# How the Prior Information Shapes Neural Networks for Optimal Multisensory Integration

He Wang[1(✉)], Wen-Hao Zhang[1,2,3], K.Y. Michael Wong[1], and Si Wu[2]

[1] Department of Physics, Hong Kong University of Science and Technology, Hong Kong, China
{hwangaa,phkywong}@ust.hk, wenhaoz1@andrew.cmu.edu

[2] State Key Laboratory of Cognitive Neuroscience and Learning, and McGovern Institute for Brain Research, Beijing Normal University, Beijing, China
wusi@bnu.edu.cn

[3] Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, USA

**Abstract.** Extensive studies suggest that the brain integrates multisensory signals in a Bayesian optimal way. In this work, we consider how the couplings in a neural network model are shaped by the prior information when it performs optimal multisensory integration and encodes the whole profile of the posterior. To process stimuli of two modalities, a biologically plausible neural network model consists of two modules, one for each modality, and crosstalks between the two modules are carried out through feedforward cross-links and reciprocal connections. We found that the reciprocal couplings are crucial to optimal multisensory integration in that their pattern is shaped by the correlation in the joint prior distribution of sensory stimuli. Our results show that a decentralized architecture based on reciprocal connections is able to accommodate complex correlation structures across modalities and utilize this prior information in optimal multisensory integration.

**Keywords:** Recurrent neural networks · Multisensory processing · Bayesian inference

## 1 Introduction

Extracting information reliably from ambiguous environments is crucial for the survivorship of organisms. The brain solves this problem by exploiting multiple sensory modalities to gather, from different aspects, as much information as possible about the same entity of interest. It has been reported in a large number of psychophysical and neurobiological studies that the brain can integrate sensory cues in an optimal way, as predicted by Bayesian inference [1–3].

Despite the accumulated behavior evidence, exactly how the brain implements optimal multisensory integration remains largely unknown. In the present


F. Cong et al. (Eds.): ISNN 2017, Part II, LNCS 10262, pp. 128–136, 2017.
DOI: 10.1007/978-3-319-59081-3_16

study, we adopt a theoretical approach to address this challenging issue. We formulate multisensory integration as a mathematical problem of optimizing network structure under the constraint that for a given prior distribution of stimuli, the network's output matches the posterior distribution. This is equivalent to requiring that the network realizes Bayesian inference when the sensory cues are sampled form their prior over many trials. We introduce different prior distributions of the multisensory stimuli and investigate how network structures depend on the choice of priors. We look for evidence to see where information about the prior and that about the likelihood are represented in the network consisting of recurrent and reciprocal connections, cross-links and direct links. These results generate predictions about the structural pre-requisites for multisensory integration. They can be tested in future experiments and shed light on our understanding of how the brain can achieve multisensory integration optimally.

## 2 Optimal Multisensory Inference with a Composite Prior

Utilizing prior information is important for multisensory information processing. A variety of studies have suggested that the prior distributions are taken into account when animals make perceptual decisions [4–6]. Specifically, multisensory processing relies on the experience about correlations among sensory cues, which usually benefits us in forming a unified and coherent perception of the external world [7], yet sometimes evokes interesting illusions [8,9].

