A dynamic neural field model of multimodal merging: application to the ventriloquist effect

Simon Forest, Jean-Charles Quinton, Mathieu Lefort

A dynamic neural field model of multimodal merging: application to the ventriloquist effect

Simon Forest[1, 2]**, Jean-Charles Quinton**[1]**, Mathieu Lefort**[2]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, UMR 5224, F-38000, Grenoble, France

[2]Université de Lyon, Université Claude Bernard Lyon 1, CNRS, LIRIS, UMR 5205, F-69621, Villeurbanne, France

Keywords: Multimodal merging; dynamic neural fields; superior colliculus; selective attention

Abstract

Multimodal merging encompasses the ability to localize stimuli based on imprecise information sampled through individual senses such as sight and hearing. Merging decisions are standardly described using Bayesian models that fit behaviors over many trials, encapsulated in a probability distribution. We introduce a novel computational

model based on Dynamic Neural Fields able to simulate decision dynamics and generate localization decisions, trial by trial, adapting to varying degrees of discrepancy between audio and visual stimulations. Neural fields are commonly used to model neural processes at a mesoscopic scale, for instance neurophysiological activity in the superior colliculus. Our model is fit to human psychophysical data of the ventriloquist effect, additionally testing the influence of retinotopic projection onto the superior colliculus, and also providing a quantitative performance comparison to the Bayesian reference model. While models performs equally on average, a qualitative analysis of free parameters in our model allows insights into the dynamics of the decision and the individual variations in perception caused by noise. We finally show that the increase in the number of free parameters does not result in overfitting, and that the parameter space may either be reduced to fit specific criteria or exploited to perform well on more demanding tasks in the future. Indeed, beyond decision or localization tasks, our model opens the door to the simulation of behavioral dynamics as well as saccade generation driven by multimodal stimulation.

1 Introduction

Humans have versatile and diverse ways of perceiving the world around them. Senses provide a dense and continuous flow of data, yet our ability to process information is limited, so we need to select a subset of all available data in order to engage in adequate interactions with the environment. Performing relevant selection involves processes pertaining to (selective) attention.

Focusing on visual attention, human vision is constrained by the heterogeneous disposition of sensors on the retina, with a denser distribution near the center of the visual field (called fovea). As a consequence, humans will tend to gaze at objects of interest, in order to see them better. One outcome of this kind of overt attention is that it may trigger visual saccades towards objects located in the periphery of the retinotopic space. Because of its weaker resolution, saccades are less precise and more likely to be disturbed by artifacts.

That issue can be circumvented with the use of additional information from other modalities (Calvert et al., 2004). For example, a sound congruent to a visual stimulus may guide saccades to this particular target (Frens et al., 1995; Kapoula and Pain, 2020). Generally speaking, it is common to merge sensory data coming from multiple modalities. They might enhance each other (Meredith and Stein, 1986), complement one another (Newell et al., 2001), or even compete together to form an interpolation of different sensory inputs (McGurk and MacDonald, 1976; Alais and Burr, 2004). These mechanisms depend on the relative reliability of the modalities, with factors including stimulus noisiness (Ernst and Banks, 2002), sensor precision (Witten and Knudsen, 2005), and possible top-down interference (such as selective attention; Driver and Spence, 2004). Studies on this topic vary from macroscopic (at a behavioral level) to microscopic (neurological) scale, but it is common for such insights to be shared across these two domains (Calvert et al., 2004; Alais et al., 2010).

Our aim is to build a computational model of multisensory integration that can be embedded in attention processes. We will focus on audiovisual merging especially.

1.1 Biological inspiration

One source of inspiration for our computational model is the superior colliculus (SC). It has been reported to integrate cues from multiple modalities, including visual, auditory and somatosensory (Wallace and Stein, 1996; Calvert et al., 2004), which makes it a relevant neural structure to be used as a reference for our model. It is also involved in the generation of motor commands such as saccades (Gandhi and Katnani, 2011). However, please note that our purpose is not to build a biologically-accurate simulation of the SC, but rather get inspiration from the brain workflow, for which mesoscopic scale models of multisensory integration are available. Such scale should allow us to remain neurally plausible, as we later turn our attention to macroscopic observations and directly model behavioral data.

In previous works, the SC has already been used as a target of computational models of visual (Taouali et al., 2015) and multimodal (Casey et al., 2012; Bauer et al., 2015) perception. A common representation of a visual map in the SC is given by Ottes et al. (1986), where the retinotopic space is mapped to the collicular space using a logpolar transformation. That transformation has been suggested to lie at the core of complex mechanisms of visual attention (Taouali et al., 2015), including saccades (Manfredi et al., 2009).

1.2 Computational model

Computational neural models of the SC exist in various forms, both for multisensory integration (Bauer, 2015, chapter 3) and for saccade generation (Girard and Berthoz, 2005). One frequently used theoretical paradigm that encompasses both aspects, and

that has been predominant when it comes to visual processing in the SC, is that of dynamic neural fields (DNF) (Marino et al., 2012; Taouali et al., 2015; Quinton and Goffart, 2018). It originated as a mathematical model of neural dynamics (Amari, 1977), and has been used to model neural activity in sensorimotor maps at a mesoscopic scale (Schöner et al., 2015). DNF describe the evolution of mean field potential over a continuous domain (usually simply called a map), for instance the average membrane potential of neurons in the intermediate layers of the SC (Trappenberg et al., 2001; Wilimzig et al., 2006). While interactions at the microscopic scale may be of interest for many neural processes, focusing on neural fields at a mesoscopic scale helps to bridge the gap with behavioral data. This is not only useful to better understand adaptive functions found in living systems (Schöner et al., 2015), but also makes it possible to build artificial systems able to reproduce them (including decision-making and attentional capabilities based on noisy sensor data) and to implement them on robots (with topologies of sensors that differ from humans). Depending on their parametrization, DNF may for instance achieve selection or interpolation between several conflicting signals (Taouali et al., 2015), robust selective attention in presence of noise and distractors (Fix et al., 2011), working or long term memory of stimuli (Sandamirskaya, 2014).

DNF have long been used as models of visual attention (Fix et al., 2011) and (visuo)motor control (Wilimzig et al., 2006; Sandamirskaya, 2014; Quinton and Goffart, 2018). However, the literature is scarcer when it comes to using DNF for multimodal fusion (Schauer and Gross, 2004; Ménard and Frezza-Buet, 2005; Lefort et al., 2013). Schauer and Gross (2004) have shown promising results with a bio-inspired DNF-based model of audiovisual integration. With very little preprocessing, they achieved a sig-

nificant response enhancement when exposed to congruent visual and auditory signals, although they did not draw connections to known psychophysical phenomena.

1.3 Psychophysical reference

In this paper, we will show that applications of DNF go as far as to account for well known psychophysical effects of multisensory integration. As an illustration of such possibilities, we will use the ventriloquist effect (Alais and Burr, 2004), which is an example of audiovisual merging. From a human participant viewpoint exposed to spatially incongruent visual and audio stimuli, the position of a stimulus is shifted towards the other, depending on which modality has the highest relative precision. The effect takes its name from ventriloquist shows, where spectators have the illusion that a puppet is speaking, while the sound is actually produced by the ventriloquist holding it.

We will draw on psychophysical data reported in Alais and Burr (2004), because their experimental paradigm and protocol can easily be replicated in silico, they provide extensive results in all conditions, and their paper is a seminal contribution to the field, with results that have not yet been challenged. One might notice that in their experiment, only the visual precision varied. However, by manipulating the relative precision between the two modalities, they showed the multiple sides of the ventriloquist effect (either vision capturing audition, the reverse, or an interpolation between both). We want our computational model to exhibit the diversity of behaviors linked to multimodal fusion, so this experiment constitutes an interesting showcase.

In addition to empirical data, we will also compare the performance of our model to optimal Bayesian integration, usually considered as the golden standard among formal

and computational models of multisensory integration (Ernst and Bulthoff, 2004; Rohde et al., 2016). However, note that we do not strive for a perfect quantitative fit of our model to the data. Indeed, even though optimization and sensitivity analysis will be combined to assess the ability of our model to robustly converge with behavioral data, our model enables a broad set of perspectives by building on past DNF models, of which the ventriloquist effect is only one illustration.

The remainder of this article is structured as follows. In section 2, we describe our computational model and its evaluation criteria in the context of the ventriloquist effect. We present the results in section 3, and discuss further on the capabilities of our model in section 4.

2 Method

2.1 General model

From a neurophysiological standpoint, the (deep) SC has been reported to receive projections from different modalities on a series of multimodal neural maps (King, 2004). In this section, we first described how these maps are modeled, before turning to the projections they receive. An overview of our general model is given in figure 1.

