Multisensory Research

Explaining the Effect of Likelihood Manipulation and Prior through a Neural Network of the Audiovisual Perception of Space

Mauro Ursino [1,*], Cristiano Cuppini [1], Elisa Magosso [1], Ulrik Beierholm [2] and Ladan Shams [3]

[1] Department of Electrical, Electronic and Information Engineering, University of Bologna, Bologna, Italy

[2] Department of Psychology, Durham University, United Kingdom

[3] Department of Psychology, Department of BioEngineering, Interdepartmental Neuroscience Program, University of California, Los Angeles, CA, USA

Received 18 April 2018; accepted 4 January 2019

[*] To whom correspondence should be addressed. E-mail: mauro.ursino@unibo.it

Abstract

Results in the recent literature suggest that multisensory integration in the brain follows the rules of Bayesian inference. However, how neural circuits can realize such inference and how it can be learned from experience is still the subject of active research. The aim of this work is to use a recent neurocomputational model to investigate how the likelihood and prior can be encoded in synapses, and how they affect audio-visual perception, in a variety of conditions characterized by different experience, different cue reliabilities and temporal asynchrony. The model considers two unisensory networks (auditory and visual) with plastic receptive fields and plastic crossmodal synapses, trained during a learning period. During training visual and auditory stimuli are more frequent and more tuned close to the fovea. Model simulations after training have been performed in crossmodal conditions to assess the auditory and visual perception bias: visual stimuli were positioned at different azimuth (±10° from the fovea) coupled with an auditory stimulus at various audio-visual distances (±20°). The cue reliability has been altered by using visual stimuli with two different contrast levels. Model predictions are compared with behavioral data. Results show that model predictions agree with behavioral data, in a variety of conditions characterized by a different role of prior and likelihood. Finally, the effect of a different unimodal or crossmodal prior, re-learning, temporal correlation among input stimuli, and visual damage (hemianopia) are tested, to reveal the possible use of the model in the clarification of important multisensory problems.

Keywords

Neural networks, audio-visual integration, Hebbian mechanisms, synapse learning, Bayesian inferences, priors, likelihoods

Supplementary Material

A. Mathematical Description of the Neural Network

1. Basal Structure of the Network

The neural network model consists of two chains of *N* unisensory neurons (Fig. 1, upper panel). Each neuron codes for a particular spatial position in its modality. Moreover, each chain is topologically organized, i.e., proximal neurons code for proximal positions. In the following, we will denote with a first subscript the particular area (auditory or visual) and with a second subscript, after a comma, the neuron position within the area.

Each neuron receives three different kinds of inputs: a sensory input from the environment (say *u*), a lateral input from neurons of the same modality (say *l*) and a cross-modal input from neurons of the other modality (say *c*). The global input (equal to the sum of the previous three contributions) is then passed through a sigmoidal relationship, $\varphi(\;)$, which accounts for the presence of a lower threshold and upper saturation in neuron activity, and a first-order low-pass filter with time constant τ, which accounts for the neuron integrative capacity.

Hence, for the generic *k*-th neuron in the modality *S* (*S* = *A* or *V* for the auditory and visual modalities, respectively) we can write

$$\tau \frac{dy_{S,k}}{dt} = -y_{S,k} + \phi\left(u_{S,k} + l_{S,k} + c_{S,k}\right) \tag{S1}$$

Where $y_{S,k}$ represents the neuron output, and the sigmoidal relationship is described by the following equation

$$\phi(x) = \frac{1}{1 + \exp\left(-s(x - x_0)\right)} \tag{S2}$$

s and *x₀* are parameters, which set the slope and the position of the sigmoidal relationship. According to Eq. (2), the neuron output activity is normalized between 0 and 1 (zero means a silent neuron, one a maximally activated neuron).

It is worth noting that, for the sake of simplicity, we used the same parameters (τ, *s* and x_0) for all neurons independently of their modality. This choice was adopted to minimize the number of model assumptions.

The expression for the sensory input is computed as the scalar product of the sensory representation of the stimulus (i.e., the vector $I_s = [i_{S,1}\ i_{S,2} \cdots i_{S,k} \cdots i_{S,N}]^T$) and the neuron receptive field (i.e., the vector $R_{s,k} = [r_{S,k1}\ r_{S,k2} \cdots r_{S,kj} \cdots r_{S,kN}]^T$) :

$$u_{S,k} = \left\langle R_{s,k}\, ,\ I_s \right\rangle = \sum_{j=1}^{N} r_{S,kj}\ i_{S,j} \tag{S3}$$

We assumed that the neuron receptive field, $R_{S,k}$, has initially a large extension, described with a Gaussian function, and then progressively shrinks during training, to fit the width of the external input (see section “Training the network”).