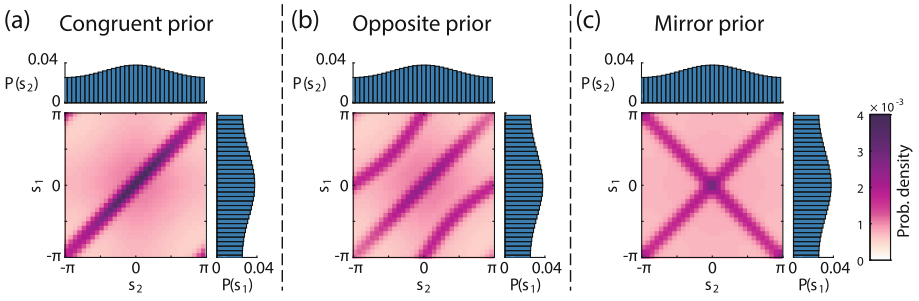


**Fig. 1.** Three types of the prior. (a) The joint prior distribution constructed from the congruent copula $c_1$. The marginal priors, which are the same for $s_1$ and $s_2$, are plotted to the sides of (a). (b) The joint prior distribution constructed from the opposite copula $c_2$. (c) The joint prior distribution constructed from the mirror copula $c_3$. The color code for (a)–(c) are the same and shown to the right of (c). Parameters: for all three cases, $\kappa_s = 0.2, \kappa_p = 11.6, p_c = 0.246$. For the opposite prior in (b) and the mirror prior in (c), $\alpha = 0.5$. (Color figure online)

Different specific forms of the prior distribution have been brought up to characterize different perceptual tasks (see [10] for a review). In general, the

joint prior should be composed of an independent part and a correlated part. Suppose $s_1$ and $s_2$ are two sensory stimuli in different modalities, whose marginal prior densities are $p(s_1)$ and $p(s_2)$, respectively. The joint prior can be described as $p(s_1, s_2) = (1 - p_c)p(s_1)p(s_2) + p_c q(s_1, s_2)$. Here, $q(s_1, s_2)$ is a correlated distribution and $p_c \in [0, 1]$ describes how often $s_1$ and $s_2$ are originated from that distribution. Ideally, $q(s_1, s_2)$ should only affect the correlation between the two underlying stimuli without changing their marginal distributions. This requirement can be satisfied by using a copula, which is a multivariate probability distribution, whose marginal distribution of each variable is uniform [11]. Consider a two-dimensional copula $c(\xi_1, \xi_2)$, satisfying the property that its marginals over $\xi_1$ or $\xi_2$ are equal to 1. According to the Sklar's theorem [12], $q(s_1, s_2)$ can be constructed as $q(s_1, s_2) = c(F(s_1), F(s_2))p(s_1)p(s_2)$, where $F(s_i)$ is the cumulative distribution function of $p(s_i)$. It can be verified that the marginal distributions of $q(s_1, s_2)$ are exactly $p(s_1)$ and $p(s_2)$.

In the present work, we consider stimuli such as heading direction residing on a circular space $[-\pi, \pi)$. We use the von Mises distribution as the marginal prior distribution, $p(s_i) \propto e^{\kappa_s \cos s_i}$, $i = 1, 2$, where $\kappa_s$ is the concentration parameter, and $\propto$ indicates proportionality. For simplicity, we consider the case that the marginal priors are the same for the two modalities, and centered at the origin. In order to observe the dependence of network structure on the prior, three forms of copulas are chosen due to their distinctive profiles:

1. Congruent copula $[c_1(\xi_1, \xi_2) \propto e^{\kappa_p \cos 2\pi(\xi_1 - \xi_2)}]$, which is derived from the von Mises distribution. Similar forms of such prior are widely applied in describing a pair of correlated sensory cues when they are originated from a common cause [8,13]. Larger $\kappa_p$ indicates higher correlation between the stimuli in the two modalities.
2. Opposite copula $[c_2(\xi_1, \xi_2) = \alpha c_1(\xi_1, \xi_2) + (1-\alpha)c_1(\xi_1, \xi_2 + 1/2)]$. The second term in $c_2$ indicated that $s_1$ and $s_2$ may come in opposition directions.
3. Mirror copula $[c_3(\xi_1, \xi_2) = \alpha c_1(\xi_1, \xi_2) + (1 - \alpha)c_1(\xi_1, -\xi_2)]$. The second term in $c_3$ indicates that $s_1$ and $s_2$ might be the mirror image of each other.

Examples of three different kinds of joint prior distributions $p(s_1, s_2)$ are shown in Fig. 1(a)–(c).

