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

Short title: effect of prior and likelihood on AV integration

Corresponding author:

Mauro Ursino, Department of Electrical, Electronic and Information Engineering, University of Bologna, Viale Risorgimento 2, I40136, Bologna, Italy –
email: mauro.ursino @unibo.it

Other author email:

cristiano.cuppini@unibo.it elisa.magosso@unibo.it
ulrik.beierholm@durham.ac.uk lshams@psych.ucla.edu

KEY-WORDS: Neural Networks, Audio-visual integration, Hebbian mechanisms, Synapse learning, Bayesian inferences, Priors, Likelihoods

Summary

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.

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 cues reliability and temporal asynchrony.

The model considers two unisensory networks (auditory and visual) with plastic receptive fields and plastic cross-modal 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 cross-modal conditions to assess the auditory and visual perception bias: visual stimuli were positioned at different azimuth (±10 deg from the fovea) coupled with an auditory stimulus at various audio-visual distances (±20 deg). 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 variety of conditions characterized by a different role of prior and likelihood. Finally, the effect of a different unimodal or cross-modal 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.

Introduction

Several findings in the recent neuroscience literature suggest that the brain adopts a Bayesian approach to develop its representation of a noisy external world (Pouget *et al*., 2013). This means that, among different possible choices, the brain favors the scenario with higher posterior probability, in order to minimize the chance of error. The "Bayesian brain" has been intensely studied in the domain of multisensory integration, to better understand how information from different sensory modalities can be fused into a single optimal percept (Alais and Burr, 2004, Battaglia *et al*., 2003, Beierholm *et al*., 2009, Ernst and Banks, 2002, Fetsch *et al*., 2012, Pouget *et al*., 2013, Shams *et al*., 2005, Ursino *et al*., 2014, Wallace *et al*., 2004).

To realize Bayesian multisensory inference (for instance, to infer the spatial position of an external stimulus) a neural circuit must incorporate and combine two different pieces of information, both statistically extracted from the environment: the so-called likelihood probability, and the prior probability. The first reflects the characteristics of the present inputs: such as, in case of spatial inference, the spatial width of the stimuli, the superimposed noise, the stimulus strength; more generally, likelihood represents the full probability distribution of the stimulus. This can change from one trial to the next (for instance, by blurring a stimulus or altering the superimposed noise). Conversely, the prior probability reflects a belief on the stimulus characteristics before any present stimulus perception, and, in a stationary environment, is stable and invariant to sensory conditions.

Various experimental and behavioral results confirm that the brain performs a near-optimal Bayesian inference in the sensory space, by weighting cross-modal cues according to their reliability, but also encoding prior experience (generally resulting in a bias toward the more probable events). This has been observed in experiments testing audio-visual integration (Alais and Burr, 2004, Battaglia *et al*., 2003, Beierholm *et al*., 2009, Shams *et al*., 2005, Wallace *et al*., 2004), visuo-tactile integration (Ernst and

Banks, 2002, Magosso *et al*., 2010), sensory auditory-visual-tactile integration (Wozny *et al*., 2008) and in studies which use within-sensory cues, such as texture and motion (Jacobs, 1999) or stereo and shading (Bülthoff and Mallot, 1988).

Despite these important contributions, however, the problem of how cue reliability and prior experience can be optimally merged in a population of neurons is still crucial in computational neuroscience and is the topic of intense research (Pouget *et al*., 2013, Ursino *et al*., 2014, Cazettes *et al*., 2016, Ma *et al*., 2006, Patton and Anastasio, 2003, Pouget *et al*., 2003). A fundamental idea is that probability distributions can be encoded in the activity of an entire population of neurons, where generally each neuron codes for a particular value of the estimated parameter. In this regard, Deneve *et al*. (1999), Ma *et al*. (2006) and Pouget *et al*. (2013) demonstrated that a population of neurons can compute the likelihood function and that cue reliability is reflected in the variations of the population activity. A further theoretical account of how the brain can represent stimulus distribution/cue reliability is the sampling model by Fiser et al. (2010). Moreover, Fischer and Peña (2011) and Cazettes *et al*. (2016) proposed that cue reliability is represented in the shape of the tuning curves. This result has been theoretically supported by our recent work (Ursino *et al*., 2017b): using a Gaussian distribution of the inputs, we showed that the inner product of the neuron's receptive field and the external stimulus is proportional to the likelihood.

The problem of how this likelihood can be combined with prior experience, to implement a true posterior probability, is more controversial. At which levels of a neural circuitry is prior information stored? And how can it be merged with likelihood in a near-optimal way?

Let us consider the problem of inferring a spatial position in case of audio-visual integration, which will be the subject of the present work. Two different aspects of the prior experience should be encoded: i) the individual prior of the single unisensory stimuli (for instance, the probability that a visual stimulus is more frequently located close to the fovea than at the periphery); ii) the joint probability of the two stimuli occurring together (simultaneous visual and auditory stimuli often originate from proximal spatial positions, since they are produced by a common cause). In the presence of a unisensory input, just the first aspect is of value. In case of multisensory inputs, the joint prior probability of the two inputs is the product of the unisensory probability of one stimulus and the conditional cross-modality probability.

According to Cazettes *et al*. (2016) and Ursino *et al*. (2017a) the first aspect (i.e., the unisensory prior) can be encoded in the density distribution of the receptive fields (i.e., on the density of the neuron tuning functions). Events that are more frequent are associated with a denser distribution of neurons. For what concerns the conditional prior, Ursino and collaborators proposed that it could be encoded in cross-modal synapses, linking neurons of different modalities that are often simultaneously engaged in a multisensory percept (Cuppini *et al*., 2014, Magosso *et al*., 2012, Ursino *et al*., 2017a, b). All these aspects have been summarized in a recent comprehensive neural network model by our group (Ursino *et al*., 2017a), where we showed that all previous terms can be extracted from the statistics of the environment and stored in synapses using a Hebb rule with a forgetting factor. In particular, by using this learning rule in a network of two populations of visual and auditory neurons, we demonstrated that: i) the width of neuron receptive fields progressively shrinks during training, to reflect the spatial reliability of the external stimuli; ii) the barycenter of the RFs moves during training so that population density reflects the unisensory prior; iii) cross-modal synapses between the two areas progressively develop to reflect a conditional prior on multisensory co-occurrence. By merging all these aspects together, the network was able to perform optimal inference of auditory and visual spatial positions, in a variety of unisensory and multisensory conditions, in satisfactory agreement with the theoretical Bayesian estimator. In particular, the model predicts a ventriloquism effect in cross-modal conditions, which depends on the azimuthal coordinate, and a visual shift toward the fovea in unimodal conditions.

However, several aspects are still insufficiently clear and the model necessitates a more exhaustive validation on the basis of real behavioral data and some aspects, not tested before, deserve a thorough computational analysis by means of new simulations.

In particular, what aspects of behavioral results depend on the prior characteristics of the environment (i.e., on past experience of the individual subject) and what aspects are affected by the characteristics of the current stimuli reliability? Are prior and likelihood really merged in our brain as predicted by the model?

Previous studies have shown that auditory-visual spatial perception closely follows Bayesian causal inference (Kording *et al*., 2007; Wozny *et al*., 2010; Odegaard *et al*., 2015; Wozny and Shams, 2011; Odegaard *et al*., 2017; Odegaard and Shams, 2016). Beierholm *et al*. (2009) using the same spatial task has also demonstrated that a radical change in stimulus noisiness which results in a significant change in likelihoods, does not lead to a change in prior probabilities, neither the unisensory priors nor the binding prior. This finding suggests that Bayesian causal inference is also a good process model for spatial perception. However, how these distributions and computations are implemented by the neural machinery of the nervous system is still unclear, and hence the present study. Shams and Beierholm (2010) have discussed how causal inference can be carried out by the utility of heavy-tailed likelihoods or priors. However, the specific neural correlates of these mechanisms have not been investigated, and the role of various aspects of lateral connectivity, network architectures and dynamics on the emergent computational properties remain unclear.

Moreover, several aspects of multisensory integration, which have a great relevance in Neuroscience, have not been tested with the model yet. Can a mature network re-learn a new prior in a not stationary environment, by partially forgetting the previous one? Furthermore, various recent studies suggest that multisensory integration crucially depends on the temporal aspects of the stimuli, being stronger in case of highly correlated stimuli and weaker in case of poor correlation (Parise *et al*., 2013; Denison *et al*., 2013; Parise and Ernst, 2016): can the model reproduce a similar behavior? Finally, the model might have a clinical impact. In particular, studies in hemianopic patients (Leo *et al*., 2008; Magosso *et al*., 2016) show a loss of audio-visual integration (as measured via the auditory ventriloquism) in the lesioned hemifield compared with the spared one, a condition that should be tested with the model.

The goal of this study is to investigate the neural mechanism underpinning Bayesian causal inference in spatial perception by exploring a) the roles of past experience (prior) and of present cue reliability (likelihood) in affecting neural network model behavior in presence of multisensory inputs, and b) comparing model behavior with behavioral data, in conditions where the reliability of the stimuli or the prior are manipulated. In particular, we analyze the differences in spatial audio-visual integration when the subject experiences two cross-modal stimuli as a single percept (C = 1) or two separate events (C = 2). It is worth noting that the behavioral data used in the present work were never employed to build the model nor to assign its internal parameters. Indeed, to simulate these data we have not modified any parameter of the previous network (Ursino *et al*., 2017a), but just assumed different characteristics of the inputs (reflecting a different likelihood or a different past experience). In particular, in the present work we improve the description of the priors. In contrast with the previous work, we now assume that not only the visual but also the auditory stimulus has a non-uniform unisensory prior, being more precise and more frequent at the center, and we use a more realistic heavy-tailed prior in cross-modal conditions.

Furthermore, we analyze the model's capacity to simulate various additional aspects summarized above, i.e., re-learning in a non-stationary environment, the effect of temporal asynchrony on the integration, and the effect of a lesioned visual hemifield.

The results confirm that the proposed model can grasp many aspects of audio-visual spatial integration, providing a plausible insight on how a trained neural net can learn the statistics of the external environment, and combine present reliability and past experience quite optimally to infer a Bayesian estimate. Furthermore, the results shed lights on how differences in behavior depend on differences in past experience (i.e., on the prior characteristics of the stimuli) and on the testing conditions of the experiment (i.e., on the reliability of the stimuli used at the moment of perception and on their temporal asynchrony) and may be exploited to analyze the plasticity in non-stationary conditions and the effect of pathological lesions as well.

Method

Qualitative model description

All model equations are reported in the Supplementary Material part I. In the following, only a qualitative summary is given.

The model includes two chains of unisensory neurons (the first devoted to localization of the auditory input, the second to the visual one) topologically organized (see Fig. 1). The activity of each neuron is simulated by means of a static sigmoidal relationship and a first-order dynamics, with time constant τ. (Eqs. S1 and S2 in Supplementary Material part I). According to the sigmoid relationship, the neuron exhibits no appreciable activity when it receives negligible input (below a given threshold) and maximal saturation activity in case of high excitatory input. In this model, the upper saturation is assumed equal to 1, i.e., all activities are normalized. The time constant describes the time required for the neuron to integrate its input and produce the response.

Each neuron codes for a different portion of space in its specific modality (either auditory or visual), although this position can be modified by experience (see below). In particular, each neuron filters the external input of its modality, by performing the convolution with its receptive field, and we assume that the preferred position can be computed as the barycenter of the neuron receptive field. In the initial (pre-training) configuration, all neurons have the same receptive field, with identical shape characterized by large width. This is realized with a Gaussian function with SD = 30 deg. Moreover, we assume that the barycenter of the receptive fields before training is uniformly distributed in space, reflecting the absence of any prior information. The model uses 180 neurons for each layer, coding for the overall azimuthal coordinates (i.e., the vertical coordinate is not considered for the sake of simplicity). Hence, the RF's center for two consecutive neurons initially differs by 1 deg. However, the preferred position of each neuron is not fixed, but it can shift as a result of the sensory training, to incorporate the statistics of the unisensory inputs. In particular, after training (see section Results) the RFs of all neurons shrink (to reflect the likelihood of the external inputs) and their preferred position moves (to reflect the input prior probability). Of course, this is possible since the connections to the sensory environment initially cover a large portion of space, and the gain adjusts automatically to reflect the mean amplitude of the input. This is warranted by the learning rule adopted (see Supplementary Material part I and Ursino *et al*, 2017b for more details).

Furthermore, neurons in the same modality interact via a competitive mechanism, which is typical of cortical layers. This is realized through lateral synapses arranged with a Mexican Hat spatial disposition: each neuron receives excitation from proximal neurons and inhibition from more distal ones. Consequently, in response to a single input of a given modality, a bubble of neurons is excited within the layer, approximately centered at the position of the external input, surrounded by an annulus of inhibited neurons. In the present work, for simplicity, we assumed that lateral synapses are not subject to training. Plasticity of lateral synapses may become important in case of a constant audio-visual shift during training, a condition not tested in the present work, which induces to the so-called "ventriloquism aftereffect" (Bertelson et al., 2006; Magosso et al., 2012).

Finally, according to the recent neurophysiological literature (Driver and Noesselt, 2008, Ghazanfar and Schroeder, 2006) neurons also receive a cross-modal input from neurons of the other modality, thus realizing multisensory integration. In fact, an implicit assumption of our model is that integration between different modalities can be realized directly within the initial layers, before information reaches a downstream multisensory layer. Cross-modal synapses are initially set at zero, since we do not have any prior information on how visual and auditory stimuli co-occur. Then, these synapses are progressively created during training in presence of a multisensory environment, to incorporate a prior probability on the audio-visual relationship.

Model training procedure

As anticipated above, both the receptive fields synapses, and the cross-modal synapses are plastic.

To assign their value, the network was trained during a training period, starting from the initial synapse condition described above (large and uniformly distributed RFs, equal for the auditory and the visual nets; cross-modal synapses initially at zero). We used a Hebbian learning rule with a forgetting factor. A synapse is strengthened if the pre-synaptic and the post-synaptic activities are high; however, in order to avoid an indiscriminate synapse potentiation, a portion of the previous synapse is lost if the post-synaptic activity is high. The same learning rule, with identical learning factors, was adopted for training both the synapses in the RFs and the cross-modal synapses between the two areas (see equations S8 and S9 in Supplementary Material part I).

The training procedure consisted of 100 epochs. During each epoch, we presented 900 inputs with a given ratio “unisensory visual”: “unisensory auditory” : “cross-modal”. Hence, the total number of trial was 90000.

Description of the inputs

During all simulations, we used auditory and/or visual inputs centered at the position θ_V and θ_A, with a Gaussian shape and superimposed Gaussian white noise with zero mean value and assigned standard deviation. In particular, the input strength is assigned during training to have significant excitation of visual and auditory neurons a little below saturation, while the standard deviation of noise (parameters υ_A and υ_V in Table I) is equal to ¼ of the input strength to set a good signal to noise ratio and so, to facilitate the creation of synapses. Conversely, noise is increased during the testing phase to mimic the Report of Unity observed in Beierholm *et al*. (2009) at higher and lower contrast. Hence, by denoting with $i_S(\theta)$ the input that excites the unisensory net of modality S ($S = V$ or A) at the azimuthal coordinate θ, in response to a stimulus centered at the position θ_S, we can write

$$i_S(\theta) = \frac{i_{S,Strength}}{\sqrt{2\pi\sigma_S^2}} \exp\left(-\frac{(d(\theta_s,\theta))^2}{2\sigma_S^2}\right) + n_s(\theta) \quad S = A \text{ or } V \tag{1}$$

where $i_{S,strength}$ is the area of the Gaussian function (which can be considered as the strength of the stimulus), σ_S is the spatial standard deviation of the stimulus, θ_s is the stimulus position (equal to the mean value of the Gaussian function) and $n_s(\theta)$ is a Gaussian white noise term (zero mean value and assigned standard deviation υ_S). Finally, $d(\theta_s,\theta)$ represents the distance between the position θ, and the central position θ_S.

