Multisensory Bayesian inference depends
on synapse maturation during training:
theoretical analysis and neural modelling implementation

Mauro Ursino, Cristiano Cuppini, Elisa Magosso

Department of Electrical, Electronic and Information Engineering

University of Bologna, Bologna, Italy

Short title: modelling multisensory Bayesian inference

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

ABSTRACT

Recent theoretical and experimental studies suggest that, in multisensory conditions, the brain performs a near-optimal Bayesian estimate of the external events, laying more weight to the more reliable stimuli. However, the neural mechanisms responsible for this behavior, and its progressive maturation in a multisensory environment, are still insufficiently understood.

Aim of this work is to analyze this problem with a neural network model of audio-visual integration, based on the "probabilistic population coding", i.e the idea that a population of neurons can encode probability functions to perform Bayesian inference.

The model consists of two chains of unisensory neurons (auditory and visual) topologically organized: they receive the corresponding input through a plastic receptive field, and reciprocally exchange plastic cross-modal synapses, which encode the spatial co-occurrence of visual-auditory inputs. A third chain of multisensory neurons performs a simple sum of auditory and visual excitations.

The work includes a theoretical part and a computer simulation study. In the first we show how a simple rule for synapse learning (consisting of Hebbian reinforcement and a decay term) can be used during training to shrink the receptive fields and encode the unisensory likelihood functions. Hence, after training each unisensory area realizes a maximum likelihood estimate of stimulus position (auditory or visual). In cross-modal conditions, the same learning rule can encode information on prior probability into the cross-modal synapses.

Computer simulations confirm the theoretical results and show that the proposed network can actually realize a maximum likelihood estimate of auditory (or visual) positions in unimodal conditions, and a Bayesian estimate, with moderate deviations from optimality, in cross-modal conditions. Furthermore, the model explains the ventriloquism illusion and, looking at the activity in the multimodal neurons, explains the automatic reweighting of auditory and visual inputs, on a trial-by-trial basis, according to the reliability of the individual cues.

1. INTRODUCTION

The problem of how the brain integrates inputs from different sensory modalities to achieve an optimal recognition of external events and produce accurate object reproduction is fundamental in cognitive neuroscience. Numerous recent behavioral works suggest that the brain combines cross-modal cues in a near-optimal statistical way, by weighting cues of different modalities according to their reliability (as assumed in the classical Bayesian inference). This occurs when integration concerns visual and acoustic stimuli (Alais & Burr, 2004; Battaglia, Jacobs, & Aslin, 2003; Shams, Ma, & Beierholm, 2005; Wallace et al., 2004), texture and motion cues (Hillis, Watt, Landy, & Banks, 2004; Jacobs, 1999), visual and tactile inputs (Ernst & Banks, 2002) as well as trimodal cues (Wozny, Beierholm, & Shams, 2008). A typical effect of this near-optimal combination is the occurrence of illusory phenomena in the presence of a conflict between two stimuli of different sensory modality: in these cases, the observer favors the more reliable cue, producing a final estimate, which is more affected by the more reliable stimulus. By limiting our attention to visual-acoustic integration, common illusions include the ventriloquism effect in the spatial domain (Bertelson & Radeau, 1981; Hairston et al., 2003; Wallace et al., 2004), the fission and the fusion effects in the temporal domain (Andersen, Tiippana, & Sams, 2004; Shams, Kamitani, & Shimojo, 2000; Shams et al., 2005), and the McGurk effect concerning integration of phonemes with lips movements (McGurk & MacDonald, 1976).

A fundamental question, raising increasing attention recently, is how neural circuits can realize this near optimal probabilistic inference. Some recent studies shed some lights on this issue (Ma, Beck, Latham, & Pouget, 2006; Ma & Rahamati, 2013; Pouget, Beck, Ma, & Latham, 2013; Pouget, Dayan, & Zemel, 2003). A basic idea, named "probabilistic population coding", is that the activity of a population of neurons can automatically encode a probability distribution. Then, some form of metrics can extract the statistically optimal estimate from this distribution. In particular, Ma, Beck, Latham, and Pouget (Ma et al., 2006), and Ma and Rahmati (Ma & Rahamati, 2013) have shown that a population of neurons having Poisson-like variability, can codify the likelihood function of the

attribute using the probabilistic population code. Moreover, they investigated how simple neural circuits can perform Bayesian inference. In particular, Ma, Pouget et al., demonstrated that Bayesian inference can be realized by simply using a linear combination of population activity (Ma et al., 2006; Ma & Rahamati, 2013; Pouget et al., 2013). Subsequent experimental results in a multisensory task, involving visual and vestibular cues, confirmed this prediction in monkeys, showing that multimodal neurons in the medial superior temporal area combine visual and vestibular inputs linearly with sub-additive weights (Morgan, Deangelis, & Angelaki, 2008), and place more weights on the more reliable cue (Fetsch, Pouget, DeAngelis, & Angelaki, 2012). However, the authors also observed that these weights should be readjusted rapidly, on a trial-by-trial basis, to follow the actual reliability of the stimuli (Fetsch et al., 2012). A form of divisive normalization, acting on the multisensory representation (Ohshiro, Angelaki, & DeAngelis, 2011) has been advocated to explain this weight readjustment with reliability.

However, despite the previous important contributions, the problem of how a network of neurons can encode probabilities, and exploit population activity to realize a near-optimal estimate in a rapidly varying environment, is still far from being completely understood. In particular, several problems deserve attention: what kind of topological network (feedforward, feedback, or a combination of feedback and feedforward), trained by experience, is more suitable to realize a near-optimal cue combination? Which mechanisms for synapse learning can be used during training, to incorporate prior information and to mimic the developmental period in which a network encodes the statistics from the environment? How can likelihood probabilities and prior probabilities be combined in a near-optimal fashion, to compute posterior distributions, as required by the Bayesian estimator? Finally, how can a network readjust weights for cue integration, on a trial-by-trial basis, using biologically inspired neural mechanisms?

In order to address the previous points, two experimental observations deserve attention.

First, it is well known that the receptive fields (RFs) of sensory neurons are not fixed, but develop during a maturation phase according to the spatial reliability of external inputs. This has been clearly

observed in some sub-cortical areas (such as the cat's superior colliculus (Wallace & Stein, 1997)), but it is also evident in cortical auditory and visual neurons (Froemke & Jones, 2011; Pecka, Han, Sader, & Mrsic-Flogel, 2014). Such plasticity of the RFs may represent a suitable way to encode the likelihood probability of the inputs, based on experience.

Second, numerous recent experimental works emphasize the idea that cortical areas, traditionally deemed as purely unisensory (such as the visual, auditory or somatosensory cortex) exhibit some kinds of cross-talk, and exchange reciprocal information (Driver & Noesselt, 2008; Ghazanfar & Schroeder, 2006). For instance, many results show that the auditory neurons exhibit some multisensory behavior, and are affected by concomitant visual inputs and somatosensory influences (Musacchia & Schroeder, 2009; Schroeder & Foxe, 2005). Similarly, studies on the visual cortex report that activity of many visual neurons can be affected by auditory stimuli (Morrell, 1972) and that auditory inputs can activate part of the inferotemporal cortex (Ghazanfar & Schroeder, 2006). A few studies have also analysed the effect of audio-visual stimuli on the somatosensory cortex (Zhou & Fuster, 2000, 2004).

A challenging hypothesis, that we wish to analyze in the present work, is that these cross-modal synapses may mature under the pressure of the external inputs (for instance, to reflect the co-occurrence of cross-modal stimuli) and constitute an appropriate way to encode prior information on the environment statistics into network topology.

In recent years, we developed neural network models of audio-visual integration, which incorporate cross-modal synapses between unisensory areas (Cuppini, Magosso, Bolognini, Vallar, & Ursino, 2014; Magosso, Cona, & Ursino, 2013; Magosso, Cuppini, & Ursino, 2012; Ursino, Cuppini, & Magosso, 2014). Typical assumptions of these models were that neurons of the two modality have different receptive fields (the visual ones more accurate in the spatial domain; the auditory ones more accurate in the temporal domain), and that the cross-modal synapses mainly link neurons with proximal spatial preference. With these models, we were able to simulate some typical audio-visual illusions (such as the ventriloquism effect (Magosso et al., 2013; Magosso et al., 2012),

and the fusion and fission effects (Cuppini et al., 2014)). In these previous works, however, both cross-modal synapses and receptive fields were assigned a priori, i.e., neither of them were trained on the basis of the statistical input experience.

Aim of the present work is to introduce training mechanisms in a previous model, to test whether the information necessary to realize a near-optimal Bayesian estimator (i.e., the likelihood probability of the individual stimuli, and the prior probability) can be encoded in the synapses using biologically realistic learning rules. The trained model is subsequently exploited (using some population code metrics) to realize a near-optimal Bayesian estimate.

The network includes two layers of unisensory neurons, which receive the external inputs (either auditory or visual) via a plastic receptive field, and can connect reciprocally via initially latent cross-modal synapses. Moreover, competitive mechanisms are implemented within each layer. In the immature phase the receptive fields are wide (equal for all neurons in both layers) and cross-modal synapses are null. Both kind of synapses are subsequently shaped using a Hebbian mechanism of potentiation with a decay term.

The paper includes both a preliminary theoretical analysis of the network (based on some ideas from Ma et al. (Ma et al., 2006) and also from Patton and Anastasio (Patton & Anastasio, 2003) and Sato et al. (Sato, Toyoizumi, & Aihara, 2007)), to demonstrate which kind of probabilities can be incorporated during training, and numerical computer simulations. In a first stage, the network is used to infer the positions of the auditory and of the visual stimuli separately. Results are compared with those obtainable with the maximum likelihood estimate, in case of unimodal stimuli, and with the Maximum Posterior Probability estimate (a kind of Bayesian estimator which exploits known prior probability) in case of cross-modal stimuli. Simulations of the ventriloquism effect are also performed, to compare model performance with behavioral results. In a second stage, the network is enriched with a further multisensory layer, which receives input from the previous two layers (auditory and visual) to compute a single position for both stimuli. This is used to assess the single causal inference problem, i.e. a problem where the observers are asked to envisage each presentation

as a single event (Alais & Burr, 2004). In this condition, we analyze whether the proposed model can automatically re-weight the cues, according to their reliability.

Results demonstrate that the proposed network, with the implemented training rule, can realize a near-optimal Bayesian estimate (exploiting both likelihood and prior information) and can automatically combine cues, taking into account their reliability on a trial-by-trial basis.

2. MATERIAL AND METHODS

2.1 THEORY

2.1.1 Bayesian estimate: the general problem

Let us assume that the brain processes two sensory inputs of different modalitites (for instance an acoustic and a visual one). In the following, a quantity belonging to a given sensory modality will be represented with the superscripts A or V, respectively. Moreover, upper case letters will be used to represent vectors or arrays, while lower case letters (with a subscript) will be used to represent scalar component of vectors.

Each sensory input consists of a vector with N component (named I^A and I^V, respectively) which describes the spatial distribution of the stimulus. For instance, the scalar quantity i_j^A denotes the jth component of the acoustic input. We assume that each component j (j = 1, 2, … N) codes for a particular spatial position, ϑ_j. Hence, a vector (Nx1) of spatial positions is defined:

$$\Theta = \begin{bmatrix} \vartheta_1 & \vartheta_2 & \cdots & \vartheta_j & \cdots & \vartheta_N \end{bmatrix}^T$$

The two vectors I^A and I^V are the auditory and visual sensory representations of the external world reaching the brain. Both are affected by the spatial position of the input stimulus, by the blurring properties of the sensory transduction, and by noise.

Let us denote with ϑ^A and ϑ^V the positions of the acoustic and the visual stimuli, respectively, which generate the sensory representations. In terms of Bayesian estimates, the problem consists in inferring the positions ϑ^A and ϑ^V starting from knowledge of the two sensory representations, I^A and

I^V, blurred and affected by noise. The problem is completely defined, from a statistical point of view, if one knows the likelihood probabilities of the two sensory representations, and the prior probability of the positions.

In the following, we will assume that the sensory representations, I^A, is only a function of the position of the acoustic stimulus, ϑ^A, while I^V is only a function of ϑ^V. Moreover, they also depend on the strength of the stimulus and on the presence of noise, both assumed independent for the two stimuli.

Hence, we can write the following expression for the likelihood probability of the overall sensory inputs:

$$p\left(I^A, I^V \middle| \vartheta^A, \vartheta^V\right) = p\left(I^A \middle| \vartheta^A\right) p\left(I^V \middle| \vartheta^V\right) \tag{1}$$

Let us assume that the prior probability of the two positions $p\left(\vartheta^A, \vartheta^V\right)$ is also known. It is worth noting that the two sensory representations are not independent [i.e., we generally have $p\left(I^A, I^V\right) \neq p\left(I^A\right)p\left(I^V\right)$] since the two positions ϑ^A and ϑ^V are not independent [generally $p\left(\vartheta^A, \vartheta^V\right)$ is not uniform, and $p\left(\vartheta^A, \vartheta^V\right) \neq p\left(\vartheta^A\right)p\left(\vartheta^V\right)$].

According to the Bayesian rule, and using Eq. (1), we can write the following expression for the a posteriori probability

$$p\left(\vartheta^A, \vartheta^V \middle| I^A, I^V\right) = \frac{p\left(\vartheta^A, \vartheta^V\right) p\left(I^A, I^V \middle| \vartheta^A, \vartheta^V\right)}{p\left(I^A, I^V\right)} = \frac{p\left(\vartheta^A, \vartheta^V\right) p\left(I^A \middle| \vartheta^A\right) p\left(I^V \middle| \vartheta^V\right)}{p\left(I^A, I^V\right)} \tag{2}$$

In order to have the better estimate, we have to maximize the numerator of Eq. (2), starting from knowledge of the sensory representations I^A and I^V. In other words, the estimates (say $\hat{\vartheta}_A$ and $\hat{\vartheta}_V$ respectively), must satisfy the following rule

$$\left[\hat{\vartheta}_A, \hat{\vartheta}_V\right] = \arg\max\left\{p\left(\vartheta^A, \vartheta^V \middle| I^A, I^V\right)\right\} = \arg\max\left\{p\left(\vartheta^A, \vartheta^V\right) p\left(I^A \middle| \vartheta^A\right) p\left(I^V \middle| \vartheta^V\right)\right\} \tag{3}$$

In order to maximize Eq. (2), we need the knowledge of the likelihood probabilities and of the prior probability.

2.1.2 The likelihood probabilities

Let us now assume that we know a given realization of the sensory input (in the following we will consider a generic sensory input S, with either $S = A$ or $S = V$). Hence, we know:

$$I^S = [i_1^S \ i_2^S \cdots i_j^S \cdots i_N^S]^T,$$

which represents a Nx1 array. This is the input stimulus reaching the brain. Moreover, let us assume that the prior distribution is uniform (or simply we do not know the prior). In this case, Eq. (3) simplifies,

$$\hat{\vartheta}_S = \arg\max\{p(I^S|\vartheta^S)\} \tag{4}$$

that is, one needs to maximize the likelihood probability of the sensory representation to obtain the sensory estimate $\hat{\vartheta}_S$.

In the following, we will assume that the sensory input consists of a deterministic term (say, $M^S(\vartheta^S)$) on which some Gaussian white noise with zero mean value is superimposed (say N^S). Hence, we can write the following expression for the random vector I^S

$$\underline{I}^S = M^S + \underline{N}^S \tag{5}$$

or in scalar form

$$\underline{i}_j^S = m_j^S + \underline{n}_j^S \qquad\qquad j = 1,2,...,N \tag{5'}$$

where the underscore signifies that the corresponding quantity is random. If the noise terms are independently generated, then $\underline{i}_j^S$ are also independent variables, and so the following expression holds for the likelihood probability

$$p(I^S|\vartheta^S) = \prod_{j=1}^{N} p(i_j^S|\vartheta^S) \tag{6}$$

The deterministic term in Eq. (5) is a function of the strength of the stimulus (indeed, the higher the strength, the higher m_j^S), and of the stimulus position. In particular, this term is maximal when

$\vartheta_j = \vartheta^S$, and progressively decreases with distance. We used a Gaussian function to represent the deterministic spatial properties of the input. We can write

$$m_j^S(\vartheta^S) = i_{Max}^S \exp\left\{-d(\vartheta^S, \vartheta_j)^2 / (2\sigma^{S2})\right\} \qquad j = 1, 2, ..., N \qquad (7)$$

where $d(\vartheta^S, \vartheta_j)$ represents the distance between the position of the stimulus ϑ^S and the actual position ϑ_j, σ^S is the standard deviation of the Gaussian function, here representing the spatial accuracy of the input (i.e., the greater σ^S, the more blurred the stimulus) and i_{Max}^S accounts for the strength of the external stimulus. According to Eq. (7), when $d(\vartheta^S, \vartheta_j) = 0$ the sensory input is maximal (apart from the effect of noise) while the sensory input progressively decreases with distance.