2.1.1 Dynamic neural fields

Our model of a SC map activity is based on dynamic field theory (Schöner et al., 2015). DNF model the evolution of the neural activity over time on each point of a topological space $\mathbf{X}$ that maps a portion of the brain. The mean field potential U at position $\mathbf{x} \in \mathbf{X}$

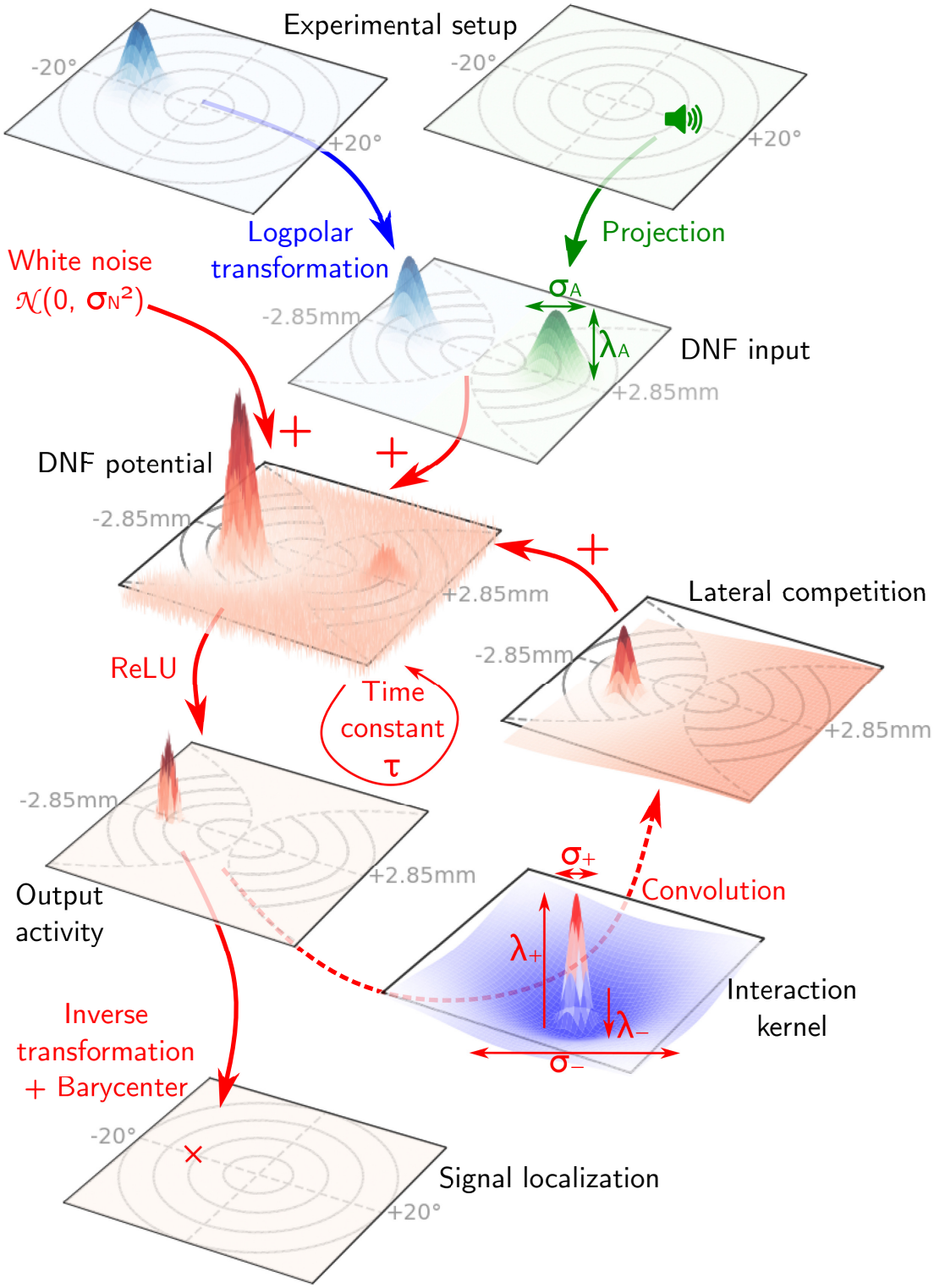

Figure 1: Visual representation of the audiovisual merging DNF model. Each rectangle represents a map, either in retinal space (shown with concentric circles) or SC (hourglass shape, obtained by performing a logpolar transformation on the visual map). The blue arrow and text relate to visual preprocessing, green to auditory. Steps and parameters from the model, other than preprocessing, are shown in red.

and time t is described by the following stochastic integro-differential equation:

$$\tau \frac{\partial U}{\partial t}(\mathbf{x}, t) = -U(\mathbf{x}, t) + I(\mathbf{x}, t) + \int_{\mathbf{x}' \in \mathbf{X}} W(\|\mathbf{x} - \mathbf{x}'\|)\, f(U(\mathbf{x}', t))\, \mathrm{d}\mathbf{x}' + \varepsilon \qquad (1)$$

where τ is the time constant which determines the response timescale of the entire field, I is the input stimulation over the field and f is a non-linear activation function; as often chosen to simplify numerical simulations, we will use a ReLU function to approximate the mean firing rate of neurons (Quinton and Goffart, 2018). The last term ε represents noise which, like the entire dynamic neural fields, can be interpreted at either a neurological (a sum of numerous variations of activity induced by external neurons) or psychophysical level (e.g. perceptual noise) (Schöner et al., 2015, box 1.4, p. 36). Due to the variations being summed over a large population of neurons, white noise is often used, and ε is therefore sampled from a normal distribution $\mathcal{N}(0, \sigma_N)$.

Finally, the kernel approximating lateral interactions within the continuous population of neurons is defined by:

$$W(\Delta\mathbf{x}) = \lambda_+ \exp\left(-\frac{\Delta\mathbf{x}^2}{2\sigma_+^2}\right) - \lambda_- \exp\left(-\frac{\Delta\mathbf{x}^2}{2\sigma_-^2}\right) \qquad (2)$$

with $\lambda_+ > \lambda_-$ and $\sigma_+ < \sigma_-$, thus giving rise to local excitation and more diffuse inhibition. In the case of visual attention models, with such constraints on parameters, and spatially coherent input stimulation reflecting the presence of localized objects within the visual field, the numerical simulation of the DNF equation will converge to a stereotypical peak of activity, filtering out noise (Fix et al., 2011; Quinton, 2010). In the case of overt attention, it is then possible to directly project the DNF activity to control eye movements (Quinton and Goffart, 2018), in agreement with visual fixations being correlated with a balance of activity in the SC (Gandhi and Katnani, 2011). In our

numerical simulations, we will simply estimate the stimulus position within the field as the barycenter of the field output $f(U)$ (Rougier, 2006).

The time course of field activity before convergence will not be the focus of this article, since we are mostly interested in the location of peaks after stabilization. Readers interested in activity evolution over time will find extensive insights in Schöner et al. (2015) and an illustration of SC dynamics simulation in (Taouali et al., 2015, figure 5).

2.1.2 Projections to the neural field

Empirical evidence supports that signals emanating from a common location in the environment, even through different modalities, will project to nearby locations in the SC (Wallace and Stein, 1996). At the same time, the structure of the SC can be linked back to retinotopic space (Ottes et al., 1986). Given these neurophysiological findings, we decompose the input I defined at each point of the DNF as the sum of a visual input I_V and an auditory input I_A. Although summing projections from different modalities introduces a strong assumption into the model, it is frequent in the literature (Sandamirskaya, 2014; Schöner et al., 2015).

The projection of visual stimuli from the retina to the SC has been modeled mathematically in the form of a logpolar transformation (Ottes et al., 1986). Formally, a visual signal at a position (u, v) in the retinotopic space will be mapped to the SC at a position $\mathbf{x} = (x, y)$ given by:

$$\begin{cases} x = B_x \log\left(\dfrac{\sqrt{(u+A)^2+v^2}}{A}\right) \\ y = B_y \arctan\left(\dfrac{v}{u+A}\right) \end{cases} \tag{3}$$

A, B_x and B_y are constant parameters that originate from the literature (Ottes et al.,

1986). Their values are given in table 1.

As for the auditory inputs and to our knowledge, there is no mathematical formulation of their projection onto the SC. To avoid introducing additional model parameters or uninformed constraints, we thus simply aligned the audio stimuli to their spatially congruent visual counterparts, since we do not aim at modeling the learning of sensory maps in the current research work. As projections to the SC through complex neural pathways are usually quite distant from raw sensory stimulation, we generate population coded auditory inputs as gaussian blobs of amplitude λ_A and width (standard deviation) σ_A. While the gaussian blob associated to the auditory stimulation is directly projected without distortion to the SC neural map, a similar gaussian blob is generated for the visual stimulation yet transformed through equation (3) during its projection on the SC. Amplitude and width of the audio stimuli are added to the list of free parameters of the model, while visual amplitude is fixed (since redundant with λ_A) and visual width is driven by the experimental setup.

2.2 Application to the ventriloquist effect

Even with constraints imposed on projections to the DNF, the model of the SC presented in the previous section and recapped in figure 1 is designed to accomplish a variety of tasks related to audio-visual perception, attention or memory, building upon existing works on neural fields (Schauer and Gross, 2004; Sandamirskaya, 2014; Taouali et al., 2015). In order to validate its capabilities for multimodal fusion, we here apply and test this generic model using an experimental paradigm associated with the ventriloquist effect, this effect being largely documented, and human data available. We use the

seminal work by Alais and Burr (2004), using human performance as ground truth for the evaluation of audio-visual fusion in our model. In their article, they reported detailed psychophysical results aggregated over hundreds of trials per condition and participant, with psychometric functions estimated in both unimodal and bimodal blocks of trials. For the latter, they relied on a fully crossed experimental design, manipulating various fusion-relevant parameters of the stimuli. Among other things, this makes their study particularly fit to replication using their data as a ground truth for computer simulations.