A crucial aspect of training is the statistics for the position and the width of the inputs. As in the previous work (Ursino *et al*., 2017a) we assumed that the visual input statistics (that is, the visual unisensory prior) depends on the azimuthal coordinate. In particular, visual inputs are more frequently close to the fovea, as a consequence of fixation head and eye movements; moreover, visual inputs are spatially more tuned at the center and have a reduced accuracy close to the periphery, reflecting a physiological behavior. Moreover, at odd with the previous version (Ursino *et al*., 2017a), in the present model we further assume that the auditory stimuli are also more precise and more frequent close to the head center than at the periphery, although they remain less precise and less focused than the visual stimuli.

The previous ideas correspond to the use of the following prior probabilities during training, where θ_V and θ_A represent the positions of the visual and auditory stimuli in the azimuthal coordinate (with $-90 \leq \theta_V \leq 90$ and $-90 \leq \theta_A \leq 90$, see Eq. (1)):

Unisensory visual: A Gaussian prior with mean values at 0 deg (the fovea in our model) and standard deviation s_V. This signifies that visual stimuli have a much greater probability near the fovea than at the periphery. Hence we have

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

Unisensory auditory: A Gaussian prior, but with $s_A > s_V$

$$p(\theta_A) = \frac{1}{\sqrt{2\pi\, s_A^2}} \exp\left(-\frac{\theta_A^2}{2s_A^2}\right) \qquad (3)$$

Cross-modal: 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. Moreover, we assume a certain probability (say β) that the two stimuli are independent, and a higher probability (say $1 - \beta$) that, in cross-modal conditions, the auditory and visual inputs originate from proximal positions.

We have

$$p(\theta_V, \theta_A) = 0.5\, p(\theta_V)\, p(\theta_A | \theta_V) + 0.5\, p(\theta_A)\, p(\theta_V | \theta_A) \qquad (4)$$

where we used equations (2) and (3) for the visual and auditory priors, and the following expression for the conditional probability

$$p\left(\theta_A \middle| \theta_V\right) = \beta\, p\left(\theta_A\right) + \left(1-\beta\right)\frac{1}{\sqrt{2\pi\, s_{AV}^2}}\exp\left(-\frac{d\left(\theta_A,\theta_V\right)^2}{2\, s_{AV}^2}\right) \tag{5}$$

$$p\left(\theta_V \middle| \theta_A\right) = \beta\, p\left(\theta_V\right) + \left(1-\beta\right)\frac{1}{\sqrt{2\pi\, s_{AV}^2}}\exp\left(-\frac{d\left(\theta_A,\theta_V\right)^2}{2\, s_{AV}^2}\right) \tag{6}$$

The first term in Eqs. (5 and 6) represents the case of independent cross-modal stimuli (with probability β) while the second term (with probability $1-\beta$) represents the case when the auditory and visual events are originated from the same source, hence with very small distance (s_{AV} = 1 deg). In this work, at odd with the previous one (Ursino *et al*., 2017a, Ursino *et al*., 2017b), we used a higher probability ($\beta = 0.2$) of independent sources in cross-modal conditions. Moreover, a sensitivity analysis has been performed on this parameter. It is worth noting that the use of a heavy-tailed prior to describe cross-modal correspondence has been suggested by various authors in recent years to distinguish cue integration vs. cue segregation. The reader can find important contributions in Ernst and Di Luca, 2011; Ernst, 2012; Körding *et al.,* 2007; Roach *et al*., 2006 or for a review see van Dam *et al*., 2014.

A visual 2D example of the cross-modal prior (Eq. 4) computed with $\beta = 0.5$ is shown in Fig. S1 of Supplementary Material part II.

Behavioral data

The acquisition of the behavioral data has previously been described (Beierholm *et al.,* 2009), here we briefly summarize it.

Nineteen naive observers participated in the experiment across two sessions, separated by one week (high visual contrast, then low visual contrast session). Subjects were seated at a viewing distance of 54 cm from a 21-inch CRT monitor. Visual and auditory stimuli were presented independently, but temporally synchronized, for 35 ms at one of five locations ([-10, -5, 0, 5, 10] degrees relative to fovea) extended along a horizontal line 5° below the fixation point. Visual and auditory stimuli were thus congruent on 20 percent of trials. Visual stimuli consisted of Gabor wavelets extending 2° on a background of visual noise. Using Gabor wavelets allowed us to increase task difficulty without excessively increasing the size of the visual stimulus. The size of visual stimulus used here was comparable to previous experiments (Alais and Burr 2004). The visual contrast was adjusted on an individual basis so that subjects' unimodal performance was 90% correct for the high contrast session and 40% correct for the low contrast session. Auditory stimuli were presented through a pair of headphones (Sennheiser HD280) and consisted of white noise filtered through an individually assessed Head Related Transfer Function (HRTF), and simulated sounds originating from the five spatial locations in the frontoparallel plane where the visual stimuli were presented. The task of the observer was to report the location of the visual stimulus as well as the location of the sound in each trial using the keyboard, with five keys mapped directly to the five possible locations. Auditory and visual stimuli were presented alone or simultaneously, leading to a total of 35 conditions (5x5+5+5), repeated 15 times each. No feedback was given.

Simulations

Training - During training we used quite strong inputs, with an elevated signal to noise ratio (i.e., the ratio $i_{S,Strength}/\upsilon_S$) both for the auditory and visual stimuli (see the upper portion of Table 1). In particular, these inputs were chosen so that, in unisensory conditions, the activity of the maximally excited neurons are quite close to saturation, both for the auditory and the visual inputs. As in the previous paper (Ursino *et al*., 2017a), this choice allows quite a rapid and efficient training.

Since one of the objectives of the present work is to investigate the effect of a change in past experience (i.e., prior probability) on network performance, the training was repeated with different percentages of cross-modal vs unimodal inputs (i.e., by changing the role of the conditional prior probability) and with a different distribution of the visual inputs close to the fovea (i.e., by changing the unisensory visual prior probability). The characteristics of the three different trainings are illustrated below, and will be named *Training1*, *Training2* and *Training3*, respectively. Furthermore, we investigated the role of independent cross-modal inputs (i.e., parameter β in Eqs. (5) and (6)), in *Trainings 4* to *6*.

Training1 (used to simulate behavioral data) makes use of unisensory visual inputs very close to the fovea. This is obtained by using $s_V = 7\,\text{deg}$ in Eq. (2). The percentage of unisensory and cross-modal inputs during training is: 40% auditory (36/90), 40% visual (36/90), 20% cross-modal (18/90). The probability of independent visual and auditory stimuli in cross modal conditions is $\beta = 0.2$.

Training2 makes use of a wider distribution of unisensory visual inputs. This is obtained by using $s_V = 30\,\text{deg}$ in Eq. (2). The percentage of unisensory and cross-modal inputs is the same as in Training1: 40% auditory, 40% visual, 20% cross-modal. Moreover, $\beta = 0.2$

Training3 makes use of the same unisensory arrangement of visual inputs close to the fovea as Training1 (i.e., $s_V = 7\,\text{deg}$ in Eq. (2)) and the same β, but a lower percentage of cross-modal inputs: 46.66% auditory (42/90), 46.66% visual (42/90), 6.66% cross-modal (6/90).

Trainings4-6 are identical as Training 1, but we assumed a different probability that auditory and visual stimuli are independent in cross-modal conditions. We used $\beta = 0$, 0.5 and 0.7, respectively.

Testing -Testing has been performed on the trained network (separately after *Training1*, *Training2* , *Training3 and Training4-6*) to mimic the ventriloquism effect with different audio-visual distances and at different visual positions eccentricity. In particular, we stimulated the trained network with a visual stimulus placed at the positions −10, −5, 0, +5, +10 deg (where 0 deg means the fovea); at each visual position, an auditory stimulus was joined, with an audio-visual distance in the range −20 ÷ +20 deg (here a positive shift means that the auditory stimulus is located on the left of the visual stimulus, and vice versa). 200 trials were then repeated per each combination of stimuli to calculate the perception bias (i.e., the perceived position of the stimulus minus the real position of the stimulus). The perceived position of each stimulus was computed as the barycenter of network activity (separately for the visual and the auditory net) using the after-training barycenter of the neuron receptive field as the preferred position for each neuron (see the previous paper (Ursino *et al*., 2017a) for a more complete equation set).

The input values used during testing (strength and noise level of the stimuli) are reported in the second part of Table 1. These allow the behavioral data by Beierholm *et al*. (2009) to be simulated quite well, using the network after *Training1*. Briefly, the input strengths and noise were assigned to have a Report of Unity (percentage of cross-modal inputs ascribed to a single cause) at small audio-visual distance (a few degree) proximal to that in the behavioral data (about 40% in the high-contrast case, about 20% in the low contrast case). Since these data have been obtained in presence of much noise, and exhibit quite a small report of unity, we used higher noise for the auditory inputs compared with the training phase (i.e., we increased the standard deviation υ_A in Eq. (1)). Moreover, as in Beierholm *et al*. (2009), we used two different contrasts for the visual inputs: a higher visual contrast first, and a smaller visual contrast thereafter (obtained by changing the ratio $i_{V,Strength}/\upsilon_V$ in Eq. (1)).

To compute the report of unity, two stimuli were ascribed to the same cause (C=1) if their perceived distance was below 2 deg, and they were ascribed to two independent causes (C=2) if their perceived distance was greater than 2 deg. The same trials were also repeated with the network obtained after *Training2 – Training6*, to point out the effect of prior on model results.

Furthermore, data were also analyzed assuming that casual inference (i.e., C = 1 vs. C = 2) is performed by a downstream multisensory layer. Two alternative strategies were adopted: i) by computing the number of peaks in a multisensory layer, as already done in Cuppini et al. (see Cuppini *et al*., 2017 for more details); ii) by computing the cross-correlation between the activity in the visual layer and the activity in the auditory layer. It is worth-noting that this cross-correlation may be estimated by a downstream multisensory layer, for instance by using the logarithm of neuron activity so that the sum of logarithms can be used to compute correlations. Due to space limitations, we did not describe this third layer in the present work, but we just briefly summarized the results.

Results

The effect of training on model synapses

During training the receptive fields progressively shrink, to match the reliability of the external cues, and their preferred positions shift to have denser RFs in the zones with higher unisensory prior. An example is shown in Fig. 2, where we show the progressive changes in the RFs of two auditory neurons and two visual neurons, with initial preferred positions at -40 deg from the fovea and at the fovea respectively. The visual RFs become more tuned than the auditory ones, due to higher spatial reliability of visual stimuli. Moreover, the visual RFs are sharper close to the fovea, where we assumed more precise visual inputs. Finally, it is evident that the RF of the visual neuron initially preferring -40° position, progressively moves closer to the fovea; the auditory neuron at the same position also exhibits an evident shift. All the previous examples refer to Training1.

Some examples of visual and auditory RFs, and some examples of cross-modal synapses linking the auditory and visual nets after Training1 are shown in Fig. 3. It is evident that visual RFs are sharper and denser close to the fovea. The auditory RFs are also denser in the proximity of the fovea. The increased density of the visual and auditory RFs is a consequence of the unisensory prior (Eqs. 2 and 3).

Looking at the bottom panel of Fig. 3, we can see that auditory neurons receive strong cross-modal synapses from visual neurons in the central azimuthal field. Theses synapses are mainly responsible of the ventriloquism effect. Hence, as also demonstrated in the previous work (Ursino *et al*., 2017a) and supported by behavioral data (Charbonneau *et al*., 2013, Hairston *et al*., 2003) the ventriloquism auditory bias decreases with the azimuth. Conversely, visual neurons receive strong cross-modal synapses from auditory neurons at positions 30-40 and 140-150 (corresponding to a barycenter of the Receptive field at about ±30 ±40 deg from the fovea), where indeed visual inputs are very rare according to our training procedure, but auditory inputs are still moderately frequent. Hence, a testable future prediction of the model is that, at an eccentricity of about 30-40 deg from the fovea, the bias of the visual stimulus by an auditory stimulus should be stronger, whereas the bias of auditory localization should be weaker. This may be tested by providing stimuli in this spatial range.

Comparison between model results and behavioral data

Simulations performed with the network after the *Training1* (i.e., with 20% of cross-modal inputs and with visual inputs very close to the fovea) lead to results in acceptable agreement with the behavioral data.

First, Fig. 4 compares the Report of Unity (RoU: percentage of cases with C = 1, computed on the basis of perceived audiovisual distance) vs. the real audiovisual distance in the cases of low contrast and high contrast (where behavioral data represent when subjects report same location for A and V). The agreement between model predictions and behavioral data is quite good at moderate AV distances. These are the only curves for which a manual fitting was performed: i.e., we chose the strength of the visual and auditory inputs, and the level of the superimposed noise to obtain the values of RoU at zero AV distances. In particular, the report of unity is quite low (about 40%). This could be obtained with the model by using high values of noise. If lower noise levels were used, the RoU turned out close to 1 at small audio-visual distances, as actually observed in many other behavioral data (see for instance (Wallace *et al*., 2004) and (Rohe and Noppeney, 2015)).

As well expected, the RoU is greater in case of higher visual contrast, and decreases with the visual contrast. Moreover, it decreases with the audio-visual distance. However, we can observe some significant differences between model and experimental data at larger audio-visual distances. In particular, it is difficult to understand why, in the experimental data, the RoU is higher in the low contrast case compared with the high contrast case when the audio-visual distance exceeds 10 deg. One hypothesis may be that noise affects the perceived position more significantly in the low-contrast case compared with our model so that, even at an audio-visual discrepancy as large as 20 deg, a certain amount of audio-visual inputs are casually assumed as coincident.

A comparison between model and behavioral perception biases is presented in Figs 5-6 for the high contrast case, and in Figs. 7-8 for the low contrast case.

As well evident in Figs. 5-8, behavioral data use only a portion of the azimuthal space around the fovea (from -10 to + 10 deg) for the positions of the visual and the auditory stimuli. As a consequence, only five audio-visual distances have been experimentally tested (since the auditory stimuli were located at -10, -5, 0, +5 and +10 deg for each value of the visual position). For instance, when the visual input was located at 10 deg, the audio-visual distances could range only from -20 deg to 0 deg; when the visual stimulus was located at 0 deg, the audio-visual distances could range only from -10 deg to 10 deg, and so on). Conversely, when using the model, we tested a larger range for audio-visual distances, spanning from -20 deg to + 20 deg, to have a more comprehensive understanding of model behavior.

Figure 5 shows the auditory perception bias plotted vs. the audio visual distance, computed at different positions of the visual input in the higher contrast case (model simulations are in the upper panels, behavioral data are in the bottom panels). Results are shown separately by including all trials (left panels), the C = 1 cases only (central panels) , and the C = 2 cases only (right panels). The same results for the visual perception bias are shown in Fig. 6.