The two stimuli $s_1$ and $s_2$ give rise to sensory observations $x_1$ and $x_2$, respectively. The sensory observations are corrupted by independent noises in different sensory pathways. We use the von Mises distribution to represent the likelihood functions, $p(x_i|s_i) \propto e^{\kappa_i \cos(x_i - s_i)}, i = 1, 2$, where $\kappa_i$ is the concentration parameter, which can be understood as the reliability of the sensory input in the corresponding modality.

These uni-sensory observations are supposed to be fed into higher level multisensory regions, where optimal multisensory estimates $\hat{s}_1$ and $\hat{s}_2$ are made. According to the Bayes' theorem, the marginal posterior distribution is given by

$$p(s_1|x_1, x_2) \propto \int p(x_1|s_1)p(x_2|s_2)p(s_1, s_2)\,\mathrm{d}s_2. \tag{1}$$

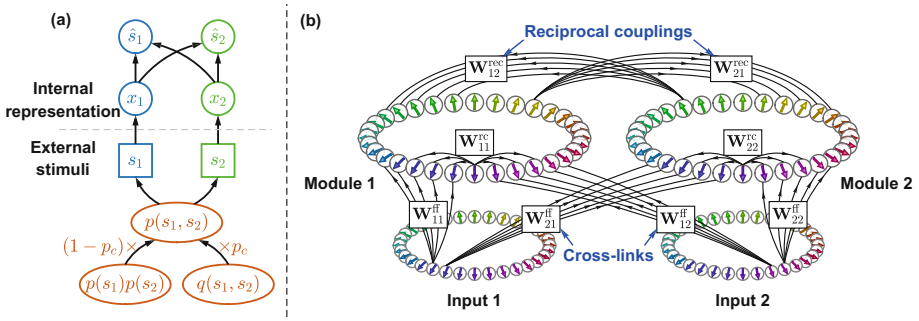


**Fig. 2.** The multimodal Bayesian inference problem and the recurrent neural network model. (a) A graphical illustration of the Bayesian inference problem. (b) Each small circle portraits one neuron, with the attached arrow indicating the neuron's preferred stimulus. Besides being recurrently connected to each of themselves, the two modules of the network model interact with each other through feedforward cross-links ($\mathbf{W}_{12}^{\mathrm{ff}}$ and $\mathbf{W}_{21}^{\mathrm{ff}}$) and reciprocal couplings ($\mathbf{W}_{12}^{\mathrm{rec}}$ and $\mathbf{W}_{21}^{\mathrm{rec}}$). The inputs of the two modules represent uni-sensory observations, corresponding to $x_1$ and $x_2$ in (a). The outputs are multisensory representations, corresponding to $\hat{s}_1$ and $\hat{s}_2$ in (a).

Usually the expected value of $s_1$ from the posterior distribution is chosen as a Bayesian optimal estimate $\hat{s}_i$ for the underlying stimulus, which minimizes a mean squared error cost function [10]. For circular random variables considered in this work, the Bayesian estimates for the stimuli in two modalities are given by, $\hat{s}_i = \arg\left[\int p(s_i = \phi|x_1, x_2)e^{j\phi}\,\mathrm{d}\phi\right]$, for $i = 1, 2$, where $j \equiv \sqrt{-1}$ is the imaginary unit. This Bayesian inference framework for multisensory processing is shown schematically in Fig. 2(a).

