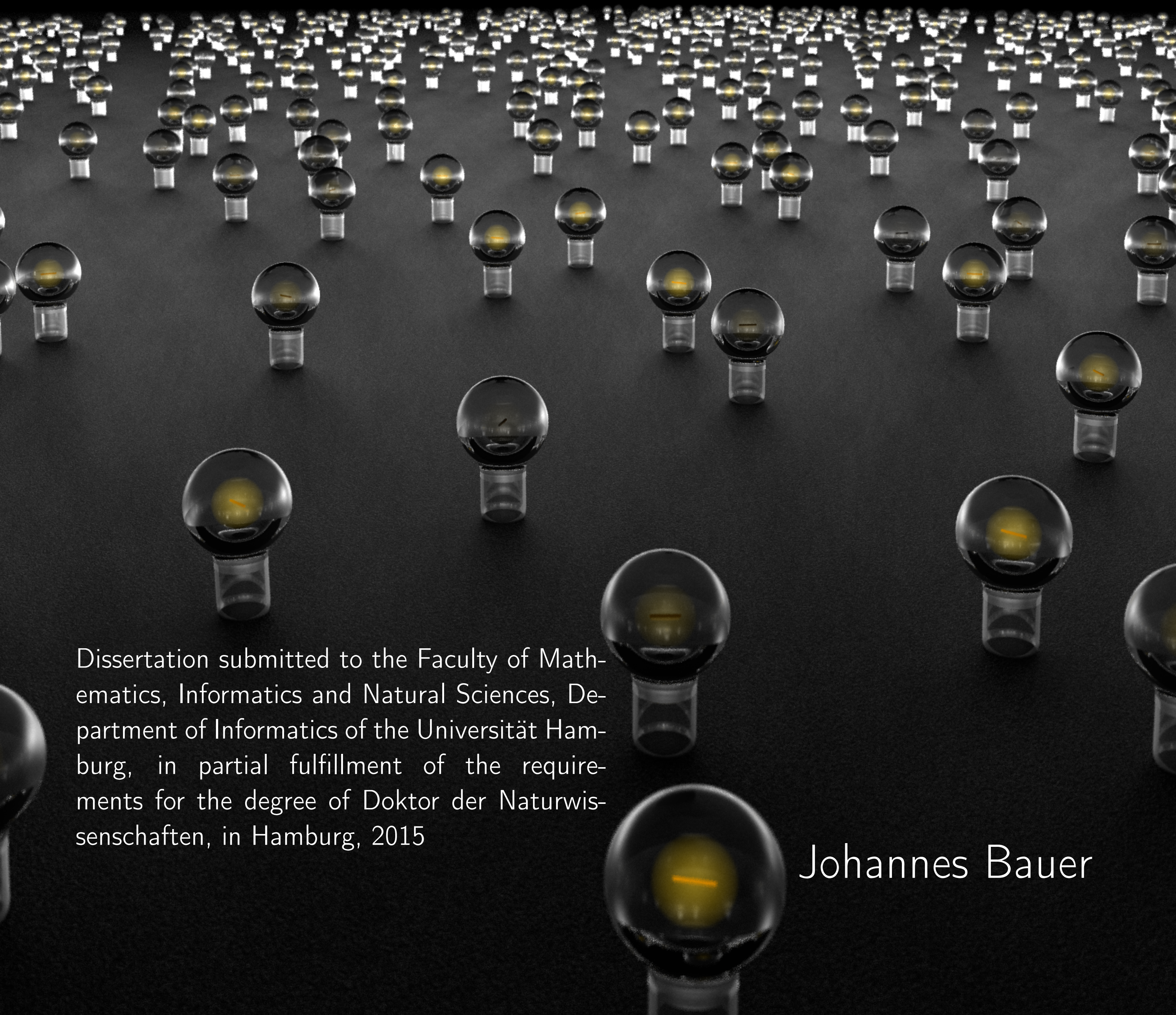
ONE COMPUTER SCIENTIST'S (DEEP) SUPERIOR COLLICULUS