Model results are in qualitative agreement with behavioral data. In particular, a ventriloquism effect is well evident: the auditory perception exhibits *a bias toward the visual position*. This is much higher in the C = 1 cases (where ventriloquism may rise almost to 20 deg) and is much smaller in the C = 2 cases. Furthermore, one can observe that the auditory curve exhibits a leftward shift as the visual input moves from 10 deg to – 10 deg, and this shift is especially evident in the C = 2 cases, but is quite negligible in the C = 1 cases.

The visual bias (Fig. 6) depends on the azimuthal position of the visual input, but is quite independent of the audio-visual distance. In particular, looking at the upper panels in Fig. 6, we can observe that, in the model, the visual perception does not exhibit any appreciable bias toward the auditory input, but exhibits a constant bias toward the fovea; this is almost the same in C = 1 and C = 2 cases.

For what concerns the behavioral data, when C = 1 (bottom middle panel) the visual bias exhibits a certain attraction toward the auditory position when the visual input is eccentric, but this is evident only at the extreme boundary of the $\theta_V - \theta_A$ range (±20 deg). It is worth noting that, in our model, C = 1 never occurred at these distances and at this visual eccentricity, in case of high visual contrast (we have no data in the upper mid panel) and that these points were actually extremely rare also in behavioral data. The presence of a visual bias at large eccentricity may be a further testing conditions for the model, in agreement with the arrangement of cross modal synapses depicted in Fig. 3.

The results, in the simulations with lower visual contrast, are shown in Figs. 7 and 8. In this case too, the agreement between model predictions and behavioral data is satisfactory. Two main differences are evident comparing the low-contrast and the high-contrast cases. First, the auditory ventriloquism is smaller when low-contrast visual inputs are used: these differences are evident especially in the C = 1 cases. Second, and more important, the visual bias increases in the low-contrast case and, when C = 1, it is significantly affected by the auditory input. This means that, in case of low contrast, the visual perception poses more weight on the prior compared with the likelihood. In the C = 1 cases, in particular, the visual position reflects both a shift toward the fovea (unisensory prior) and an appreciable shift toward the auditory input (a kind of "visual ventriloquism", conditional prior). When the visual input is placed at the left of the auditory one ($\theta_V - \theta_A < 0$), it exhibits a rightward shift; when the visual input is at the right of the auditory input ($\theta_V - \theta_A > 0$), it exhibits a leftward shift; this is superimposed on constant shift toward the center. This "visual ventriloquism" almost disappears in the C = 2 case, where the shift toward the fovea prevails.

The agreement between model results and behavioral data has been assessed by computing the correlation coefficient. The values (reported in Supplementary Material part II, together with the correlation curves, Figs. S2 and S3) are always higher than 0.86, with the only exception of the visual bias in the high contrast C = 1 case (r = 0.786). In this condition, however, the bias is always very small (less than 1 deg).

Results similar to those in Figs. 4-8 can be obtained also using a third multisensory layer, to discriminate C = 1 and C = 2. The agreement between the model and behavioral data is still quite satisfactory in the low-contrast case; however, in the high contrast case, we observed that the report of unity computed with a multisensory layer remains too high at an audio-visual distance as large as ±10 deg (i.e., it only moderately decreases with distance) and the auditory bias in the C = 1 case exhibits some differences (see also Discussion).

The effect of a different training

We compared model results (concerning both the auditory perception bias and the visual perception bias) in the alternative training conditions, to elucidate the role of past experience. For the sake of brevity, only results of the high-contrast simulations are reported, without a distinction between C = 1 and C = 2. Results, summarized in Fig. 9, strongly confirm that the auditory ventriloquism (i.e., a progressive shift in the auditory perceived position toward the visual one) is strongly affected by the percentage of cross-modal inputs used during training. In fact, if the percentage of cross-modal inputs during training is reduced (*Training3*), the auditory ventriloquism is drastically reduced. The reason is that, in these conditions, the cross-modal synapses are weak. This underlines the relationship between cross-modal synapses, auditory bias, and cross-modal conditional prior in our model.

Moreover, simulations confirm that both the auditory and the visual bias depend on the unisensory prior (i.e., Eq. (2)); in particular, an increase in parameter s_V in Eq. (2) (that is, assuming a larger distribution of visual inputs around the fovea) almost completely abolishes the visual bias, at least in the azimuthal space examined here (± 10 deg around the fovea) and reduces the auditory bias.

The previous analysis was repeated in the low-contrast case (see Fig. S4 in Supplementary Material part II). Results show that the auditory bias is further reduced in the low-contrast case, especially when training was performed with a reduced number of cross-modal inputs. The visual bias is quite independent on training, and shows a significant bias toward the center.

Finally, we evaluated the effect of a different cross-modal prior on the results, by varying the probability that the auditory and visual stimuli come from independent sources in the cross-modal case (this is parameter β in equations (5) and (6)). In our previous papers (Ursino *et al.*, 2017a,b) this probability was close to zero, while in the simulations of Figs. 2 -9 we used $\beta = 0.2$. Fig. 10 shows the effect of a different probability (0, 0.2, 0.5 and 0.7). As well expected, increasing the probability of cross-modal independence reduces the auditory bias, with a moderate effect on the visual bias too; these effects are negligible when the visual stimulus is at the fovea, but becomes evident when the visual stimulus is located at ± 10 deg.

Re-learning

An important future possible application of the model consists in the study of re-learning, i.e. a condition when the network, starting from a mature configuration, must be able to modify its multisensory integration capacity to match a new environment with different priors. To test this possibility, we re-trained the network assuming a change in the unimodal visual prior after the mature stage was reached. First, the network was trained with a standard deviation for the visual prior as large as $s_V = 30$ deg, starting from the initial naïve condition (i.e., null cross-modal synapses and large receptive fields). This is the same situation shown in the second column of Fig. 9. Then, starting from the mature configuration, the network was trained again (100 epochs) assuming a smaller standard deviation of the visual prior ($s_V = 7$ deg, i.e., the visual stimuli are now more focused close to the fovea). Fig. 11 compares results on the auditory and visual bias obtained: i) by training the network with a visual prior SD as large as 30 deg from the naïve initial conditions; ii) by training the network with a visual prior SD as low as 7 deg, from the naïve initial condition (first column in Fig. 9); iii) by re-learning the prior from SD = 30 deg to SD = 7 deg. As it is clear form Fig. 11, the network can re-learn its multisensory integration characteristics quite well, to reach a final mature stage that approximates the required one.

To better explain this result, an example of how cross-modal synapses change during re-calibration is shown and commented in Fig. S5 of the Supplementary Material part II.

Temporal asynchrony

Various studies in recent years demonstrated that multisensory integration is significantly affected by the temporal correlation between the cross-modal stimuli (Parise *et al*., 2013; Denison *et al*., 2013; Parise and Ernst, 2016; Odegaard *et al*., 2017). Hence, a future important development of the model concerns the study of the temporal aspects of integration. A preliminary result is presented in Fig. 12, where we show the multisensory auditory shift (i.e., the difference between the auditory bias in multisensory and unisensory conditions) computed as a function of the stimulus onset asynchrony (SOA), that is the temporal distance between the start of visual and auditory stimuli. In these simulations we assumed two visual and auditory impulses, with a 50 ms duration each; moreover, we used different values for the strength of the visual stimulus to analyze its impact. For briefness, simulations were performed in noiseless condition. The time constants of the neuron dynamics was 30 ms. Results show that integration decreases with the SOA, and significantly depends on the strength of the stimuli. Moreover, integration is better preserved when the visual stimulus precedes the auditory one, but is more fragile when the auditory stimulus comes first. This asymmetry in the SOA agrees with results shown in van Eijk *et al*., (2008) and Stevenson *et al*., (2014).

Briefly, the asymmetry of the SOA can be explained as follows: a visual representation is spatially much more narrowly tuned, and so, thanks to the presence of cross-modal synapses, induces a significant sub-threshold activation in the auditory net at its spatial location. This sub-threshold auditory bias lasts for the overall duration of the visual activity + about two time constants (the time necessary to decay). Therefore, even 100 ms after the visual stimulus, an auditory stimulus works on an auditory net which is still sub-threshold polarized around the visual location. The opposite condition (auditory stimulus coming first) is not so influential, since the auditory representation, which is broadly tuned, cannot affect the visual net at a particular well-defined position.

Comparison with neurological patients

Another important function of neurocomputational models consists in the simulation of lesions in neurological patients. To this end, we simulated the effect of a damage in one hemifield of the visual net, by silencing a given proportion of the visual neurons coding for the right hemifield (i.e., neurons occupying ordinal position >90 in the visual net and thus coding positive degrees with respect to the fovea). This situation resembles the condition occurring in hemianopic patients, characterized by lateralized damage in the primary visual cortex and consequent loss or reduction of visual responses in the left or right hemifield. Under these conditions, we replicate the same audiovisual simulations as in Figure 5 by computing the auditory bias vs audiovisual distance, for different positions of the visual inputs. For briefness, simulations were performed in noiseless condition and using only one value of visual strength (= 12 as in high contrast condition). Results are shown in Figure 13 for the intact condition and for different levels of network damage: the proportion of silenced visual neurons in the right hemifield was increased from 60% to 100% of their total number (ninety). As the percentage of the damaged neurons increases, the influence that a right visual stimulus may exert on a simultaneous auditory stimulus decreases proportionally. Indeed, the auditory bias induced by the visual inputs at +10° and at +5° gradually declines and eventually disappears as the severity of lesion increases up to 100%.

This is the consequence of the reduced cross-modal synaptic input reaching the auditory area since silenced visual neurons provide no output signal. The network predictions are in line with real patients data (Leo *et al*., 2008, Magosso *et al*., 2016), which show that mislocalization of an auditory stimulus by a spatially disparate visual stimulus is strongly reduced in the hemianopic field. These aspects are further emphasized in the Supplementary Material part II where additional simulations with the lesioned network were performed and paralleled with in vivo data (see Fig. S6).

Discussion

The present results underline that the model can simulate conditions characterized by a variety of inputs (different visual contrasts; differences between C = 1 and C = 2; changes in audio-visual distance; changes in stimulus eccentricity; alterations in unisensory and cross-modal priors during training, including re-learning; temporal asynchrony between the stimuli). It is worth noting that all these aspects have been simulated with a single model, without any change in its internal parameters, but only varying the position, amplitude and statistical occurrence of the inputs.

A significant aspect of our model, compared with previous ones (Cazettes *et al*., 2016, Deneve *et al*., 1999, Fischer and Peña, 2011, Ma *et al*., 2006, Pouget *et al*., 2003), is that synapses can be trained on the basis of past experience, to incorporate the prior probability of past events. This past experience produces two main effects, both clearly visible in behavioral data. First, it induces a ventriloquism effect (basically, an auditory perception shift in the direction of the visual input; but, in case of low visual contrast, also a visual shift in the auditory direction). This cross-modal effect is stored in cross-modal synapses (a feature of our model, not incorporated in previous theoretical works). Second, the past experience produces a significant visual bias, independent of the auditory position, which moves the visual perception toward the fovea, and reflects the prior probability of visual unisensory experience. This is stored in the density of visual receptive fields. A similar but less evident bias occurs in the auditory unisensory perception. By including these aspects into model synapses, via a biological learning rule, we were able to simulate many aspects of behavioral data, by modifying the inputs to the model only.

Input quantities in the model - It is worth noting that, in order to simulate behavioral data, we did not modify any internal parameter in the model (all parameters have exactly the same value as in the former theoretical papers (Ursino *et al*., 2017a, b)) but we only acted on the characteristics of the inputs. In particular, in the previous paper, to show the effect of the azimuthal coordinate, we used a standard deviation of the visual prior as large as 30 deg (i.e., we assumed that unisensory visual inputs become very rare only at the extreme periphery, ±90 deg). However, behavioral data suggest that the effect of eccentricity is already fully evident at a coordinate as low as ±10 deg. These data could be reproduced quite well by our model, but this required the assumption that unisensory visual inputs are almost entirely close to the fovea (i.e., we used a value for the parameter s_V as low as 7 deg, i.e., unisensory visual inputs are almost entirely at a distance ± 20 deg from the centre). This can be justified thinking that head and eye movements almost always move new visual stimuli close to the fovea, and that peripheral visual stimuli are probably of small attentive interest. A similar but less accentuated bias has been used for the auditory inputs too, assuming that head movements can move auditory stimuli toward the center of the head.

Furthermore, during our testing trials we used a poor signal to noise ratio for the auditory and visual inputs (but with two different visual contrast levels). In particular, when simulating the low and high contrast levels, we did not modify the synapses reflecting past experience, but just the present experimental conditions. This agrees with data by Beierholm *et al*. (2009). These authors investigated the independence of the priors from the likelihood, by manipulating the inputs, and confirmed that the estimated prior probabilities are independent of the immediate trials. Our results support this point: we used the network in Training1 to simulate all data by Beierholm *et al*. (2009), i.e., we used just a single prior, but two different likelihoods (with variation as to their strength and noise). The use of large noise in our test phase is justified by the low report of unity observed in the behavioral data, and by the high noise level used in the experimental preparation. If a lower noise level were used (or alternatively, higher input strength were given) the model furnishes values of report of unity much closer to 100% at small A-V distances, in agreement with many other behavioral data (see (Wallace *et al*., 2004) and (Rohe and Noppeney, 2015)).

Finally, we wish to remark that behavioral data used in this work cover only a small portion of stimulus condition (± 20 deg distances) while the model makes predictions also at larger spatial disparity. Hence, the model deserves further validation in future more extensive behavioral studies.

Causal Inference - An important aspect concerns the causal inference problem. In the present paper we assumed that cross-modal stimuli are recognized as a single cause if their perceived distance is less than 2 deg, whereas they are ascribed to two separate causes if their distance is larger. This is substantially the same strategy used in the behavioral data. In fact, in the behavioral data used in the present work, the subject did not respond to whether he/she perceived one cause or two causes, but just indicated the perceived positions of the visual and auditory cues separately.

However, we also tested a different strategy assuming that C = 1 or C = 2 estimation depends on the activity of a third multisensory layer (results are not shown for briefness) . In particular, two different rules were implemented: i) evaluation of the number of peaks in the multisensory layer (see also Cuppini *et al.,* 2017 for more details); ii) computing the cross-correlation between the activities in the auditory and visual unisensory layers (in fact, cross-correlation can be implemented via a multisensory layer, using logarithms of inputs activities to convert products or divisions into sums or subtractions). We observed that the results of Figs. 4-8 can be simulated quite well implementing causal inference with a third layer too, but with some discrepancies from behavioral data: in particular, in the high-contrast case, the report of unity only scarcely decreases with the audio-visual distance.

Hence, the C = 1 or C = 2 cases in the examined behavioral data can be better reproduced using a simple index of the perceived spatial separation, rather than a thorough causal inference based on a more complex multisensory layer.

Bayesian inference - Some comments on why the present results support Bayesian inference may be of value. The Bayesian estimate depends on two contributions: the likelihood, which encompasses the reliability of the present inputs, and the prior, which incorporates previous experience.