## 3 A Bi-modular Recurrent Neural Network Model

Bi-modular recurrent neural network models have been applied in many studies on the multisensory integration to explain experimental findings and provide insights into the functional roles of connections between brain areas [14,15]. We will explore the capability of such bi-modular recurrent network models in encoding an arbitrary prior distribution and optimally integrating multisensory information based on that prior. Consider a bi-modular recurrent neural network model with its dynamical equation [16],

$$\tau_s \frac{\partial}{\partial t}\begin{bmatrix}\boldsymbol{u_1}\\ \boldsymbol{u_2}\end{bmatrix} = -\begin{bmatrix}\boldsymbol{u_1}\\ \boldsymbol{u_2}\end{bmatrix} + \begin{bmatrix}\mathbf{W}_{11}^{\mathrm{rec}} & \mathbf{W}_{12}^{\mathrm{rec}}\\ \mathbf{W}_{21}^{\mathrm{rec}} & \mathbf{W}_{22}^{\mathrm{rec}}\end{bmatrix}\begin{bmatrix}\boldsymbol{r_1}\\ \boldsymbol{r_2}\end{bmatrix} + \begin{bmatrix}\mathbf{W}_{11}^{\mathrm{ff}} & \mathbf{W}_{12}^{\mathrm{ff}}\\ \mathbf{W}_{21}^{\mathrm{ff}} & \mathbf{W}_{22}^{\mathrm{ff}}\end{bmatrix}\begin{bmatrix}\boldsymbol{I_1}\\ \boldsymbol{I_2}\end{bmatrix}. \quad (2)$$

Here, $\boldsymbol{u_i}$ is a $N$-element vector, whose $m^{\mathrm{th}}$ element $u_{i,m}$ is the synaptic input of the $m^{\mathrm{th}}$ neuron in module $i$. The $m^{\mathrm{th}}$ element of the vector $\boldsymbol{r_i}$, $r_{i,m}$, is the firing rates of the $m^{\mathrm{th}}$ neuron in module $i$. The firing rate is related to the synaptic input $\boldsymbol{u_i}$ through an activation function $\boldsymbol{r_i} = f(\boldsymbol{u_i})$. $\boldsymbol{I_i}$ is the external inputs

applied on module $i$. $\mathbf{W}_{ij}^{\mathrm{ff}}$ is the feedforward weight matrix from module $j$ to module $i$. $\mathbf{W}_{ij}^{\mathrm{rec}}$ is the recurrent weight matrix from module $j$ to module $i$. The preferred stimuli of neurons in each module are supposed to be evenly distributed over a circle, $\phi_m = (2\pi m)/N - \pi$. In the following results, both modules consist of $N = 32$ neurons. The architecture of this network model is illustrated in Fig. 2(b).

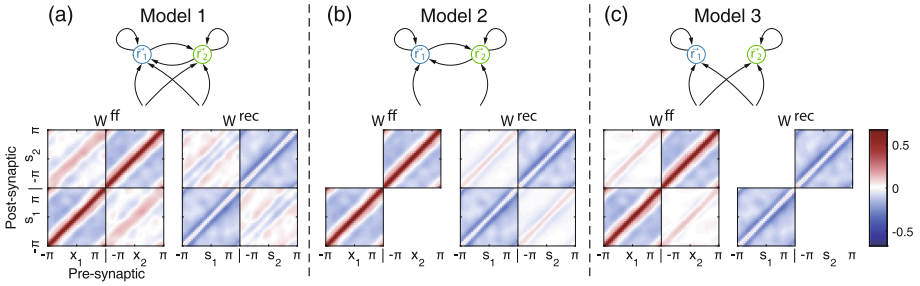


**Fig. 3.** Comparison of model structures. Network architecture and connection weights for: (a) model 1, a fully connected model; (b) model 2, where feedforward cross-links are cut; and (c) model 3, where reciprocal couplings are cut. All three structures are optimized for congruent copula $c_1$. Parameters: $\kappa_s = 0.2, \kappa_p = 11.6, p_c = 0.246$, $\widetilde{\kappa}_1 = \widetilde{\kappa}_2 = 10.7$.