2.2.1 Experimental data

For each bimodal trial, participants were exposed to a sequence of two presentations of audio-visual stimuli (conflicting and non-conflicting, in random order), and had to report which of them was perceived more leftward. In the non-conflict presentation, auditory information (1.5 ms sound click with position determined by the interaural time difference) and visual information (15 ms low-contrast Gaussian blob of controlled width, with standard deviation $\sigma_V \in \{2°, 16°, 32°\}$) were perfectly aligned with each other, but their eccentricity relative to the center of the participant's field of view was manipulated (from $-20°$ to $+20°$, as depicted on the horizontal axis of figure 1 of Alais and Burr, 2004). In the conflict presentation, stimuli were still aligned on the azimuthal axis, but an horizontal spatial discrepancy was introduced between the two, with the visual stimulus moving of $\Delta \in \{-5°, -2.5°, 0°, 2.5°, 5°\}$ (from left to right) and the auditory stimulus moving of $-\Delta$ (horizontal positions in figure 2).

As a consequence, we aim at replicating the psychometric curves (proportion of conflict stimuli perceived rightward as a function of eccentricity of the non-conflict

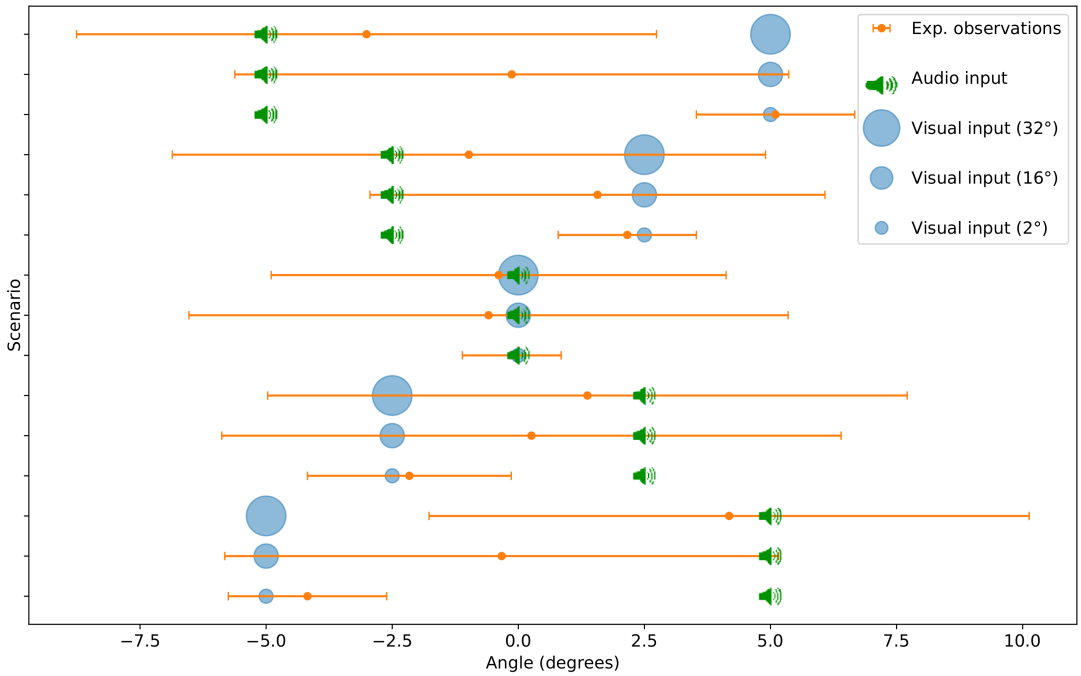

Figure 2: List of scenarios and experimental measures from Alais and Burr (2004). In each line: The green speaker symbol gives the position of the auditory stimulus in the conflicting presentation. The blue circle of growing size gives the position of the visual stimulus, of width $\sigma_V = 2°$, 16° or 32° (not to scale). The measures of bimodal localization are represented by an orange error bar (mean $\pm$ SD).

stimuli) obtained in the 15 scenarios of the original study (3 visual precisions $\times$ 5 spatial distances). These psychometric curves were approximated by cumulative Gaussian functions (sigmoids with near-logistic shape; Bowling et al., 2009), thus reducing them to two parameters: median (also named point of subjective equality, equal to the mean for a Gaussian distribution) and standard deviation (accuracy). The Gaussian distributions associated to the unimodal and bimodal psychometric functions from Alais and Burr (2004) are reproduced on figure 2.

As a synthesis of their results, a thin visual stimulis ($\sigma_V = 2°$) captures the location of the merged signal given its high accuracy. When it is very wide (32°), the auditory stimulus does. In-between (16°), the merging is located between both. In addition, the higher the precision of the inputs (e.g. 2° visual stimulus), the lower the standard deviation of the human localization distribution after fusion, reflecting that auditory and visual information were taken into account in a statistically optimal manner (Rohde et al., 2016).

2.2.2 Model constraints and simulation

For this specific operationalization of the ventriloquist effect, all presentations happen on a single azimuthal axis: $y = 0$. While the version of our DNF model presented in section 2.1.1 could be used as a suitable model of two-dimensional maps in the SC, it introduces parameters that are not directly supported by empirical data from the selected study, and would simply make optimization and interpretation more complex. Committing to the principle of parcimony, we have therefore chosen to restrict our model to a unidimensional projection of the SC, reducing the computational cost of the

simulations.

Whereas asking which stimuli were perceived as more leftward made sense experimentally to reduce task difficulty and prevent biases in responses, numerical simulations allow to directly estimate localization probability density functions. Yet given the noise and non-linearities from equation (1), we rely on the Monte Carlo method to sample the localization distribution under each condition through repeated simulation, and estimate summary statistics (mean and standard deviation of the empirical Gaussian distribution) for the conflict presentation alone. This means that the (static) inputs used in our model always consist of a bimodal signal, having a median location set at the fovea, and made of two unimodal components located opposite from each other. The non-conflict presentation is no longer necessary in this numerical setting. Since there is no generic analytical solution to this class of stochastic integro-differential equations, we rely on numerical resolution, which makes simulations computationally intensive and parameter estimation complex.

To correctly model the spatial distribution of stimuli used in the ventriloquist experiment, the simulated neural field covers angles from $-20°$ to $20°$ in retinal space (which, after the transformation of equation (3), corresponds to ± 2.85 mm in SC) with a spatial resolution of 100 points ($\Delta x = 0.057$ mm). Similarly, to ensure a correct approximation of the temporal dynamics of the multimodal fusion and guarantee convergence to a stable localization, we solve equation (1) using the Euler scheme with a temporal resolution of 100 iterations per second ($\Delta t = 0.01$ s). All simulation constants are recapitulated in table 1. Algorithmically, the mean field potential (vector U) is initialized

to zero and updated by applying the following equation:

$$\begin{aligned}\forall k \in K, U(k\Delta x, t+\Delta t) = U(k\Delta x, t) \\ + \frac{\Delta t}{\tau}\Big(- U(k\Delta x, t) \\ + I(k\Delta x, t) \\ + \sum_{k' \in K} W\big(|k\Delta x - k'\Delta x|\big)\, f\big(U(k'\Delta x, t)\big) \\ + \varepsilon \Big)\end{aligned} \tag{4}$$

where $K = \{-50, -49, \ldots, 50\} = \{\frac{-2.85}{\Delta x}, \frac{-2.85+\Delta x}{\Delta x}, \ldots, \frac{+2.85}{\Delta x}\}$ and I can be decomposed according to section 2.1.2:

$$I(k\Delta x, t) = I_V(k\Delta x, t) + I_A(k\Delta x, t) \tag{5}$$

Table 1: Constant settings for all simulations. The values and descriptions of A, B_x and B_y are taken from Ottes et al. (1986). High spatial and temporal resolutions were chosen to prevent any qualitative impact on the results.

Constant	Value	Unit	Description
B_x	1.4	mm	x-axis scaling for the SC map
B_y	1.8	mm/°	y-axis scaling for the SC map
A	3	°	Shape of the mapping, relatively to $\frac{B_x}{B_y}$
Δt	0.01	s	Simulation time step
$\mathbf{X}$	[-2.85, 2.85]	mm	Spatial domain in SC
Δx	0.057	mm	Spatial discretization step

Given that we model a forced decision task (i.e. where human participants were asked to always answer even if they needed to guess), adequate parameters should always lead to the (quick) emergence of a stable activity pattern in presence of stimuli,

usually under the form of a stereotyped peak of activity on the neural field. An example of this output is given in figure 3, using artificial inputs and zero noise for the demonstration. We can see that, given two similar but conflicting stimuli, the DNF will in any case generate a prototypical peak of activity (an attractor in the dynamical system modelled by the set of differential equations), from which the barycenter can be used as the bimodal stimulus localization estimate, as developed at the end of section 2.1.1. The ensuing decision will either correspond to an interpolation between unimodal signals, or to the selection of the strongest one (barring random fluctuations not shown here). The choice between these two behaviors will depend on both the distance between the stimuli (as in this figure) and their relative precision (illustrated in the result section, with much lower stimuli precision).