Modeling, understanding, and learning from a multisensory midbrain structure



Dissertation submitted to the Faculty of Mathematics, Informatics and Natural Sciences, Department of Informatics of the Universität Hamburg, in partial fulfillment of the requirements for the degree of Doktor der Naturwissenschaften, in Hamburg, 2015

Johannes Bauer

ONE COMPUTER SCIENTIST'S (DEEP) SUPERIOR COLLICULUS

Modeling, understanding, and learning from a multisensory midbrain structure

Dissertation submitted to the Faculty of Mathematics, Informatics and Natural Sciences, Department of Informatics of the Universität Hamburg, in partial fulfillment of the requirements for the degree of Doktor der Naturwissenschaften, in Hamburg, 2015

Johannes Bauer

Day of oral defense: 29. 10 .2015

The following evaluators recommend to admit this dissertation:

Prof. Dr. Stefan Wermter

Prof. Dr. Andreas K. Engel

Prof. Dr. Wolfgang Menzel

This thesis is dedicated to my wife Sylvia, who keeps me sane and drives me crazy.

My sincere gratitude goes to my thesis supervisor, Stefan Wermter, and all of my colleagues both in the WTM group and the CINACS graduate school, for guidance, inspiration, and criticism.

I am particularly grateful to my family for supporting me and especially to my parents, for asking until I could explain.

Contents

Preface

Shortly after starting my research project, I found myself at a friend's wedding reception talking to a post-doctoral researcher in another field. He asked me what people usually ask in such situations: what was my research about? I verbosely and hand-wavingly explained to him about the superior colliculus, about robots and neural networks. He listened, patiently and with polite interest, and, when I was done, he said something I have been thinking about ever since: he said the moment I was able to explain, in one sentence, what my work was about, I would be ready to hand in my thesis.

Over three years later, I am finally submitting my dissertation for review and this is what it is about:

> I am modeling the superior colliculus, a structure in the vertebrate midbrain which receives and integrates multi-sensory information to localize objects and events in the world and generate orienting actions. To model the superior colliculus, I use artificial neural networks. The goal is to provide interpretation for psychological and neuro-scientific findings, on the one hand, and to develop new algorithms exploiting solutions found in nature, on the other.

Abstracts

Abstract—English

Every natural organism embodies solutions to a host of ecological problems, found through eons of evolution. The study of these solutions and their applications in technical settings is called biomimetics and it has been a driving force in many areas of research. Biomimetic approaches at various levels are attractive especially in robotics due to the similarity of the challenges in robotics to those faced by living organisms. Neurorobotics is the application of biomimetics to robotic applications at the level of neural information processing. It is a promising direction for improving the speed, flexibility, robustness, generality, and adaptivity of robotic systems, the lack of which often limits the practicability of such systems.

But it is not only computer science and robotics which can profit from the study of natural organisms. Neurorobotics can also provide validation for neuroscientific theories by testing them in real, controlled, and highly observable sensorimotor settings. Another possible contribution is an algorithmic, information processing view on biological neural physiology and phenomenology. Such a view may provide interpretation, inspire research questions, and help form theories from biological observation.

This thesis is dedicated to the study of the deep superior colliculus, a region of the vertebrate midbrain. The deep superior colliculus integrates visual, auditory, and other sensory input to localize stimuli and guide motor behavior. Its evolutionary preservation, its role in sensorimotor processing, and the amount of available biological data about it makes it an ideal candidate for neurorobotic and computational inquiry. At the same time, the study of the deep superior colliculus and the multisensory integration it performs has largely been descriptive: for the most part, models have either aimed to describe its neural biology or its system-level behavior mathematically. This is true especially of the development of multisensory integration in the young animal. An algorithmic view is required to describe the way system-level behavior is implemented in neural biology and in order to allow for an evaluation and for transfer of the solutions implemented in the deep superior colliculus to technical problems.