The lateral input is computed as follows

$$l_{S,k} = \sum_{j=1}^{N} \lambda_{kj} y_{S,j} \tag{S4}$$

where λ_{kj} represents a lateral intra-area synapse connecting the presynaptic neuron *j* to the post synaptic neuron *k* in the same area. Here we used the classical Mexican-hat arrangement: a neuron is excited by proximal neurons in the same area, and inhibited by more distal ones

$$\lambda_{kj} = \lambda_{ex} \exp\left(-\frac{d(\theta_j,\theta_k)^2}{2\sigma_{ex}^2}\right) - \lambda_{in} \exp\left(-\frac{d(\theta_j,\theta_k)^2}{2\sigma_{in}^2}\right) \qquad \text{(S5)}$$

where $\lambda_{ex}, \lambda_{in}, \sigma_{ex}, \sigma_{in}$ are parameters which set the strength and width of the excitatory and inhibitory portions of the Mexican hat. In particular, we have $\lambda_{ex} > \lambda_{in}$ and $\sigma_{ex} < \sigma_{in}$. Moreover, $d(\theta_j, \theta_k)$ represents the distance between neurons' preferred positions, i.e.

$$d(\theta_j,\theta_k) = |\theta_j - \theta_k| \qquad \text{(S6)}$$

It is worth noting that we used the same expression of lateral synapses (Eq. (S5)) in both the auditory and visual areas, to limit the number of model assumptions.

Finally, the cross-modal term in Eq. (1) is computed as the convolution of the vector of cross modal synapses and the activity in the other unisensory area, i.e.

$$c_{S,k} = \sum_{j=1}^{N} w_{SQ,kj}\, y_{Q,j} \qquad \text{with } \; S = A \,\text{or}\, V \quad Q = A \,\text{or}\, V \; \text{ with } \; S \neq Q \qquad \text{(S7)}$$

where $w_{SQ,kj}$ represents a cross-modal synapse from the pre-synaptic neuron *j* in the area *Q* to the post-synaptic neuron *k* in the area *S*. We assumed that the cross-modal synapses are initially ineffective and are progressively reinforced during the training phase.

2. Training the Network

Starting from the initial basal value of synapses, the network has been trained during a training period in which the sensory input representations (i.e., I_A and I_V) have been given with a random distribution.

The synapses describing the receptive field, $r_{S,kj}$, and those describing the cross-modal link between the two areas, $w_{SQ,kj}$, have been trained using a learning rule with a classical Hebbian potentiation factor and a decay term. We can write, in scalar form

$$\Delta r_{S,kj} = \gamma\, y_{S,k}\left(i_{S,j} - r_{S,kj}\right) \qquad \text{with} \quad S = A,\ V \tag{S8}$$

$$\Delta w_{SQ,kj} = \gamma\, y_{S,k}\left(y_{Q,j} - w_{SQ,kj}\right) \qquad \text{with} \quad S = A,\ V \quad Q = A,\ V \quad Q \neq S \tag{S9}$$

Eqs. (8) and (9) have been applied, at each step, using the final steady state values of the neuron output (i.e., when transient phenomena are exhausted).

At the beginning of training all cross-modal synapses are assumed equal to zero. Conversely, the receptive-field synapses have a broad spatial extension, and moderate amplitude, identical for the two modalities, i.e.,

$$r_{S,kj} = r_0 \exp\left(-\frac{\left(d\left(\theta_j, \theta_k\right)\right)^2}{2\sigma_R{}^2}\right) \text{ with } S = A,V \tag{S10}$$

where *r₀* sets the initial strength of the receptive field, and σ_R establishes its initial spatial extension (we assume $\sigma_R > \sigma_A$ and $\sigma_R > \sigma_V$ i.e., a wide initial receptive fields) . Of course, Eq. (10) holds only at the first step of training.

3. Probability Distribution and Spatial Accuracy of the Inputs

According to the previous section, we assumed that the sensory inputs are composed of a deterministic term, which represents the spatial distribution of the input, centered on the stimulus spatial position, and a Gaussian white noise term (zero mean value and assigned standard deviation). Hence

$$i_{S,k}(\theta) = \frac{i_{S,Strength}}{\sqrt{2\pi\sigma_S^2}} \exp\left(-\frac{\left(d\left(\theta_S, \theta_k\right)\right)^2}{2\sigma_S^2}\right) + n_{S,k}(\theta) \quad S = A \text{ or } V \tag{S11}$$

where θ_S represents the spatial position of the stimulus, $i_{S,Strength}$ is the stimulus strength in the absence of noise, and σ_S is the standard deviation of the spatial representation. According to physiology, we assumed that the visual inputs are spatially more accurate than the auditory ones, hence we set $\sigma_V < \sigma_A$. Conversely, we assumed that the standard deviation of noise (say υ_S) is a given fraction of the input strength, to set the signal to noise ratio (see Table 1 in the text).