The inputs of the neural network are the neural population representation of the uni-sensory observations of the external stimuli. Due to the uncertain nature of the external world and noisy neuronal firings, the neural population representation is constantly fluctuating, hypothetically sampling the likelihood function. In a similar way, the outputs of the multisensory neural population should sample the posterior distribution. If we consider a time scale that is much longer than this sampling process, the temporal average of the neuronal inputs and outputs should resemble the likelihood function and the network's estimate of the posterior distribution, respectively. Therefore, the external input vector $\boldsymbol{I_i}$ is set to be the same as the likelihood functions $p(x_i|s_i = \phi_m)$, and the stationary firing rates $\boldsymbol{r_i}^*$ will eventually approach the network's estimate of the marginal posterior $\boldsymbol{p_i}$, whose $m^{\mathrm{th}}$ element is $p(s_i = \phi_m|x_1, x_2)$, during the network optimization described below.

We use a divisive normalization function as the activation function. Recently, due to its success in accounting for important features of multisensory integration, such as the principle of inverse effectiveness and the spatial principle [17], the divisive normalization model was proposed to be a canonical integration operation [18,19]. Here, we follow the form of divisive normalization in a continuous attractor neural network model [20], $r_{i,m} = [u_{i,m}]_+^2 / \{1 + k_{\mathrm{I}} \sum_n [u_{i,n}]_+^2\}$, for $i = 1, 2$, and $m, n = 1, 2, \ldots, N$. Here, $[x]_+ \equiv \max(x, 0)$, and $k_{\mathrm{I}}$ is the strength of global inhibition. The performance of the network is the best with divisive normalization function, compared with sigmoid or piece-wise linear functions (data

not shown here). In the present work, we fix $k_{\mathrm{I}} = 0.1$, while small changes in $k_{\mathrm{I}}$ does not affect the results very much.

### 3.1 Optimize the Connection Weights Through Stochastic Gradient Descent

We optimize the connection weights in order to minimize the mean squared error $L$ between the stationary network activity $\boldsymbol{r}^*$ and the marginal posterior distribution $\boldsymbol{p}_i$, $L \equiv \langle \sum_{i=1,2} \|\boldsymbol{r}_i^* - \boldsymbol{p}_i\|^2 \rangle_{p(x_1,x_2)}$. Usually a recurrent neural network is trained using back-propagation through time [21,22]. Since only the steady state is relevant in this work, we use a simple stochastic gradient descent algorithm to optimize the steady state. Samples of training inputs $(\boldsymbol{I}_1, \boldsymbol{I}_2)$ and training outputs $(\boldsymbol{p}_1, \boldsymbol{p}_2)$ are generated in the following way. Given the prior distribution $p(s_1, s_2)$ and the mean reliabilities of sensory inputs $\widetilde{\kappa}_1$ and $\widetilde{\kappa}_2$, we first draw the true value of external stimuli $s_1$ and $s_2$ from the prior distribution and draw the reliabilities for each sensory input $\kappa_1$ and $\kappa_2$ independently from log-normal distributions $\ln \mathcal{N}(\widetilde{\kappa}_i, \sigma_\kappa^2)$. In this work, we always set $\sigma_\kappa$ to be 0.5. Secondly, draw the sensory input $x_i$ from the von Mises distribution $p(x_i|s_i) \propto e^{\kappa_i \cos(s_i - x_i)}$. Then, the training inputs and the training outputs can be calculated according to the Bayes' theorem in Eq. (1).

## 4 Results

### 4.1 Model Comparison

Crosstalks between different sensory areas may happen at different levels. In general, we consider two types of communication across modalities: the feedforward cross-links ($\mathbf{W}_{ij}^{\mathrm{ff}}$ for $i \neq j$), and the reciprocal couplings ($\mathbf{W}_{ij}^{\mathrm{rec}}$ for $i \neq j$). By forcing either of them to be zero, we tested three different model structures (Fig. 3). Model 1 is the fully connected model (Fig. 3(a)). In model 2, the interaction between the two modules are limited to the reciprocal connections, with the feedforward cross-links forced to be zero (Fig. 3(b)). In model 3, the reciprocal connections are set to be zero (Fig. 3(c)). We also tested a purely feedforward network structure, model 4, to see if recurrent connections are essential for optimal multisensory integration. We found that model 1, 2 and 3 are almost indistinguishable in their performances, while the purely feedfoward structure, model 4, is obviously worse than the others (data not shown here). Examples of the connection weights for model 1, 2 and 3 are shown in the lower parts of Fig. 3(a)–(c). In the following part of this work, we will focus on model 2, while general results are similar for model 1 and model 3.