Therefore, the strategy of this thesis is as follows: it first closes the gap between models of physiology and system-level behavior of multisensory processing in the deep superior colliculus, and especially of the development thereof. To this end, a new model of the deep superior colliculus based on self-organized statistical learning is proposed. As an indication of its adequacy, the model is shown to replicate a variety of important neurophysiological and behavioral phenomena, including experience-dependent learning, topographic mapping, the spatial principle and the principle of inverse effectiveness, as well as effects of spatial and feature attention. These effects are demonstrated in computer simulations and in a neurorobotic experiment.

Next, the model is studied functionally, described in mathematical terms, and refined into a practical machine learning algorithm. Finally, that algorithm is applied to a practical problem in robotics: binaural sound-source localization. We show that, given input related to interaural time and level differences in binaural recordings, our algorithm can learn to perform sound-source localization. Our system competes with state-of-the-art sound-source localization systems in terms of localization accuracy. What is more, our system improves on the state of the art, being an unsupervised learning system, capable of online learning, and producing not only a best estimate but a probabilistic population code for different hypotheses.

Thus, this thesis contributes to computational neuroscience a new model of the deep superior colliculus which explains a unique set of phenomena and provides a functional interpretation to its biology. It contributes to general machine learning an unsupervised learning algorithm which learns a topographic latent variable model of its input. And it contributes to robotics a system for unsupervised, online learning of binaural sound-source localization.

Zusammenfassung—Deutsch

Jeder natürliche Organismus trägt in sich Lösungen zu einer Vielzahl von ökologischen Problemen, die die Evolution im Verlauf von Äonen entwickelt hat. Die Beschäftigung mit diesen Lösungen und mit deren Anwendung im technischen Bereich nennt man Biomimetik. In vielen Forschungdisziplinen hat sie sich im Laufe der Zeit als bedeutender Antrieb erwiesen. Besonders in den unterschiedlichen Teilbereichen der Robotik sind biomimetische Ansätze attraktiv, da die Herausforderungen, die sich Robotern und damit Robotikern stellen, häufig denen ähneln, mit denen auch lebende Organismen zu kämpfen haben. Die Neurorobotik ist die Anwendung der Biomimetik im Bereich der Informationsverarbeitung in Nervensystemen. Sie ist ein vielversprechender Ansatz, von dem sich Verbesserungen in der Geschwindigkeit, Flexibilität, Robustheit, Allgemeinheit, und Adaptivität robotischer Systeme erhoffen lassen.

Es ist aber nicht so, dass nur die Informatik und die Robotik von der Erforschung natürlicher Organismen profitieren können. Andersherum kann auch die Neurorobotik helfen, Theorien der Neurowissenschaften unter realen, kontrollierten und genau beobachtbaren Bedingungen zu testen. Weiterhin kann die Informatik einen algorithmischen Blickwinkel auf biologische Neurophysiologie und Phänomenologie beisteuern, der die Informationsverarbeitung im Fokus hat. Ein solcher Blickwinkel bietet Interpretation, inspiriert Forschungsfragen und kann helfen, Theorien aus biologischen Beobachtungen zu formen.

Die vorliegende Arbeit beschäftigt sich mit dem tiefen Colliculus Superior, einer Region des Mittelhirns der Wirbeltiere. Der tiefe Colliculus Superior integriert visuellen, auditorischen und anderen sensorischen Input, um Stimuli zu lokalisieren und Motorik zu steuern. Die Tatsache, dass er evolutionär hoch stabil ist, seine Rolle in der sensomotorischen Verarbeitung und die schiere Menge vorhandenen biologischen Wissens über ihn machen ihn ideal für die Betrachtung aus neurorobotischer und algorithmischer Sicht. Trotzdem ist seine bisherige Erforschung hautpsächlich deskriptiv geblieben: Bestehende Modelle bemühen sich weitgehend, entweder seine Neurobiologie oder sein Wirken auf Verhaltensebene detailliert abzubilden. Besonders für den Aspekt der Entwicklung multisensorischer Integration im jungen Tier ist das der Fall. Um die Lösungen sensomotorischer Probleme, die im tiefen Colliculus Superior implementiert sind, evaluieren und für technische Systeme nutzbar zu machen, ist es notwendig diese Lücke zwischen Neurobiologie und Verhaltensebene durch Modelle zu schließen, die sich des algorithmischen Blickwinkels bedienen.