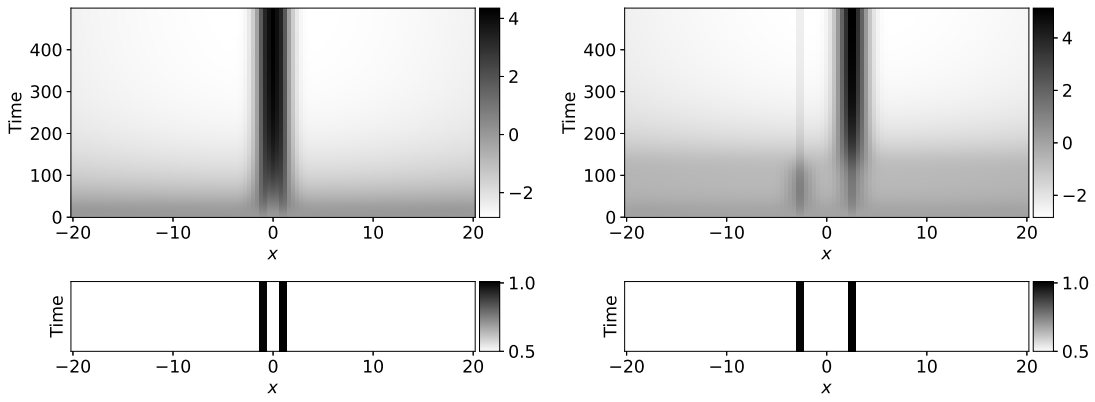

Figure 3: Evolution of DNF potential U on neural field (x) over time (top), using two different custom-made static inputs I (bottom). Parameters are taken from the "Selected" column of DNF+id in table 2, except noise is reduced to zero for explanatory purpose (on this figure only). To break the symmetry, in both subfigures, the right stimulus is 1% stronger than the left. Left and right subfigures differ by the distance between the stimuli.

2.3 Evaluation

While our task is not limited to a quantitative fit to empirical data, we will use the differences between model outputs and psychophysical results as a performance metric, which allows an indirect comparison of numerical models using human behavior as ground truth. As all (human and simulated) localization distributions roughly follow a Gaussian profile, performance will be computed based on estimated means and standard deviations on all scenarios from figure 2.

2.3.1 Compared models

The seminal experimental results on which we rely were already accompanied by a mathematical model (Alais and Burr, 2004). It is based on Bayesian modeling using maximum a posteriori estimates on localization distributions, which remains the dominant paradigm for multisensory integration (Rohde et al., 2016) to which we will compare. It explicitely relies on the hypothesis that the psychometric functions of visual and auditory stimuli are Gaussian cumulative distribution functions. The mean estimate and derived variance for their Bayes optimal combination are given by:

$$\hat{S}_{AV} = \frac{1/\sigma_V^2}{1/\sigma_V^2 + 1/\sigma_A^2}\hat{S}_V + \frac{1/\sigma_A^2}{1/\sigma_V^2 + 1/\sigma_A^2}\hat{S}_A \tag{6}$$

$$\sigma_{AV}^2 = \frac{\sigma_V^2\sigma_A^2}{\sigma_V^2 + \sigma_A^2} \tag{7}$$

where $\hat{S}_V$ and $\hat{S}_A$ are the mean estimates of the visual and auditory signals positions respectively (assumed to coincide with the actual position of the sources), and σ_V^2 and σ_A^2 their variances (derived from the unimodal psychometric functions, as described in Rohde et al., 2016). The Bayesian model differs by design from ours, insofar that it uses

the unimodal performance to predict the bimodal behavior, whereas we fit our model directly on the bimodal scenarios, without prior knowledge of the unimodal variances.

In the case of our DNF model, for a given set of parameters allowing convergence to a stable localization decision through numerical resolution, each simulation should generate a single scalar output (between $-20°$ and $20°$ after projecting back to the visual space). By replicating such simulations, the Monte Carlo method therefore produces an approximate localization distribution in each condition. As the 15 generated distributions (one per condition) are expected to be roughly Gaussian and were tested against extreme observations (to prevent biaises in mean and standard deviation estimates due to statistical outliers), 50 simulations per condition were assessed as sufficient to extract accurate distribution parameters, and used as indices of model performance.

To test the usefulness of the logpolar transformation to correctly explain the experimental results for different eccentricities (confounded with varying degrees of audio-visual discrepencies), as well as to test the robustness of the DNF model to distortions in inputs projections, we will use two versions of our model: one where visual inputs go through a logpolar transformation following equation (3) (referenced as DNF+log in tables and figures); another where the transformation is replaced by an identity function (DNF+id), meaning $x = u$ and $y = v$. In the latter case, the DNF will operate directly on a visual map, i.e. $\mathbf{X} = [-20°, 20°]$, $\Delta x = 0.2°$, and the auditory inputs need no realignment.

2.3.2 Model parametrization

Following previous definitions and constraints, our model has eight free parameters (see table 2): six from the DNF equation, and two from our modeling of auditory inputs in the SC as a Gaussian blob. This is true for both versions (DNF+log and DNF+id), since the logpolar transformation parameters are constant and derived from the literature. The behavior of a DNF depends mostly on the shape of its interaction kernel W. Therefore, fusion performance can mainly be correlated to the four parameters λ_+, λ_-, σ_+ and σ_-. The dynamic and nonlinear nature of the DNF equation can make the dependencies very hard to comprehend, with strong interactions between parameters, especially when related to the kernel. Since we will also measure the variance of the model localization output, σ_N, which controls the amount of noise in the equation, will also play an important role; as well as τ, which controls the integration rate, and thus the weight of the noise compared to stimuli. Finally, while λ_A and σ_A do not intervene in the inner dynamics of the DNF, they can also be tweaked as part of the audio preprocessing of the model. They do have some interaction with the other parameters, as the shape of the interaction kernel determines which shape of input signals will be favored.

To ensure a fair comparison of models, free parameters had to be adjusted to the multimodal merging task. Within the high-dimensional parameter space, meta-heuristics that were already applied to the optimization of DNF parameters (such as Quinton, 2010) did not prove to be robust enough in the case of our multimodal fusion scenarios and evaluation procedure. Indeed, we could not easily combine into a single optimization criteria our two metrics: mean multimodal localization and localization variance. Trying to tackle this multicriteria optimization problem on stochastic integro-

differential equations also did not lead to acceptable Pareto-optimal sets of solutions.

Therefore, after a review of articles in the DNF literature, and extended preliminary simulations, we extracted for each parameter an interval in which suitable behavior was possible, and simply relied on an iterative and partial grid-search approach. Similarly to Jenkins et al. (2021), we started by picking some expertise-driven parameter values, then analyzed model performance as a function of one or two parameters at a time. Keeping the best values found, we iterated over sets of parameters until convergence. In a way similar to a simplex algorithm, we obtained the parameter values in column "Selected" of table 2. We have found that a change in σ_A was sufficient at first sight to compensate most of the distortion of visual inputs by the logpolar transformation. Consequently, it is possible to switch between DNF+log and DNF+id and obtain results of the same order of magnitude, by tweaking σ_A and leaving other parameters intact.

3 Results

Relying on the (locally) optimal parameters from table 2, this section first shows qualitative and illustrative behaviors of the DNF, before comparing performance between the different models described in section 2.3.1 (Bayesian, DNF+id, DNF+log), and then turning to a sensitivity analysis of the DNF model performance, studying the impact of pairs of parameters when keeping the others fixed. The objectives are to show that good performance from either DNF model versions cannot be attributed to over-parametrization (and thus overfit to the experimental data), and to study the effect of parameters on the DNF behavior.

Table 2: Model parameters. When one is fixed, its value is given in the "Selected" column. When one varies, either for exploration or visualization, it takes its values in the specified interval, discretized uniformly into 20 values. For DNF+log, values in italics have to be rescaled by a factor $\frac{2.85}{20}$ to accommodate for the change in field size from $[-20, 20]$ degrees to $[-2.85, 2.85]$ millimeters: while the transformation in the model is not linear, we use this field-wide rescaling to express all width and SD values in the same unit, opting for degrees. After the input is transformed, the DNF always operates on a regular space. σ_A has two different values for DNF+id and DNF+log respectively.

Parameter	Min.	Max.	Selected	Description
τ	0.05	0.5	0.15	Time constant
λ_+	0.1	1	0.425	Amplitude of lateral excitation
λ_-	0.05	0.2	0.15	Amplitude of lateral inhibition
σ_+	*0.2*	*2*	*0.85*	Width of lateral excitation
σ_-	*2*	*100*	*40*	Width of lateral inhibition
σ_N	0.5	5	2.8	Standard deviation of noise distribution
λ_A	0.1	2	1.1	Amplitude of auditory input
σ_A	*2*	*64*	20 | *26*	Standard deviation of auditory input

3.1 Evolution of field potential

As a way to showcase the behaviors of our models, we start by observing their dynamics in realistic experimental conditions, complementing the illustration of qualitative differences in DNF outputs based on stimuli distance in section 2.2.2. For this subsec-

tion, we will make tests using the DNF+id model, as its output can be directly read and easily interpreted in the topological space of the source stimuli. We use the parameters from the "Selected" column of table 2. The inputs in the second experimental scenario ($\Delta = -5°$, $\sigma_V = 16°$) and related model activity are given in figure 4.