In the trials with *higher visual contrast*, the visual cues exhibit higher reliability compared with the auditory cues, and so, according to a maximum likelihood strategy, the network gives more confidence to the visual estimate than to the auditory one. Two major priors affect these results. A conditional prior establishes that, in cross-modal conditions, auditory and visual cues often originate from proximal positions (at least when C = 1), i.e., they are perceived as a single cause; a unisensory prior, which is strongly focused near the fovea for the visual cues. In both conditions (C = 1 and C = 2) the visual perception is just barely affected by the auditory one, and the visual bias only reflects a balance between a strong visual reliability (the visual likelihood) and its unisensory prior: the visual input is perceived as moderately shifted closer to the fovea. Conversely, the auditory perception, which is less reliable than the visual one, is strongly affected by the visual position and less affected by the likelihood: this is extremely evident in the C = 1 case, when the prior conditional probability plays a major role, and is less evident when C = 2. Furthermore, the model predicts that the auditory bias is larger for smaller AV disparities than for larger disparities, a result linked to causal inference (indeed, a small disparity, which suggests a common source, leads to a stronger auditory bias). This behavior is unfortunately less evident in behavioral data, but is supported by data in Wallace *et al*. (2004).

In the trials with *reduced visual contrast*, the reliability of the visual and auditory cues are more comparable (but with the auditory cue still less precise than the visual one): as a consequence, in the C = 1 case, the auditory perception exhibits a reduced shift, while the visual perception exhibits a clear shift in the direction of the auditory one (i.e., a sort of "visual ventriloquism"). The latter effect is especially evident at the higher values of the azimuthal coordinate (± 10 deg) where the accuracy of the visual RFs is poorer than at the center.

These aspects can be seen clearly both in the behavioral data and in the corresponding model simulations, which exhibit quite a satisfactory agreement in all conditions tested.

An alternative model to the present, to infer causal inference, is The Bayesian Causal Inference model (Kording *et al*. 2007 ; Beierholm *et al.*, 2009; Wozny *et al.*, 2010 ; , Odegaard *et al.*, 2015), which has been shown to provide a good account to behavioral data in spatial localization. However this is a computational model and does not address how the underlying neural mechanisms are, and how such computations can be implemented by the neural machinery of the brain. The present model is a neurophysiological model, aimed precisely to shed light on the neural mechanisms involved in this process. The parameters of the current model were not fitted to the behavioral data, nor were they modified compared with the model published previously (Ursino *et al.,* 2017a), other than the distribution of the inputs. An improved agreement to the data could be potentially achieved by fitting some parameters to the data, as well as by simulating individual observer's data.

The effect of a different training (*Training2*, *Training3* and *Trainings 4-6*) is also in line with Bayesian ideas. If the percentage of cross-modal inputs is reduced, the network poses less weight on the C = 1 hypothesis than on the C = 2, resulting in a strongly reduced auditory ventriloquism. A larger distribution of visual inputs around the fovea, in turn, also reduces the auditory and visual bias. Finally, a greater probability of independent cross-modal inputs (i.e., parameter β in Eq. 4) moderately reduces the auditory bias, especially when an eccentric visual stimulus is used.

Re-learning – A new aspect of this work, never tested before, concerns the capacity of the model to re-learn a new prior, starting from a previous mature configuration. In particular, Fig. 11 shows that model behavior after re-learning approximates the behavior of a network that was trained with the second prior from the very beginning (i.e., from the immature configuration). This is made possible by the learning rule adopted, which includes a forgetting factor to progressively dissipate those aspects of the environment statistics that are no longer occurring.

Hence, in perspective the network could be used in non-stationary environments too, to investigate the exploitation/exploration trade off. Of course, the simulations of Fig. 11 are just preliminary. More complex non-stationary scenarios should be tested in future work, to further challenge the adopted learning rule against non-stationarity and perhaps to improve this rule.

Another possibility, in the future, is to compare the multisensory development in the model with behavioral data acquired from early infancy to adultness. In fact, some recent studies suggest that the capacity to fuse different sensory information emerges only quite late during development (typically after 10-11 year old, Dekker *et al*. (2015), or after 12 year old, Nardini *et al*. (2010)) and that the brain circuits that merge senses take very long time to mature (Dekker *et al*., 2015). The present model may be worthwhile to provide a quantitative framework for the analysis of these developmental scenarios.

Temporal aspects – A further important novelty concerns the study of temporal differences among the inputs. In particular, we tested the onset asynchrony between two cross-modal impulse stimuli. Results confirm that integration (tested as the difference between the multisensory and unisensory auditory bias) is strongly affected by the temporal discrepancy, and suggest that integration is better preserved when the visual stimulus precedes the auditory one. The latter result finds some support in experimental works which evaluated the multisensory temporal binding window (see van Eijk *et al*., 2008 and Stevenson *et al*., 2012).

However, we wish to stress that the particular temporal window shown in Fig. 12 critically depends on the duration of the impulses used as input (50 ms) and the time constant employed (30 ms). Indeed, it is difficult to establish correct values for these parameters, since they reflect not only properties of the network, but also the overall pre-processing of the sensory inputs, from the retina and the cochlea to the cortex via the thalamic pathways.

More generally, various recent pivotal papers (Parise *et al*., 2012; Denison *et al*., 2013; Parise *et al*., 2013) stressed that multisensory integration is affected by the cross-correlation among the stimuli, and that strong correlation results in stronger integration. A recent model by Parise and Ernst (2016) also incorporates a correlation detector to replicate human perception data. Analysis of the effect of complex temporal correlations may be a future application of this model, maybe including a downstream layer which detects correlation explicitly to infer causal inference and affect multisensory integration.

Clinical aspects – In this work we also presented an example of possible model use in a clinical setting. In particular, the model can replicate some aspects of hemianopia, characterized by a progressive loss of ventriloquism. Indeed, the study of a lesioned network can provide further important elements, both to validate the model, to reach a deeper understanding of the neurophysiological mechanism implicated in pathological behavior, as well as to delineate a possible model use in neurological rehabilitation procedures.

Limitations of the present work - Finally, it is important to point out some limits in the present work and lines for future studies. A first limitation consists in the representation of the auditory network. In both networks, we assumed a topological spatial organization of neurons. While a topological organization is well documented for what concerns the visual primary and secondary areas in the cortex, this is not documented in the auditory cortex. Indeed, the primary auditory cortex is not spatially organized, and spatial information is calculated indirectly from interaural time difference or interaural phase difference, even though a simpler spatial organization is present at the hemispheric level (i.e., the left primary auditory cortex prefers right auditory stimuli and viceversa, see Ortiz-Rios *et al.,* 2017). Nevertheless, we think that the basic idea of our model, i.e., that conditional priors can be realized via cross modal synapses linking elements of the visual and auditory nets participating to the same task, is still valid as a direct consequence of the Hebb rule (see Ursino *et al*., 2015 and Zhang *et al*., 2016). In other words, this major assumption is quite independent of the real positions of neurons in the auditory cortex, but it is especially affected by neuron activation during perception. A more physiological description of the auditory processing stage will be the subject of subsequent extended versions of the model.

A further limitation concerns realignment between auditory and visual cues during head and eye movements. In our model, we assumed that the auditory and visual maps are always aligned, not only during testing (which may be a consequence of fixed head and eyes), but also in the previous training phase. The problem is extremely complex and would require an additional model that works upstream the present, to align maps as a function of retinal and head motion.

A last limitation concerns a kind of adaptation to the ventriloquism effect, named “aftereffect” (Bertelson et al., 2006; Wosny and Shams, 2011). In this adaptation, following a training period in which the auditory and visual stimuli have a constant discrepancy, the perceived location of even a unisensory auditory stimulus is shifted toward the visual side. It is worth noting that this phenomenon cannot be simulated by the two mechanisms included in the present work, i.e., recalibration of the receptive fields and adjustments in cross-modal synapses. In fact, the first mechanism moves the receptive field toward the most frequent unisensory stimuli, whereas the second is efficacious only during bi-sensory stimulation. In a previous work (Magosso et al., 2012) we explained the aftereffect trough a change in *lateral synapses*. Indeed, this mechanism can describe the unisensory aftereffect, since excitatory synapses become stronger toward the position stimulated by the visual input (assuming a constant AV distance during training) and then can subsequently affect the auditory perception in the *unisensory case* too. Since in the present work we never used a constant bias between the auditory and visual inputs during training (i.e., audio-visual stimuli were either coincident or randomly placed), we did not train lateral synapses for the sake of simplicity. Nevertheless, we claim that plasticity of lateral synapses may allow the simulation of experiments by Bertelson et colleagues (2006), when the audio-visual discrepancy is fixed, by causing a shift in the likelihood (see Magosso et al., 2012). This agrees with some ideas in Wozny and Shams (2011) suggesting that the aftereffect shift in the perceived auditory locations is associated with a shift in the mean of the auditory likelihood functions in the direction of the experienced visual offset.

Legends to figures

Fig. 1 – Neural network used in the present work. r_{kj} (red lines) represent receptive fields entering into auditory and visual neuros. λ_{kj} (blue lines) are lateral synapses with a Mexican Hat disposition, connecting neurons in the same modality. w_{kj} (green lines) are cross-modal synapses connecting neurons of different modalities. Synapses in the Receptive Fields and Cross-modal synapses are modified by the experience, using a Hebb rule with a forgetting factor.

Fig. 2 – Some examples of the receptive field (RF) training. Here, the azimuthal space on x-axis has been mapped from -90 deg to +90 deg, with 0 deg representing the fovea. The upper panels represent the RFs of two auditory neurons occupying two different ordinal positions in the net: the neuron at the ordinal position 50, i.e. with initial preferred position at -40 deg from the fovea, and the neuron at position 90, i.e. with initial preferred position at 0 deg in the fovea. The bottom panels represent the RFs of two visual neurons, with the same initial preferred positions. The RFs shrink during training, to meet the same accuracy as the average inputs. In particular, the visual RFs become sharper near the fovea than at the periphery (compare right vs left bottom panel). Moreover, the RF of the visual and auditory neurons, initially located at -40 deg from the fovea (left panels), exhibit a significant shift toward the fovea.

Fig. 3 – Arrangement of the synapses at the end of Training1. These synapses have been used during the testing phase. The upper panels show the final Receptive Fields (RFs) of some auditory (left) and some visual (right) neurons. In particular, we focus attention on 17 neurons occupying ordinal positions within the net ranging from 10 to 170, with a step of 10. The bottom panels show the cross-modal synapses entering into auditory (left) and visual (right) neurons (at the same positions as the in the upper panels), from all neurons of the other modality.

Fig. 4 – Report of Unity (fraction of trials with C = 1 based on the perceived audio-visual distance) plotted as a function of the real audio-visual distance, obtained with the model (left) and from behavioral

data (right). Continuous lines refer to the trials with high-contrast visual inputs; the dashed lines with low-contrast visual inputs.

Fig. 5 – Bias in the perception of an *auditory* stimulus (perceived position minus real position), simulated with the model (upper panels) and obtained from behavioral data (bottom panels). The figure refers to the trials with *high-contrast visual inputs*. The first column considers all results. The second column considers just the cases with C = 1 (distance between the auditory and visual perceptions less than 2 deg); the right panel considers only the cases with C = 2 (distance between the auditory and visual perceptions greater than 2 deg). During the trials, the visual stimulus was positioned at five different azimuthal coordinates (ranging from -10 deg to + 10 deg from the fovea) and, at each visual position, an auditory stimulus was superimposed, with an audio-visual distance ranging between −20 and + 20 deg in the model. Absence of points in the upper middle panel for some large values of A-V distances (especially in case of the magenta and blue lines) is due to the fact that in the model, C=1 did not occur in those circumstances in case of high visual contrast. In the lower panels (behavioral data), only five A-V distances were evaluated for each position of the visual stimulus, since, in the behavioral tests, only positions in the azimuthal space between -10 and + 10 deg with 5 deg step were used, which limits the number of possible audio-visual distances actually tested.

Fig. 6 - Bias in the perception of a *visual* stimulus (perceived position minus real position), simulated with the model (upper panels, same trials as in Figure 5) and obtained from behavioral data (bottom panels). The figure refers to the trials with *high-contrast visual inputs*. The meaning of panels and of lines is the same as in Fig. 5.

Fig. 7 - Bias in the perception of an *auditory* stimulus, simulated with the model (upper panels) and obtained from behavioral data (bottom panels). The figure refers to the trials with *low-contrast visual inputs*. The meaning of panels and of lines is the same as in Fig. 5.

Fig. 8 - Bias in the perception of a *visual* stimulus, simulated with the model (upper panels) and obtained from behavioral data (bottom panels). The figure refers to the trials with *low-contrast visual inputs*. The meaning of panels and of lines is the same as in Fig. 5.

Fig. 9 – Dependence of model results on the stimuli experienced during training (i.e., on the *prior probability*). 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 in the first column of Figure 5 (i.e., for brevity we show just results obtained by considering all trials, without a distinction between the C =1 and C = 2 cases, and by considering high-contrast visual inputs). The first column was obtained after Training1 (that is the same used in Figures 2-8). The second column was obtained after a different training (Training2) characterized by a larger spatial arrangement of visual stimuli around the fovea. In these conditions, the auditory and visual bias are reduced. The third column was obtained after another different training (Training3), characterized by a smaller percentage of cross-modal inputs. In these conditions, the auditory ventriloquism is dramatically reduced, but the constant visual bias is almost unaffected.

Figure 10 – Effect of a different cross-modal prior, obtained by changing the probability that auditory and visual stimuli come from independent positions (probability values $\beta = 0$, 0.2, 0.5 and 0.7). The upper line shows the auditory bias vs. the audio-visual distance, whereas the bottom line shows the visual bias. The visual stimulus was positioned at -10 deg from the fovea (left colun), at the fovea (central column), and at + 10 deg from the fovea (right column). The assumption of a larger probability of independent cross-modal stimuli is reflected in a moderate reduction of the auditory bias evident at the positions ± 10 deg.

Figure 11 – *Effect of re-learning*. The upper line shows the auditory bias vs. the audio-visual distance, whereas the bottom line shows the visual bias. The visual stimulus was positioned at -10 deg from the fovea (left column), at the fovea (central column), and at + 10 deg from the fovea (right column). The green lines were obtained using the mature net trained with a standard deviation of the visual prior s_V= 30 deg in Eq. (2). The blue lines were obtained using a mature net trained with a standard deviation of the visual prior s_V = 7 deg. The red line was obtained after a re-learning, starting from the mature net trained with s_V = 30 deg, and using 100 additional training epochs performed with s_V = 7 deg. The net can re-learn the new prior quite well, reaching a final configuration proximal to the expected one.

Figure 12 – Effect of the Stimulus Onset Asynchrony (SOA) on the ventriloquism effect (computed as the difference between the multisensory and unisensory auditory bias). All simulations were performed in noiseless condition using two cross-modal stimuli with a 50 ms duration each. The time constants of the auditory and visual neurons were 30 ms. The position of the auditory stimulus was at -15 deg from the fovea, while the visual stimulus was located at the fovea. The strength of the auditory input was the same as in Table 1, while three different strengths were used for the visual stimulus (12, 20 and 34 respectively) to emphasize its effect. Multisensory integration is affected by the SOA, and is more robust when the visual stimulus precedes the auditory one (positive values of SOA) than when the auditory stimulus comes first (negative values of the SOA).

Figure 13 – *Effect of lesion* - Auditory bias (perceived auditory position minus real auditory position) simulated with the intact (left upper panel) and damaged network (other panels) vs. the visual-auditory distance, computed for different positions of the visual stimulus. All simulations were performed in noiseless condition and using a visual stimulus with strength = 12 and an auditory stimulus with strength = 36. The damage consists in silencing a percentage of visual neurons coding for the right hemifield

(positive degrees), simulating conditions of right hemianopia. The positions of the silenced neurons were chosen randomly within the right hemifield. Results for three different levels of damage (60%, 80% and 100%) are reported. As the level of lesion increases, the impact of the right visual stimuli (at +5° and + 10°) on auditory bias tends to vanish.