Die vorliegende Arbeit verfolgt daher die folgende Strategie: zuerst wird die oben beschriebene Lücke zwischen Neurobiologie und Verhaltensebene der multisensorischen Integration und ihrer Entwicklung geschlossen. Dazu wird ein neues Modell vorgestellt, das auf Selbstorganisation und Lernen von Statistik beruht. Um die Angemessenheit dieses Modells zu demonstrieren, wird gezeigt, dass es wichtige Phänomene der Neurophysiologie und des Verhaltens multisensorischer Integration repliziert, darunter erfahrungsbasiertes Lernen, topographische Karten, das sogenannte *spatial principle*, und das *principle of inverse effectiveness*, sowie Effekte räumlicher und merkmalsbezogener Aufmerksamkeit. Diese Effekte werden sowohl in Computersimulationen als auch in einem neurorobotischen Experiment demonstriert.

Im nächsten Schritt wird das entstandene Modell aus funktioneller Sicht betrachtet, mathematisch beschrieben und zu einem praktisch anwendbaren Maschinenlernalgorithmus weiterentwickelt. Schließlich wird der Algorithmus an einer praktischen Anwendung in der Robotik erprobt, der binauralen Lokalisierung von Tonquellen. In einem Roboterexperiment wird gezeigt, dass der Algorithmus lernen kann, Information über interaurale Zeitdifferenz und interaurale Lautstärkedifferenz zu integrieren und so Tonquellen zu lokalisieren. Die Genauigkeit des so entwickelten Systems konkurriert mit der führender Systeme. Dabei hat es diesen führenden Systemen voraus, dass es ein unüberwacht lernendes System ist, das fortlaufend lernen kann und das nicht nur eine Lokalisierung für einen Ton produziert, sondern einen probabilistischen Populationscode für eine Reihe von Hypothesen zur Herkunft des Tons.

Die vorliegende Arbeit trägt also zur *computational neuroscience* ein neuartiges Modell des tiefen Colliculus Superior bei, das eine einzigartige Kombination von biologischen Phänomenen reproduziert und eine funktionale Interpretation für die Biologie des tiefen Colliculus Superior liefert. Sie trägt außerdem zum Maschinenlernen einen unüberwachten Algorithmus bei, der ein topographisches latentes Variablenmodell seines Inputs lernt. Schließlich trägt sie zur Robotik ein System für das unüberwachte, fortlaufende Lernen binauraler Lokalisierung von Tonquellen bei.

1 Introduction

1.1 The Superior Colliculus

Before diving into the details of precisely which question we want to answer in this thesis and why it is important for the computer scientist, and into the reasons for choosing our particular approach and methods, we want to use this section to provide a minimum amount of background information on our subject of study, the superior colliculus (SC) (Section 1.1.1), and to give the reader an idea of its importance from the point of view of neuroscience (Section 1.1.2) and computer science (Section 1.1.3). We hope to thereby give the reader orientation and to motivate, leaving a more detailed, general discussion of the SC to Chapter 2.

1.1.1 The Very Basics

The SC is a layered structure in the vertebrate midbrain which receives direct and indirect, primary and secondary sensory input from the visual, auditory, and tactile sensory modalities and others. It is crucially involved in integrating multisensory stimuli to generate motor responses—most prominently saccades, orienting movements of the eyes and the body (King 2013; Stein and Meredith 1993).