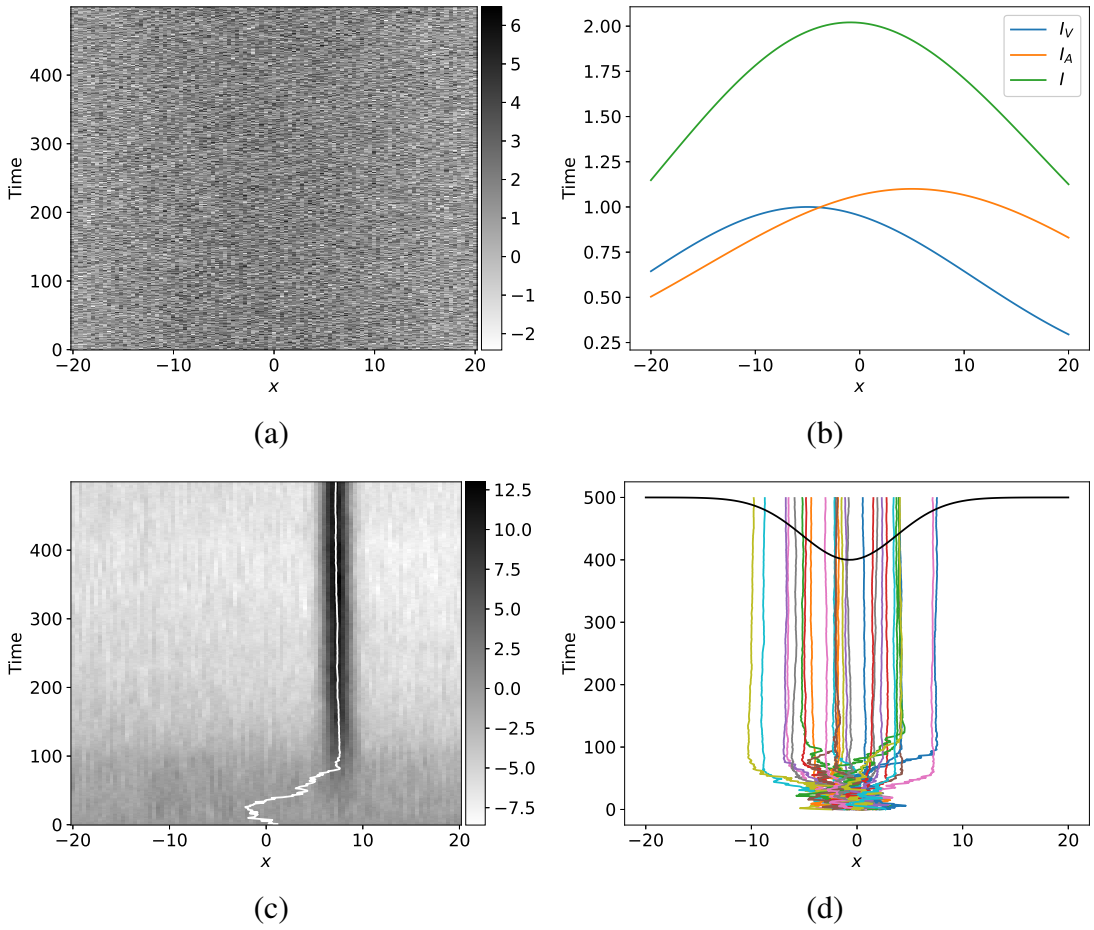

Figure 4: Evolution of DNF+id activity having $\Delta = -5°$ and $\sigma_V = 16°$. (a) Inputs summed with noise on neural field (x) over time. (b) Theoretical distribution of inputs in absence of noise. (c) Field potential U during one single run. The white line shows the evolution of the barycenter of field output $f(U)$. (d) Barycenters of DNF output for 30 other runs of the model. The black line shows the approximate Gaussian distribution obtained with the mean and SD of the final 30 positions.

As can be seen in subfigure (a), the amount of noise in the simulated data makes it almost impossible to distinguish the raw stimuli (b) with the naked eye. The evolution of DNF potential U is shown for one run of the model in subfigure (c). A peak forms at a seemingly random position, which is actually biased by the position of the stimuli. The underlying distribution of selected multimodal locations becomes apparent when the model is run multiple times (d). Some decisions do happen quite far from the source, which is consistent with stereotypical psychophysical studies, in which participants sometimes realize extreme guesses. But the distribution of selected multimodal locations shows that on average, decisions are made in between the two stimuli. The mean and variance of this DNF output distribution are the summary statistics used for model evaluation.

3.2 Model evaluation

Given the aforementioned models, we simulated the experimental scenarios to compare with the psychophysical data. The results are summarized on figure 5. As a reminder, we observe two metrics: the mean localization of a bimodal presentation (center of the intervals on figure 5) and its standard deviation (half-amplitude of the intervals). To mitigate the influence of extreme observations due to the stochasticity of the model, and thus provide accurate estimates, results presented in this section have been aggregated over 2500 runs instead of 50.

The quality of fit varies between scenarios. For example, DNF-based models achieve better fits in scenarios 6, 14 and 15, while the Bayesian model fares better in scenarios 3, 11 and 13. The distances between model and experimental outputs are summarized

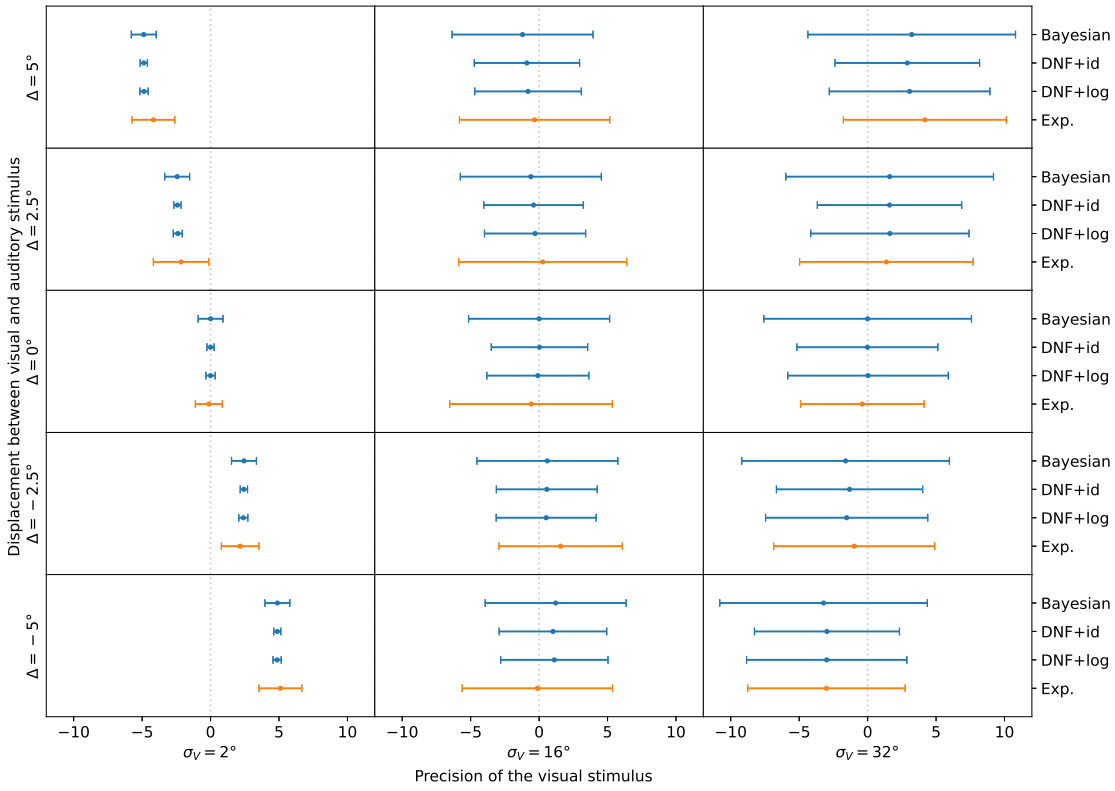

Figure 5: Experimental results of bimodal presentation (orange intervals, same as figure 2) and corresponding model outputs (in blue). For each error bar, the center dot represents the average localization, and the half-amplitude is the standard deviation.

in table 3. This shows a slight superiority of DNF+log over DNF+id, and a slight advantage of the Bayesian model when it comes to representing the localization variance only.

Table 3: Comparison between our model with logpolar transformation (DNF+log), without logpolar transformation (DNF+id), and the reference Bayesian model, using root mean square error between simulated and experimental data over the 15 scenarios.

	Error between means	Error between SD
DNF+log	**0.626**	1.33
DNF+id	0.638	1.38
Bayesian	0.677	**1.28**

Meanwhile, DNF come with the ability to model complex dynamical behaviors and are closer to known neurobiological mechanisms. So it is worth noting that our model enables a versatile point of view of multisensory integration, for a quantitative fit similar to the classical model. In particular, our model can simulate observations on a smaller scale (one run is one human decision) than Bayesian models (mostly focusing on the global distribution of the results). Our model can simulate all random variations between observations, while staying faithful to important mechanisms of multisensory integration.

