	1.0		
Excitatory Input Neuron r_{in}			
$\beta_{r_{in}}$	200.0	$\kappa_{r_{in}}$	200.0
Prior Integration Neuron r_{prior}			
$\beta_{r_{prior}}$	2.0	$\gamma_{r_{prior}}$	1.0
Integration Neuron r^{Bin}			
$\beta_{r^{Bin}}$	2.0	$\kappa_{r^{Bin}}$	1.0

Funding

This research has been conducted as part of the VA-MORPH project financed by the Baden-Württemberg foundation in the Neurorobotik program (project no. NEU012).The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

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

2. B. Bathellier, L. Ushakova, and S. Rumpel. Discrete neocortical dynamics predict behavioral categorization of sounds. *Neuron*, 76(2):435–449, 2012.

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

4. K. Binns, S. Grant, D. Withington, and M. Keating. A topographic representation of auditory space in the external nucleus of the inferior colliculus of the guinea-pig. *Brain research*, 589(2):231–242, 1992.

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

6. J. C. Boudreau and C. Tsuchitani. Binaural interaction in the cat superior olive s segment. *Journal of neurophysiology*, 31(3):442–454, 1968.

7. T. Brosch and H. Neumann. Computing with a Canonical Neural Circuits Model with Pool Normalization and Modulating Feedback. *Neural Computation*, 26(12):2735–2789, Sept. 2014.

8. Y. E. Cohen, E. I. Knudsen, Y. E. Cohen, E. I. Knudsen, Y. E. Cohen, and E. I. Knudsen. Maps versus clusters: Different representations of auditory space in the midbrain and forebrain. *Trends in Neurosciences*, 22(3):128–135, Mar. 1999.

9. K. A. Davis, R. Ramachandran, and B. J. May. Auditory processing of spectral cues for sound localization in the inferior colliculus. *JARO - Journal of the Association for Research in Otolaryngology*, 4(2):148–163, 2003.

10. I. DeWitt and J. P. Rauschecker. Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, 109(8):E505–E514, 2012.

11. R. Ege, A. J. Van Opstal, and M. M. Van Wanrooij. Accuracy-precision trade-off in human sound localisation. *Scientific reports*, 8(1):1–12, 2018.

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

13. S. Grossberg. Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological cybernetics*, 23(3):121–134, 1976.

14. S. Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural networks*, 1(1):17–61, 1988.

15. B. Grothe and U. Koch. Dynamics of binaural processing in the mammalian sound localization pathway - The role of GABA B receptors. *Hearing Research*, 279(1-2):43–50, 2011.

16. B. Grothe and D. H. Sanes. Bilateral inhibition by glycinergic afferents in the medial superior olive. *Journal of Neurophysiology*, 69(4):1192–1196, 1993.

17. F. H. Guenther, A. Nieto-Castanon, S. S. Ghosh, and J. A. Tourville. Representation of sound categories in auditory cortical maps. *Journal of Speech, Language, and Hearing Research*, 2004.

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

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

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

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

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

23. N. Kraus and M. Cheour. Speech sound representation in the brain. *Audiology and Neurotology*, 5(3-4):140–150, 2000.

24. D. McLaughlin, R. Shapley, M. Shelley, and D. J. Wielaard. A neuronal network model of macaque primary visual cortex (v1): Orientation selectivity and dynamics in the input layer 4cα. *Proceedings of the National Academy of Sciences*, 97(14):8087–8092, 2000.

25. J. G. Mellott, M. E. Bickford, and B. R. Schofield. Descending projections from auditory cortex to excitatory and inhibitory cells in the nucleus of the brachium of the inferior colliculus. *Frontiers in Systems Neuroscience*, 8:188, 2014.

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

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

28. M. Morimoto. The contribution of two ears to the perception of vertical angle in sagittal planes. *The Journal of the Acoustical Society of America*, 109(4):1596–1603, Mar. 2001.

29. A. D. Musicant, J. C. Chan, and J. E. Hind. Direction-dependent spectral properties of cat external ear: New data and cross-species comparisons. *The Journal of the Acoustical Society of America*, 87(2):757–781, 1990.

30. I. Nelken and E. Young. Two separate inhibitory mechanisms shape the responses of dorsal cochlear nucleus type IV units to narrowband and wideband stimuli. *Journal of neurophysiology*, 71(31edc418-e21d-8882-e919-5c18733e14e2):2446–2462, 1994.

31. C. Neti, E. D. Young, and M. H. Schneider. Neural network models of sound localization based on directional filtering by the pinna. *The Journal of the Acoustical Society of America*, 92(6):3140–3156, 1992.

32. T. Oess, M. Ernst, and H. Neumann. Computational investigation of visually guided learning of spatially aligned auditory maps in the colliculus. *Proceedings of the International Symposium on Auditory and Audiological Research*, 7:149–156, Apr. 2020.

33. T. Oess, M. O. Ernst, and H. Neumann. Computational principles of neural adaptation for binaural signal integration. *PLOS Computational Biology*, 16(7):1–23, 07 2020.

34. T. Oess, M. P. R. Löhr, D. Schmid, M. O. Ernst, and H. Neumann. From near-optimal bayesian integration to neuromorphic hardware: A neural network model of multisensory integration. *Frontiers in Neurorobotics*, 14:29, 2020.

35. H. F. Olson. *Music, physics and engineering*, volume 1769. Courier Corporation, 1967.

36. R. D. Patterson, M. H. Allerhand, and C. Giguere. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4):1890–1894, 1995.

37. D. C. Peterson and B. R. Schofield. Projections from auditory cortex contact ascending pathways that originate in the superior olive and inferior colliculus. *Hearing research*, 232(1-2):67–77, 2007.

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

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

40. E. Salinas. Background synaptic activity as a switch between dynamical states in a network. *Neural computation*, 15(7):1439–1475, 2003.

41. C. Schreiner and J. A. Winer. *The inferior colliculus*. Springer, 2005.

42. B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107(3):1627–1636, 2000.

43. M. Slaney. Auditory toolbox. *Interval Research Corporation, Tech. Rep*, 10(1998), 1998.

44. W. H. Slattery III and J. C. Middlebrooks. Monaural sound localization: acute versus chronic unilateral impairment. *Hearing research*, 75(1-2):38–46, 1994.

45. D. C. Somers, E. V. Todorov, A. G. Siapas, L. J. Toth, D.-S. Kim, and M. Sur. A local circuit approach to understanding integration of long-range inputs in primary visual cortex. *Cerebral cortex (New York, NY: 1991)*, 8(3):204–217, 1998.

46. E. Süli and D. F. Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003.

47. H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida. Mechanism for generating peaks and notches of head-related transfer functions in the median plane. *The Journal of the Acoustical Society of America*, 132(6):3832–3841, 2012.

48. M. M. V. Wanrooij and A. J. V. Opstal. Relearning Sound Localization with a New Ear. *Journal of Neuroscience*, 25(22):5413–5424, June 2005.

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

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

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

52. T. C. T. Yin. Neural Mechanisms of Encoding Binaural Localization Cues in the Auditory Brainstem. *Integrative Functions in the Mammalian Auditory Pathway. Berlin: Springer-Verlag*, 15:99–159, 2002.

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