Table 1

Input values for the stimuli (strength and noise), used during the training and testing phases.

Training:	$i_{A,Strength} = 36$	$i_{V,Strength} = 18$	$i_{A,Strength} / \upsilon_A = 4$	$i_{V,Strength} / \upsilon_V = 4$
Testing: high-contrast	$i_{A,Strength} = 36$	$i_{V,Strength} = 12$	$i_{A,Strength} / \upsilon_A = 1$	$i_{V,Strength} / \upsilon_V = 4$
Testing: low-contrast	$i_{A,Strength} = 36$	$i_{V,Strength} = 8$	$i_{A,Strength} / \upsilon_A = 1$	$i_{V,Strength} / \upsilon_V = 0.5$

References

Alais, D. and Burr, D. (2004) The ventriloquist effect results from near-optimal bimodal integration. *Current biology,* 14 (3), pp. 257-262.

Battaglia, P.W., Jacobs, R.A. and Aslin, R.N. (2003) Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision* 20 (7), pp. 1391-1397.

Beierholm, U.R., Quartz, S.R. and Shams, L. (2009) Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of vision,* 9 (5) 23, pp. 1-9.

Bertelson, P. et al. (2006) The aftereffects of ventriloquism: patterns of spatial generalization. Perception & Psychophysics, 68(3), pp. 428-436.

Bülthoff, H.H. and Mallot, H.A. (1988) Integration of depth modules: stereo and shading. *Journal of the Optical Society of America A,* 5 (10), pp. 1749-1758.

Cazettes, F., Fischer, B.J. and Peña, J.L. (2016) Cue reliability represented in the shape of tuning curves in the owl's sound localization system. *Journal of Neuroscience,* 36 (7), pp. 2101-2110.

Charbonneau, G. *et al.* (2013) The ventriloquist in periphery: impact of eccentricity-related reliability on audio-visual localization. *Journal of vision,* 13 (12), pp. 20-20.

Cuppini, C. *et al.* (2014) A neurocomputational analysis of the sound-induced flash illusion. *NeuroImage,* 92, pp. 248-266.

Cuppini, C. *et al.* (2017) A biologically inspired neurocomputational model for audiovisual integration and causal inference. *European Journal of Neuroscience,* 46 (9), pp. 2481-2498.

Dekker, Tessa M. *et al.* (2015) Late Development of Cue Integration Is Linked to Sensory Fusion in Cortex. *Current Biology,* 25 (21), pp. 2856-2861.

Deneve, S., Latham, P.E. and Pouget, A. (1999) Reading population codes: a neural implementation of ideal observers. *Nature neuroscience,* 2 (8), pp. 740-745.

Denison, R.N., Driver, J. and Ruff, C.C. (2013) Temporal structure and complexity affect audio-visual correspondence detection. *Frontiers in psychology*, *3*, p. 619.

Driver, J. and Noesselt, T. (2008) Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron,* 57 (1), pp. 11-23.

Ernst, M.O. (2012) Optimal multisensory integration: Assumptions and limits. In B. E. Stein (Ed.), The new handbook of multisensory processes (pp. 1084-1124). Cambridge, MA: MIT Press.

Ernst, M.O. and Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature,* 415 (6870), pp. 429-433.

Ernst, M.O. and Di Luca, M. (2011) Multisensory perception: from integration to remapping. *Sensory cue integration*, pp.224-250.

Fetsch, C.R. *et al.* (2012) Neural correlates of reliability-based cue weighting during multisensory integration. *Nature neuroscience,* 15 (1), pp. 146-154.

Fischer, B.J. and Peña, J.L. (2011) Owl's behavior and neural representation predicted by Bayesian inference. *Nature neuroscience,* 14 (8), pp. 1061-1066.

Fiser, J. *et al.* (2010) Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences,* 14 (3), pp. 119-130.

Garcia, S.E. *et al.* (2017) Auditory Localisation Biases Increase with Sensory Uncertainty. *Scientific reports,* 7.

Ghazanfar, A.A. and Schroeder, C.E. (2006) Is neocortex essentially multisensory? *Trends in Cognitive Sciences,* 10 (6), pp. 278-285.

Hairston, W.D. *et al.* (2003) Visual localization ability influences cross-modal bias. *Journal of Cognitive Neuroscience,* 15 (1), pp. 20-29.

Jacobs, R.A. (1999) Optimal integration of texture and motion cues to depth. *Vision Res,* 39 (21), pp. 3621-3629.

Körding, K.P. *et al.* (2007) Causal inference in multisensory perception. *PLoS one,* 2 (9), pp. e943.

Leo, F., *et al.* (2008) Cross-modal localization in hemianopia: new insights on multisensory integration. *Brain,* 131 (3), pp. 855-865.

Lewald, J., Dörrscheidt, G.J. and Ehrenstein, W.H. (2000) Sound localization with eccentric head position. *Behavioural brain research,* 108 (2), pp. 105-125.

Ma, W.J. *et al.* (2006) Bayesian inference with probabilistic population codes. *Nature neuroscience,* 9 (11), pp. 1432-1438.

Magosso, E., Cuppini, C. and Ursino, M. (2012) A Neural Network Model of Ventriloquism Effect and Aftereffect. *Plos One,* 7 (8), pp. e42503.

Magosso, E. *et al.* (2010) Neural bases of peri-hand space plasticity through tool-use: Insights from a combined computational–experimental approach. *Neuropsychologia,* 48 (3), pp. 812-830.

Magosso, E., *et al.* (2016) Audiovisual integration in hemianopia: a neurocomputational account based on cortico-collicular interaction. *Neuropsychologia,* 91, pp. 120-140.

Nardini, M., Bedford, R. and Mareschal, D. (2010) Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences,* 107 (39), pp. 17041-17046.

Odegaard, B. and Shams, L. (2016) The brain's tendency to bind audiovisual signals is stable but not general. *Psychological science*, *27*(4), pp. 583-591.

Odegaard, B., Wozny, D.R. and Shams, L. (2015) Biases in visual, auditory, and audiovisual perception of space. *PLoS computational biology,* 11 (12), pp. e1004649.

Odegaard, B., Wozny, D.R. and Shams, L. (2017) A simple and efficient method to enhance audiovisual binding tendencies. *PeerJ*, *5*, pp. e3143.

Ortiz-Rios, M., *et al.* (2017) Widespread and opponent fMRI signals represent sound location in Macaque auditory cortex. *Neuron,* 93 (4), pp. 971-983.

Parise, C.V., Spence, C. and Ernst, M.O. (2012) When correlation implies causation in multisensory integration. *Current Biology*, *22*(1), pp. 46-49.

Parise, C.V. and Ernst, M.O. (2016) Correlation detection as a general mechanism for multisensory integration. *Nature communications*, *7*, p.11543.

Parise, C.V., *et al.* (2013) Cross-correlation between auditory and visual signals promotes multisensory integration. *Multisensory research,* 26 (3) pp. 307-316.

Patton, P.E. and Anastasio, T.J. (2003) Modeling cross-modal enhancement and modality-specific suppression in multisensory neurons. *Neural Computation,* 15 (4), pp. 783-810.

Pouget, A. *et al.* (2013) Probabilistic brains: knowns and unknowns. *Nature neuroscience,* 16 (9), pp. 1170.

Pouget, A., Dayan, P. and Zemel, R.S. (2003) Inference and computation with population codes. *Annual review of neuroscience,* 26 (1), pp. 381-410.

Roach, N.W., Heron, J. and McGraw, P.V. (2006) Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 273, 2159-2168.

Rohe, T. and Noppeney, U. (2015) Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision,* 15 (5), pp. 22-22.

Shams, L., Ma, W.J. and Beierholm, U. (2005) Sound-induced flash illusion as an optimal percept. *NeuroReport,* 16 (17), pp. 1923-1927.

Shams, L. and Beierholm, U.R. (2010) Causal inference in perception. *Trends in cognitive sciences*, *14*(9), pp. 425-432.

Stevenson, R.A., Zemtsov R.K., and Wallace M.T. (2012) Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance,* 38 (6), pp. 1517

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

Ursino, M., Cuppini, C. and Magosso, E. (2014) Neurocomputational approaches to modelling multisensory integration in the brain: a review. *Neural Networks,* 60 141-165.

Ursino, M., Cuppini, C. and Magosso, E. (2015) A neural network for learning the meaning of objects and words from a featural representation. *Neural Networks,* 63 234-253.

Ursino, M., Cuppini, C. and Magosso, E. (2017b) Multisensory Bayesian inference depends on synapse maturation during training: theoretical analysis and neural modeling implementation. *Neural computation,* 29 (3), pp. 735-782.

van Eijk, R.L., *et al*. (2008) Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. *Perception & psychophysics*, 70(6) pp. 955-968.

van Dam, L.C.J., Parise, C.V. and Ernst M. O. (2014) Modeling multisensory integration, in: Sensory Integration and the Unity of Consciousness, Bennett, D.J. and Hill C.S. (Eds), pp. 209-229, MIT Press, Cambridge MA, USA.

Wallace, M.T. *et al.* (2004) Unifying multisensory signals across time and space. *Experimental Brain Research,* 158 (2), pp. 252-258.

Wozny, D.R., Beierholm, U.R. and Shams, L. (2008) Human trimodal perception follows optimal statistical inference. *Journal of vision,* 8 (3), pp. 24-24.

Wozny, D.R., Beierholm, U.R. and Shams, L. (2010) Probability matching as a computational strategy used in perception. *PLoS computational biology*, *6*(8), pp. e1000871.

Wozny, D.R. and Shams, L. (2011) Computational characterization of visually induced auditory spatial adaptation. *Frontiers in integrative neuroscience*, *5*, p.75.

Zhang, W.-h. *et al.* (2016) Decentralized multisensory information integration in neural systems. *Journal of Neuroscience,* 36 (2), pp. 532-547.

Figure 1

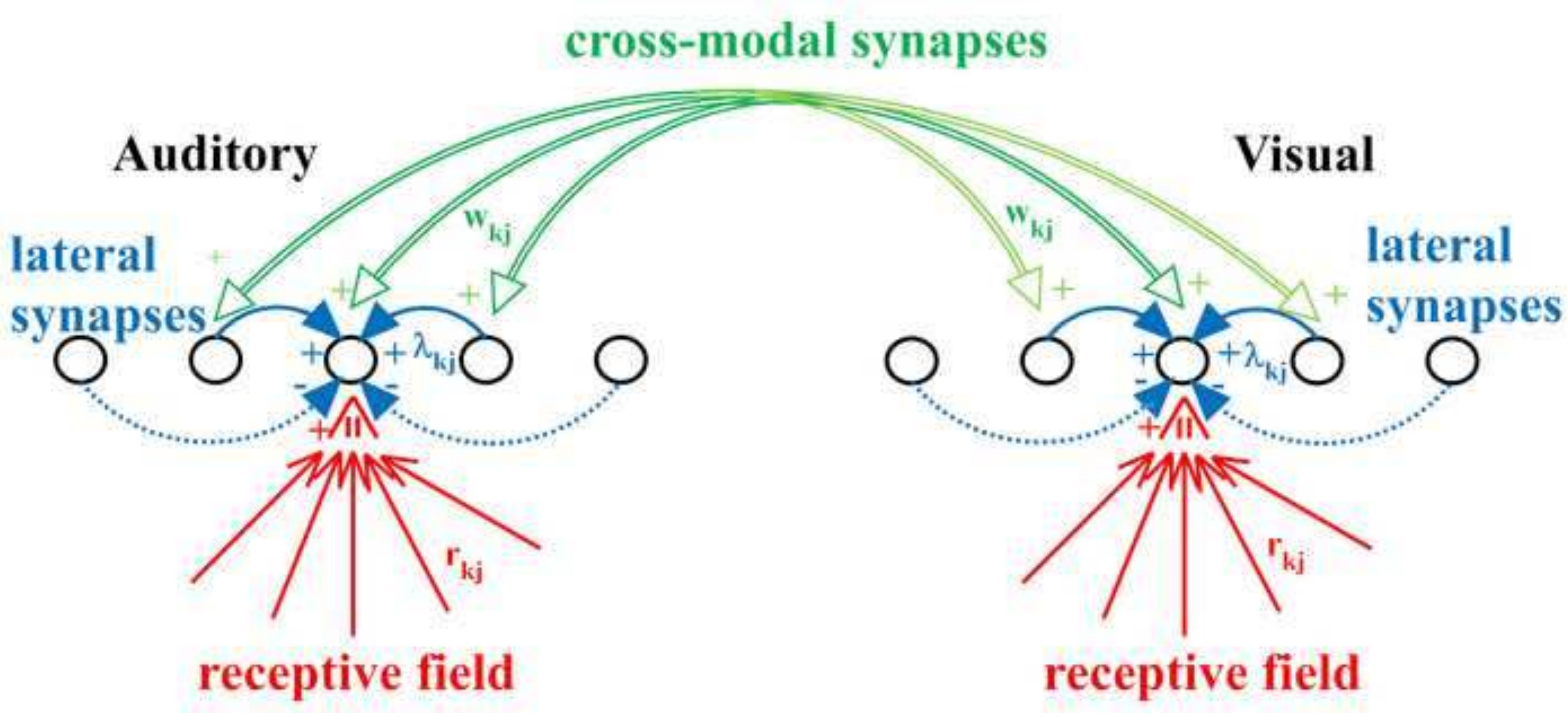
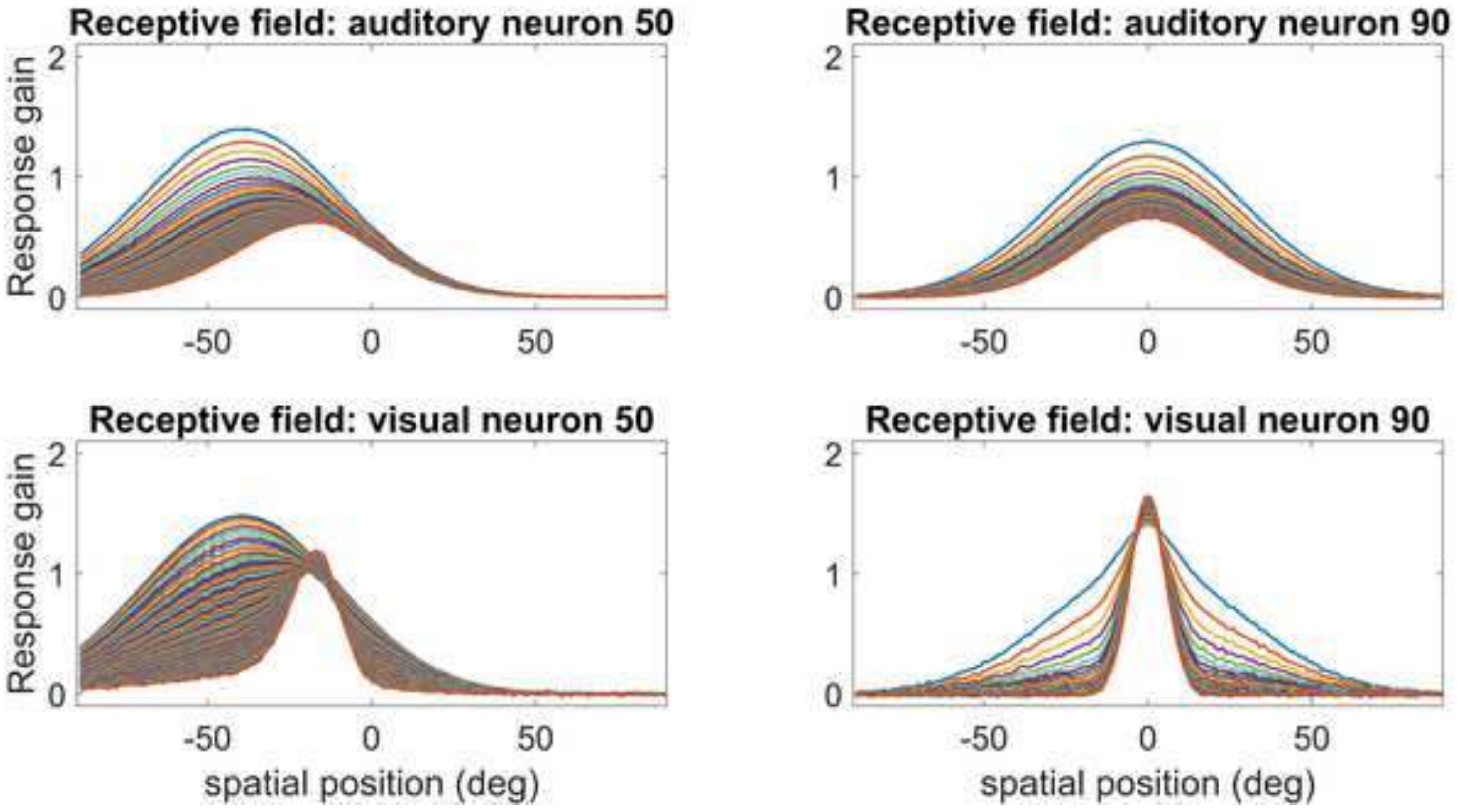

cross-modal synapses
Auditory
Visual
lateral synapses
lateral synapses
w_{kj}
w_{kj}
$+\lambda_{kj}$
$+\lambda_{kj}$
+
-
r_{kj}
r_{kj}
receptive field
receptive field