In order to simulate the presence of better acuity at the center, and reduced acuity at the periphery, we assumed that the SDs of the visual and auditory inputs increase with the eccentricity of the stimulus.

The expression of σ_V has been taken from an empirical curve on visual acuity by Dacey (1993) (see also Ursino *et al.*, 2017 for more details). By denoting with $e_V = \theta_V$ the eccentricity with respect to the fovea, we have

$$\sigma_V(e_V) = \sigma_{V0} + \frac{\varepsilon\sqrt{3}}{60}\left(0.058\, e_V + 0.022\, e_V^2 - 0.00022\, e_V^3\right) \tag{S12}$$

σ_{V0} represents the SD of the visual inputs at the fovea (i.e., at zero eccentricity). We used the same value as in the previous paper, i.e. σ_{V0} = 4 deg. Finally, we use a parameter, ε, to adapt the function so that, σ_V ranges between 4 deg, at 0 eccentricity, to about 12 deg at maximum eccentricity.

The auditory acuity also decreases from the center to the periphery, although it is difficult to quantify this effect being influenced by many factors, such as the stimulus intensity and frequency (Middlebrooks and Green, 1991; Wood and Bizley, 2015). However, this effect is less evident and of smaller entity compared with the visual one (Perrott and Saberi, 1990). Hence, we used a simpler linear relationship, assuming that σ_{A0} linearly increases from about 20 deg at the fovea to 30 deg at the periphery:

$$\sigma_A(e_A) = \sigma_{A0} + 10|e_A|/90 \tag{S13}$$

with $\sigma_0^A = 20$ and $e_A = \theta_A$ is the eccentricity of the auditory position with respect to the head center.

The positions of the two stimuli (i.e., θ_A and θ_V in Eq. (S11)) have been randomly generated from the prior probability distribution described below.

We assume that both the visual and auditory input have a greater probability close to the fovea, and smaller probability at the periphery. This corresponds to have a non-uniform prior in visual unisensory conditions. The following probabilities have been used to generate the position of the visual and auditory inputs during training.

Visual unisensory prior: the visual position follows a Gaussian distribution, centered at the fovea. Hence

$$p(\theta_V) = \frac{1}{\sqrt{2\pi\, s_V^2}} \exp\left(-\frac{\theta_V^2}{2\, s_V^2} \right) \tag{S14}$$

The standard deviation s_V (which here plays the role of a space constant) has been set at 7 deg; i.e., the visual stimuli becomes very rare at ±20 deg eccentricity.

Auditory unisensory prior: the auditory position follows a Gaussian distribution, centered at the head center.

$$p(\theta_A) = \frac{1}{\sqrt{2\pi\, s_A^2}} \exp\left(-\frac{\theta_A^2}{2\, s_A^2} \right) \tag{S15}$$

The standard deviation is assumed higher than in the visual case: we have $s_A = 30$ deg assuming that, head movements in auditory unimodal conditions are less efficient than eye movement in visual unimodal conditions to maintain the stimulus close to the center.

Cross modal prior: in the cross modal case during training, we assumed that the visual and auditory inputs originate from independent causes with a given probability (say β) but are produced by the same cause, hence originate from proximal spatial positions, with the complementary probability (1 – *β*). According to the Bayes rule, the joined prior probability can be computed from knowledge of the individual probability of one stimulus, and the conditional probability of the other. A problem is whether, in cross modal conditions, the distribution is dominated by the visual prior (more sharply close to the center) or by the auditory one (less sharply close to the center). We assumed that, in 50% of cases, the cross-modal stimuli follow the visual distribution and in the other 50% of cases follow the auditory one. Hence

$$p(\theta_V,\theta_A) = 0.5p(\theta_V)p(\theta_A|\theta_V) + 0.5p(\theta_A)p(\theta_V|\theta_A) \tag{S16}$$

where we used equations (S14) and (S15) for the visual and auditory priors, and the following expression for the conditional probability

$$p(\theta_A|\theta_V) = \beta\, p(\theta_A) + (1-\beta)\frac{1}{\sqrt{2\pi\, s_{AV}^2}}\exp\left(-\frac{d(\theta_A,\theta_V)^2}{2\,s_{AV}^2}\right) \tag{S17}$$

$$p(\theta_V|\theta_A) = \beta\, p(\theta_V) + (1-\beta)\frac{1}{\sqrt{2\pi\, s_{AV}^2}}\exp\left(-\frac{d(\theta_A,\theta_V)^2}{2\,s_{AV}^2}\right)$$