In the present work, the distance has been computed through a circular structure, so that every sensory input receives similar excitation, independently of being close to the center or to the border. Hence, the following equation has been used to compute the distance:

$$d(\vartheta^S, \vartheta_j) = \begin{cases} |\vartheta_j - \vartheta^S| & \text{if } |\vartheta_j - \vartheta^S| \le D/2 \\ D - |\vartheta_j - \vartheta^S| & \text{if } |\vartheta_j - \vartheta^S| > D/2 \end{cases} \qquad (8)$$

where D represents the overall spatial distance (i.e., $0 \le \vartheta_j \le D$). By way of example, assuming $D = 180°$, the position $\vartheta^S = 1°$ is equally distant from position $\vartheta_j = 180°$ and from position 2°; moreover, it is equally distant from the position 179° and from position 3°, etc.

Since we assumed that noise has a Gaussian distribution with zero mean value, the likelihood probability $p(i_j^S | \vartheta^S)$ assumes the following expression

$$p(i_j^S | \vartheta^S) = \frac{1}{\sqrt{2\pi\upsilon^{S2}}} \exp\left\{-[i_j^S - m_j^S(\vartheta^S)]^2 / (2\upsilon^{S2})\right\} =$$

$$= \frac{1}{\sqrt{2\pi\upsilon^{S2}}} \exp\left\{-\frac{\left[i_j^S - i_{Max}^S \exp\left(-(d(\vartheta^S, \vartheta_j))^2 / (2\sigma^{S2})\right)\right]^2}{2\upsilon^{S2}}\right\} \qquad j = 1, 2, ..., N \qquad (9)$$

where υ^S represents the standard deviation of the noise (hence, the greater υ^S the greater the effect of noise).

Finally, using Eqs. (6) and (9) together, we can write the overall expression of the *likelihood probability* for the sensory input as a function of the stimulus position:

$$p\left(I^S\middle|\vartheta^S\right)=\prod_{j=1}^{N} p\left(i_j^S\middle|\vartheta^S\right)=\prod_{j=1}^{N}\frac{1}{\sqrt{2\pi\upsilon^{S2}}}\exp\left\{-\frac{\left[i_j^S-i_{Max}^S\exp\left(-\left(d\left(\vartheta^S,\vartheta_j\right)\right)^2\middle/\left(2\sigma^{S2}\right)\right)\right]^2}{2\upsilon^{S2}}\right\}\tag{10}$$

It is worth noting that Eqs. (7) and (10) are not only a function of the stimulus position, ϑ^S, but also of the input strength (i.e., parameter i_{Max}^S). However, for simplicity the latter dependence has not been explicitly shown in the left hand member.

The *likelihood function* is given by the expression (10), in which a specific value of I^S (a single realization of the random vector) is used. Hence:

$$l\left(\vartheta^S\right)=p\left(I^S\middle|\vartheta^S\right)\quad\text{with}\quad I^S\text{ known.}\tag{11}$$

Let us compute the natural logarithm of the likelihood function. From Eq. (11) we have

$$\ln\left(l\left(\vartheta^S\right)\right)=-\sum_{j=1}^{N}\ln\left(\sqrt{2\pi\upsilon^{S2}}\right)-\frac{1}{2\upsilon^{S2}}\sum_{j=1}^{N}\left\{i_j^S-i_{Max}^S\exp\left[-\frac{\left(d\left(\vartheta^S,\vartheta_j\right)\right)^2}{2\sigma^{S2}}\right]\right\}^2\tag{12}$$

The Maximum Likelihood Estimate (MLE) is obtained by computing the value (say $\hat{\vartheta}^S$) that maximizes Eq. (12). Similar equations hold for the visual ($S = V$) and the auditory ($S = A$) likelihood functions.

2.1.3 A Neural Network model for maximum likelihood estimate

Our problem is now to find a simple neural network, consisting of N neurons, which can be trained with a realistic rule to automatically find the maximum of Eq. (12), from knowledge of the input vector I^S.

In the following, each neuron will be represented through the subscript k. Let us consider that each neuron in the chain is more sensitive to a specific position, ϑ_k ($k = 1, 2, …, N$), i.e., we use the same

positions previously used for the sensory vector; this can be obtained using a receptive field, for each neuron, centered at the preferred position. We will denote each receptive field as R_k^S ($k = 1, 2, ..., N$); this is a vector with dimension $N \times 1$. The input to the k-th neuron with a preferred position ϑ_k in the sensory modality S (say u_k^S) is then computed as the scalar product between the sensory input and its receptive field. We can write

$$u_k^S = \sum_{j=1}^{N} r_{kj}^S \, i_j^S \tag{13}$$

where r_{kj}^S is the j-th component of vector R_k^S.

Each neuron then computes its output activity (say y_k^S) by passing the input through a monotonically increasing non linear function (which mimics the presence of a lower threshold and upper saturation for neurons activity). By denoting this monotonic function with $\varphi(u)$, we can write

$$y_k^S = \varphi(u_k^S) = \varphi\left(\sum_{j=1}^{N} r_{kj}^S \, i_j^S \right) \tag{14}$$

During numerical simulations, we will use a sigmoidal function, as habitually done in neural network modeling. However, for the present considerations, we just need that $\varphi(u)$ is monotonically increasing.

We can now demonstrate that, in order to compute the maximum likelihood estimate, we need that

 i) all neurons have an identical receptive field, which differs just for the preferred position;

ii) after training, the receptive field reproduces the spatial distribution of the sensory input, i.e.

$$r_{kj}^S = r_{max}^S \exp\left[-\frac{\left(d(\vartheta_k, \vartheta_j) \right)^2}{2\sigma^{S2}} \right] \qquad j = 1,2,...,N \tag{15}$$

In the next paragraph we will analyze how Eq. (15) can be approximately realized using a physiological training rule (i.e., a Hebbian rule with a decay term).

If we make use of Eq. (15) within Eq. (12), we obtain the following general expression for the logarithmic likelihood function, at the particular position ϑ_k ($k = 1, 2, \ldots, N$):

$$\ln(l(\vartheta_k)) = -\sum_{j=1}^{N} \ln\left(\sqrt{2\pi\upsilon^{S2}}\right) - \frac{1}{2\upsilon^{S2}} \sum_{j=1}^{N} \left\{ i_j^S - i_{Max}^S \exp\left[-\frac{\left(d\left(\vartheta_k, \vartheta_j\right)\right)^2}{2\sigma^{S2}}\right] \right\}^2 =$$

$$= -\sum_{j=1}^{N} \ln\left(\sqrt{2\pi\upsilon^{S2}}\right) - \frac{1}{2\upsilon^{S2}} \sum_{j=1}^{N} \left[i_j^S - \frac{i_{Max}^S}{r_{max}^S} r_{kj}^S \right]^2 = \qquad (16)$$

$$= -\sum_{j=1}^{N} \ln\left(\sqrt{2\pi\upsilon^{S2}}\right) - \frac{1}{2\upsilon^{S2}} \sum_{j=1}^{N} i_j^{S2} - \frac{1}{2\upsilon^{S2}} \left(\frac{i_{Max}^S}{r_{max}^S}\right)^2 \sum_{j=1}^{N} r_{kj}^{S2} + \frac{1}{\upsilon^{S2}} \frac{i_{Max}^S}{r_{max}^S} \sum_{j=1}^{N} i_j^S r_{kj}^S$$

Of course, the first two terms in the right hand member of Eq. (16) are independent of k. However, in virtue of the assumption i), we can claim that the quantity $\sum_{j=1}^{N} r_{kj}^{S2}$ is also independent of k (in particular, we use a circular distance in writing the expression of the receptive fields, to avoid any border effect). Therefore, we can write

$$\ln(l(\vartheta_k)) = \alpha + \frac{1}{\upsilon^{S2}} \frac{i_{Max}^S}{r_{max}^S} \sum_{j=1}^{N} r_{kj}^S i_j^S \qquad (17)$$

where α represents the sum of the first three terms in the right hand member of (16), which does not depend on the particular value of ϑ_k. Therefore, in order to maximize the logarithmic likelihood function, one simply needs to maximize the quantity $u_k^S = \sum_{j=1}^{N} r_{kj}^S i_j^S$ (see Eq. (13)). Finally, by remembering that $\varphi(u)$ is an increasing monotonic function, we have

$$\hat{\vartheta}^S = \arg\max\{\ln(l(\vartheta_k))\} = \arg\max\{u_k^S\} = \arg\max\{y_k^S\} \qquad (18)$$

Eq. (18) says that the neuron with maximal activity signals the stimulus position, according to a maximum likelihood estimate (of course, with a resolution provided by the distance between two consecutive ϑ_k). For this reason, in the following we will assume that *the stimulus position is coded by the neuron with maximal activity.* Alternatively, in order to improve the resolution, we will also

use a barycenter metrics (which, due to the network circular symmetry, provides a value quite close to the maximum).

In order for Eq. (18) to be verified, the assumptions i) and ii), concerning the receptive fields, must be true. These require a few comments. Assumption i) implies that the preferred position of the neurons has a uniform distribution (i.e., positions are equally represented in the network) and that, during training, all positions are stimulated from the external input in the same way (i.e., no position receives a stronger input than another or a more frequent input). Assumption ii) implies that the receptive field, after training, reproduces the average spatial distribution of the input at the given preferred position (i.e., $r_{kj}^S \propto m_j^S(\vartheta_k)$). The latter requirement can be attained using the following training rule, which in vectorial form can be written as:

$$\Delta R_k^S = \gamma \, y_k^S \left(I^S - R_k^S \right) \tag{19}$$

where ΔR_k^S represents the *variation* in the receptive field entering a neuron in the sensory modality S with preferred position ϑ_k. Or, in scalar form

$$\Delta r_{kj}^S = \gamma \, y_k^S \left(i_j^S - r_{kj}^S \right) \tag{19'}$$

which is substantially an Hebb rule with a decay term (see Neural Network textbooks for more details on this rule, for instance (Hertz, Krogh, & Palmer, 1991).

According to Eq. (19), a neuron with a high output activity modifies its receptive field, shaping it based on the actual input, and pruning the distal portion of its previous configuration. Conversely, silent neurons with poor output activity do not appreciably modify their receptive field. As demonstrated in neural network textbooks (Hertz et al., 1991), after a long training, the receptive field vector R_k^S will be positioned close to the *mean value* of the inputs which significantly activated that neuron. In our case, this value is proportional to $M^S(\vartheta_k)$: in scalar form, this is Eq. (7) with $\vartheta^S = \vartheta_k$. Hence, we have

$$R_k^S \propto M^S(\vartheta_k) \, . \tag{20}$$

which is the vector form of Eq. (15)

It is worth noting that, to improve the effect of the adopted training rule, so that Eq. (20) is actually satisfied, one can benefit from a winner takes all dynamics for the neurons, so that in response to an external input just a few neurons win the competition, while most neurons are silent. In this way, a neuron modifies its synapses only when the input is close to its preferred spatial position. For this reason, as usually done when working with this kind of networks, we introduced lateral synapses implementing a competition between neurons (see the section "Description of the Neural network").

2.1.4 A cross-modal network for the posterior probability (Bayesian) estimate

Let us now consider the case in which the two sensory inputs (I^A and I^V) are not independent, as in the previous case, but are linked through the prior probability $p(\vartheta^A, \vartheta^V)$. In this case, one has to maximize the numerator of Eq. (2), that can be rewritten in logarithmic form. We have

$$
\begin{aligned}
\left[\hat{\vartheta}_A, \hat{\vartheta}_V\right] &= \arg\max\left\{p(\vartheta^A, \vartheta^V)p(I^A|\vartheta^A)p(I^V|\vartheta^V)\right\} = \\
&= \arg\max\left\{\ln(p(\vartheta^A, \vartheta^V)) + \ln(p(I^A|\vartheta^A)) + \ln(p(I^V|\vartheta^V))\right\}
\end{aligned}
\tag{21}
$$

Of course, in case of uniform prior probability, Eq.(21) is the same as the likelihood estimate, performed separately on ϑ^A and ϑ^V .

In the following, we will denote with $\psi(\vartheta^A, \vartheta^V)$ the function to be maximized, i.e.,

$$
\psi(\vartheta^A, \vartheta^V) = \ln(p(\vartheta^A, \vartheta^V)) + \ln(p(I^A|\vartheta^A)) + \ln(p(I^V|\vartheta^V)) = \ln(p(\vartheta^A, \vartheta^V)) + \ln(l(\vartheta^A)) + \ln(l(\vartheta^V)) \quad \text{(21')}
$$

with I^A and I^V known.

Let us now consider that each of the log likelihood function $\ln(l(\vartheta^S))$ in Eq. (21') has expression (12), but with different parameters in the two modalities (S=A, S=V); in particular, we assume that σ^A is different from σ^V, i.e. the two stimuli have different spatial accuracy and $i^A_{Max} \neq i^V_{Max}$ (i.e., the two stimuli may have a different strength). Hence, Eq. (21') can be written as:

16

$$\psi\left(\vartheta^A, \vartheta^V\right) = -\sum_{j=1}^{N} \ln\left(\sqrt{2\pi\upsilon^{A2}}\right) - \frac{1}{2\upsilon^{A2}} \sum_{j=1}^{N} \left\{ i_j^A - i_{Max}^A \exp\left[-\frac{\left(d\left(\vartheta^A, \vartheta_j\right)\right)^2}{2\sigma^{A2}}\right] \right\}^2 +$$

$$-\sum_{j=1}^{N} \ln\left(\sqrt{2\pi\upsilon^{V2}}\right) - \frac{1}{2\upsilon^{V2}} \sum_{j=1}^{N} \left\{ i_j^V - i_{Max}^V \exp\left[-\frac{\left(d\left(\vartheta^V, \vartheta_j\right)\right)^2}{2\sigma^{V2}}\right] \right\}^2 + \ln\left(p\left(\vartheta^A, \vartheta^V\right)\right)$$

(22)

where υ^V and υ^A represent the standard deviation of the noise in the visual and auditory sensory representation. The maximum posterior probability Bayesian Estimator (BE) is obtained by computing the values ($\hat{\vartheta}^A$, $\hat{\vartheta}^V$) that maximize Eq. (22), using a known prior probability.

Let us now consider a network of N neurons for each modality implementing the corresponding likelihood function according to the expression (16) used above. By using Eq. (16) within Eq. (22), we can compute the following expression for the function $\psi\left(\vartheta^A, \vartheta^V\right)$, evaluated at two different positions (a position ϑ_k^A for the k-th auditory neuron and position ϑ_h^V for the h-th visual neuron):

$$\psi\left(\vartheta_k^A, \vartheta_h^V\right) = -\sum_{j=1}^{N} \ln\left(\sqrt{2\pi\upsilon^{A2}}\right) - \frac{1}{2\upsilon^{A2}} \sum_{j=1}^{N} i_j^{A2} - \frac{1}{2\upsilon^{A2}} \left(\frac{i_{Max}^A}{r_{max}^A}\right)^2 \sum_{j=1}^{N} r_{kj}^{A2} + \frac{1}{\upsilon^{A2}} \frac{i_{Max}^A}{r_{max}^A} \sum_{j=1}^{N} i_j^A r_{kj}^A$$

$$-\sum_{j=1}^{N} \ln\left(\sqrt{2\pi\upsilon^{V2}}\right) - \frac{1}{2\upsilon^{V2}} \sum_{j=1}^{N} i_j^{V2} - \frac{1}{2\upsilon^{V2}} \left(\frac{i_{Max}^V}{r_{max}^V}\right)^2 \sum_{j=1}^{N} r_{hj}^{V2} + \frac{1}{\upsilon^{V2}} \frac{i_{Max}^V}{r_{max}^V} \sum_{j=1}^{N} i_j^V r_{hj}^V + \ln\left(p\left(\vartheta_k^A, \vartheta_h^V\right)\right)$$

(23)

Let us now consider that all neurons in a given modality have identical receptive fields. In this condition, the third and seventh terms in Eq. (23) are independent of the particular values of k and h. Hence, the maximization of Eq. (23) corresponds to the maximization of the following new function (by also neglecting all other terms that do not depend on k and h):

$$\widetilde{\psi}\left(\vartheta_k^A, \vartheta_h^V\right) = \frac{1}{\upsilon^{A2}} \frac{i_{Max}^A}{r_{max}^A} \sum_{j=1}^{N} i_j^A r_{kj}^A + \frac{1}{\upsilon^{V2}} \frac{i_{Max}^V}{r_{max}^V} \sum_{j=1}^{N} i_j^V r_{hj}^V + \ln\left(p\left(\vartheta_k^A, \vartheta_h^V\right)\right)$$

(24)

It is worth noting that the first two terms in the right hand member of Eq. (24) are proportional to the external inputs to the auditory and visual neurons respectively (u_k^S, Eq. (13)), and implement the likelihood functions in the two modalities. Hence, if we assume that the neuron output depends only on the term $u_k^S = \sum_{j=1}^{N} r_{kj}^S i_j^S$ with $S = A$ or V, we cannot account for the prior probability (i.e., the third

term in Eq. (24)). As it is clear from Eq. (24), estimation of optimal values of $\vartheta_k^A, \vartheta_h^V$ requires that the inputs to the auditory and visual neurons are modified, to account for the prior probability. In other words, both the auditory and visual neurons must receive a "cross-modal" term from the other modality, reflecting the prior knowledge.