3.3 Parameter exploration

Our model already shows quantitative results comparable to the most standard modeling paradigm, but there are other useful properties that can be displayed. In this section,

we will verify that performance is indeed consequent to our design choices, and not of overfitting. We will also show that there is still room for finetuning if one were to target some more specific criteria (such as a maximal fit of localization variance).

In order to emphasize parameter interactions in the most readable way, we have chosen to display the effects of two parameters at a time. In figures 6 to 8, six parameters keep the selected values mentioned in section 2.3.2, and two vary on a regular grid within the bounds given by table 2. We will only consider the DNF+log model from now on, our original and most complete version (even though similar analyses could be obtained with DNF+id).

We have found that depending on parameters, model behavior could fall into one of the following four categories. Only the first one is relevant to our simulation, the others will be masked in following figures.

1. For all scenarios, one single peak of activity emerges and stabilizes (often called a "bubble" in DNF literature). The rest of the field is inhibited thanks to lateral inhibition.

2. One bubble emerges but does not stabilize. The maximum potential increases indefinitely because of self-excitation. This is clearly implausible on a neural level.

3. No bubble emerges by lack of interaction, i.e. the term factored by W in equation (4) is negligible compared to the others. So the potential U will converge to an approximation of I. Two peaks will be observable when the stimuli are spatially discrepant, but they do not correspond to a bubble enhanced by self-

excitation. The outcome is that the decision-making role of DNF goes missing, which falls far away from our objectives.

4. In scenarios where stimuli are far apart, two distinct bubbles emerge. This happens when there is not enough long-range inhibition for one bubble to take over the other. Psychophysically, that would account for an observer explicitely noticing that there are two distinct stimuli. Alais and Burr (2004) do not report this happening in their experiments.

3.3.1 Pairwise variations

Our first step is to make all 8 parameters of our model vary by pairs. The results are compiled in two triangular matrices (one for each error measure) in figure 6 (means bottom left, SD top right), of which each element contains a 2D regular grid. The bounds of each parameter are listed in table 2.

First, we can see that τ and σ_N have a strong effect on the localization standard deviation, and a slight effect on the mean localization. In general, increasing σ_N or decreasing τ would give moderately less reliable localization means, but more plausible standard deviations. This is coherent with our simulation paradigm: increasing σ_N means adding more noise, and decreasing τ means a quicker integration of new information through time, both increasing the weight of the noise relatively to the stable audio and visual stimuli. We can also see that the mean localization is not completely smooth, and even less so for higher σ_N or lower τ. As a reminder, our results are by default aggregated over 50 runs for each parameter combination, for the purposes of smoothing the graphics. Fluctuations caused by extreme values are still expected, so it

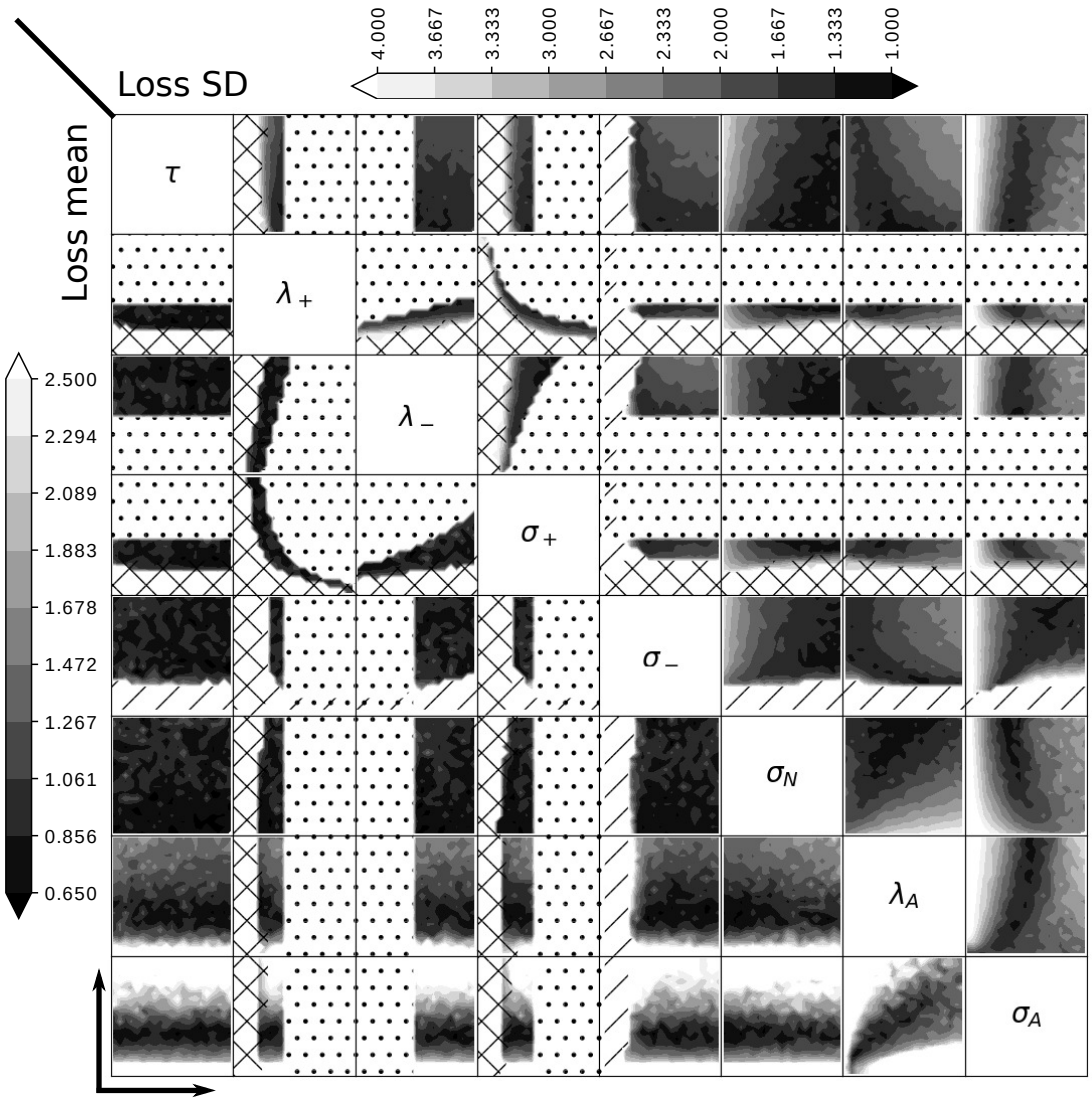

Figure 6: RMSE obtained by the DNF+log model depending on pairs of parameters. The bottom left triangular matrix is based on errors in mean localization of bimodal presentations, the top right one on their standard deviations. For each entry, the parameter labeled in row increases from bottom to top, and the parameter labeled in column increases from left to right. The blank areas filled with geometrical shapes designate parameter sets that fall out of scope of our simulation plan (cf. section 3.3). Dotted: no convergence, or overflowing activity (case 2). Hatched: more than one peak (cases 3 and 4). Crossed: no interaction (case 3).

is consistent that they become more apparent when the amount of noise in the system is increased.

There is some predictible interaction between λ_A and σ_A. The graphs outline a parabola-shaped ridge, along which these parameters can evolve with little impact on the results. It is worth noting that an increase of σ_A can be compensated by an increase of λ_A. That is a characteristic of the DNF. The model is designed to select in priority stimuli whose profile match the positive part of the interaction kernel, which is very thin in the case of the selected parameters ($\sigma_+ = 0.85°$, or 0.12 mm after rescaling). When σ_A augments, the auditory stimulus strays further away from the thin template, and loses weight in the DNF integration. This loss of importance can be artificially compensated by an increase of λ_A.

Interaction kernel parameters λ_+, λ_-, σ_+ and σ_- have clear bounds. In a DNF, when a peak forms due to self-excitation, a minimum amount of inhibition is necessary for the system to stabilize. Too much excitation or too little inhibition will cause the peak to increase in amplitude indefinitely, which does not fit plausibly to any neural mechanism. On the contrary, too little excitation and no peak will form, no interaction will happen and the model will simply replicate its inputs as outputs. This is out-of-scope because it is impossible to generate a saccade or focus for fine-grained processing two stimuli that lie in different locations of the visual field. It is worth noting that λ_+ has an impact on the thresholds for λ_- and σ_+, and vice versa. That means that any of these parameters can be tweaked largely, as long as some ratios of excitation or inhibition are maintained. Interestingly enough, σ_- is less affected by the other three. The main use of this parameter is to ensure the presence of long-range inhibition, so it primarily