In writing Eq. (S17) we assumed that the conditional probability is computed as the weighted sum of the prior unimodal distribution, reflecting the moderate possibility that the two stimuli are independent, and a second term, $\frac{1}{\sqrt{2\pi\, s_{AV}^2}}\exp\left(-\frac{d(\theta_A,\theta_V)^2}{2s_{AV}^2}\right)$, reflecting the probability that the auditory and visual events are originated from the same source. As in the previous work, we used a value of space constant s_{AV} = 1 deg, assuming a small audio-visual distance when the two stimuli originate from the same source.

References

Dacey, D.M. (1993) The mosaic of midget ganglion cells in the human retina. *Journal of Neuroscience*, *13*(12), pp. 5334-5355.

Middlebrooks, J.C. and Green, D.M. (1991) Sound localization by human listeners. *Annual review of psychology*, *42* (1), pp. 135-159.

Perrott, D.R. and Saberi, K. (1990) Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, *87*(4), pp. 1728-1731.

Ursino, M. *et al.* (2017a) Development of a Bayesian Estimator for Audio-Visual Integration: A Neurocomputational Study. *Frontiers in computational neuroscience,* 11, 89.

Wood, K.C. and Bizley, J.K. (2015) Relative sound localisation abilities in human listeners. *The Journal of the Acoustical Society of America*, *138*(2), pp. 674-686.