A typical condition occurs when the two stimuli (the auditory and visual one) frequently originate from a same spatial event, hence the two positions are correlated. We can assume that the prior probability is given by the weighted sum of a uniform distribution, $p'(\vartheta^A, \vartheta^V)$, reflecting the possibility that a visual and an auditory stimulus are produced by different events, and a second term, $p''(\vartheta^A, \vartheta^V)$ reflecting the (higher) probability that the auditory and visual events are originated from the same source:

$$p(\vartheta^A, \vartheta^V) = \beta_1 p'(\vartheta^A, \vartheta^V) + \beta_2 p''(\vartheta^A, \vartheta^V) \tag{25}$$

We can write:

$$p'(\vartheta^A, \vartheta^V) = \frac{1}{D^2} \qquad \text{(uniform distribution)} \tag{26}$$

Moreover

$$p''(\vartheta^A, \vartheta^V) = p(\vartheta^A) p''(\vartheta^V | \vartheta^A) = \frac{1}{D} \frac{1}{\sqrt{2\pi\sigma^{AV2}}} \exp\left(-\frac{d(\vartheta^A, \vartheta^V)^2}{2\sigma^{AV2}}\right) \tag{27}$$

Eq. (27) has been written assuming that a single position (for instance the auditory one) has a uniform distribution (i.e., $p(\vartheta^A) = 1/D$); the probability of the second position, in case of a single source, dramatically decreases with the distance. The parameter σ^{AV} reflects the spatial accuracy for the superimposition of the two stimuli, when they originate from the same source. Of course, Eq. (27) integrated on the entire space of possible positions (i.e., between 0 and D), must satisfy the fundamental axiom of probability. To this end, the following constraint must be used: $\beta_1 + \beta_2 = 1$, hence $\beta_2 = 1 - \beta_1$.

We thus obtain

$$p\left(\vartheta^{A},\vartheta^{V}\right)=\beta_{1}\frac{1}{D^{2}}+\left(1-\beta_{1}\right)\frac{1}{D}\frac{1}{\sqrt{2\pi}\sigma^{AV2}}\exp\left(-\frac{d\left(\vartheta^{A},\vartheta^{V}\right)^{2}}{2\sigma^{AV2}}\right) \tag{28}$$

Parameter β_{1} represents the fraction of cross-modal stimuli coming from independent sources. Conversely, $1-\beta_{1}$ represents the fraction of cross modal stimuli coming from a single source.

It is worth noting that Eq. (28) can also be used to implement the causal inference problem: $\beta_{1}=1$ signifies that the subject considers the two sources as completely independent, while $\beta_{1}=0$ considers the two sources as coming from the same event.

Finally, let us consider the inclusion of Eq. (28) into Eq. (24) in the two extreme cases. If $\beta_{1}=1$, maximization of Eq. (24) simply returns the maximum likelihood estimate, separately performed on the auditory and the visual network (to separately maximize the first and second terms in the right hand member). Conversely, if $\beta_{1}\neq1$ the prior probability must be taken into account. Let us consider the typical case where $\beta_{1}\ll1$ and so the first term in Eq. (28) is negligible. Then, Eq. (24) furnishes:

$$\widetilde{\psi}\left(\vartheta_{k}^{A},\vartheta_{h}^{V}\right)=\frac{1}{\upsilon^{A2}}\frac{i_{Max}^{A}}{r_{\max}^{A}}\sum_{j=1}^{N}i_{j}^{A}r_{kj}^{A}+\frac{1}{\upsilon^{V2}}\frac{i_{Max}^{V}}{r_{\max}^{V}}\sum_{j=1}^{N}i_{j}^{V}r_{hj}^{V}-\ln\left(D\sqrt{2\pi}\sigma^{AV2}\right)-\frac{d\left(\vartheta_{k}^{A},\vartheta_{h}^{V}\right)^{2}}{2\sigma^{AV2}} \tag{29}$$

Since the third term is constant, the quantity to be maximized now becomes:

$$\widetilde{\widetilde{\psi}}\left(\vartheta_{k}^{A},\vartheta_{h}^{V}\right)=\frac{1}{\upsilon^{A2}}\frac{i_{Max}^{A}}{r_{\max}^{A}}\sum_{j=1}^{N}i_{j}^{A}r_{kj}^{A}+\frac{1}{\upsilon^{V2}}\frac{i_{Max}^{V}}{r_{\max}^{V}}\sum_{j=1}^{N}i_{j}^{V}r_{hj}^{V}-\frac{d\left(\vartheta_{k}^{A},\vartheta_{h}^{V}\right)^{2}}{2\sigma^{AV2}} \tag{30}$$

This signifies that we need for a cross-modal term, which reduces the function to be maximized, the greater the distance between the two positions, ϑ_{k}^{A} and ϑ_{h}^{V}

Let us now assume that the last term in Eq. (30) is equally divided into two equal contributions. We can write

$$\widetilde{\widetilde{\psi}}\left(\vartheta_{k}^{A},\vartheta_{h}^{V}\right)=\frac{1}{\upsilon^{A2}}\frac{i_{Max}^{A}}{r_{\max}^{A}}\left[\sum_{j=1}^{N}i_{j}^{A}r_{kj}^{A}-\frac{\upsilon^{A2}}{4\sigma^{AV2}}\frac{r_{\max}^{A}}{i_{Max}^{A}}d\left(\vartheta_{k}^{A},\vartheta_{h}^{V}\right)^{2}\right]+\frac{1}{\upsilon^{V2}}\frac{i_{Max}^{V}}{r_{\max}^{V}}\left[\sum_{j=1}^{N}i_{j}^{V}r_{hj}^{V}-\frac{\upsilon^{V2}}{4\sigma^{AV2}}\frac{r_{\max}^{V}}{i_{Max}^{V}}d\left(\vartheta_{k}^{A},\vartheta_{h}^{V}\right)^{2}\right] \tag{31}$$

We can interpret the term inside the first bracket as the input to the auditory neuron at position k, and the term in the second bracket as the input to the visual neuron at position h. Both are then

transformed into neuron output through a monotonic function. If this assumption holds, maximization of Eq. (31) is the same as maximization of the activity in the two neural chains (the auditory one, to obtain ϑ_k^A and the visual one, to obtain ϑ_k^V). We can conclude that each neuron must receive an additional input, besides that coming from its receptive field, signaling the distance from the winner in the other modality. We claim this can be easily realized with cross-modal synapses.

These ideas are implemented via the neural network described below.

2.2 CALCULATION: IMPLEMENTATION OF THE NEURAL NETWORK

2.2.1 Basal structure of the network

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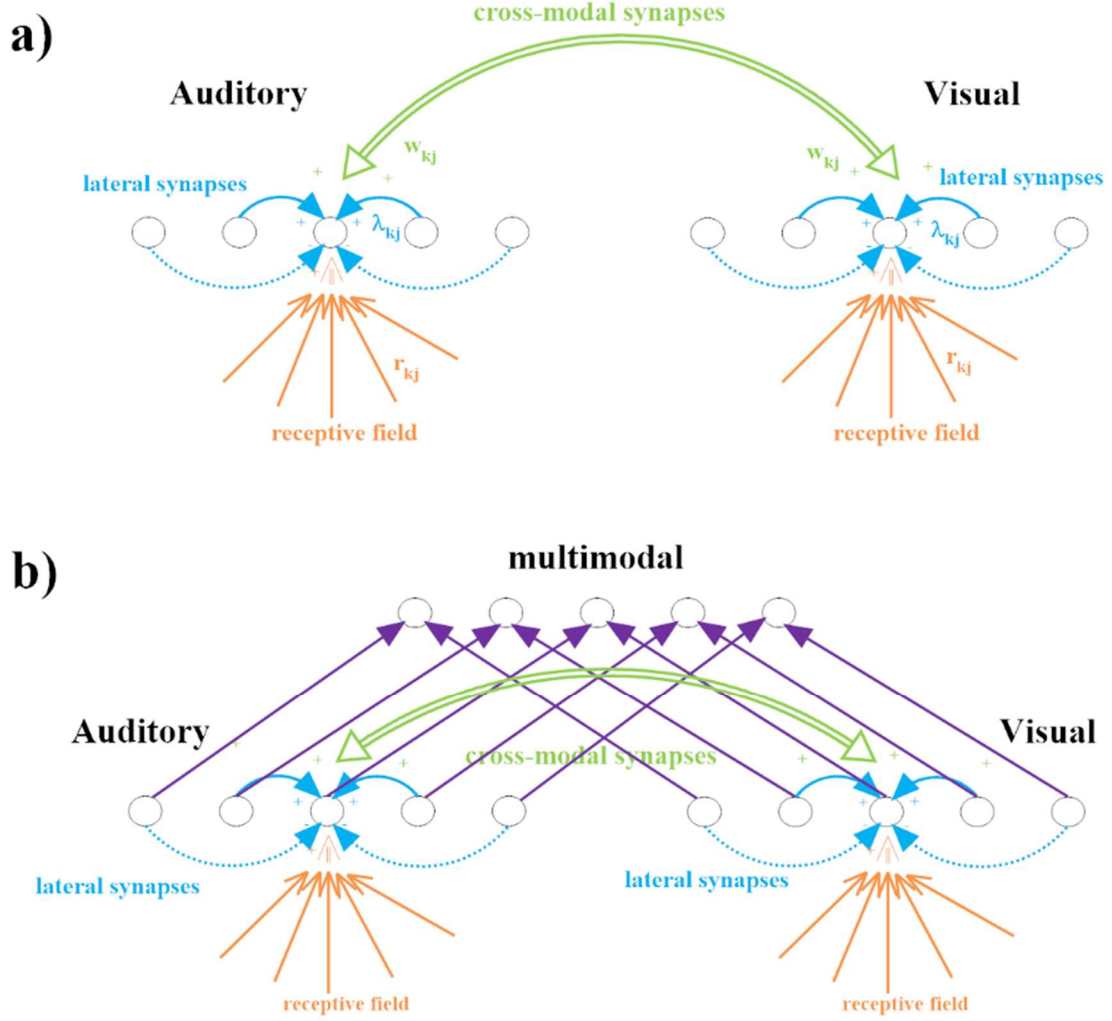
Fig. 1 – Neural network model

Panel a: The neural network used in the present work to address the problem of two separate estimates for the positions of the auditory and visual stimuli. Each neuron accomplishes the scalar product of the external stimulus and its receptive field (r_{kj}), but also receives lateral synapses (λ_{kj}) from other neurons of the same modality, and cross-modal synapses (w_{kj}) from neurons of the other modality. Synapses r_{kj} and w_{kj} are trained with the adopted learning rule. After training they acquire a campanular shape (see Figure 4). *Panel b*: modification of the previous network to address the single causal inference problem, when the observer is envisaged to treat the two stimuli as a single event (see Section 2.2.3). Neurons in the downstream multimodal network receive excitatory inputs from the two unisensory (auditory and visual) neurons located at the same position, with identical weights.

Each neuron receives three different kinds of inputs: a sensory input from the environment (say *u*), a lateral input from neurons of the same modality (say *l*) and a cross-modal input from neurons of the other modality (say *c*). The global input (equal to the sum of the previous three contributions)

is then passed through a sigmoidal relationship, $\varphi(\ )$, which accounts for the presence of a lower threshold and upper saturation in neuron activity, and a first-order low-pass filter with time constant τ, which accounts for the neuron integrative capacity.

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

$$\tau \frac{dy_k^S}{dt} = -y_k^S + \phi\left(u_k^S + l_k^S + c_k^S\right) \tag{32}$$

where y_k^S represents the neuron output, and the sigmoidal relationship is described by the following equation

$$\phi(x) = \frac{1}{1 + \exp\left(-s(x - x_0)\right)} \tag{33}$$

s and x_0 are parameters, which set the slope and the position of the sigmoidal relationship. According to Eq. (33), the neuron output activity is normalized between 0 and 1 (zero means a silent neuron, one a maximally activated neuron).

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

The expression for the sensory input is computed as the scalar product of the sensory representation of the stimulus ($I^S = [i_1^S\ i_2^S \cdots i_k^S \cdots i_N^S]^T$) and the neuron receptive field ($R_k^S = [r_{k1}^S\ r_{k2}^S \cdots r_{kj}^S \cdots r_{kN}^S]^T$) :

$$u_k^S = \sum_{j=1}^{N} r_{kj}^S i_j^S \tag{34}$$

which is the same as Eq. (13).

We assumed that the neuron receptive field, R_k^S, has initially a large extension, described with a Gaussian function, and then progressively shrinks during training, to fit the width of the external input (see section "Training the network").

22

The lateral input is computed as follows

$$l_k^S = \sum_{j=1}^{N} \lambda_{kj} y_j^S \tag{35}$$

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

$$\lambda_{kj} = \lambda_{ex} \exp\left(-\frac{d(\vartheta_j, \vartheta_k)^2}{2\sigma_{ex}^2}\right) - \lambda_{in} \exp\left(-\frac{d(\vartheta_j, \vartheta_k)^2}{2\sigma_{in}^2}\right) \tag{36}$$

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

$$d(\vartheta_j, \vartheta_k) = \begin{cases} |\vartheta_j - \vartheta_k| & if \ |\vartheta_j - \vartheta_k| \le D/2 \\ D - |\vartheta_j - \vartheta_k| & if \ |\vartheta_j - \vartheta_k| > D/2 \end{cases}.$$

It is worth noting that we used the same expression of lateral synapses (Eq. 36) in both the auditory and visual areas, to limit the number of model assumptions.

Finally, the cross-modal term in Eq. (32) is computed as the convolution of the vector of cross modal synapses and the activity in the other unisensory area, i.e.

$$c_k^S = \sum_{j=1}^{N} w_{kj}^{SQ} y_j^Q \qquad with \ \ S = A \ or \ V \quad Q = A \ or \ V \ \ with \ \ S \ne Q \tag{37}$$

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

2.2.2 Training the network

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

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

$$i_k^S = i_{Max}^S \exp\left(-\frac{\left(d\left(\vartheta^S, \vartheta_k\right)\right)^2}{2\sigma^{S2}}\right) + n_k^S \quad \text{with } S = A, V \tag{38}$$

where ϑ^S represents the spatial position of the stimulus, i_{Max}^S is the stimulus strength (equal to the value of the stimulus at the central position in the absence of noise) and σ^S is the standard deviation of the spatial representation. According to physiology, we assumed that the visual inputs are spatially more accurate than the auditory ones, hence we set $\sigma^V < \sigma^A$. Conversely, we assumed that the standard deviation of noise (say υ^S) is a given fraction of the input strength. In the subsequent simulations we assumed $\upsilon^S = i_{Max}^S/3$

The positions of the two stimuli (i.e., ϑ^A and ϑ^V in Eq. 38) have been randomly generated from the prior probability distribution in Eq. (28), using a very small value for parameter β_l (see section results) i.e. we assumed that the two stimuli, when simultaneously present (cross-modal condition) are always generated from the same source.