The SC maintains sensory and motor maps which are in retinotopic register: neurons in different columns along the sagittal and coronal axes respond to stimuli from different directions. Stimuli tend to evoke activity in the same columns of the SC regardless of their sensory modality. Stimulation (natural or artificial) of neurons in the deep superior colliculus (dSC) can evoke gaze shifts. After a gaze shift initiated by dSC stimulation, the eyes tend to look in the direction from which a stimulus would have naturally evoked this stimulation (King 2013; Sparks 1988; Stein and Meredith 1993).

A more exhaustive discussion of the general biology of the SC can be found in Chapter 2. Specific details will also be provided wherever necessary throughout the thesis.

1.1.2 The Superior Colliculus as a Subject of Neuroscientific Study

The SC has been an active focus of research for quite some time and the connection between its sensory input and the overt behavior it generates are comparatively well-understood (Stein 2012a). One of the reasons for the continued and intense interest in this brain region is its importance as a first site of multisensory convergence. While the brain areas projecting to the SC are mostly unisensory (King 2013), new findings keep emerging about the complex patterns of neural activity in the SC in response to stimuli in different modalities (see e.g. Brang et al. 2013; Gutfreund and King 2012; Stein and Stanford 2008).

Another interesting feature of the SC is its evolutionary stability. It is present and its basic functionality is similar in all vertebrates (being called the optic tectum (OT) in non-mammals) (Sparks 1988). On the one hand, this speaks to the claim of evolutionary optimality which will become important in Section 1.2.2. On the other hand, it allows us to study the same or similar processes in different species which are amenable to different experimental designs answering different questions.

The SC is important for a host of different brain functions: according to King (2013), the principal function of the SC is initiating gaze shifts. However, it also generates other orienting behavior, like turning the ears towards a sound in those species with movable pinnae (Stein 2012a). Moreover, the SC has been found to be involved in forelimb motor actions like reaching in humans (Linzenbold and Himmelbach 2012; Reynolds and Day 2012; Stuphorn et al. 2000) and in tongue-snapping in toads (Satou et al. 1985). Deactivating or stimulating certain parts of the dSC induces arousal, freezing and escape behavior as well as a raise in blood pressure, heart rate, and respiration. Reactions can be as complex as running and jumping in rats, for example (Brandão et al. 1994).

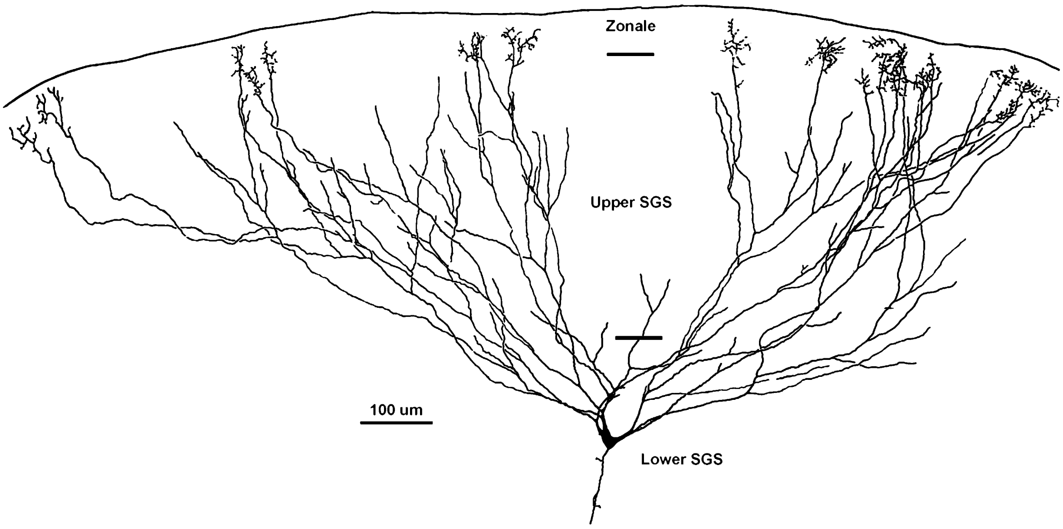
The SC is plastic and adapts to the specific sensory world of the individual (Stein 2012a; Wallace and Stein 2007), and it seems to be important in the post-natal development of other brain regions (Xu et al. 2012). It is part of a sub-cortical visual pathway which may be responsible for aspects of blindsight (Nakano et al. 2013) and may drive face detection and orienting towards faces, especially in newborns (Johnson 2005). It may thus play an important part in social development (Maior et al. 2012).