### 4.2 Coupling Weights for Different Priors

To reveal the impact of the prior information on the recurrent neural network model, we compare the coupling weights of networks optimized with different

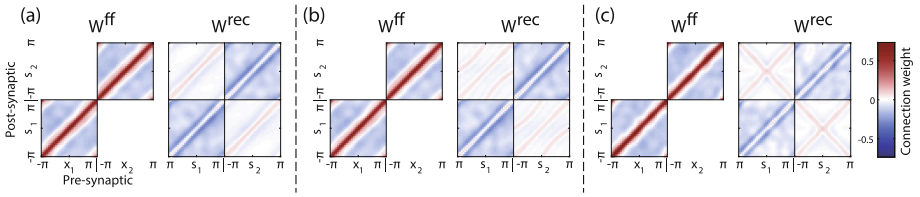


**Fig. 4.** The optimized coupling weights for three types of the prior. (a) The connection weights of model 2 trained with the congruent prior in Fig. 1(a). (b) The connection weights of model 2 trained with the opposite prior in Fig. 1(b). (c) The connection weights of model 2 trained with the mirror prior in Fig. 1(c). Parameters: for all three cases, $\widetilde{\kappa}_1 = \widetilde{\kappa}_2 = 10.7$.

prior distributions. Three examples are shown in Fig. 4. The prior distributions share the same marginal distributions, and are then constructed using three different types of copulas: the congruent copula $c_1$ (Fig. 1(a)), the opposite copula $c_2$ (Fig. 1(b)) and the mirror copula $c_3$ (Fig. 1(c)). Coupling weights of the networks trained with the three priors are shown in Fig. 4(a)–(c). The same-side connection weights ($\mathbf{W}_{11}^{\mathrm{ff}}, \mathbf{W}_{22}^{\mathrm{ff}}, \mathbf{W}_{11}^{\mathrm{rec}}$ and $\mathbf{W}_{22}^{\mathrm{rec}}$) are nearly identical for the three cases. However, the reciprocal couplings ($\mathbf{W}_{12}^{\mathrm{rec}}$ and $\mathbf{W}_{12}^{\mathrm{rec}}$) exhibit patterns resembling the corresponding prior distribution. This result strongly suggests that the reciprocal connections, as a bridge between different sensory modules, are able to encode the information of the joint prior distribution, taking the correlation structure between sensory stimuli into account when performing multisensory integration.

## 5 Conclusion

We have developed a framework to link the network structure of the multisensory processing brain region to the statistical structure of Bayesian inference. We found that a recurrent network structure appears to be necessary for implementing optimal multisensory integration. Furthermore, we have studied the dependence of the network structure for multisensory information processing on the choice of the priors and likelihoods. We found clear evidence that information about the prior is encoded in the indirect couplings (reciprocal connections and cross-links). This can be seen from the correspondence between the profiles of the indirect couplings and the correlation pattern in the joint prior of the stimuli. In the present models, the priors can be encoded in either cross-links or reciprocal connections or both. In the future, we can consider how biological constraints can narrow down these possibilities for realistic architecture exploited by the neural system.

Multisensory integration is not limited to biological systems. In other artificial intelligence applications, such as computer vision and robotics, integrating signals optimally from multiple sensors is also a fundamental technique. The optimal structure we found has implications to the decentralized architecture for

multisensory information processing. It demonstrates that composite prior distributions can be encoded in a decentralized fashion in the reciprocal connections.