B. Further Results

1. Cross-modal Prior

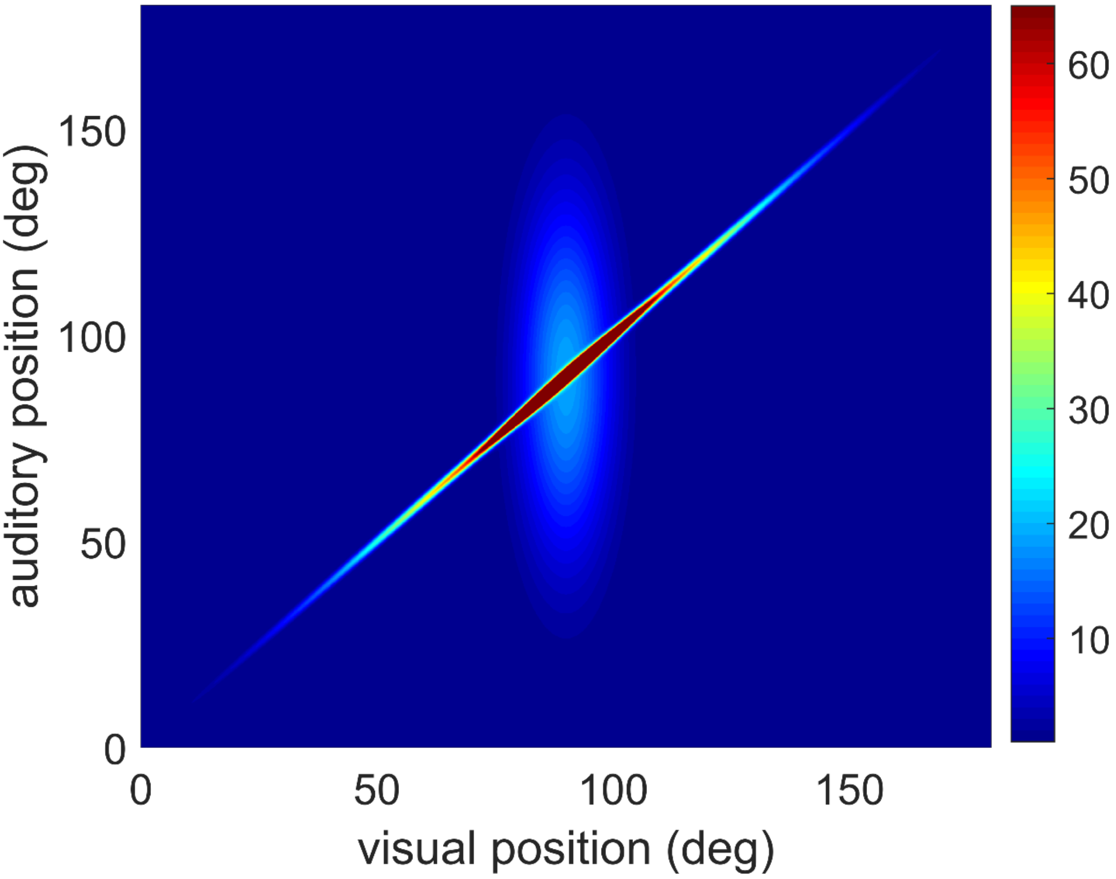

Figure S1. 2D color map of the joint cross-modal probability (i.e., Supplementary Eq. (16)) obtained using $\beta = 0.5$ (50% probability of independent inputs in cross-modal conditions), $s_V = 7$ deg (standard deviation of the visual unisensory prior), $s_A = 30$ deg ((standard deviation of the visual unisensory prior) and $s_{AV} = 1$ deg (standard deviation of the conditioned probability in case of stimuli originating from the same cause).

2. Correlation among Experimental and Model Data

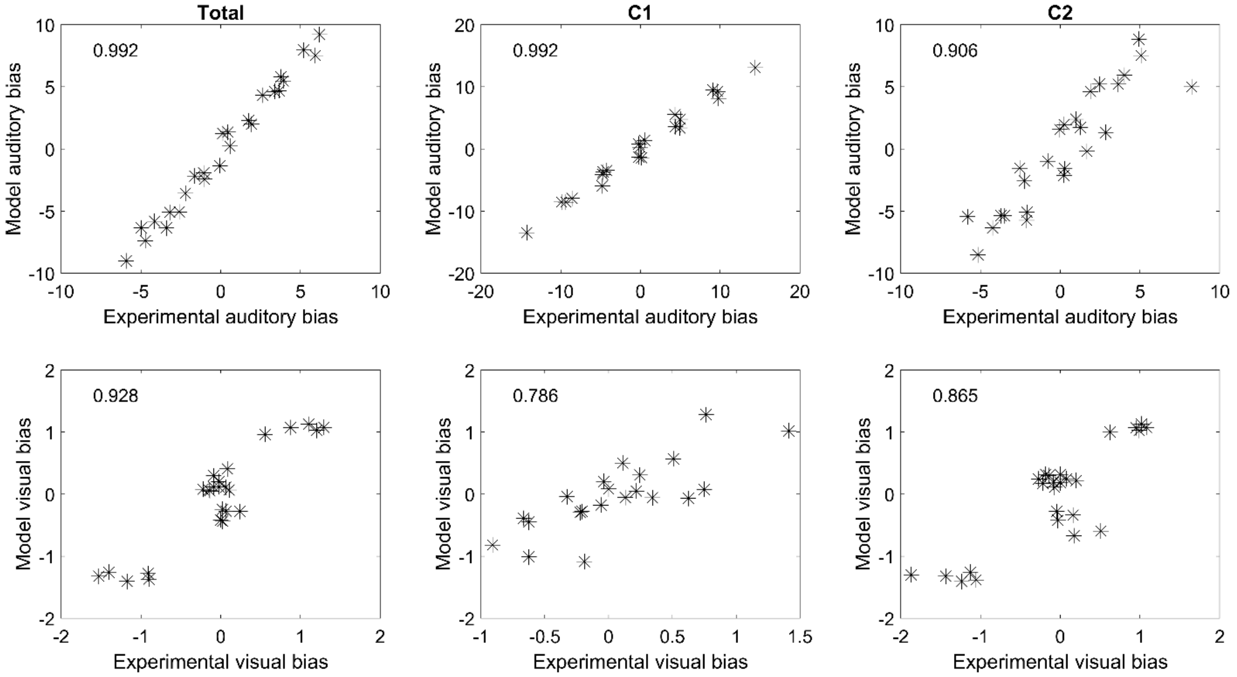

Figure S2. Correlation among the experimental and model values of the auditory bias (upper line) and the visual bias (bottom line), evaluated in the *high contrast condition* using all data available (first column) and considering just the cases with C = 1 (second column) and C = 2 (third column). In the figures, only data simultaneously available from both the experimental conditions and model simulations are reported. These data are taken from the six panels in Fig. 5 (for what concerns the auditory bias) and from the six panels in Fig. 6 (for what concerns the visual bias). The value of the correlation coefficient is reported in each panel.