Figure 2



Receptive field: auditory neuron 50

Receptive field: auditory neuron 90

Receptive field: visual neuron 50

Receptive field: visual neuron 90

Figure 3

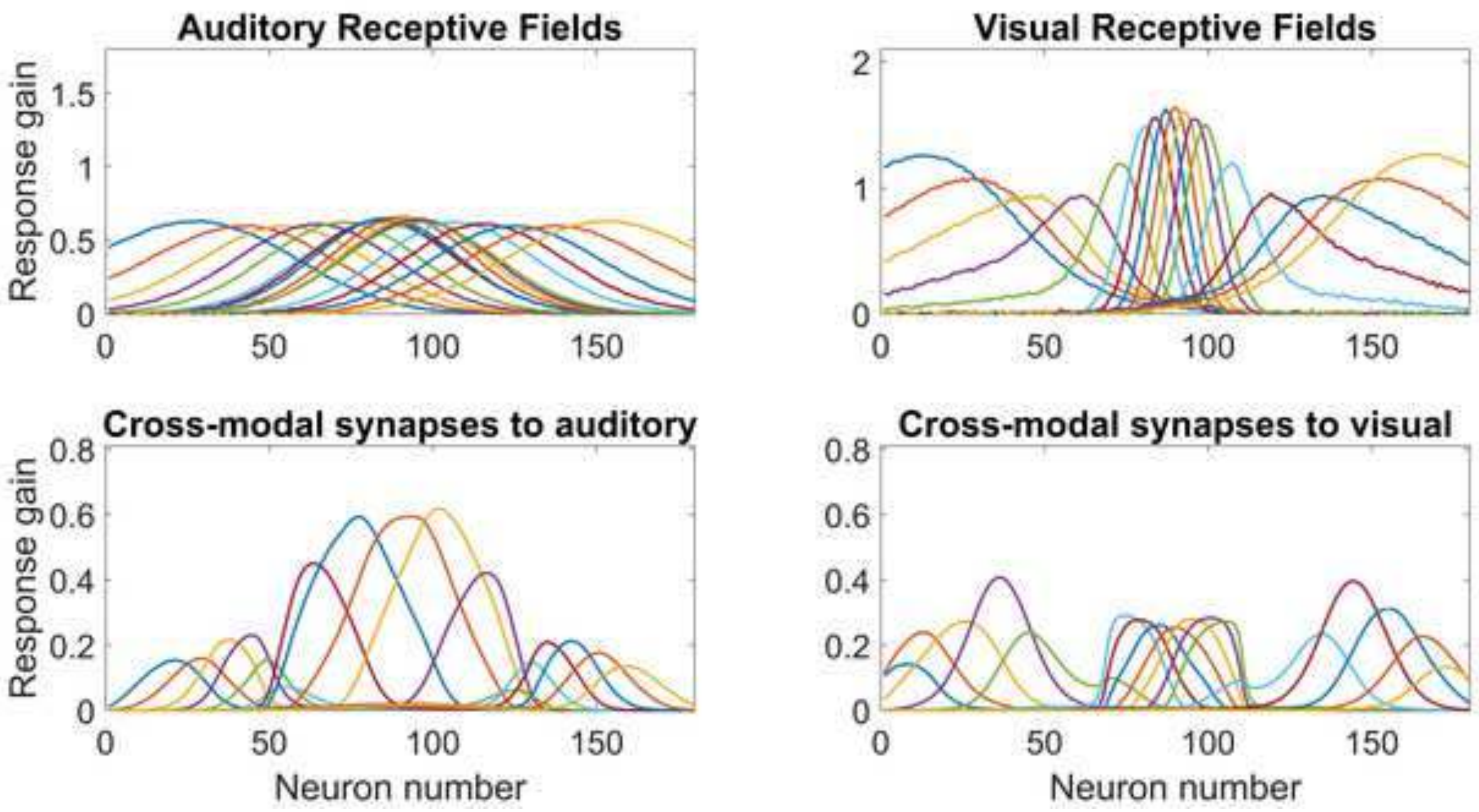

Auditory Receptive Fields
Response gain
1.5
1
0.5
0
0
50
100
150
Visual Receptive Fields
2
1
0
0
50
100
150
Cross-modal synapses to auditory
Response gain
0.8
0.6
0.4
0.2
0
0
50
100
150
Neuron number
Cross-modal synapses to visual
0.8
0.6
0.4
0.2
0
0
50
100
150
Neuron number

Figure 4

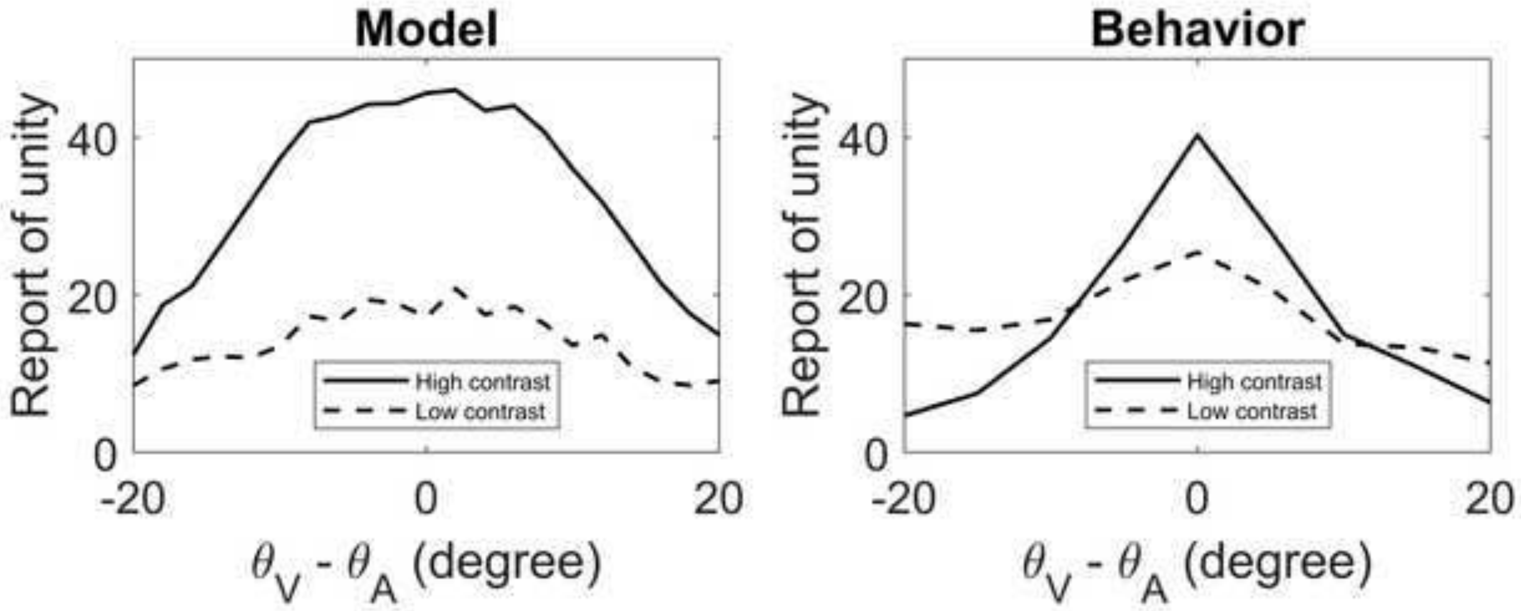

Model
Behavior
Report of unity
Report of unity
0
20
40
0
20
40
-20
0
20
-20
0
20
$\theta_V - \theta_A$ (degree)
$\theta_V - \theta_A$ (degree)
High contrast
Low contrast
High contrast
Low contrast

Figure 5

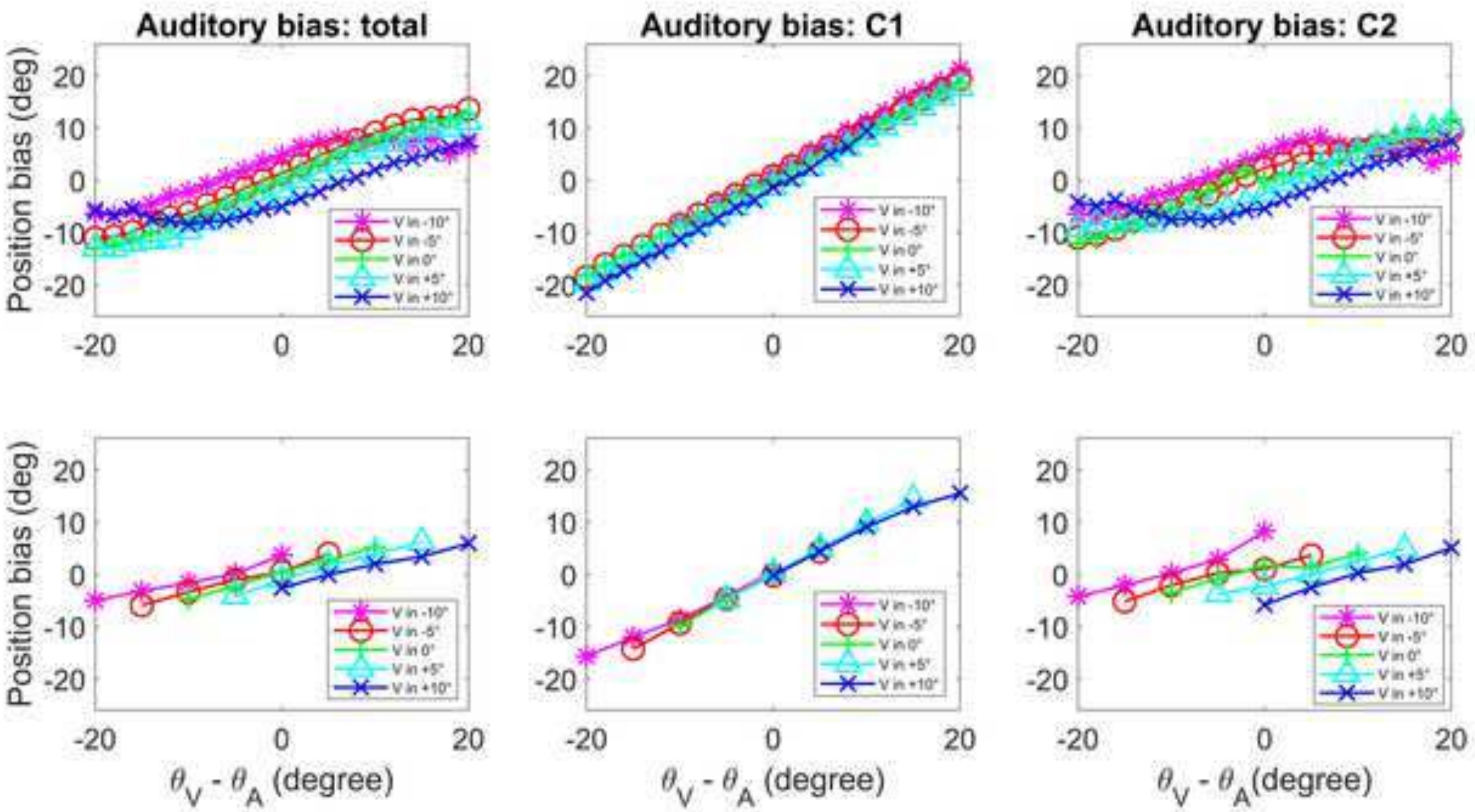

Auditory bias: total
Auditory bias: C1
Auditory bias: C2
Position bias (deg)
Position bias (deg)
20
10
0
-10
-20
-20
0
20
V in -10°
V in -5°
V in 0°
V in +5°
V in +10°
θ_V - θ_A (degree)
θ_V - θ_A (degree)
θ_V - θ_A(degree)