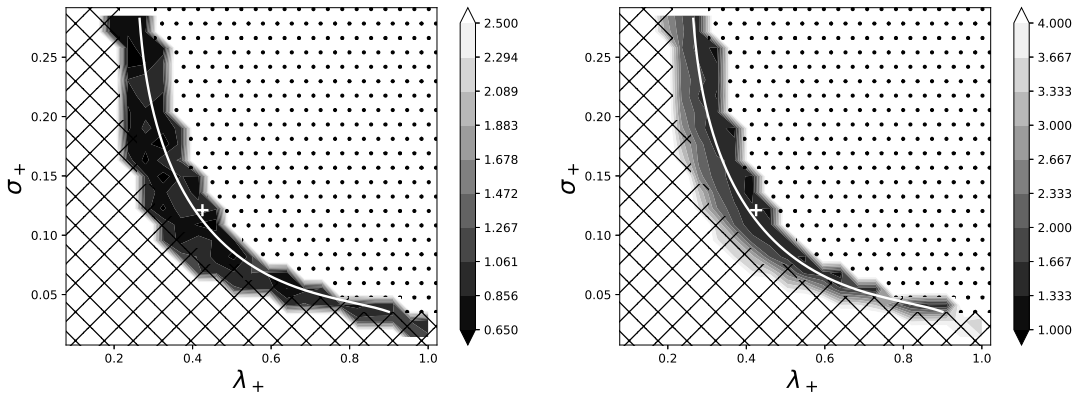

Figure 7: RMSE obtained by the DNF+log model depending on λ_+ and σ_+ (expanded from figure 6). The left graph is based on mean localization of bimodal presentations, the right graph on their standard deviations. The white cross indicates the default values used in the previous section. The white line shows the parametric curve that will be used for parameter reduction. The blank areas filled with geometrical shapes designate parameter sets that fall out of scope of our simulation plan. Dotted: no convergence, or overflowing activity. Crossed: no interaction (U replicates I).

needs to be sufficiently high. That is consistent with alternative implementations of DNF in the literature, where local inhibition in W is replaced by a constant global inhibition parameter, in situations where only one stimulus should be selected in the entire field (Schöner et al., 2015; Taouali et al., 2015). This can be seen as a reduction of equation (2) with σ_- tending to infinity. Our model does not make this restriction: while a multi-selection is irrelevant in our application to the ventriloquist effect, we did not make the assumption of a unique selection in the entire SC.

3.3.2 Reducing the dimensionality of the parameter space

Some regular grids present ridges along which the two parameters vary while the model error stays approximately constant. This is particularly clear for the pair (λ_+, σ_+), allowing us to define a parametric curve on the optimal performance ridge which covers the whole range of parameter values. This curve is defined as a function of an abstract parameter p_+, with the grids and curves for the localization mean and standard deviation reproduced on figure 7. The use of p_+ allows us to check for interaction with other parameters, with one less dimension, and to cancel the effect of the local excitation parameters on the model error. The new grids made with p_+ are given in figure 8.

We can see that there are no interaction effects left, including between p_+ and λ_-. This confirms that the model behavior remains approximately invariant to its excitation parameters as long as as certain ratio is kept. Consequently, the number of parameters in our model could be decreased: for each value of σ_+ within a certain range, there is a value of λ_+ that achieves a similar fit.

The representation of figure 8 also makes clear the tolerable range of certain parameters, and the latitude in their tuning. Inhibition parameters have to exceed a certain threshold ($\lambda_- > 0.11$, $\sigma_- > 5°$), otherwise the self-excitation of the DNF will not be compensated, and the membrane potential U will increase endlessly. In addition, σ_- must be high enough (above approximately 30°) to ensure that only one peak is selected. We can see that a better fit in localization standard deviation can be attained by either decreasing τ or increasing σ_N, but at the detriment of the fit in mean localization. Similarly, λ_A and σ_A show vertical strips where the fit is maximal, but these strips do not coincide between both error measures. Given our goal of reproducing in general

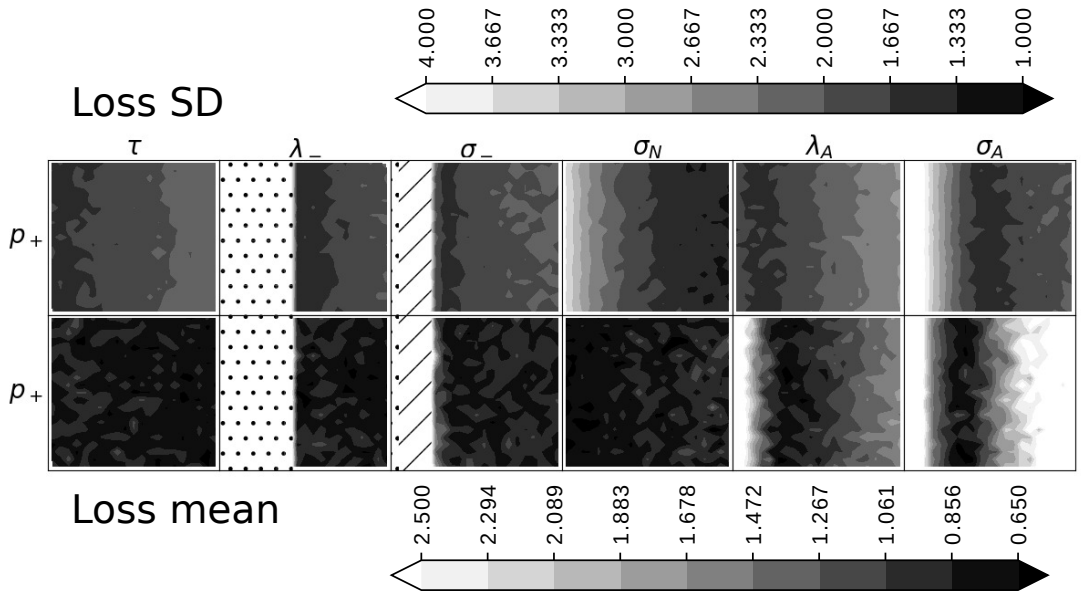

Figure 8: RMSE obtained by the DNF+log model depending on p_+ (from the parametric curve of figure 7) and other parameters. The bottom row is based on mean localization of bimodal presentations, the top row on their standard deviations. In each entry, the parameter labeled on the top increases from left to right. The bottom of a square corresponds to a low λ_+ and high σ_+, the top corresponds to a high λ_+ and low σ_+. See figure 6 for the rest of the legend.

aspects a psychophysical experiment, we have had to settle for a good quantitative fit in both criteria. But as we can see, if our objective was to fit either the mean localization or its standard deviation, performance could be increased substantially. There are no sharp ridges or spikes, and the local optima (see darkened areas on figure 8) are quite wide, so the parameter fitting would be relatively smooth, and the results we obtained in table 3 do not rely exclusively on finetuned values of many parameters.

In summary, there are several ways the number of parameters can be decreased. We have seen earlier that changes in λ_A and σ_A can compensate each other, so λ_A

could be fixed arbitrarily, and some finetuning would be feasible with σ_A alone. σ_A determines, together with the kernel parameters, the relative weight each stimulus will have in the DNF. For an estimation of the mean localization of the bimodal signal, if we assume that λ_- and σ_- always remain above a necessary threshold, and that λ_+ and σ_+ are restricted to the parametric curve in figure 7, then we are left with only two free parameters: p_+ and σ_A. Remaining parameters intervene in the dynamic capabilities of our model (e.g. to predict response times) and its ability to explain some of the inter-observational variations.

4 Conclusion

Models of multimodal merging in psychophysics come predominantly from the Bayesian paradigm. We have shown, using the ventriloquist effect as an illustrative example, that it is possible to model such a task using a neurally-inspired, population-based dynamical system. The model we created conciliates known characteristics of the superior colliculus and the paradigm of dynamic field theory, reaching a quantitative fit comparable to the classical paradigm. The difference between the two models has to be examined at a more theoretical level, given that they operate at different levels of abstraction. DNF are meant to model neural dynamics (Amari, 1977). While they do not constitute an exact simulation of neurons at a microscopic level, the behaviors that emerge from the dynamic system echo physically observable neural patterns at a larger scale, aggregating over thousands ot neurons. Bayesian models of multimodal fusion, on the contrary, were not derived to accurately relate to biological mechanisms (although fine-grained

Bayesian models may be perfectly fit to model such mechanisms), but rather to estimate subjects' decision distributions at coarser spatiotemporal scales. Using the terminology from Marr (1982), the Bayesian model operates at the level of the computational theory, in that it describes the logic by which information coming from different sensory modalities will be integrated, without delving into the ways the inputs are represented or the algorithm is implemented. DNF models could be placed in the other two levels: either representation-algorithm, when the way inputs are transformed into a decision is described through mathematical equations; or hardware implementation, when we consider the discretized field where each neuron acts as a processing unit. Note that these levels are not mutually exclusive, and previous works have hinted at perspectives to analyze either Bayesian modeling (Ma et al., 2006) or DNF (Gepperth and Lefort, 2016) at the level of the other. In any case, this different positioning does not preclude the ability of any of these paradigms to generalize to a wide range of tasks and mechanisms. Both make sense at their own level, although it can be argued that Bayesian modeling might be too broad to capture some of the most subtle behaviors that may emerge from neural interaction (Jenkins et al., 2021). That additional precision of DNF comes at the cost of an extended parameter space.

It is worth noting that our choice of parameters is not detrimentally constraining. There is some latitude in the parameter tuning, thus our modeling hypotheses do not particularly weaken the value of our results. In particular, there is flexibility in the shape of auditory inputs (the model does not rely on one specific pair of values (λ_A, σ_A)), and quantitative fit did not discriminate against the use of the logpolar transformation.