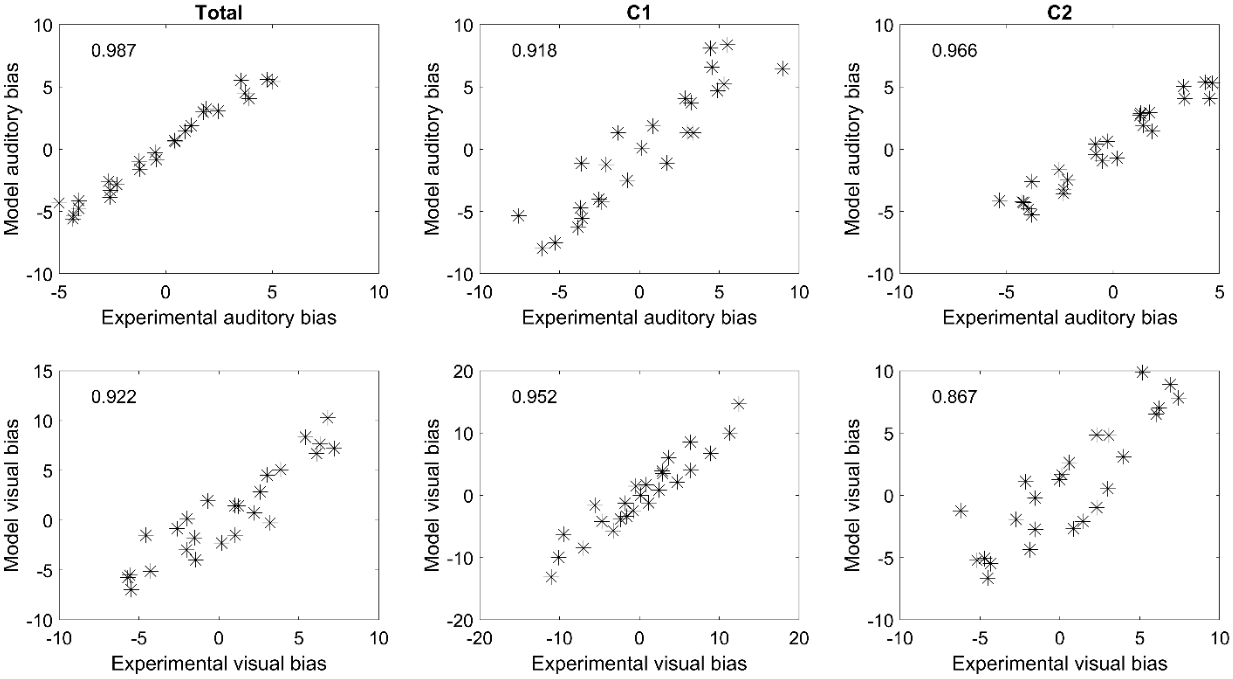

Figure S3. Correlation among the experimental and model values of the auditory bias (upper line) and the visual bias (bottom line), evaluated in the *low contrast conditions* using all data available (first column) and considering just the cases with C = 1 (second column) and C = 2 (third column). In the figures, only data simultaneously available from both the experimental conditions and model simulations are reported. These data are taken from the six panels in Fig. 7 (for what concerns the auditory bias) and from the six panels in Fig. 8 (for what concerns the visual bias). The value of the correlation coefficient is reported in each panel.