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

$$\Delta r_{kj}^S = \gamma \, y_k^S\left(i_j^S - r_{kj}^S\right) \quad \text{with } S = A, V \tag{39}$$

$$\Delta w_{kj}^{SQ} = \gamma \, y_k^S\left(y_j^Q - w_{kj}^{SQ}\right) \quad \text{with } S = A,V \quad Q = A,V \quad Q \neq S \tag{40}$$

Eqs. (39) and (40) have been applied, at each step, using the final steady state values of the neuron output (i.e., when transient phenomena are exhausted).

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

$$r_{kj}^{S} = r_0 \exp\left(-\frac{\left(d\left(\vartheta_j, \vartheta_k\right)\right)^2}{2\sigma_R^2} \right) \text{ with } S = A, V \tag{41}$$

where r_0 sets the initial strength of the receptive field, and σ_R establishes its initial spatial extension (we assume $\sigma_R > \sigma_A$ and $\sigma_R > \sigma_V$ i.e. a wide initial receptive fields). Of course, Eq. (41) holds only at the first step of training.

2.2.3 Inference of a single position

The previous network can provide a separate estimation for the auditory and visual positions (using the maximum activity or the barycenter in each layer separately). However, in a complex environment, the observer is asked to infer a single position for a multisensory event. To assess this problem, we included a further downstream layer of multimodal neurons (see Fig. 1, bottom panel). The single position is then computed using the maximum activity (or the barycenter) in this layer.

For the sake of simplicity, we assumed that each neuron in the multimodal layer receives excitation just from the two unisensory neurons (auditory and visual), located at the same spatial position, via identical weights. This network arrangement agrees with the assumption by others (see (Ma et al., 2006) and (Fetsch et al., 2012)) that the sum of population codes can be used to realize an optimal Bayesian inference. The output of the multimodal neurons is then computed using the same sigmoidal characteristic, $\phi(x)$ see Eq. (33), used for unisensory neurons. Hence, we can write:

$$y_k^M = \varphi\left(u_k^M\right) = \varphi\left(w^{MA} y_k^A + w^{MV} y_k^V\right) \tag{42}$$

where y_k^M is the activity of the multimodal neuron at position k, and $w^{MA} = w^{MV}$ are the two synapse weights from the unisensory neurons to the multimodal neuron (see Table II and the Appendix for

the parameter choice). Temporal dynamics has not been used in Eq. (42), since this is superfluous in a feedforward schema.

All model parameters are shown in Table 1. Parameter assignment is discussed in the Appendix.

Table 1

Parameters values

$\sigma^A = 20\,\mathrm{deg}$	$\sigma^V = 4\,\mathrm{deg}$ (20 deg, 40 deg)	$i^A_{Max} = 36$	$i^V_{Max} = 20$ (basal)	$i^V_{Max} = 47$ (moderately blurred)	$i^V_{Max} = 86$ (strongly blurred)
$v^A = i^A_{Max}/3$	$v^V = i^V_{Max}/3$	$N = 180$	$D = 180$	$x_0 = 0.7$	$s = 0.7$
$\tau = 5$ ms	$\lambda_{ex} = 1.9$	$\lambda_{in} = 1.85$	$\sigma_{ex} = 12$ deg	$\sigma_{in} = 24$ deg	$r_0 = 1.5$
$\sigma_R = 30\,\mathrm{deg}$	$w^{MA} = 16$	$w^{MV} = 16$	$\gamma = 0.04$	$\sigma^{AV} = 1.5\,\mathrm{deg}$	$\beta_1 = 10^{-14}$

3. RESULTS

In order to test the concepts developed in the method section, we simulated the neural network behavior and compared the results with those obtainable with the theoretical estimators (the maximum likelihood estimator (MLE) in unimodal conditions, and the Bayesian maximum posterior probability estimator in cross-modal conditions).

The network was subjected to 18000 epochs of training. Each epoch consisted of five trials: 2 unisensory auditory stimuli, 2 unisensory visual stimuli and 1 cross-modal stimulus. At the end of each individual trial (duration 120 ms), when the neurons are in steady state conditions, the synapses were trained using the learning rules described in Method section (i.e., a Hebb rule with a decay factor). Hence, the total number of trials was 90000. The position of the stimuli was randomly chosen between 1 and 180 (uniform distribution). In case of cross modal stimuli, the visual stimulus was always "almost coincident" with the auditory one: i.e., we used Eq. (28), with $\sigma^{AV} = 1.5$ deg and $\beta_1 \cong 0$ (hence, the two positions differs less than 3 deg, with 90% probability). This choice reproduces a situation in which the two stimuli originate from the same source, and so

have almost identical location. The same cross-modal probability was used, as a prior probability, in the implementation of the theoretical Bayesian estimator.

All parameters characterizing the training phase (learning rate, superimposed noise, ecc…) are reported in the method section.

Receptive fields – First, we analyzed how the RFs are progressively affected by training. To this end, Fig. 2 shows how the RFs of two representative auditory and visual neurons change: they are initially large, and progressively shrink, until they reach a final stable configuration. Worth noting, the RFs of auditory neurons remain quite large, whereas the visual ones become much more tuned, reflecting the standard deviation (SD) of the external stimuli.
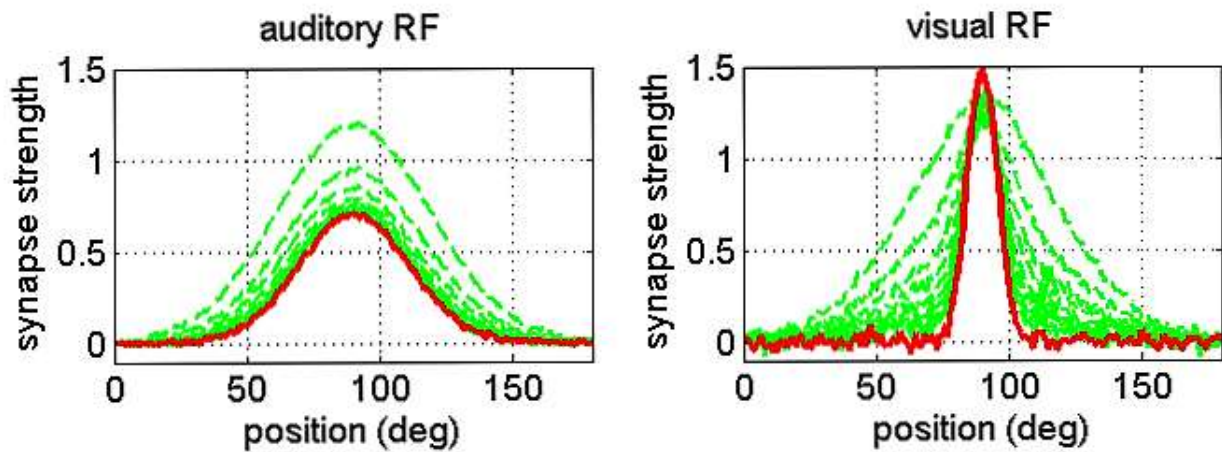

Fig. 2 – The receptive fields shrink during training

Examples of the progressive shrinking of the receptive fields (RFs) during training with the adopted rule. The figures illustrate the RFs of two exemplary neurons (one auditory, left panel, the other visual, right panel) with preferred position at 90 deg. The behavior of the other RFs is similar, with small statistical differences due to noise and random input presentations. The dashed green lines are the RFs at different steps during training. The continuous red lines are the final RFs. It is worth noting that, at the end of training, the visual RFs are more tuned than the auditory ones, reflecting the more precise spatial localization of the inputs.

An important consequence of the adopted training rule is that, after training, the RF of each neuron reflects the average input exciting that neuron. In our model, this is provided by Eq. 7 in Method, i.e.,

a Gaussian function with assigned standard deviation ($\sigma^A = 20\,\mathrm{deg}$ for the auditory inputs, and $\sigma^V = 4\,\mathrm{deg}$ for the visual ones). To check this prediction, Fig. 3 compares the final RFs of the two exemplary neurons with the Gaussian functions representing the average inputs used during training. The agreement is very good for what concerns the auditory RF; the visual RF also shows good agreement as to the spatial SD, but with a small difference in the amplitude. We ascribe this small difference to the effect of lateral synapses.

In conclusion, we can say that, after training with the specific adopted rule, the SD of the RFs carefully reflects the average SD of the input stimuli, with the visual neurons more spatially focused than the auditory ones.
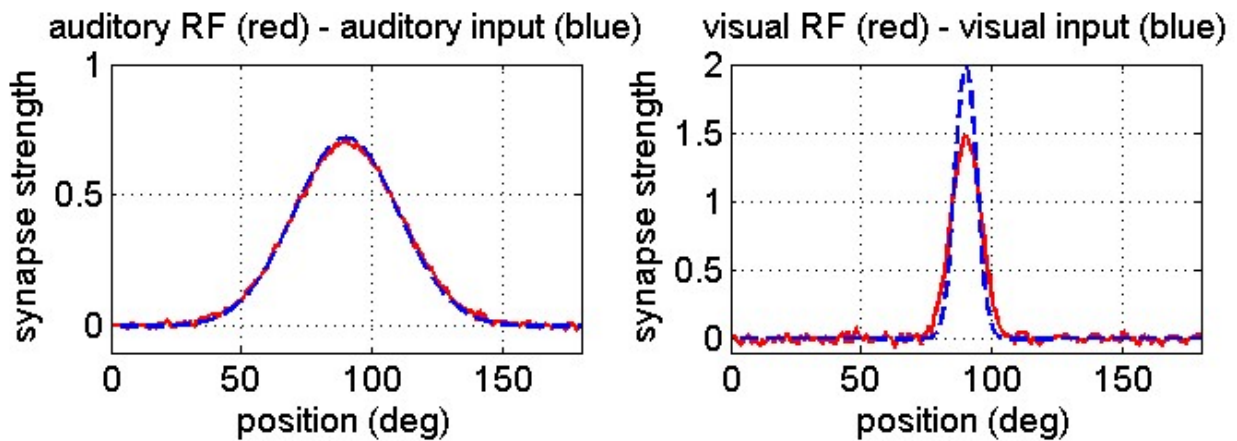

Fig. 3 - The final receptive fields match the average inputs

Comparison between the final values of the RFs (for the two same exemplary neurons as in Fig. 2) and the average inputs used during training at the given location (see Eq. 7). The RFs (continuous red lines) are very close to the average inputs (dashed blue lines), which is an essential property of the adopted learning rule.

Cross modal synapses – If all input stimuli were unimodal, no cross-modal synapses would be generated, and the two networks would behave as purely unisensory. In this condition, changes in the RFs would be the unique effect of training, and the network would implement a maximum likelihood estimate of the individual unisensory inputs (see also the Method section for a demonstration). However, thanks to the presence of a certain amount of cross-modal stimuli, cross-modal synapses

are also created between the two networks. By way of example, Fig. 4 shows the synapses entering into one exemplary auditory and one exemplary visual neuron from all neurons in the other modality. The cross-modal synapses initially are ineffective, and progressively are reinforced on the basis of the sensory experience perceived during the training. At the end of this phase, a neuron in one modality receives synapses from neurons in the other modality, which code for approximately the same position. This incorporates the multisensory experience in the cross-modal synapses and thus reflects a prior knowledge on the frequent co-occurrence of stimuli at the same position.
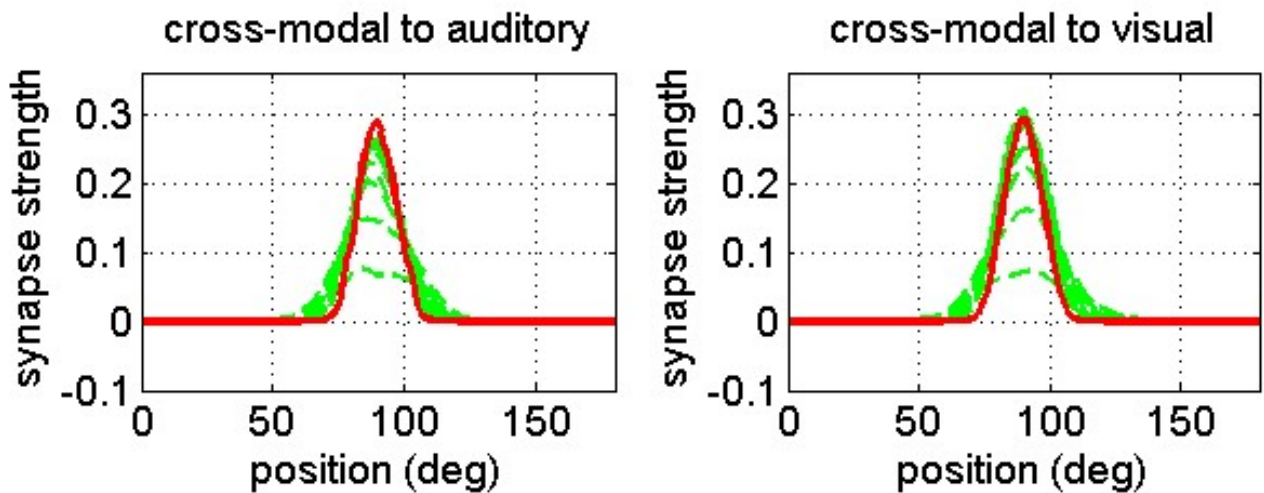

Fig. 4 – Cross-modal synapses development during training

Examples of the progressive development of cross-modal synapses during training with the adopted rule. The figures illustrate the synapses entering into two exemplary neurons (one auditory, left panel, the other visual, right panel) with preferred position at 90 deg, from all other neurons in the other modality (i.e., these are synapses W_{90j}^{AV} and W_{90j}^{VA} with j = 1,2, ..., 180). The dashed green lines are the synapse patterns at different steps during training. The continuous red lines are the final synapses. The synapses start from zero and then progressively rise up to a final steady state level, reflecting the statistics of cross-modal stimuli.

After training, we tested network behavior with a variety of input stimuli. First, we analyzed how the network provides separate estimates for the visual and auditory modality (both in unimodal and cross-modal conditions), computed as a metric (the maximum or the barycenter) in the corresponding

visual and auditory layer. We also tested a third metrics (i.e., the vector population decoder described in (Fischer & Peña, 2011)). However, it provides results almost undistinguishable from those of the barycenter metrics. Then, we examined how the network, in case of cross-modal stimuli presentation, can infer a single position by envisaging the bimodal presentation as a single event, by computing the metrics in the downstream multisensory layer.

Model response to unimodal inputs –. First, we simulated the response of the trained network to unimodal inputs. In this condition, due to the low values of cross-modal synapses, only the network in one modality is excited, while the other layer remains silent (i.e., we verified that a single unimodal stimulus does not produce any phantom effect in the other modality).

As described in the section "theoretical aspects" of Methods, in unimodal conditions the external input to neurons (i.e., the inner product between the stimulus and the receptive field) is a close approximation of the logarithm of the likelihood function (see Eq. (17)). To test this prediction, Fig. 5 shows the external input to all neurons in the network, in response to an auditory stimulus (left panel/auditory net) or a visual stimulus (right panel/visual net) placed at position 90 deg compared with the log likelihood function. The latter was computed from knowledge of the likelihood probabilities used during training (i.e. Eq. (12)). The curves in Fig. 5 are normalized to sum 1. Figure 5 requires some comments. The external input to the neuron reproduces the true log-likelihood very well, in accordance with the theoretical predictions. However, the total input is sharper (especially in the auditory case), due to the presence of lateral inhibitory mechanisms. This signifies that lateral inhibition can modify the perceived accuracy of the input. This aspect, further commented in the Discussion, can be exploited in future works to modulate the likelihood, according to some prior information on the unisensory inputs.