The relative freedom in model optimization opens up new simulation perspectives.

First, there is room for additional parameters and tuning, not included in our current simulations as a first parsimonious approximation. For instance, in our model, as in many previous DNF models (Wilimzig et al., 2006; Fix et al., 2011), white noise is used while not spatially correlated. One could expect that spatially correlated noise (as used in Taouali et al., 2015; Jenkins et al., 2021) would help fit the variance better, especially in scenarios involving a very thin visual stimulus. Then, we have seen that the parameter dimensionality could be reduced (for example by removing σ_- and using global inhibition), and that some pairs of parameters could compensate one another in an optimization task (most notably, λ_+ and σ_+, τ and σ_N, λ_A and σ_A). Consequently, we have reason to believe that our model can be used to fit more demanding tasks. A hypothetical situation would be to simulate a bimodal perception task and fit both the signal localization and an observer's response time. One could then consider locking pairs of parameters on parametric curves (as we did with λ_+ and σ_+) for localization fitting, and use the newly freed dimensions (such as p_+) to fit for the additional constraints.

Indeed, our model has room for the integration of additional functionalities, and the first novelty brought by DNF stands in its dynamic properties. DNF are fully capable of integrating any kind of time-dependant signals (so long as they can be projected onto a topological map). Moreover, their inner dynamics may account for behavioral responses of a human during the perception process. For instance, the peaks of activity in the DNF can generate population-coded motor commands for visual saccades (Wilimzig et al., 2006; Quinton and Goffart, 2018). While the experimental data we have used did not highlight any particular time-related merging effect, our model incorpo-

rates by design the groundwork for the modeling of new dynamic properties.

Additionally, we have seen that DNF are suitable when perceptive fields are not homogeneous across the map, as was showcased by the logpolar transformation. In that particular case, the expectation is that a visual stimulus that appears further away from the fovea will have an increased precedence in the audiovisual fusion. Indeed, in the periphery of the retina, the logpolar transformation will activate a smaller region of the multisensory map, and in our case the DNF matches thinner signals better. This situation is out of scope in the classical ventriloquist experiment, which centers on the fovea, with little eccentricity. This limitation in the experimental data may explain the lack of difference we found between DNF+id and DNF+log. But our simulation would still provide an interesting baseline for the modeling of eccentric audiovisual merging, especially with regards to saccade generation. A visual signal in the border of the field of view will be a likely target for a saccade, although (or, according to many models of saccade generation, because) it is seen less precisely. At the psychophysical level, how much this interferes with the general paradigm of multisensory integration (for which a less precise visual stimulus would actually be captured more easily by other modalities) is still an open question. However, on a computational level, our model reunites some of the keys to a common ground between multimodal fusion and active perception.

Acknowledgments

This work forms part of the project AMPLIFIER. It was funded by the French region Auvergne-Rhône-Alpes in the context of the “Pack Ambition Recherche” initiative.

Preparations to this work have been partially funded by the PERSYVAL-Lab LabEx (ANR-11-LABX-0025-01) under the French program *Investissement d'avenir*, as well as the ANR GAG and CNRS project APF[2].

Most the computations presented in this paper were performed using the GRICAD infrastructure (`https://gricad.univ-grenoble-alpes.fr`), which is supported by Grenoble research communities.

References

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

Alais, D., Newell, F. N., and Mamassian, P. (2010). Multisensory processing in review: from physiology to behaviour. *Seeing Perceiving*, 23(1):3–38.

Amari, S.-I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87.

Bauer, J. (2015). *One Computer Scientist's (Deep) Superior Colliculus: Modeling, understanding, and learning from a multisensory midbrain structure*. PhD thesis, University of Hamburg.

Bauer, J., Magg, S., and Wermter, S. (2015). Attention modeled as information in learning multisensory integration. *Neural Networks*, 65:44–52.

Bowling, S. R., Khasawneh, M. T., Kaewkuekool, S., and Cho, B. R. (2009). A logistic

approximation to the cumulative normal distribution. *Journal of Industrial Engineering and Management*, 2(1):114–127.

Calvert, G., Spence, C., and Stein, B. (2004). *The Handbook of Multisensory Processes*. A Bradford book. MIT Press.

Casey, M. C., Pavlou, A., and Timotheou, A. (2012). Audio-visual localization with hierarchical topographic maps: Modeling the superior colliculus. *Neurocomputing*, 97:344–356.

Driver, J. and Spence, C. (2004). Crossmodal spatial attention: Evidence from human performance. In Spence, C. and Driver, J., editors, *Crossmodal Space and Crossmodal Attention*. Oxford University Press.

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

Ernst, M. O. and Bulthoff, H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162–169.

Fix, J., Rougier, N., and Alexandre, F. (2011). A dynamic neural field approach to the covert and overt deployment of spatial attention. *Cognitive Computation*, 3(1):279–293.

Frens, M. A., Van Opstal, A. J., and Van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics*, 57(6):802–16.

Gandhi, N. J. and Katnani, H. A. (2011). Motor functions of the superior colliculus. *Annual Review of Neuroscience*, 34(1):205–231.

Gepperth, A. and Lefort, M. (2016). Learning to be attractive: Probabilistic computation with dynamic attractor networks. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 270–277.

Girard, B. and Berthoz, A. (2005). From brainstem to cortex: Computational models of saccade generation circuitry. *Progress in Neurobiology*, 77(4):215–251.

Jenkins, G. W., Samuelson, L. K., Penny, W., and Spencer, J. P. (2021). Learning words in space and time: Contrasting models of the suspicious coincidence effect. *Cognition*, 210:104576.

Kapoula, Z. and Pain, E. (2020). Differential impact of sound on saccades vergence and combine eye movements: A multiple case study. *Journal of Clinical Studies & Medical Case Reports*, 7:095.

King, A. J. (2004). The superior colliculus. *Current Biology*, 14(9):335–338.

Lefort, M., Boniface, Y., and Girau, B. (2013). SOMMA: Cortically inspired paradigms for multimodal processing. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438.

Manfredi, L., Maini, E. S., and Laschi, C. (2009). Neurophysiological models of gaze control in humanoid robotics. In Choi, B., editor, *Humanoid Robots*, chapter 10. IntechOpen, Rijeka.

Marino, R. A., Trappenberg, T. P., Dorris, M., and Munoz, D. P. (2012). Spatial interactions in the superior colliculus predict saccade behavior in a neural field model. *J. Cognitive Neuroscience*, 24(2):315–336.

Marr, D. (1982). *Vision. A Computational Investigation Into the Human Representation and Processing of Visual Information*. MIT Press.

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

Meredith, M. A. and Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, 56(3):640–662.

Ménard, O. and Frezza-Buet, H. (2005). Model of multi-modal cortical processing: Coherent learning in self-organizing modules. *Neural Networks*, 18(5):646–655. IJCNN 2005.

Newell, F. N., Ernst, M. O., Tjan, B. S., and Bülthoff, H. H. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science*, 12(1):37–42. PMID: 11294226.

Ottes, F. P., Gisbergen, J. A. V., and Eggermont, J. J. (1986). Visuomotor fields of the superior colliculus: A quantitative model. *Vision Research*, 26(6):857–873.

Quinton, J.-C. (2010). Exploring and optimizing dynamic neural fields parameters using genetic algorithms. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Quinton, J.-C. and Goffart, L. (2018). A unified dynamic neural field model of goal directed eye movements. *Connection Science*, 30(1):20–52.

Rohde, M., van Dam, L. C., and Ernst, M. O. (2016). Statistically optimal multisensory cue integration: A practical tutorial. *Multisensory Research*, 29(4-5):279–317.

Rougier, N. P. (2006). Dynamic neural field with local inhibition. *Biological Cybernetics*, 94(3):169–179.

Sandamirskaya, Y. (2014). Dynamic neural fields as a step toward cognitive neuromorphic architectures. *Frontiers in Neuroscience*, 7:276.

Schauer, C. and Gross, H. M. (2004). Design and optimization of Amari neural fields for early auditory-visual integration. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 4, pages 2523–2528.

Schöner, G., Spencer, J., and DFT Research Group (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford Series in Developmental Cognitive Neuroscience. Oxford University Press.

Taouali, W., Goffart, L., Alexandre, F., and Rougier, N. P. (2015). A parsimonious computational model of visual target position encoding in the superior colliculus. *Biological Cybernetics*, 109(4):549–559.

Trappenberg, T. P., Munoz, D. P., and Klein, R. M. (2001). A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *Journal of Cognitive Neuroscience*, 13(2):256–271.

Wallace, M. T. and Stein, B. E. (1996). Chapter 21: Sensory organization of the superior colliculus in cat and monkey. In Norita, M., Bando, T., and Stein, B. E., editors, *Extrageniculostriate Mechanisms Underlying Visually-Guided Orientation Behavior*, volume 112 of *Progress in Brain Research*, pages 301–311. Elsevier.

Wilimzig, C., Schneider, S., and Schöner, G. (2006). The time course of saccadic decision making: Dynamic field theory. *Neural Networks*, 19(8):1059–1074. Neurobiology of Decision Making.

Witten, I. B. and Knudsen, E. I. (2005). Why seeing is believing: merging auditory and visual worlds. *Neuron*, 48(3):489–496.