3. Sensitivity Analysis at Low-contrast

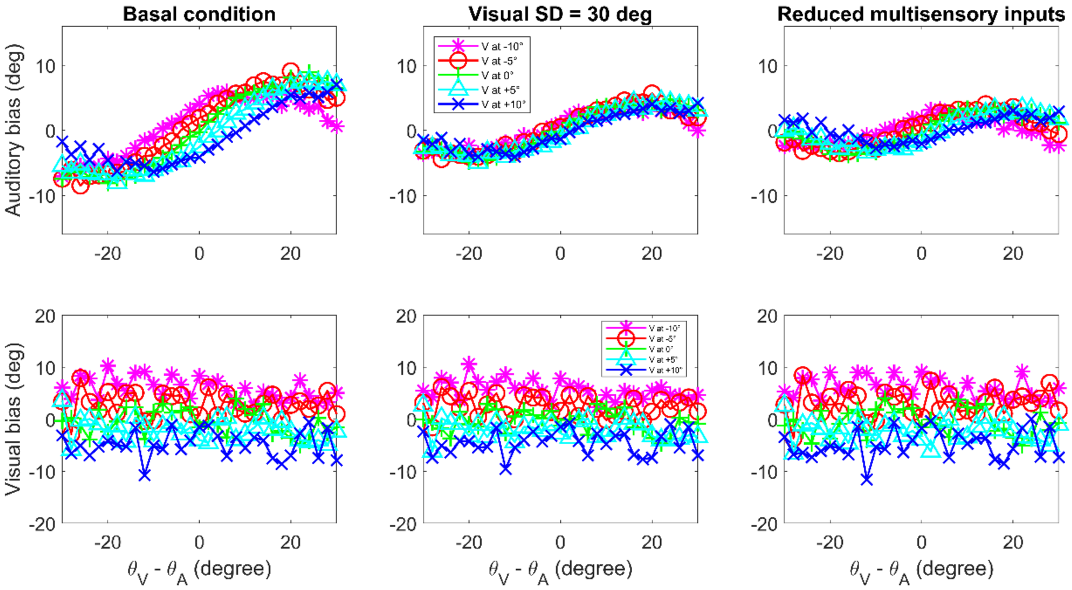

Figure S4. Dependence of model results on the stimuli experienced during training (i.e., on the *prior probability*) in *low-contrast conditions*. The upper panels show the bias in the perceived position of the auditory stimulus; the bottom panels the bias in the visual perception. The meaning of lines is the same as in Fig. 9. The first column was obtained after *Training1* (that is the same used in Figs. 2-8). The second column was obtained after a different training (*Training2*) characterized by a larger spatial arrangement of visual stimuli around the fovea. The third column was obtained after *Training3*, characterized by a smaller percentage of cross-modal inputs.

4. Synapse Changes during Re-learning

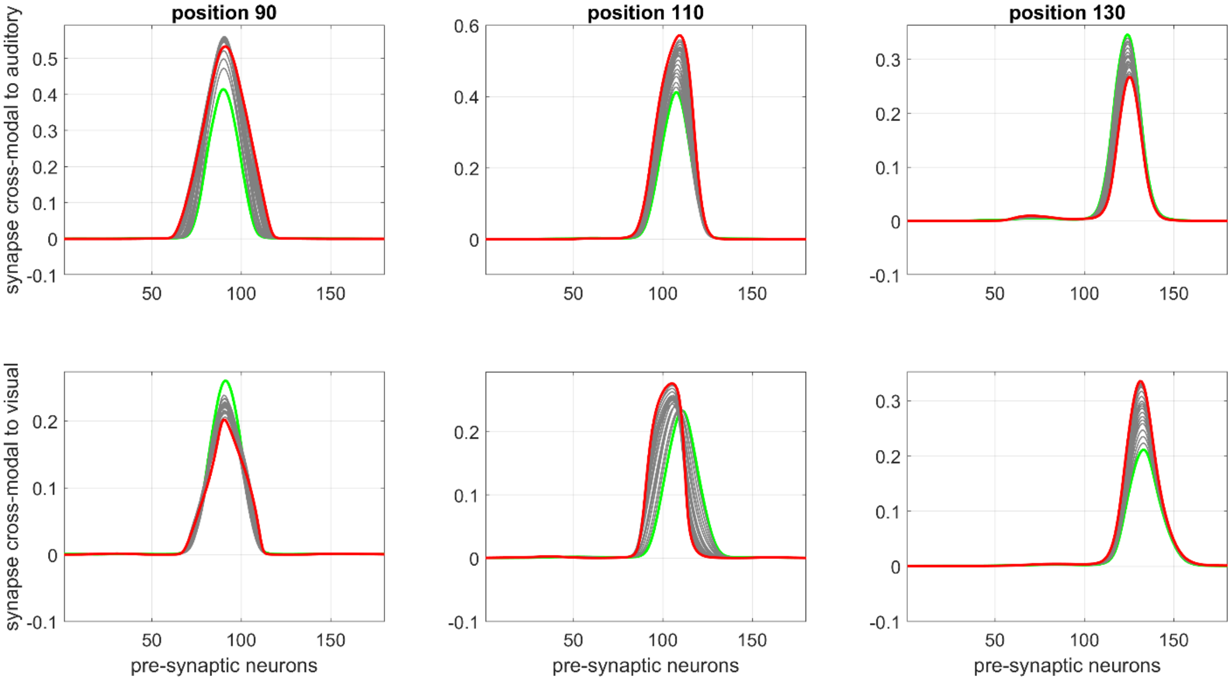

Figure S5. An example of how cross-modal synapses change during re-learning from a condition with a standard deviation of the visual unisensory prior as large as s_V = 30 deg, to a condition with s_V = 7 deg (that is the same re-learning illustrated in Fig. 11 of the text). The upper line represents synapses entering into an auditory neuron from all visual neurons in the visual net. The bottom line represents cross modal synapses entering into a visual neuron from all auditory neurons in the auditory net. The green line represents the synapse distribution in the mature configuration before re-learning, whereas the red line is the synapse distribution after re-learning. Gray lines are examples of iterations during the re-learning. As for cross-modal synapses entering auditory neurons, it is evident a reinforcement of the cross-modal input close to the fovea. As for cross-modal synapses entering visual neurons, it is evident a shift of the cross-modal input toward the fovea at intermediate azimuthal positions, and a reinforcement of the visual cross-modal input at more peripheral azimuthal locations.