Moreover, we can observe that the network auditory likelihood is larger than the visual one, reflecting the poorer spatial accuracy of the auditory stimuli. This signifies that a visual stimulus placed at a certain distance may affect the auditory response (due to the broader neural input of auditory neurons) but not viceversa. As shown later, this phenomenon is at the origin of

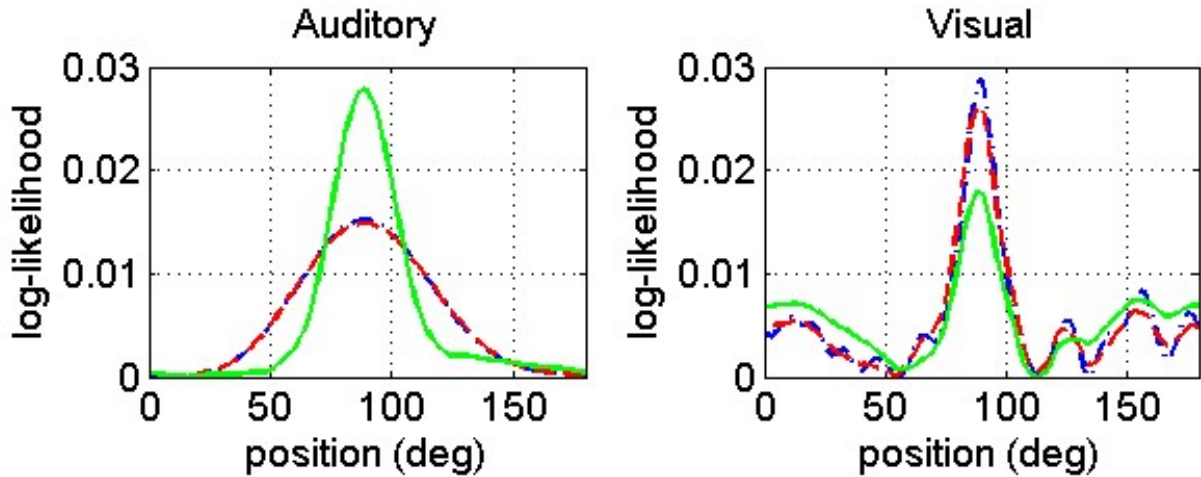ventriloquism, i.e., the capacity of the more accurate stimulus to affect perception of the less accurate one.

Fig. 5 – The external input in a unimodal layer matches the likelihood function in unimodal conditions

Comparison between the external input (inner product between the stimulus and receptive field) reaching the 180 neurons in a unisensory layer during a unimodal stimulation (dashed red line), and the logarithm of the likelihood function (dot-dashed blue line). All figures refer to a single unimodal stimulation (left panel auditory, right panel, visual), with the noisy external stimulus placed at position 90 deg. Per each neuron, the external input was computed as the scalar product of the external stimulus and the receptive field. The continuous green lines represent the *overall* input to the neurons, including also the contributions of lateral synapses from other neurons of the same modality. The curves are normalized to sum 1. Worth noting, the log likelihood function is very close to the external input. The contribution of the lateral synapses moderately sharpens this curve, but does not appreciably modify nor the maximum neither the barycenter.

Finally, we compared the positions estimated with the network (in *unimodal conditions*) with those of the MLE. To this end, we performed 180 different trials for each modality, by providing a noisy unimodal stimulus at each position. Results are shown in Fig. 6, where the predictions obtained with the MLE (by maximizing Eq. (12)) are compared with the predictions obtained with the network using the position of the maximally active neuron. For completeness, we also show the results obtained using the barycenter of neuron activity in the layer. The results underline the strong similarities between the MLE and the network estimate, both in the case of maximum activity or

barycenter metrics. As well expected, the estimate of the visual position is more accurate, compared with the auditory position estimate.
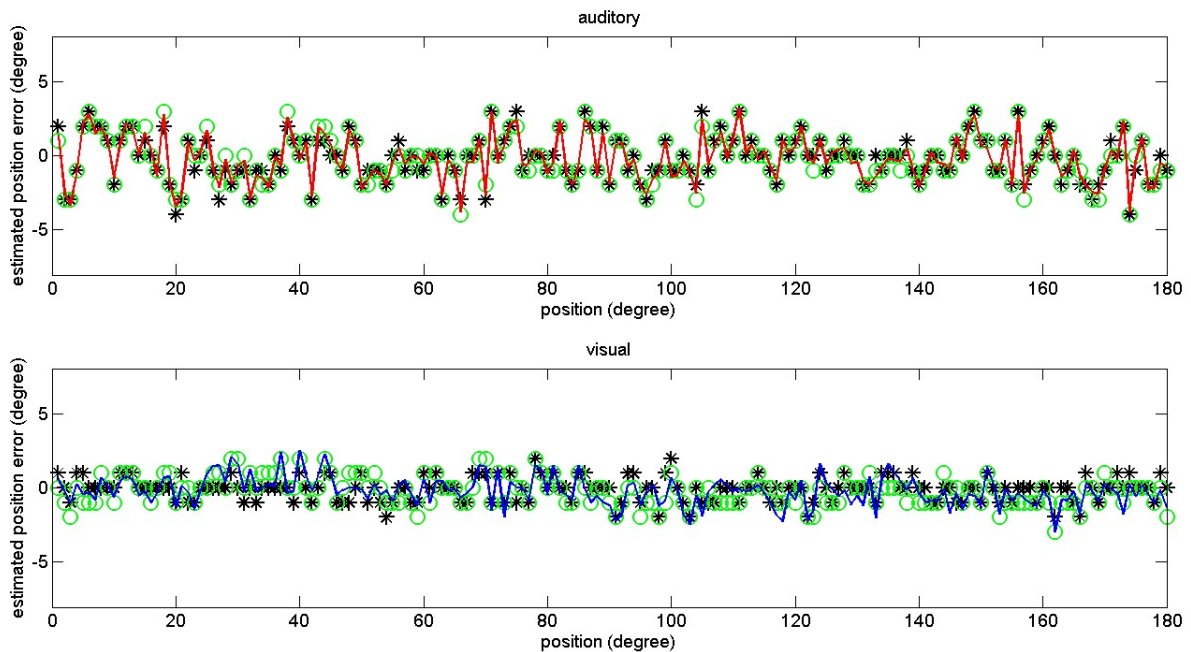

Fig. 6 – The errors in the network estimates matches the maximum likelihood errors

Comparison between the errors in the perceived positions of the auditory (upper panel) and visual (bottom panel) stimuli, obtained in unimodal trials with the maximum likelihood estimator (*black asterisks*) and with the network metrics (position of the neuron with maximal activity, *green circles*; barycenter of the activity, *continuous line*). 180 unimodal trials were performed per each modality, with stimuli at all positions. The figures display the position of the stimulus in the x-axis, and the perceived error (estimated position minus true position) in the y-axis. Worth noting, the estimates obtained with the network are close to the MLE estimates. Moreover, visual estimates are more precise than the auditory ones. Mean values and SD are reported in Table 2.

A final comparison between the different estimates (in terms of mean value and SD) is given in the upper panel of Table 2. In this table we present the mean value of the estimates, evaluated on 180 trials, to show that all estimates are almost unbiased, and the SD, as an indicator of their accuracy.

Table 2

Comparison between the performance of the Maximum Likelihood estimator, of the network barycenter estimator, and of the maximal network activity estimator, computed in unimodal conditions (upper panel). The input strengths and noise levels were the same as in Table 1. Results have been obtained within 180 independent trials per each modality, at all network positions. We report the mean values of the estimates, to show that all estimators are almost unbiased, and the SD of the estimates, as an index of their variability (hence, accuracy of the single estimate).

Unisensory

	Maximum likelihood	Barycenter	Maximum activity
Mean (auditory)	-0.21	-0.24	-0.23
SD (auditory)	1.55	1.51	1.58
Mean (visual)	-0.1	-0.21	-0.22
SD (visual)	0.81	0.98	0.97

Cross-modal

The same estimations performed in cross-modal conditions, with coincident visual and auditory stimuli. In this case, we compared network results with those of the Bayesian estimator to infer the position of the auditory and visual stimuli. Finally, the last lines represents the estimate of a single common position inferred by the network model in the downstream multisensory layer.

	Bayesian estimate	Barycenter	Maximum activity
Mean (auditory)	-0.18	-0.15	-012
SD (auditory)	0.98	1.00	0.99
Mean (visual)	-0.1	-0.22	-0.22
SD (visual)	0.78	1.00	0.95
Mean (multisensory)	-	-0.20	-0.11
SD (multisensory)	-	0.93	0.95

In conclusion, we can say that, in unimodal conditions, each trained network approaches a maximum likelihood estimate of the stimulus position quite well. In particular, it is worth noting that the SDs in Table 2 depends on two factors: the noise level added to the inputs (the higher the noise, the higher the SD), and the way noise affects the maximum in the curves of Fig.5 (the wider the

curves, the higher the noise effect). Hence, SDs are not identical to the width of the likelihood functions, but influenced by it.

Model response to cross-modal inputs – Subsequently, we simulated network behavior in response to cross-modal stimuli.

First, we repeated the estimates in Table 2 using two cross-modal stimuli at the same position. The results are presented in the bottom panel of Table 2, and compared with the predictions of the Maximum Posterior Probability Bayesian Estimator (Eq. 22). To this end, we used the same prior probability (Eq. 28) adopted during training in case of cross-modal stimuli (i.e., a uniform distribution of one stimulus and a Gaussian density with 1.5 deg standard deviation, to represent the distance between the two stimuli). Parameter β_I was chosen close to zero, to admit the remote possibility that the two-cross modal inputs originate independently (see the Method section for more details). The choice of the prior probability is further discussed in the last section. Finally, in cross-modal conditions we also evaluated the accuracy of a single position estimate for both stimuli, performed in the multisensory layer (see also the section "Inference of a single cause" below).

Several conclusions can be drawn by comparing results in cross-modal conditions with those in the unisensory cases. i) The accuracy of the auditory estimate improves compared with the unisensory case, thanks to the presence of a spatially coincident input of the other modality. ii) The accuracy of the visual estimate in multisensory condition is almost the same as in the unisensory case. iii) The accuracy of the position inferred in the multisensory layer is just a little better than position estimates in each of the unisensory layers). iv) There is good agreement between model estimates in multisensory conditions, and the Bayesian estimates, although the latter have a smaller SD for the visual case.

However, the aforementioned results were obtained using rather strong inputs, able to evoke a significant activity in the unisensory layers (more than 2/3 the maximum). Since the benefits of multisensory integration are especially evident in the presence of weak stimuli (inverse effectiveness) we repeated the previous estimates using weaker inputs (able to evoke an activity in unisensory

conditions as low as 30% of the maximum) and greater percentage noise. Now (Table 3) one can observe that the SDs in unisensory conditions are higher than in the previous case, due to the greater influence of noise. In multisensory conditions, we can observe a moderate improvement in both auditory and visual estimates, especially evident looking at the maximum activity estimator. However, the performance of the network are less accurate compared with the Bayesian estimator. A significant improvement of the estimate can be achieved looking at the position in the multisensory layer (that computes a single position for both stimuli); in particular, the barycenter metrics exhibits a very high performance. This is the consequence of inverse effectiveness, i.e. activity in the multisensory layer reveals a strong enhancement compared with the unisensory case.

These results point out the advantage of having multisensory integration vs. unisensory inputs.

Table 3

Comparison between the performance of the Maximum Likelihood (or Bayesian) estimator, of the maximal network activity estimator, and of the network barycenter estimator, computed in unimodal conditions (upper panel) and in cross-modal conditions with spatially coincident auditory and visual stimuli (bottom panel). The meaning of the results is the same as in Table 2. However, these simulations have been performed with weaker inputs ($i_{\max}^A = 26$, $i_{\max}^V = 11$, and greater absolute noise $v^A = i_{Max}^A / 2$, $v^V = i_{Max}^V / 2$) to emphasize the benefit of multisensory integration.

Unisensory

	Maximum likelihood	Barycenter	Maximum activity
Mean (auditory)	0.17	0.19	0.16
SD (auditory)	2.49	2.57	2.57
Mean (visual)	-0.39	-0.68	-0.56
SD (visual)	3.44	2.72	3.42

Cross_modal

	Bayesian estimate	Barycenter	Maximum activity
Mean (auditory)	-0.02	0.12	0.08
SD (auditory)	1.17	2.09	1.84
Mean (visual)	-0.07	-0.63	-0.33

SD (visual)	1.01	2.48	1.41
Mean (multisensory)	-	-0.13	-0.17
SD (multisensory)	-	0.95	1.36

Subsequently, we investigated the ventriloquism effect (Bertelson & Radeau, 1981; Hairston et al., 2003; Wallace et al., 2004). To this end, we provided two cross-modal inputs at different positions: the auditory stimulus was placed from position 1 to 180 deg and, at each auditory position, we added a second visual stimulus with a distance from – 40 deg to + 40 deg from the auditory one. A fundamental aspect of these trials is that both networks are simultaneously excited and, due to the presence of cross-modal synapses, the activity in one network is affected by the activity in the other. Hence, the perceived position of the stimuli (both computed with the maximum activity or the barycenter metrics) is shifted compared with the real one.

A first summary of the results is presented in the upper panels of Fig. 7, where we show the shift between the true position of the stimulus and the perceived position (in the corresponding unimodal layers), plotted vs. the distance between the auditory and the visual stimuli. Each point in the figure is the average over 180 trials (i.e., we averaged all trials with the same distance between the two stimuli), computed both with the maximum and barycenter metrics. As it is clear from the figure, the perceived position of the auditory stimulus is significantly shifted in the direction of the visual one (positive distances signify that the visual stimulus is at the right of the auditory one and viceversa); hence, we can observe a sort of "ventriloquism effect". The perception error is maximal (about 8 deg) when the audio-visual distance is at about 20 deg. The perceived position of the visual stimulus is also shifted in the direction of the auditory one, but the error is much smaller (maximum error less than 1.0 deg). The results are in good agreement with the behavioral ones (Bertelson & Radeau, 1981; Hairston et al., 2003; Wallace et al., 2004) (see bottom panel in Fig. 7). The only significant difference is that, in the model, the perception error decreases at a distance greater than 20 deg, where the

behavioral data still exhibit a large error. Possible reasons for these differences are discussed in the last section.
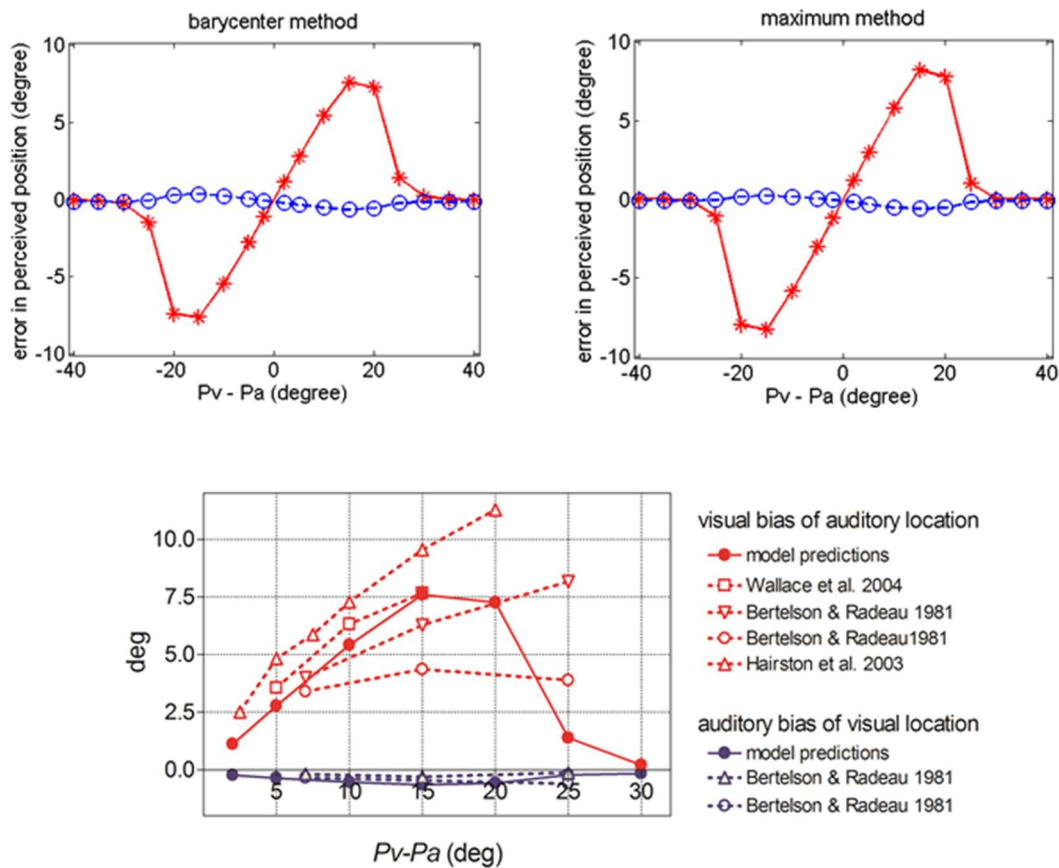

Fig. 7 – The network simulates the ventriloquism effect

Ventriloquism effect simulated with the network during cross-modal trials, and compared with data from the literature. Cross modal trials were performed, by moving the auditory stimulus from position 1 deg to position 180 deg and, at each auditory position, adding a second visual stimulus with a distance from – 40 deg to + 40 deg from the auditory one. Results are averaged over the 180 trials per each shift. The x-axis represents the audio-visual distance (where positive values indicate that the visual stimulus is placed on the right), the y-axis is the perceived error (estimated position minus true position). The upper panels show the results, either with the barycenter metrics (left) or the maximum activity metrics (right) (auditory perception: continuous red line; visual perception: dotted blue line). The bottom panel compares model results (continuous line) with data from the literature. Behavioral data are from (Bertelson & Radeau, 1981; Hairston et al., 2003; Wallace et al., 2004). The auditory perception exhibits a significant shift toward the visual one (ventriloquism). The visual perception exhibits just a minor shift toward the auditory location.