Figure 6

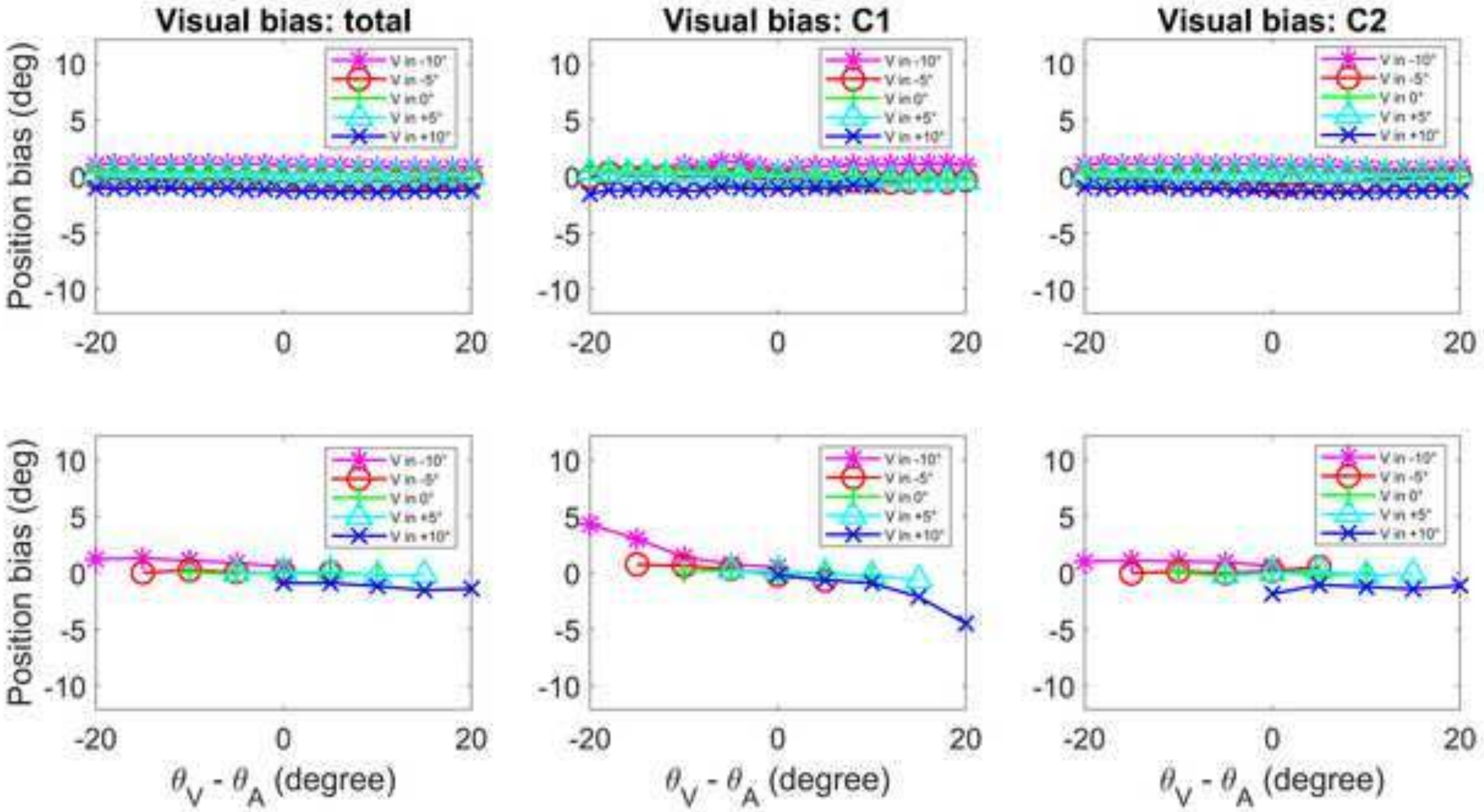

Visual bias: total
Visual bias: C1
Visual bias: C2
Position bias (deg)
10
5
0
-5
-10
-20
0
20
$\theta_V - \theta_A$ (degree)
V in -10°
V in -5°
V in 0°
V in +5°
V in +10°

Figure 7

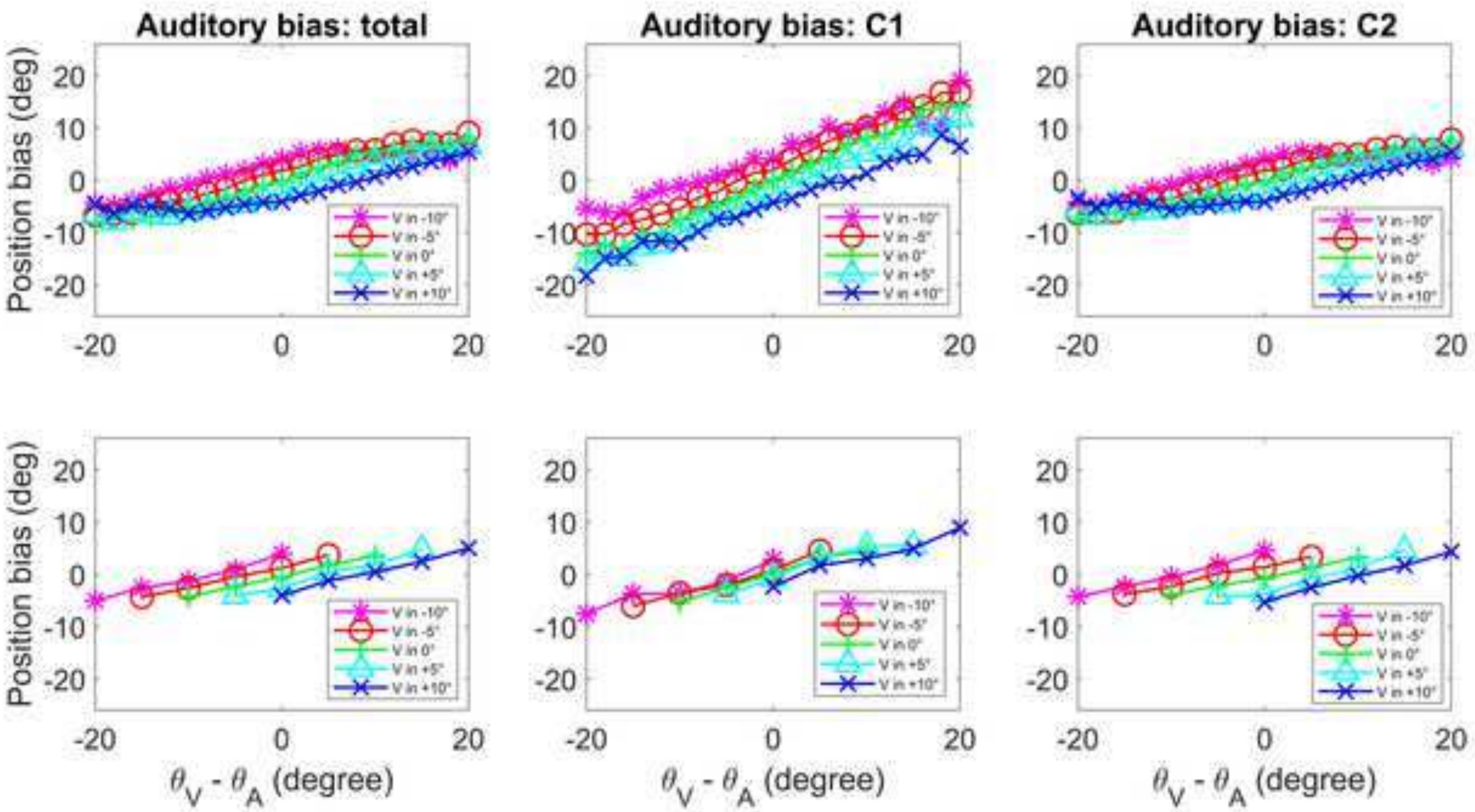

Auditory bias: total
Auditory bias: C1
Auditory bias: C2
Position bias (deg)
20
10
0
-10
-20
-20
0
20
θ_V - θ_A (degree)
V in -10°
V in -5°
V in 0°
V in +5°
V in +10°

Figure 8

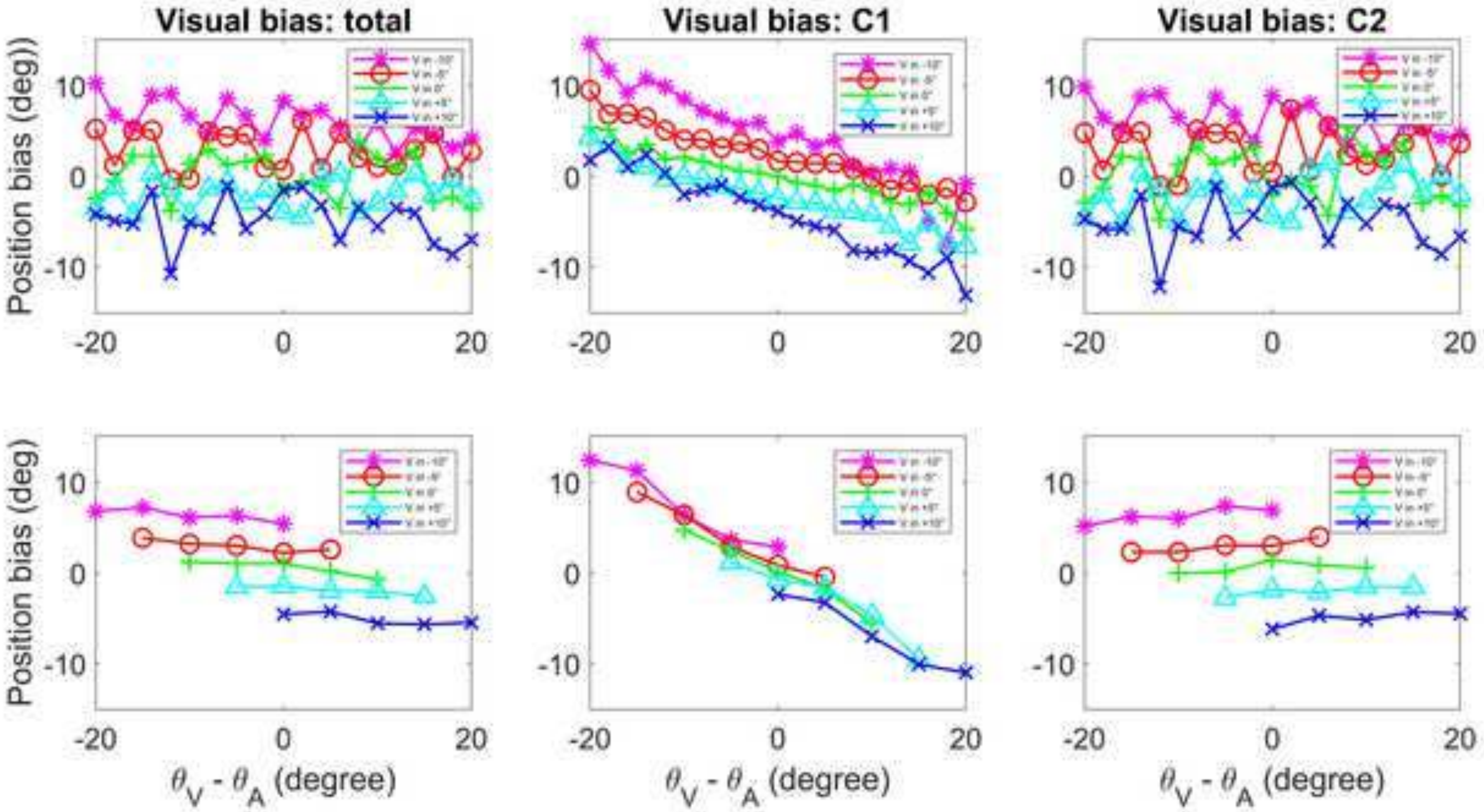

Visual bias: total
Visual bias: C1
Visual bias: C2
Position bias (deg))
Position bias (deg)
V in -10°
V in -5°
V in 0°
V in +5°
V in +10°
10
0
-10
-20
0
20
θ_V - θ_A (degree)

Figure 9

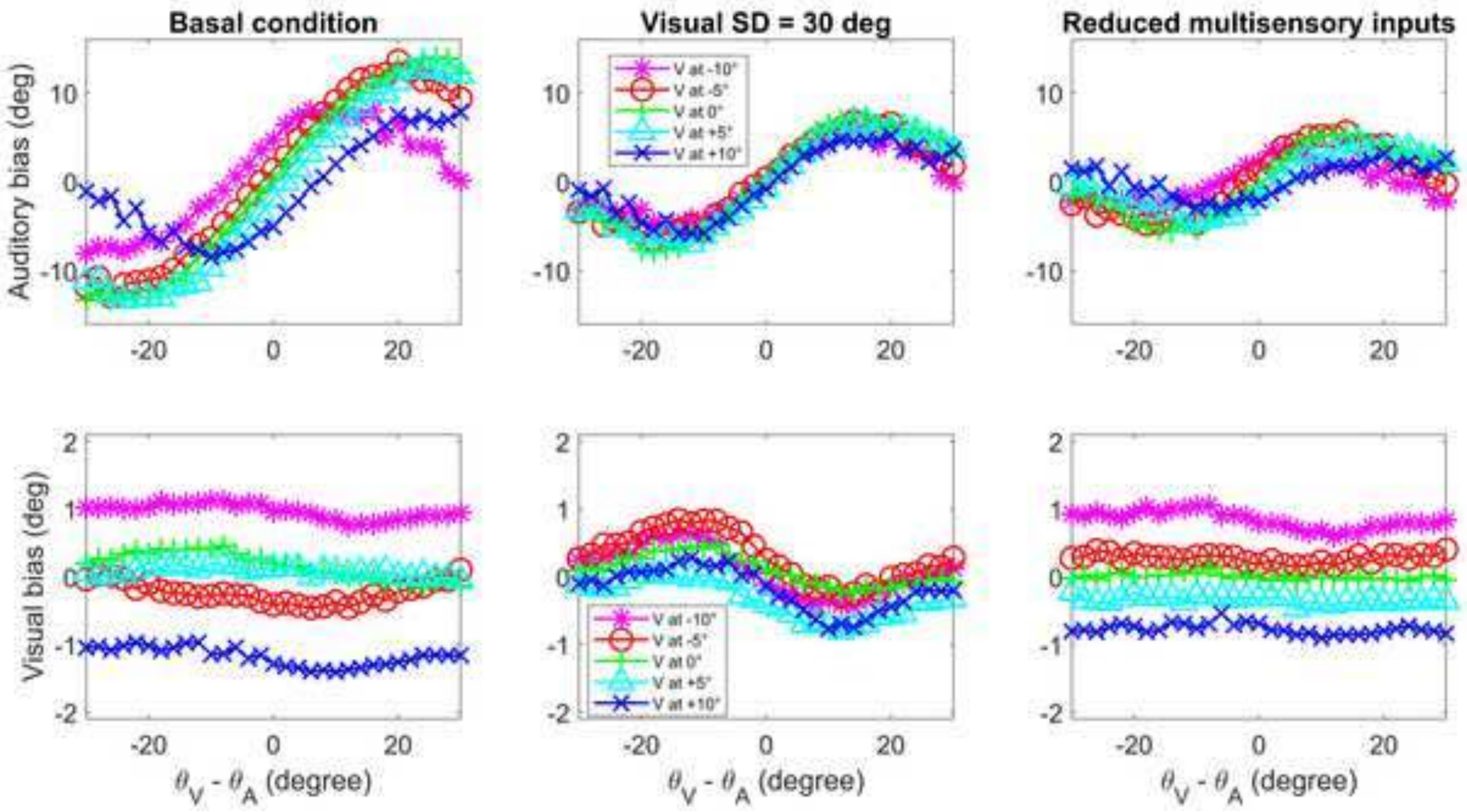

Basal condition
Visual SD = 30 deg
Reduced multisensory inputs
Auditory bias (deg)
10
0
-10
-20
0
20
Visual bias (deg)
2
1
0
-1
-2
V at -10°
V at -5°
V at 0°
V at +5°
V at +10°
$\theta_V - \theta_A$ (degree)