**Acknowledgments.** This work is supported by the Research Grants Council of Hong Kong (N_HKUST606/12, 605813 and 16322616) and National Basic Research Program of China (2014CB846101) and the Natural Science Foundation of China (31261160495).

## References

1. Alais, D., Burr, D.: No direction-specific bimodal facilitation for audiovisual motion detection. Cogn. Brain Res. **19**(2), 185–194 (2004)
2. Ernst, M.O., Banks, M.S.: Humans integrate visual and haptic information in a statistically optimal fashion. Nature **415**(6870), 429–433 (2002)
3. Gu, Y., Angelaki, D.E., DeAngelis, G.C.: Neural correlates of multisensory cue integration in macaque MSTd. Nat. Neurosci. **11**(10), 1201–1210 (2008)
4. Girshick, A.R., Landy, M.S., Simoncelli, E.P.: Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nat. Neurosci. **14**(7), 926–932 (2011)
5. Fischer, B.J., Peña, J.L.: Owl's behavior and neural representation predicted by Bayesian inference. Nat. Neurosci. **14**(8), 1061–1066 (2011)
6. Körding, K.P., Wolpert, D.M.: Bayesian integration in sensorimotor learning. Nature **427**(6971), 244–247 (2004)
7. Ghazanfar, A.A., Schroeder, C.E.: Is neocortex essentially multisensory? Trends Cogn. Sci. **10**(6), 278–285 (2006)
8. Sato, Y., Toyoizumi, T., Aihara, K.: Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. Neural Comput. **19**(12), 3335–3355 (2007)
9. Shams, L., Ma, W.J., Beierholm, U.: Sound-induced flash illusion as an optimal percept. NeuroReport **16**(17), 1923–1927 (2005)
10. Shams, L., Beierholm, U.R.: Causal inference in perception. Trends Cogn. Sci. **14**(9), 425–432 (2010)
11. Durante, F., Sempi, C.: Principles of Copula Theory. Taylor & Francis, Boca Raton (2015)
12. Sklar, A.: Random variables, joint distribution functions, and copulas. Kybernetika **9**(6), 449–460 (1973)
13. Körding, K.P., Beierholm, U., Ma, W.J., Quartz, S., Tenenbaum, J.B., Shams, L.: Causal inference in multisensory perception. PLoS One **2**(9), e943 (2007)
14. Zhang, W.H., Chen, A., Rasch, M.J., Wu, S.: Decentralized multisensory information integration in neural systems. J. Neurosci. **36**(2), 532–547 (2016)
15. Magosso, E., Cuppini, C., Ursino, M.: A neural network model of ventriloquism effect and aftereffect. PLoS One **7**(8), e42503 (2012)
16. Amari, S.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biol. Cybern. **27**(2), 77–87 (1977)
17. Ohshiro, T., Angelaki, D.E., DeAngelis, G.C.: A normalization model of multisensory integration. Nat. Neurosci. **14**(6), 775–782 (2011)
18. Carandini, M., Heeger, D.J.: Normalization as a canonical neural computation. Nat. Rev. Neurosci. **13**(1), 51–62 (2012)

19. Van Atteveldt, N., Murray, M.M., Thut, G., Schroeder, C.E.: Multisensory integration: flexible use of general operations. Neuron **81**(6), 1240–1253 (2014)
20. Fung, C.C.A., Wong, K.Y.M., Wu, S.: A moving bump in a continuous manifold: a comprehensive study of the tracking dynamics of continuous attractor neural networks. Neural Comput. **22**(3), 752–792 (2010)
21. Williams, R.J., Zipser, D.: Gradient-based learning algorithms for recurrent networks and their computational complexity. In: Back-Propagation: Theory, Architectures and Applications, pp. 433–486 (1995)
22. Seung, H.S.: Learning continuous attractors in recurrent networks. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems 10, pp. 654–660. MIT Press, Cambridge (1998)