In order to clarify the nature of the ventriloquism effect, Fig. 8 shows some snapshots obtained during a single trial (i.e., the activity in the two layers frozen at different instants during an exemplary simulation). This figure shows how the activity in the auditory layer, initially positioned around the true auditory stimulus (85 deg), progressively shifts in the direction of the visual one, as a consequence of the cross-modal inputs, to reach a final position at about 92 deg.
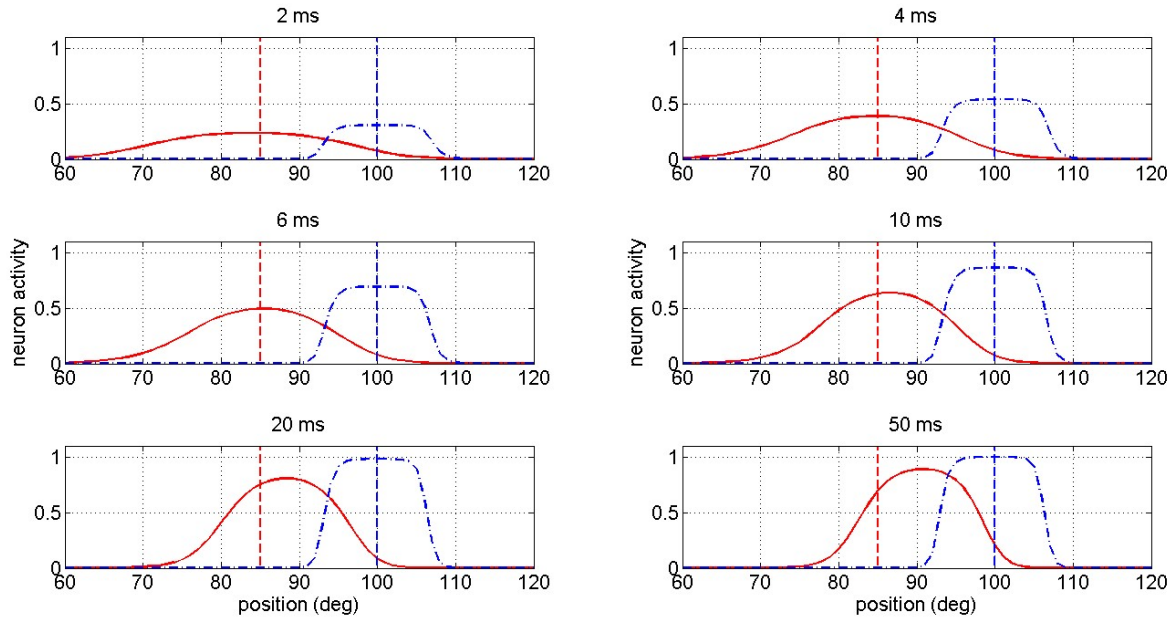

Fig. 8 – The temporal development of ventriloquism effect

Snapshots of the network activity, during the initial 50 ms of a cross-modal stimulation. The figures represent the activities of some central neurons (between position 60 and 120) in the auditory layer (continuous red lines) and in the visual network (dash-dotted blue lines) at successive instants, in response to a cross-modal stimulation. The auditory stimulus was placed at 85 deg, and the visual stimulus at 100 deg (the true positions are displayed through vertical dashed lines). The progressive shift of the auditory activity toward the visual position is evident. At the end, the auditory shift is about 7 deg.

Finally, we compared the predictions of the model, in case of cross modal inputs, with those obtained with the Maximum Posterior Probability Bayesian Estimator (i.e., Eq. (22)) in the same cross-modal conditions.

Results, illustrated in Fig. 9, show a satisfactory agreement between the position errors, obtained with the model, and those provided by the Bayesian Estimator. However, we can observe a certain

discrepancy between model and Bayesian estimates, especially in the range 20-30 deg for the auditory position, and 15-20 deg for the visual one. In these ranges, in fact, the Bayesian estimator predicts a smaller ventriloquism for the auditory position and a greater visual shift than the model prediction. To understand these differences we remind that, during training, we never used independent cross-modal inputs, but inputs at strongly correlated positions. Accordingly, in order to realize the Bayesian estimator, we assumed a prior probability with a very low value of parameter β_1 (close to zero) in Eq. 28. However, even very small changes in the value of β_1 are reflected in significant changes in the Bayesian estimate. Moreover, when realizing the Bayesian estimator, we assumed that the standard deviation of correlated inputs in the prior probability (i.e., parameter $\sigma^{AV} = 1.5$ deg) is perfectly known. This is not the case of a true estimator, that should also perform an estimation of this parameter on real data. For instance, with a value $\sigma^{AV} = 1.0$ deg, the results of the Bayesian estimator exhibit a smaller ventriloquism range for the auditory perception (up to about 15 deg) while the ventriloquism range increases if one uses $\sigma^{AV} = 2.0$ deg. Conversely, the model behavior is only marginally affected by moderate changes in the parameter σ^{AV} [indeed, we repeated the training procedure with σ^{AV} in the range 1-2 deg, and observed that the network results are only moderately affected].

In conclusion, we can say that the Bayesian estimates are strongly affected by the values of parameters in the prior probability, while model predictions are less sensitive to a change in these parameters in the training phase.
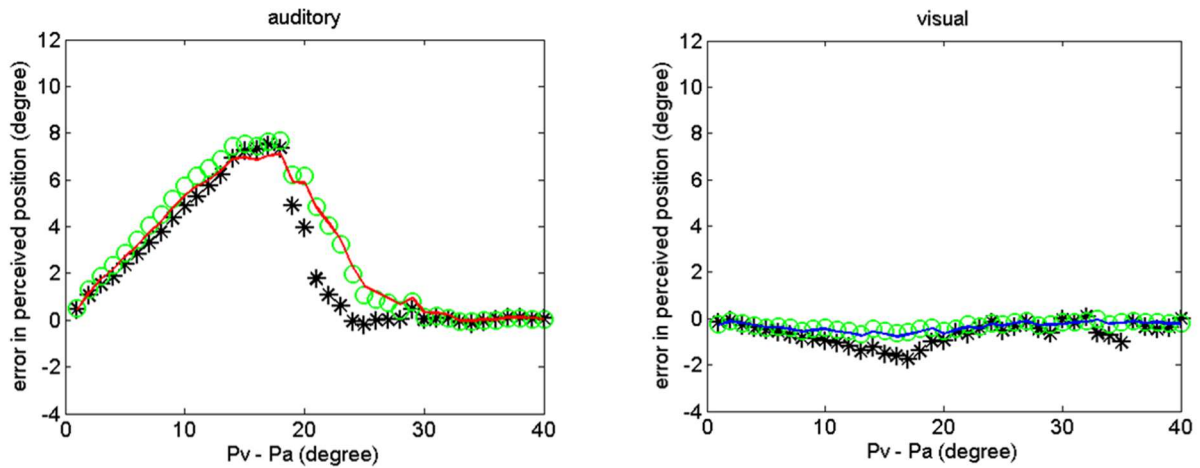
Fig. 9 – The network produces a near-optimal Bayesian estimate in cross-modal conditions

Comparison between the errors in the perceived positions of the auditory (left panel) and visual (right panel) stimuli, obtained in cross-modal trials with the maximum posterior probability (Bayesian) estimator (*black asterisks*) and with the network metrics (position of the neuron with maximal activity, *green circles*; barycenter of the activity, *continuous red line*). Cross-modal trials were performed with the auditory stimulus varying from position 1 to position 180 deg and, at each auditory position, adding a second visual stimulus with a distance from – 40 deg to + 40 deg from the auditory one. The figures display the audio-visual distance in the x-axis, and the perceived error (estimated position minus true position) in the y-axis. Worth noting, the estimates of the auditory shift obtained with the network are close to the Bayesian estimates in the range 0-20 deg, and above 30 deg.

Inference of a single cause – In all previous simulations, the network provided two separate estimates for the visual and the auditory positions (computed as the barycenter or maximum activity in the corresponding layer), i.e., we mimicked trials where subjects are asked to infer the two positions distinctly. Conversely, in the last portion of this work we simulated behavioral experiments, in which subjects are asked to infer a unique position, from the combination of the stimuli, treated as a single multisensory event. To perform this behavioral task, we exploited the third layer of multimodal neurons downstream the unisensory layers, as illustrated in Fig. 1b.

The single position estimates, in case of spatially coincident cross-modal stimuli are shown in the Tables 2 and 3. In particular, Table 3 emphasizes the strong benefit of a multisensory estimate in case of poor signal to noise ratio.

Subsequently, we repeated the cross-modal spatial localization tasks by Alais and Burr. In this case, each trial consists of a cross-modal presentation, with the visual and auditory stimuli in conflict: the visual input was shifted Δ degrees rightwards and the auditory stimulus Δ degrees leftwards of the central target position. These trials were repeated at all 180 target positions, by varying the shift in the range $[-6 \text{ deg } + 6 \text{ deg}]$; in each trial, the barycenter of the multimodal network (or, alternatively, the maximum) in the final steady state condition was computed as a metrics for the estimated position. The results were then averaged on all 180 positions for each shift. The trials were repeated in three different conditions: i) basal parameter values, where the visual input is much more spatially accurate than the auditory one ($\sigma^V = 4 \text{deg}; \sigma^A = 20 \text{deg}$). ii) the two inputs with the same spatial accuracy ($\sigma^V = 20 \text{deg}; \sigma^A = 20 \text{deg}$), iii) blurred conditions, where the visual stimulus is less accurate than the auditory one ($\sigma^V = 40 \text{deg}; \sigma^A = 20 \text{deg}$).

The results are summarized in Fig. 10 and compared with the results by Alais and Burr. The network automatically accounts for the different reliability of the stimuli. In basal conditions, the estimated position follows the visual shift. Conversely, in blurred conditions, the estimated position is mainly affected by the position of the auditory input.

Finally, in order to better understand the previous results, Fig. 11 shows an example of the global activity in the multimodal layer at three different positions of the auditory and visual stimuli, (in the basal case, moderately blurred and strongly blurred) and Fig. 12 summarizes the activity of two exemplary neurons in the multimodal layer. This has been computed by using a central position at 90 deg for the multisensory combination (hence the visual stimulus shifts to the right from 84 deg to 96 deg and, simultaneously, the auditory stimulus shifts to the left from 96 deg to 84 deg). The first neuron has preferred position at 85 deg (i.e., to the left compared with the central position), the second preferred position at 94 deg (to the right).
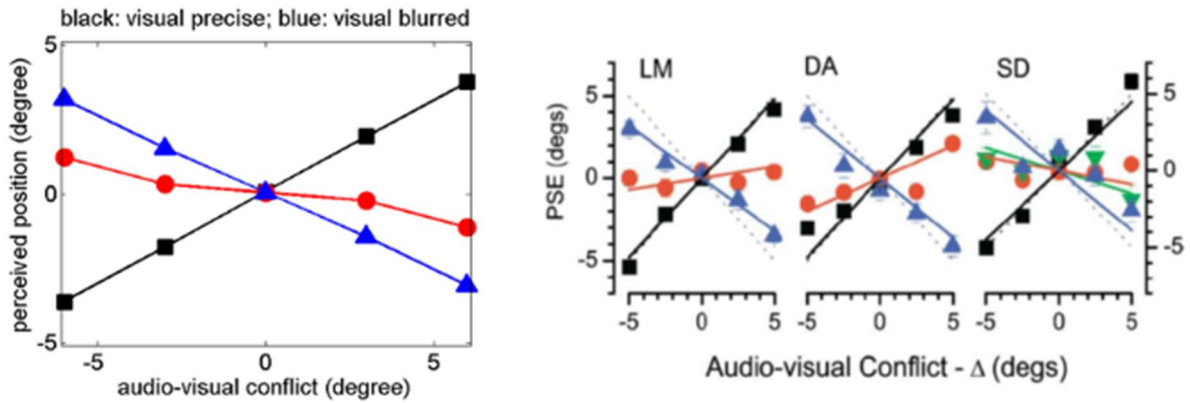
Fig. 10 – The multimodal layer provides more weights to the more reliable stimulus

Model simulations of the experiments by Alais and Burr (2004) (Alais & Burr, 2004) where the observer is asked to treat the two cross-modal stimuli as a single perception. The position was estimated using the barycenter of the activity in the multimodal network of Fig. 1b. Cross-modal trials were performed by varying a target position from 1 deg to 180 deg and, at each target position, using a visual stimulus shifted by $+\Delta$ deg from the center, and an auditory stimulus shifted by $-\Delta$ deg from the center (hence, positive value of Δ mean that the visual stimulus is shifted rightwards). At each target position, the "conflict" parameter Δ was varied from -6 deg to $+6$ deg. Results shows the distance between the perceived position and the target position (i.e., the position error), averaged over the 180 cases, plotted vs. the audio-visual conflict. Three cases are presented: i) basal parameter values, with the visual stimulus much more spatially accurate than the auditory one ($\sigma^V = 4\,\text{deg}; \sigma^A = 20\text{deg}$, black line and squares). ii) the two inputs with the same spatial accuracy ($\sigma^V = 20\text{deg}; \sigma^A = 20\text{deg}$, red line and closed circles).), iii) blurred condition, where the visual stimulus is less accurate than the auditory one ($\sigma^V = 40\text{deg}; \sigma^A = 20\text{deg}$, blue line and triangles).). Results by (Alais & Burr, 2004) are reported in the right panel, for comparison. In basal conditions, the visual stimulus dominates the response. In blurred conditions, the response follows the auditory shift.