5. Additional Simulations with the Lesioned Network

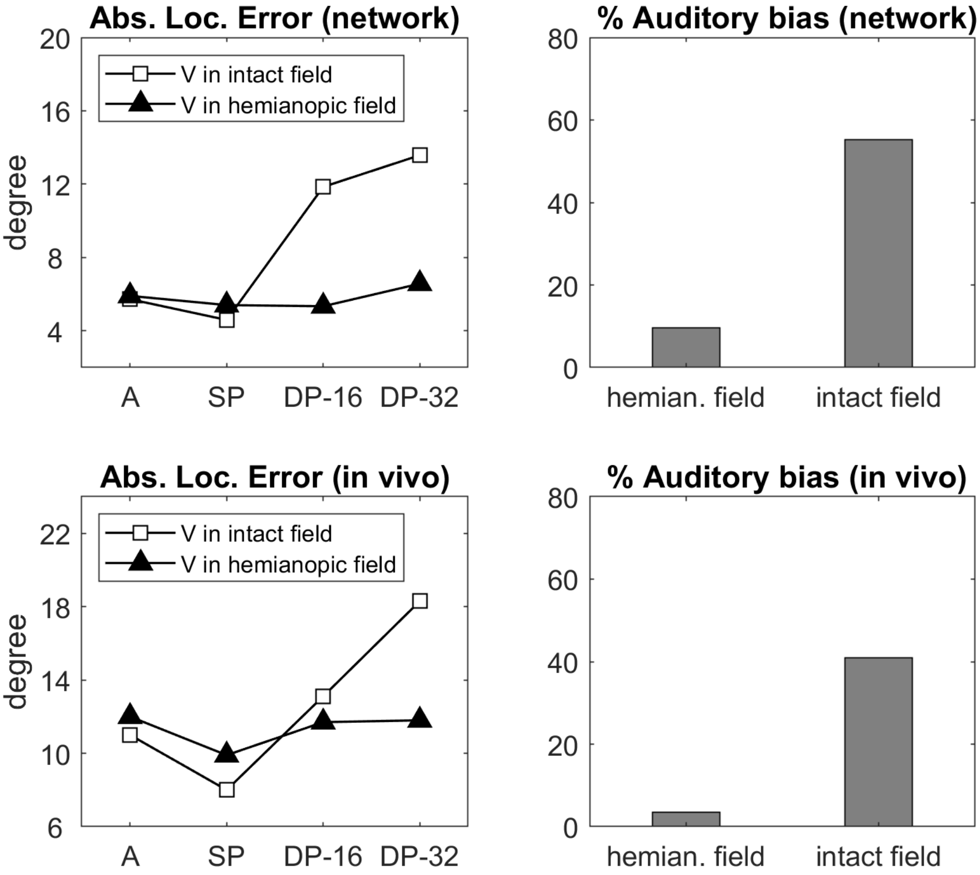

Figure S6. The lesioned network (90% of damaged neurons in the right visual hemifield) was used to replicate an experiment similar to that performed in hemianopic patients (Leo *et al.*, 2008). Simulated results are shown in the upper plots and in vivo data are redrawn in the lower plots. In the network, a visual stimulus was applied either at –10° (intact hemifield) or at +10° (in the lesioned hemifield) and paired with an auditory stimulus applied at the same spatial position (SP) or at 16° and 32° of spatial disparity (DP16, DP32). The auditory stimuli were presented in unimodal conditions (A), too. The simulations were performed using a visual stimulus with strength 18, an auditory stimulus with strength 36, and in noisy condition (average values are displayed). Plots in the left column show the absolute localization error (absolute difference between the perceived auditory location and the real auditory location) computed in each condition (A, SP, DP16 and DP32) separately for the visual stimulus in the intact and damaged hemifield. Plots in the right column show the percentage of auditory bias [100*(perceived auditory location minus the real auditory location) / (actual visual-auditory disparity)] in DP16 and DP32 conditions (collapsed together) for the visual stimulus in the intact and damaged hemifield. According to the network (upper plots), a visual stimulus in the intact hemifield slightly reduces the auditory localization error in SP condition and strongly increases auditory mislocalization in DP condition, producing a high ventriloquism effect; conversely, a visual stimulus in the lesioned hemifield has only a weak impact on auditory localization error, and the ventriloquism effect radically declines. These network outcomes display good agreement with the in vivo data (lower plots).