Figure 10

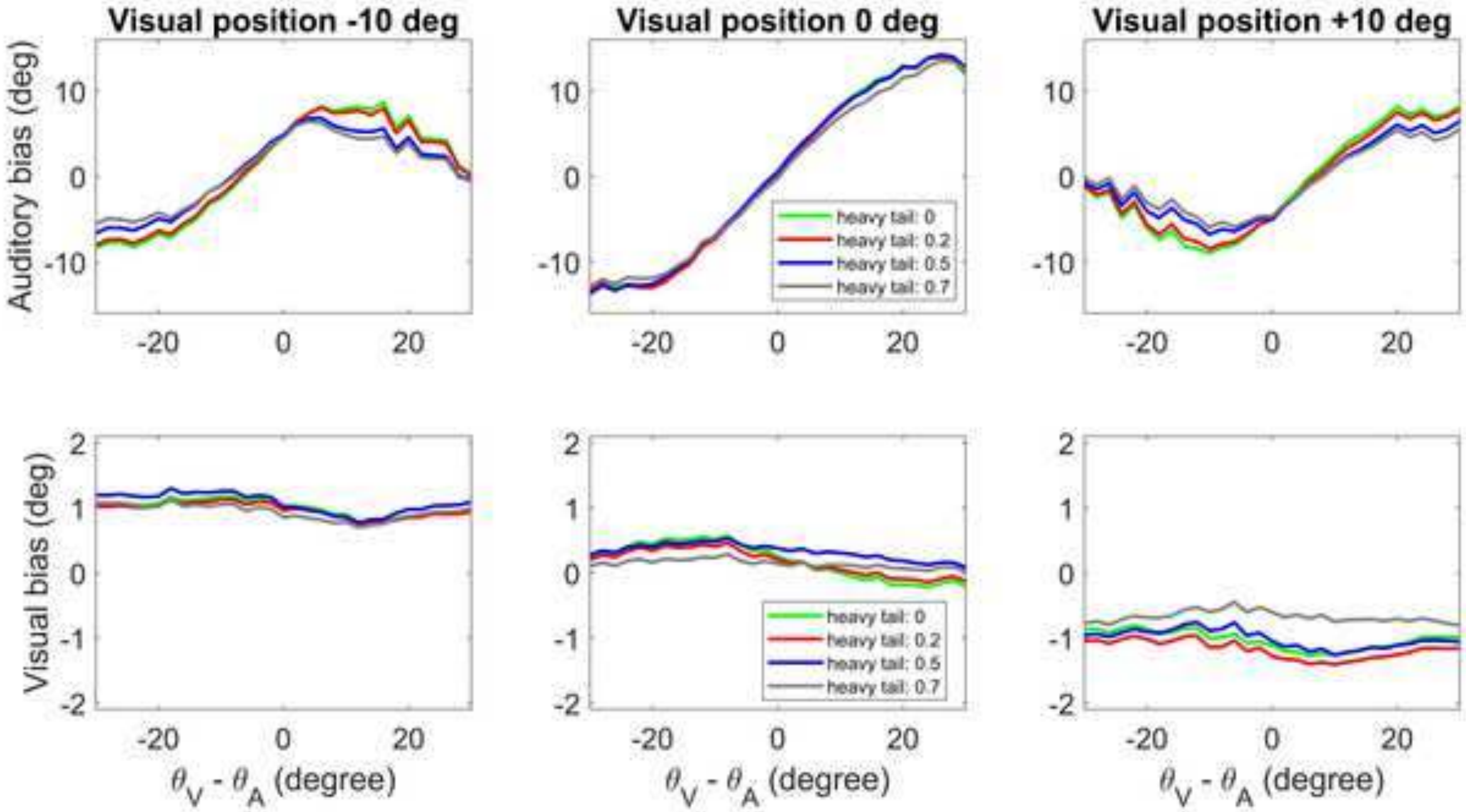

Visual position -10 deg
Visual position 0 deg
Visual position +10 deg
Auditory bias (deg)
Visual bias (deg)
10
0
-10
2
1
0
-1
-2
-20
0
20
heavy tail: 0
heavy tail: 0.2
heavy tail: 0.5
heavy tail: 0.7
$\theta_V - \theta_A$ (degree)

Figure 11

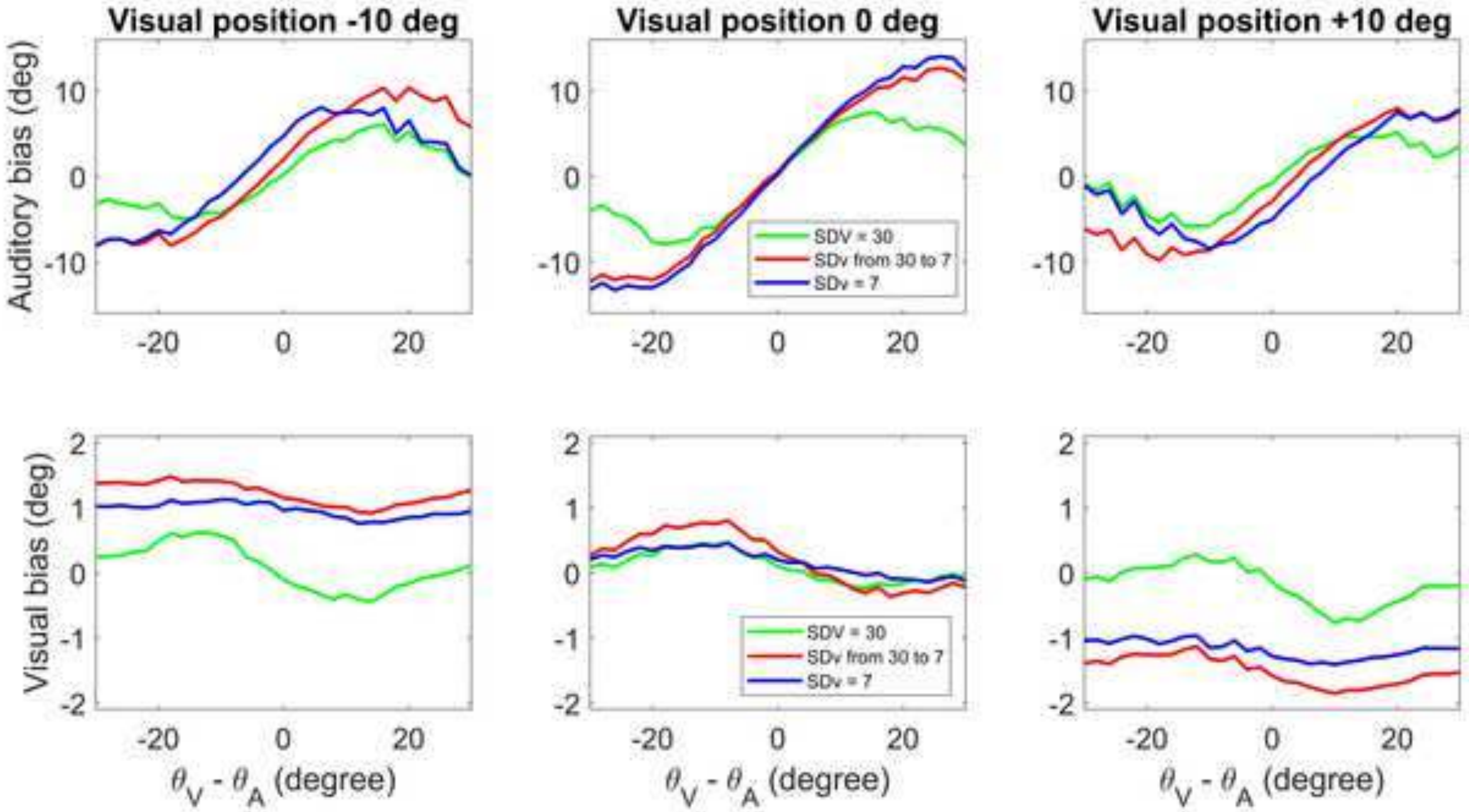

Visual position -10 deg
Visual position 0 deg
Visual position +10 deg
Auditory bias (deg)
Visual bias (deg)
10
0
-10
2
1
0
-1
-2
-20
0
20
θ_V - θ_A (degree)
SDV = 30
SDv from 30 to 7
SDv = 7
SDV = 30
SDv from 30 to 7
SDv = 7

Figure 12

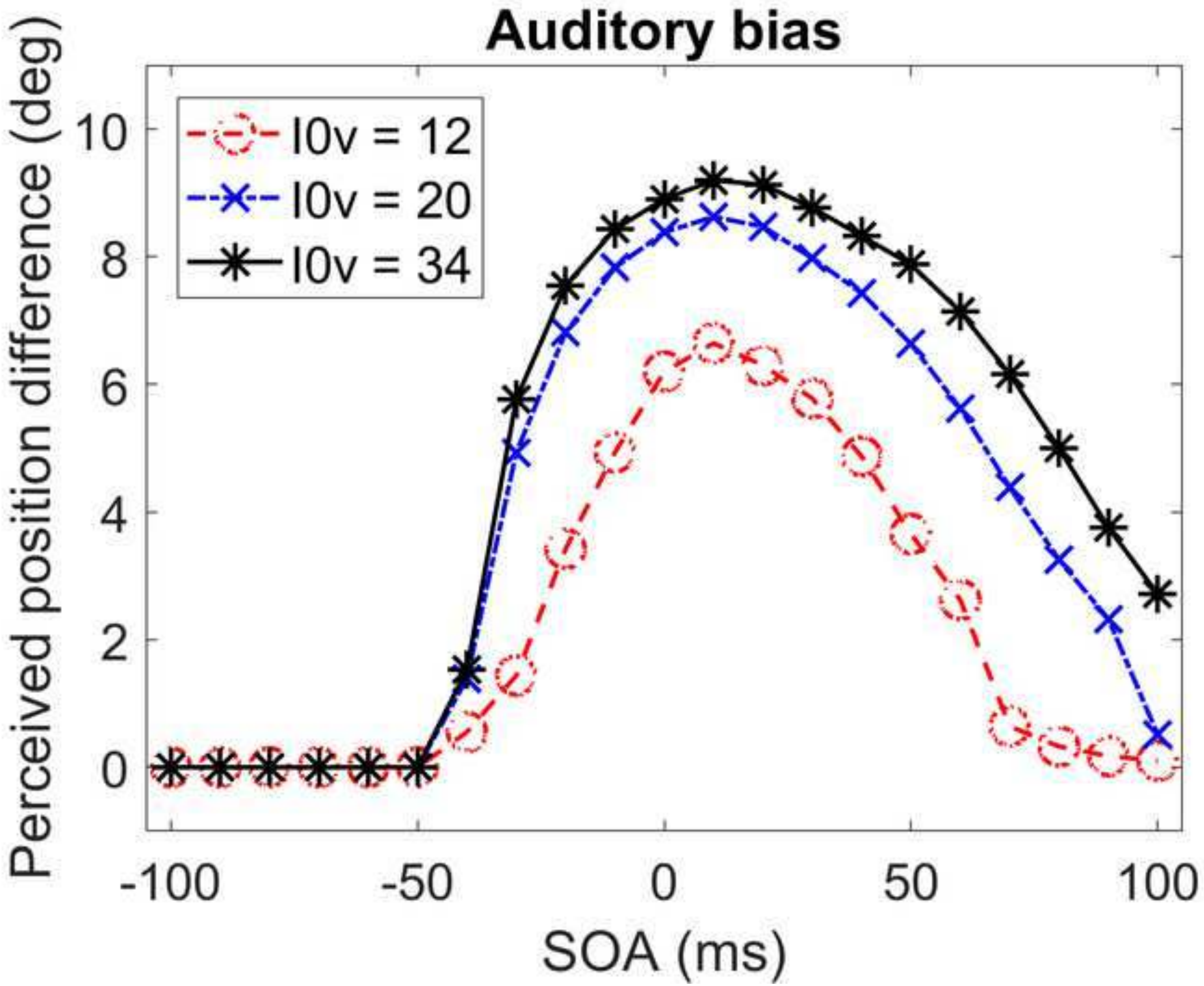
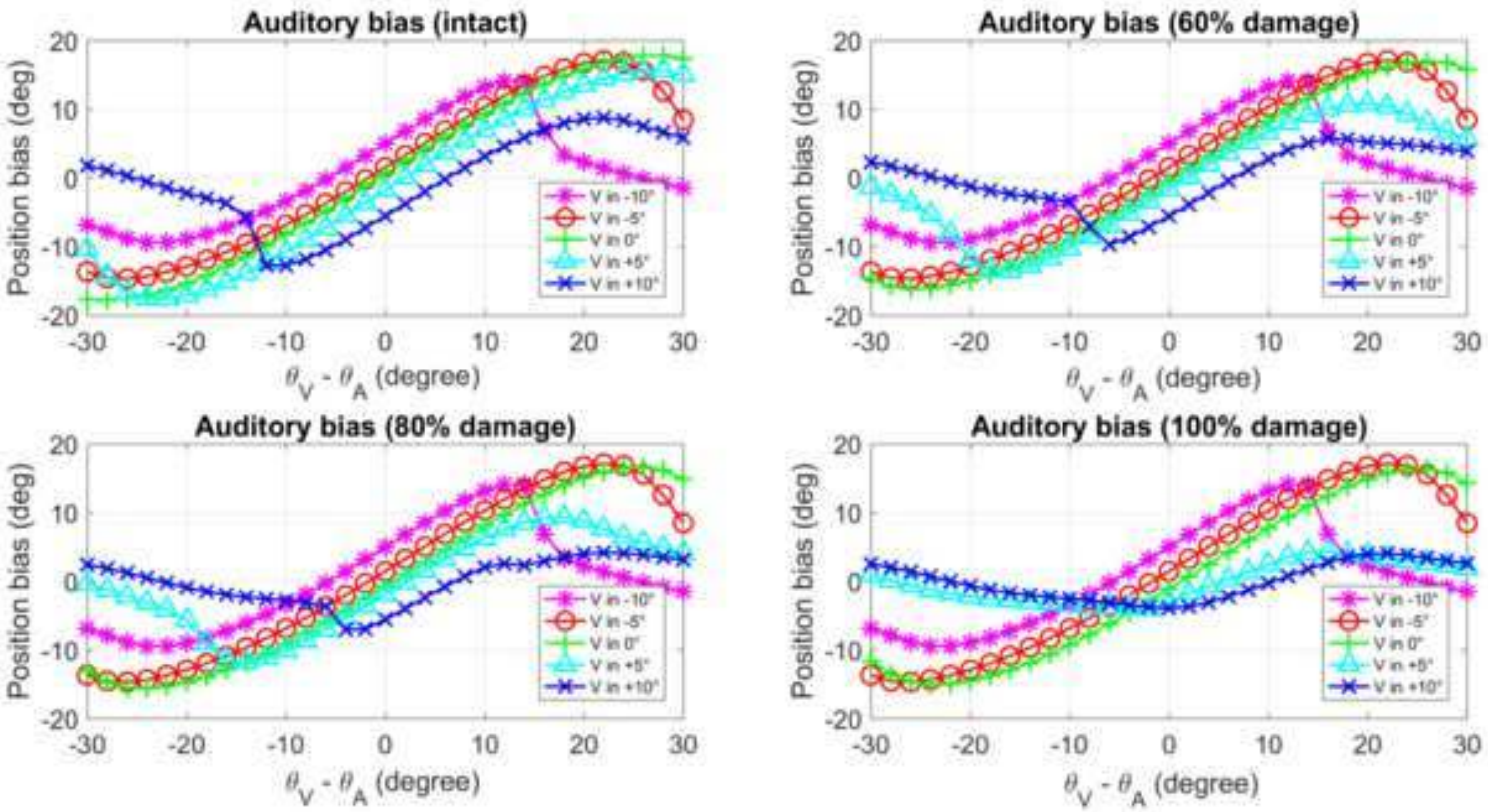

Auditory bias
I0v = 12
I0v = 20
I0v = 34
Perceived position difference (deg)
10
8
6
4
2
0
-100
-50
0
50
100
SOA (ms)

Figure 13


Auditory bias (intact)
Auditory bias (60% damage)
Auditory bias (80% damage)
Auditory bias (100% damage)
Position bias (deg)
20
10
0
-10
-20
-30
-20
-10
0
10
20
30
$\theta_V - \theta_A$ (degree)
V in -10°
V in -5°
V in 0°
V in +5°
V in +10°

Supplementary Material part I

Supplementary Material part II