Fig. 11 shows that the activity in the multimodal layer follows the more reliable input. Fig. 12 displays the same behavior at the level of the individual neuron. In basal conditions, both neurons maximally respond when the visual stimulus is close to their preferred position, and reduce their response when the visual stimulus moves away from the preferred position, despite the simultaneous proximity of an auditory stimulus. This signifies that these neurons give more relevance to the visual input than to the auditory one, and their tuning curves are prevalently visual. Conversely, in blurred

conditions, the same neurons maximally respond to the presence of an auditory stimulus at their preferred position, and reduce their response when the auditory inputs leave the position, despite the presence of a visual stimulus, i.e., the neurons gave more emphasis to the auditory modality. Hence, the *appearance* is that of a shift of the tuning curves, and of an automatic readjustment of the corresponding weights.
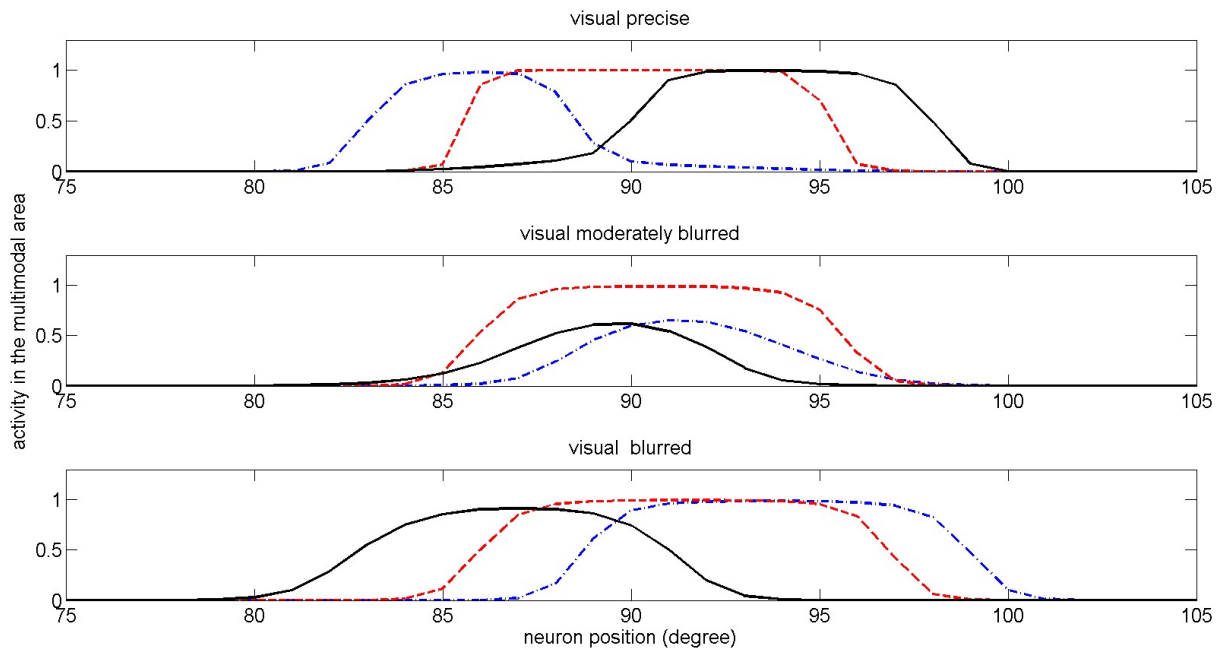

Fig. 11 – The activity in the multimodal layer shifts with the stimulus reliability

Activity in the multimodal layer, during the trials described in Fig. 10, when the target (i.e., the central) position for the stimuli was at 90 deg. In these simulations, in order to improve the aspect of the tuning curves with less saturation, we slightly reduced the inputs (auditory input: i_{Max}^{A} = 35; visual input i_{Max}^{V} = 14 (normal), 35 (moderately blurred), 70 (blurred) . The dot-dashed blue lines are the responses when the visual stimulus is at position 84 deg and the auditory stimulus at position 96 deg. The dashed red lines are the responses when both stimuli are at position 90 deg (note the significant multisensory enhancement). The continuous black lines are the responses when the visual stimulus is at 96 deg and the auditory stimulus at 84 deg. Upper panel - normal conditions: the activity is dominated by the visual input. Bottom panel - blurred condition: the activity is dominated by the auditory input.
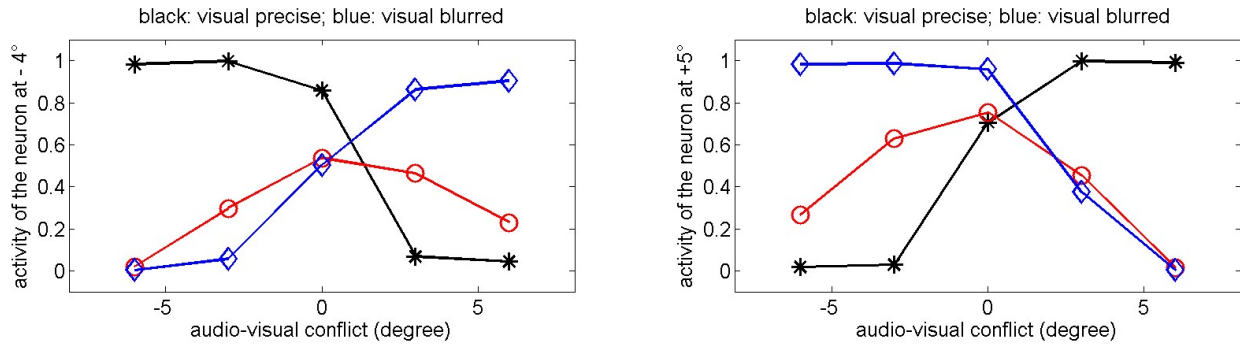
Fig. 12 – The tuning curves of multimodal neurons shift with the reliability of the stimulus

The responses of two exemplary neurons in the multimodal network, during the trials described in Fig. 11, when the target (i.e., the central) position for the stimuli was at 90 deg. The left panel refers to the neuron with preferred position 85 deg; the right panel to the neuron with preferred position 94 deg. The x-axis reports the audio-visual conflict, the y-axis displays the neuron activity in the three experimental conditions (black line and asterisks: basal, i.e., visual more accurate; red line and open circles: same accuracy for the audio and visual stimuli; blue line and diamonds: blurred visual stimulus). Worth noting, in basal condition each neuron gives more weight to the visual input (i.e., it maximally responds when the visual input is close to its preferred position); in blurred conditions, each neuron provides more weight to the auditory input (i.e., it maximally responds when the auditory input is close to its preferred position).

4. DISCUSSION

The aim of this work was to analyze whether a simple neural network, consisting of two layers of unimodal neurons, connected via plastic cross-modal synapses, can be trained to realize a near-optimal Bayesian estimator. This approach is inspired by some pivotal ideas recently developed by Ma, Pouget et al (Ma et al., 2006; Ma & Rahamati, 2013; Pouget et al., 2013; Pouget et al., 2003), usually referred to with the term "probabilistic population coding". In particular, we exploit the idea that a chain of neurons can encode the probability functions. However, the present approach also introduces some original elements within this framework.

Innovative aspects and significance of the results

First, the synapses are not assigned a priori, but learned during a training period, in order to encode the statistics of the environment, i.e., the network learns the likelihood probabilities and the prior

probability functions from a repetition of random stimuli. To this end, we used a biologically plausible rule, consisting of the classic Hebb potentiation, based on the correlation of the presynaptic and postsynaptic activities, joined with a forgetting factor, to avoid an unbounded synapse increase. The significant achievement of this rule is that, after a sufficient training, it can encode the average input vector that excites the neuron (Hertz et al., 1991).

As it is well known, in order to realize a Bayesian estimator with maximal posterior probability, two kinds of complementary pieces of information must be encoded: the likelihood probability and the prior probability. The first summarizes the process of data formation, i.e., how the input data statistically depend on the unknown parameters (in our case the unknown parameters are the positions of the auditory and of the visual stimuli). The second encodes information on the statistics of the parameter themselves.

Our theoretical analysis (see section 2.1), and the subsequent simulation trials, demonstrate that, under the assumption of Gaussian white noise, the likelihood probabilities can be encoded in the neuron receptive fields, while the prior knowledge on the co-occurrence of the stimuli can be encoded in the cross-modal synapses. The possibility of encoding other forms of prior probability will be discussed later, in the section devoted to "future developments".

Perhaps the most innovative aspect of the present work consists in the role assigned to cross-modal synapses to represent the prior. An indirect validation (or rejection) of this aspect may consist in comparing the strength of these synapses after training, with results taken from physiology. Although many data in the literature support the existence of cross-modal links between the auditory and visual areas (Driver & Noesselt, 2008; Falchier, Clavagnier, Barone, & Kennedy, 2002; Ghazanfar & Schroeder, 2006; Rockland & Ojima, 2003; Schroeder & Foxe, 2005), at present it is quite difficult to infer numerical values for the synapse strength. However, a few considerations can be made. These synapses must be weak enough to avoid a phantom effect (i.e., a unisensory stimulus in one modality does not produce activation of neurons in the other unisensory area). Moreover, they must be strong enough to significantly reduce the response time in one part of the cortex (auditory or visual) in cross-

modal conditions, compared with the unisensory case. We verified that the present values of cross-modal synapses satisfy both conditions (unpublished simulations).

As evident from Fig. 9, some differences can be observed between the Bayesian estimates and those provided by the network. As discussed above, a reason is that the Bayesian estimator is significantly affected by small changes in the prior, whereas the model exhibits lesser variability. It is worth noting, however, that, with our network, we do not aspire to simulate a Bayesian estimator perfectly, but rather to simulate behavioral data, and to provide a tentative explanation on how these can be reproduced via neural mechanisms.

Although the present realization of a near-optimal Bayesian estimator works satisfactorily, as shown in Fig. 9, the posterior probability estimator can be realized also with a feedforward neural network, in which three layers of neurons (two encoding the likelihood functions for each modality, the other the prior probability) converge toward a superior layer (see also Ma et al, 2006; Pouget et al., 2013). This idea exploits the property of the logarithmic function (Pouget et al., 2013), i.e., the product of probabilities can be converted into the sum of their logarithms, and so computed in a downstream network, which realizes this sum. A problem with this feedforward schema, however, is how a chain of neurons can learn the prior probability and, moreover, where this is realized in the brain.

Orhan and Ma (Orhan & Ma, 2016) recently used feedforward networks with hidden units, trained with an error based learning rule as a model of probabilistic inference (however, this training algorithm is not biologically plausible). Ohshiro et al. (Ohshiro et al., 2011) used a model consisting of two layers of primary neurons, each sensitive to a different sensory modality, which send feedforward synapses to a layer of integrative multisensory neurons. Hence, their schema resembles our network in Figure 1b, but without cross-modal synapses. The main dissimilarity with our model is that they use divisive normalization instead of lateral inhibition to simulate suppression, and assume different weights for the feedforward synapses, in order to account for a different reliability of multisensory cues. Conversely, the different reliability of the stimuli in our model does not derive

from a difference in feedforward synapses, but from the different likelihood, encoded in the neuron input, and from the cross-modal interaction (so that the more reliable input affects the less reliable one). This was achieved with a biologically realistic rule for synapse reinforcement. In particular, it is worth noting that we used exactly the same rule for adjusting both the receptive fields and the cross-modal terms. In our opinion, a limitation of the Oshiro model is that the feedforward synapses must be modified ad hoc, to provide a different dominance of one cue on the other, whereas in our model this reweighting occurs automatically, as a consequence of input changes, as demonstrated in the simulations of Figures 10-12 (see also below for further comments on this automatic reweight).

The cross-modal synapses developed during training can account for the presence of some well-known illusory phenomena of multisensory integration. In the present work we simulated the ventriloquism effect, where a flash attracts a spatially neighboring beep. In a previous work (Cuppini et al., 2014), using a more sophisticate temporal dynamics for synapses, we demonstrated that cross-modal terms are able to account for the fusion and fission effects too, where beeps cause illusions on the number and time of perceived flashes (Andersen et al., 2004; Shams et al., 2000). Recently, we also analyzed the emergence of the well-known McGurk effect as result of a crossmodal training, and, conversely, whether a delayed maturation of cross-modal synapses can account for the poor integration of phonemes and lip movements observed in children with autism spectrum disorders (Cuppini et al., 2015).

Various authors, recently, underlined the strict relationship that exists between Bayesian inference (exploiting prior probabilities) and illusory phenomena, in a variety of multisensory tasks (Kilteni, Maselli, Kording, & Slater, 2015; O'Reilly, Jbabdi, & Behrens, 2012; Samad, Chung, & Shams, 2015). Moreover, the same Bayesian inference principles that govern perception of objects in the environment can govern perception of one's own body (Kilteni et al., 2015; Samad et al., 2015). Hence, we expect that encoding prior knowledge via maturation of cross-modal synapses may represent a quite diffuse mechanism able to unify many different illusory phenomena (both

concerning the external world and the self) and Bayesian inference into a single neural network topology.

Finally, in the last part of this study, using a further layer of multimodal neurons, we investigated the "single causal inference problem", i.e., a problem in which the observer *must* provide a single inference (in our case, a single spatial position) for conflicting stimuli. This problem can be encountered in many practical cases (let us consider, for instance, a classic multisensory structure, the superior colliculus, whose final inference is a single spatial position to drive a saccadic movement). It is worth noting that, while the problem of inferring separate positions for the two (auditory and visual) stimuli is dealt by the two unisensory networks separately (which, however, exhibit a cross-modal talk, hence acquire a kind of multisensory behavior) the problem of computing just one position to drive behavior requires a chain of multisensory neurons receiving stimuli previously elaborated by the respective unisensory regions. Hence, according to our schema in Fig. 1b, we can distinguish between two kinds of multisensoriality: a) the neurons in the auditory and visual layers do not respond to a unisensory stimulus of the other modality (indeed, we verified that, after learning, a single unisensory stimulus does not induce a phantom activity in the other modality). Here, the multisensory behavior emerges only in cross-modal conditions, in the form of a modification in the activity of one network induced by the activity in the other. b) Conversely, the neurons in the downstream layer are intrinsically multimodal, i.e., they actually respond to unisensory stimuli in each modality.

Our network suggests that both kinds of multisensoriality are useful in practice, but with different aims. The first kind of "weak" multimodal behavior may be typical of perceptual areas traditionally deemed as unisensorial; indeed, in these areas some cross-modal links with regions of other modalities have been recently demonstrated, both anatomically and functionally (Driver & Noesselt, 2008; Ghazanfar & Schroeder, 2006; Morrell, 1972; Musacchia & Schroeder, 2009; Schroeder & Foxe, 2005). We hypothesize that the aim is to implement the prior knowledge on the co-occurrence of the stimuli. The second "strong" multisensoriality is typical of downstream associative areas (such

as superior temporal sulcus, ventral intraparietal area, ventral premotor cortex (Ghazanfar & Schroeder, 2006)), where multimodal neurons have been known for a long time. Their aim is to identify a single response, common to both stimuli, to drive behavior.

As mentioned previously, an interesting result of our model is that the downstream multimodal neurons apparently reweight the individual inputs according to their reliability: the more reliable the input, the stronger its influence on multimodal neurons. This kind of behavior agrees with the predictions of Bayesian inference models (Pouget et al., 2013) and has been recently confirmed both in behavioral tasks (Alais & Burr, 2004) and in neurophysiological recordings on multimodal neurons (Fetsch et al., 2012; Morgan et al., 2008). However, the true neural mechanisms through which neurons can reweights the cues on a trial-by-trial basis was still unclear (Fetsch et al., 2012).

In our model, this result is obtained without modifying the real synaptic weights from the unimodal to the multimodal layers (which have been taken equal for both modalities), but is simply the consequence of the automatic remodulation of neuron activities in the two unimodal layers. The latter, in turn, depends on two concurrent biological mechanisms: i) the external input to the neuron, which is the inner product of the stimulus and the neuron receptive field. Although the receptive fields are trained on the basis of the experience, to encode the past reliability of the stimuli (and so the likelihood probability), hence they do not change significantly from one trial to the next, this inner product changes on a trial-by-trial basis, reflecting the characteristic of the present input (i.e., more precise or more blurred). ii) The cross-modal connections, which move the barycenter from the position of the less reliable stimulus toward the position of the more reliable one. The mechanism can be understood looking at Fig. 8. In the location of the more reliable stimulus (that is position $\vartheta^V = 100$ deg in Fig. 8) the neurons in the other modality (i.e., the auditory ones) still exhibit a not-negligible activity (or their activity is just sub-threshold); hence these neurons can be significantly excited by the cross-modal inputs coming from the other modality. The positive loop occurring between the two activities (visual and auditory) then produces even a stronger excitation at the given position, which is detected by the downstream multimodal layer as a reinforcement of the more

precise modality influence. Conversely, in the position where the less reliable stimulus is centered (that is position $\vartheta^A = 85$ deg in Fig. 8) the activity of the neurons in the other modality (i.e., the visual ones) is significantly below threshold; hence these neurons cannot be appreciably excited by the cross modal connections. This effect may be further emphasized by the competitive mechanisms, implemented via lateral Mexican-hat synapses, which further modulate the activity in the unimodal networks, to favor the stronger and depress the weaker.

Limitations and future studies

Of course, the present study exhibits some limitations, which should be clearly recognized and may become the target of subsequent investigation.

First, in our theoretical analysis, and in the following training procedure, we assumed that the receptive fields of all neurons in the same modality are equal (or almost equal, apart from minor statistical fluctuations). Accordingly, during training we used a uniform distribution to generate the unimodal inputs and to shrink the RFs. Indeed, in these conditions, each single unisensory network realizes the likelihood function of the stimulus, and a metrics extracts the MLE, which is the best estimator in case of uniform prior. The only non-uniform prior information exploited in this work is on the co-occurrence of cross-modal stimuli, encoded in the reciprocal cross-modal synapses.

Hence, in the present study we did not consider the possibility of having a not-uniform distribution of unimodal stimuli (for instance, that some positions in space are more frequently stimulated than others, or that the reliability of the stimuli depends on position). Of course, in the latter case the MLE would not be optimal even in unimodal trials, and prior information should be incorporated directly within the unisensory nets.

As demonstrated by Fisher and Peña (Fischer & Peña, 2011), a way to account for a non-uniform prior for the unisensory stimuli is that of using a population code where the prior is encoded in the density of the preferred directions (i.e., using more neurons in the more probable spatial zones). In

our model, this may be realized using a greater number of neurons and by modulating the receptive fields, which can be automatically trained to emphasize the non-uniform distribution of the inputs.

In the present model, we used just 180 neurons (resolution 1 per degree) just to reduce the computational effort. However, the number of neurons is not constrained and can be increased, thus improving both resolution and robustness (for instance, we performed some trials using 500 neurons, as in (Cazettes, Fischer, & Pena, 2016; Fischer & Peña, 2011); the results, not shown for briefness, are quite similar as the present ones).

Furthermore, in the present study we did not train lateral synapses. This is a common choice when dealing with self-organized networks, where lateral connections are usually assigned before training (let us consider, for instance, the classic Kohonen algorithm for topological maps (Kohonen, 1995)). Actually, training lateral synapses poses serious problems for network stability (our unpublished simulations), and so we decided not to train them, to reduce model complexity. However, we can consider this aspect in future model extensions. Indeed, as shown in Figure 5, by modulating lateral synapses one can modify the unisensory likelihood, thus providing more or less accuracy to the input stimulus. We argue that training lateral synapses may be important in conditions where the prior distribution of the unisensory stimuli is not uniform, to increase excitation in the more likely spatial zones, and decrease excitation in the less probable ones. This may be alternative or complementary to a spatial modulation of the receptive fields (as described in the previous paragraph).

Another limitation concerns the way we simulated cross-modal conditions. During training we always assumed that two cross-modal stimuli originate in the same position (with a possible difference of just a few degrees). This is also reflected in the value of parameter β_1 in Eq. (28), that was chosen very close to zero. It will be of great interest to train the network with different cross-modal paradigms: for instance, assuming a consistent probability also for independent cross-modal inputs (i.e., a higher value of β_1) and/or using a different ratio of cross-modal vs. unimodal stimuli (that was fixed at 1/3 per each neuron in the present study). In particular, according to the adopted

training rule, the strength of cross-modal synapses is affected by the mean value of the inputs, and so by the ratio of cross-modal vs. unimodal stimuli: the higher this ratio, the higher the cross-modal terms which, in turn, may be reflected in a stronger ventriloquism effect (our unpublished simulations).

Another kind of prior information that may be taken into account in future works is a different frequency of auditory and visual stimuli (for instance, in speech recognition tasks, the auditory input may occur more frequently and acquire a greater value than the lips movement information). We claim that the last case may be treated using different values for the feedforward synapses from the unisensory areas to the downstream multisensory area, reflecting a different prior value of the two modalities. Actually, these synapses were not trained in the present trials, but assigned with identical values for the two modalities.

Furthermore, during training we used a given spatial reliability for the auditory input ($\sigma^A = 20$ deg) compared with visual reliability ($\sigma^V = 4$ deg). This is the only difference between the two unisensory layers. The use of auditory inputs with even smaller reliability (for instance $\sigma^A = 32$ deg as used by others (Magosso et al., 2013)) may extend the ventriloquism effect even at larger distances, thus explaining some experimental data showing cross-modal influences even at 30-40 deg. Finally, also the width and strength of lateral synapses, which control the tuning curves of individual neurons and so the excitation region in each network, may have a role in modulating cross-modal synapses and so the extension of the ventriloquism effect.

A sensitivity analysis of all these aspects is certainly worthwhile in future studies, leading to the generation of testable predictions on how the manipulation of model parameters may be reflected in changes in learning and behavior.

A limitation of our model, which may be the target of future studies, concerns the description of the auditory net. In fact, while the visual net can be considered a good replication of the primary visual areas, where a spatial topological organization is already present, the auditory net possibly mimics higher stages of the auditory pathway. In fact, the primary auditory cortex is not spatially

organized, and spatial information on the auditory stimuli is extracted only later, from interaural time difference or interaural phase difference (Recanzone & Sutter, 2008; Saberi et al., 1998). A problem here may be the assumed independence of noise, since at higher stages noise may exhibit a certain correlation due to feedforward convergence. Nevertheless, neural noise is not only due to convergence from previous stages, but may also arise from intrinsic stochasticity within neurons, i.e., neural variability at the present stage. Hence, we expect that noise in the auditory net should exhibit both a correlated component and an independent one. The effect of correlated noise can be tested in future works, both in the simulation set up and in the theoretical equations (i.e., with the use of a proper autocorrelation matrix).

Finally, an important problem deserving future study concerns the judgment on the number of independent causes, a problem often referred to as "causal inference". The reader can consult some recent pivotal papers by Kording et al. (Kording et al., 2007)), Shams and Beierholm (Shams & Beierholm, 2010) and Ma and Rahmati (Ma & Rahamati, 2013) on the main theoretical aspects of this problem. In the present study we always used a single cause during training (either unimodal or cross-modal) i.e., we voluntarily decided to ignore the possibility of two simultaneous causes, to reduce problem complexity. Hence, the two unisensory networks are used to assess the auditory and visual positions separately, and the multimodal network try to fuse them into a single event to drive behavior. A more complex scenario occurs when the observer first infers whether the inputs are produced by two independent causes or by a single cause, and then, based on this preliminary decision, infers the acoustic and visual positions. This can be reformulated saying that the estimator should use the first term in the right hand member of Eq. (28) (i.e., $\beta_1 = 1$) when two separate causes were previously recognized, and the second term (i.e., $\beta_1 = 0$) when a single cause is inferred. Looking at the simulations of the ventriloquism effect, we can say that the network partially incorporates this distinction: when the audio-visual distance is less than 20 deg, the auditory stimulus is shifted toward the visual one, in an effort to treat the two stimuli as a single event. This is indeed what is occurring in the Bayesian estimator when using a single cause for the prior (i.e., $\beta_1 = 0$).

When the distance is greater than 20 deg, the audio visual interaction disappears, as in the case of two independent causes.

We think that the causal inference problem may be better addressed by the downstream multimodal network: The latter should discriminate between two causes or a single cause, for instance on the basis of the peaks in its activity, and then provide a feedback to the individual unisensory areas. This more complex network structure may be a future ambitious target, in an effort to find a network implementation for causal inference.

Conclusions

In conclusion, this study represents an effort to implement Bayesian inference via biologically inspired neural networks. Taking inspiration from the "probabilistic population coding" theory proposed by Ma, Pouget e al (Ma et al., 2006; Ma & Rahamati, 2013; Pouget et al., 2013; Pouget et al., 2003), we exploit the idea that populations of neurons can encode probability. However, we further suggest that cross-modal synapses may represent a straightforward way to incorporate prior knowledge, and that the adopted learning rule is a powerful mechanism to encode the statistical information on the environment (both in the RFs and in cross-modal connections). A further downstream multimodal network may finally compute a single position, if required to drive behavior.

Further studies should examine cases where the stimuli haven't a uniform distribution, the effect of parameter manipulations on network performance and behavior, to formulate testable predictions, and the possibility to discriminate between one and two causes, to automatically deal with the causal inference problem.

APPENDIX: PARAMETER ASSIGNMENT

The value of all model parameters (see Table 1), has been assigned from data present in the literature according to the main criteria summarized below.

External inputs – A fundamental point in the model is that the visual inputs exhibit a better spatial resolution compared with the auditory ones (Recanzone, 2009; Recanzone, Guard, & Phan, 2000). This is mimicked in the network by setting $\sigma^A > \sigma^V$ in Eq.(7). It's worth noting that these are the only differences between the two areas; all other parameters were assumed equal for the two areas, to reduce the number of assumptions.

The strength of the external visual and auditory stimuli (parameters i_{Max}^A and i_{Max}^V in Eq. (7)) was chosen so that the overall input elicits a response in the upper portion of the linear part of the sigmoidal static characteristic (i.e., a little below saturation). However, the values of i_{Max}^V were suitably increased in blurred conditions, to counteract the presence of a wider lateral inhibition (this is also the reason why $i_{Max}^A > i_{Max}^V$, see Table 1).

The standard deviation of noise was chosen as great as 1/3 of the maximum input (i.e., $\nu^A = i_{Max}^A/3$ and $\nu^V = i_{Max}^V/3$). This is the maximal noise that can be tolerated without inducing spurious activation bubbles in the network (which should be erroneously interpreted as a phantom input).

Parameters of individual neurons – The total number of neurons, *N*, was 180; their total spatial distance, *D*, was 180 deg, assuming a spatial resolution of 1 deg between one neuron and the next. The central abscissa for the sigmoidal relationship, x_0 in Eq. (33) has been assigned to have negligible neuron activity in basal condition (i.e., when the input is zero). The slope of the sigmoidal relationship, *s*, has been assigned to have a smooth transition from silence to saturation in response to external stimuli. The time constant, τ in Eq. (32) agrees with values (a few milliseconds) normally used in deterministic mean-field equations (Treves, 1993).

Parameters of synaptic connections – Parameters which establish the width and the strength of lateral synapses in both areas (i.e., λ_{ex}, λ_{in}, σ_{ex} and σ_{in} in Eq. (36)) have been assigned to simultaneously satisfy several criteria: (1) the excitation is locally greater than inhibition, to have a Mexican Hat arrangement; (2) excitation is small enough to avoid unstable phenomena (like the spread of a large activation bubble, or the

excessive amplification of local noise); (3) inhibition must be strong enough to warrant competition between two stimuli in the same area.

Parameters r_0 and σ_R in Eq. (41) were given so that the initial receptive fields of all neurons are wider than the auditory and visual inputs, but with a smaller amplitude. This causes an initial wide activity in the networks, but avoiding instability (i.e., not all neurons become simultaneously active).

The strength of cross modal synapses, w_{kj}^{SQ} in Eq. (37),was initially set at zero, assuming ineffective cross-modal link before maturation.

The synapses from the unisensory areas to the downstream multimodal area (i.e., w^{MA} and w^{MV} in Eq. (42)) were given so that a unimodal stimulus close to saturation causes half-maximal activation in the multimodal area (hence, two congruent cross-modal activations in the unisensory areas, both quite close to saturation, lead the multimodal area to saturation too).

Training parameters – The learning factor, γ in Eqs. (39) and (40) was given small enough to avoid excessive oscillations from one learning trial to the next (i.e., to avoid that the new value of the synapses is excessively affected by the last input) but large enough to ensure a good convergence within the 100 training epochs.

The ratio between cross-modal inputs vs. unimodal inputs was 1/5 (1 cross modal, 2 visual unimodal, 2 auditory unimodal per epoch). Different ratios may be tried in future works.

Parameter σ^{AV} in the expression of the prior probability (Eq. (28)), which establishes the possible distance between two cross-modal inputs coming from the same source, was taken as low as 1.5 deg. This signifies that two inputs from the same source have 90 per cent probability to differ less than 3 deg. The effect of a different parameter choice has been commented in the main text. Parameter β_1 in Eq. (28), which establishes the percentage occurrence of independent cross-modal inputs, was set very close to zero (i.e., during training we never use independent cross-modal inputs). This value cannot be set exactly equal to zero, to allow that incongruent stimuli, during the ventriloquism trials, can be separated. This problem requires a deeper analysis, using the causal inference problem assessment (see Discussion).

REFERENCES

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

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Brain Research. Cognitive Brain Research, 21*(3), 301-308.

Battaglia, P. W., Jacobs, R. A., & 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), 1391-1397.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics, 29*(6), 578-584.

Cazettes, F., Fischer, B. J., & Pena, J. L. (2016). Cue Reliability Represented in the Shape of Tuning Curves in the Owl's Sound Localization System. *The Journal of Neuroscience, 36*(7), 2101-2110.

Cuppini, C., Magosso, E., Bolognini, N., Vallar, G., & Ursino, M. (2014). A neurocomputational analysis of the sound-induced flash illusion. *Neuroimage, 92*, 248-266.

Cuppini, C., Ursino, M., Magosso, E., Ross, L. A., Foxe, J., & Molholm, S. (2015). *Hebbian Learning Mechanisms Help Explain the Maturation of Multisensory Speech Integration in Children with Autism Spectrum Disorder (ASD) and with Typical Development (TD): a Neurocomputational Analysis.* Paper presented at the Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science, Torino, Italy. http://dblp.uni-trier.de/db/conf/eapcogsci/eapcogsci2015.html#CuppiniUMRFM15

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

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

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience, 22*(13), 5749-5759.

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience, 15*(1), 146-154.

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

Froemke, R. C., & Jones, B. J. (2011). Development of auditory cortical synaptic receptive fields. *Neuroscience and Biobehavioral Reviews, 35*(10), 2105-2113.

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

Hairston, W. D., Wallace, M. T., Vaughan, J. W., Stein, B. E., Norris, J. L., & Schirillo, J. A. (2003). Visual localization ability influences cross-modal bias. *Journal of Cognitive Neuroscience, 15*(1), 20-29.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction To The Theory Of Neural Computation*: Addison-Wesley Publishing Company.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: optimal cue combination. *Journal of Vision, 4*(12), 967-992.

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

Kilteni, K., Maselli, A., Kording, K. P., & Slater, M. (2015). Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Frontiers in Human Neuroscience, 9*, 141.

Kohonen, T. (1995). *Sel-Organizing Maps*. Berlin: Springer-Verlag.

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One, 2*(9), e943.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*(11), 1432-1438.

Ma, W. J., & Rahamati, M. (2013). Towards A Neural Implementation of causal inference in cue combination. *Multisensory Research, 26*, 159-176.

Magosso, E., Cona, F., & Ursino, M. (2013). A neural network model can explain ventriloquism aftereffect and its generalization across sound frequencies. *BioMed Research International, 2013 (Article ID 475427)*, 1-17.

Magosso, E., Cuppini, C., & Ursino, M. (2012). A neural network model of ventriloquism effect and aftereffect. *PLoS One, 7*(8), e42503.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

Morgan, M. L., Deangelis, G. C., & Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron, 59*(4), 662-673.

Morrell, F. (1972). Visual system's view of acoustic space. *Nature, 238*(5358), 44-46.

Musacchia, G., & Schroeder, C. E. (2009). Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex. *Hearing Research, 258*(1-2), 72-79.

O'Reilly, J. X., Jbabdi, S., & Behrens, T. E. (2012). How can a Bayesian approach inform neuroscience? *European Journal of Neuroscience, 35*(7), 1169-1179.

Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2011). A normalization model of multisensory integration. *Nature Neuroscience, 14*(6), 775-782.

Orhan, A. E., & Ma, W. J. (2016). Efficient Probabilistic Inference in Generic Neural Networks Trained with Non-Probabilistic Feedback. *arXiv:1601.03060*

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

Pecka, M., Han, Y., Sader, E., & Mrsic-Flogel, T. D. (2014). Experience-dependent specialization of receptive field surround for selective coding of natural scenes. *Neuron, 84*(2), 457-469.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience, 16*(9), 1170-1178.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience, 26*, 381-410.

Recanzone, G. H. (2009). Interactions of auditory and visual stimuli in space and time. *Hearing Research, 258*(1-2), 89-99.

Recanzone, G. H., Guard, D. C., & Phan, M. L. (2000). Frequency and intensity response properties of single neurons in the auditory cortex of the behaving macaque monkey. *Journal of Neurophysiology, 83*(4), 2315-2331.

Recanzone, G. H., & Sutter, M. L. (2008). The biological basis of audition. *Annual Review of Psychology, 59*, 119-142.

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology, 50*(1-2), 19-26.

Saberi, K., Takahashi, Y., Konishi, M., Albeck, Y., Arthur, B. J., & Farahbod, H. (1998). Effects of interaural decorrelation on neural and behavioral detection of spatial cues. *Neuron, 21*(4), 789-798.

Samad, M., Chung, A. J., & Shams, L. (2015). Perception of body ownership is driven by Bayesian sensory inference. *PLoS One, 10*(2), e0117178.

Sato, Y., Toyoizumi, T., & Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Computation, 19*(12), 3335-3355.

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology, 15*(4), 454-458.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences, 14*(9), 425-432.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature, 408*(6814), 788.

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

Treves, A. (1993). Mean-field analysis of neuronal spike dynamics. *Network, 4*, 259-284.

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

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research, 158*(2), 252-258.

Wallace, M. T., & Stein, B. E. (1997). Development of multisensory neurons and multisensory integration in cat superior colliculus. *Journal of Neuroscience, 17*(7), 2429-2444.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *Journal of Vision, 8*(3), 1-11.

Zhou, Y. D., & Fuster, J. M. (2000). Visuo-tactile cross-modal associations in cortical somatosensory cells. *Proceedings of the National Academy of Sciences of the United States of America, 97*(17), 9777-9782.

Zhou, Y. D., & Fuster, J. M. (2004). Somatosensory cell response to an auditory cue in a haptic memory task. *Behavioural Brain Research, 153*(2), 573-578.