The SC has also been implicated with being important for consciousness: the same processes in the dSC which initiate saccades have also been shown to have an impact on visual attention on the cortical level (Born et al. 2012; Cavanaugh et al. 2006; Krauzlis et al. 2013; Müller et al. 2005). The SC determines saccade targets and thus makes available the necessary pieces of detail of our visual environment, just when we need them, so quickly and precisely that we get the impression that we are aware of all that is there to see at once (O'Regan and Noë 2001). Finally, together with other parts of the midbrain, the SC acts as a computational bottleneck which integrates, serializes, and coordinates the access of the highly parallel cortical and sub-cortical processes to motor output (Merker 2007).

1.1.3 The Computer Scientist's View of the Superior Colliculus

After explaining in the last section why understanding the SC is important for neuroscience, this section will be dedicated to showing why modeling the SC is a worthwhile goal for a computer scientist. For that purpose, let us look at the problem from the perspective of a single neuron in the dSC.

The gross behaviors of the SC as described in Sections 1.1.2 above and 2 below is coordinated, it fulfills important functions, and, to be able to do that, it adapts to the



> The squirrel SC measures a few millimeters across in either direction (May 2006; Tigges 1970). Since retinocollicular connections are retinotopically organized, this neuron receives input from retinal neurons with receptive fields spanning a wide area in visual space. The correspondence between activity at this neuron's input synapses and events in the outside world is therefore non-trivial. Reproduced with permission from May (2006).

Figure 1.1: Stained cell from the gray squirrel SC.

organism and its environment. All these behaviors of the SC are the result of a myriad neurons transforming action potentials at their input synapses into action potentials they send along their axons to other neurons. Learning to do this correctly, i.e. to generate the correct response, implies making and severing connections and parameterizing the input-output transformation in the right way. The correct response depends on the position of the neuron in the SC.

The precise position of the neuron and the origin and meaning of the incoming action potentials are not known to the neuron when the animal is born (see Figure 1.1). Some of the connectivity is set up pre-natally through chemical gradients, but this early connectivity is coarse and far from mature in many altricial species (Stein and Stanford 2013). The rest of the organization must be learned and there are good reasons to believe that this learning is at least partially unsupervised.[1]. Lastly, the tasks performed by the SC, like all sensorimotor tasks, are probabilistic, and this is aggravated by the fact that neural responses are inherently stochastic (Tolhurst et al. 1983; Vogels et al. 1989).

We can thus describe the SC as a system performing a complex, distributed, and probabilistic information processing and unsupervised learning task. And the study of

[1] See Chapter 2.

complex, distributed, and probabilistic information processing and unsupervised learning tasks falls into the domain of computer science. A computer scientist may therefore be interested in modeling the SC for at least two reasons. The first reason is that he or she may contribute the point of view of information processing to the interdisciplinary field of computational neuroscience. Since the dSC is engaged in information processing, as we have just argued, an understanding of the dSC is not complete without an understanding of it as an information processor. Of course, such an understanding must be informed by and it must conform to biological fact. Thus, an exchange of expertise between neuroscience and computer science can further the understanding of the dSC. The other reason for a computer scientist to work on models of the dSC is that the dSC is an information processing system which was not designed by humans and the exercise of analyzing it may yield new insights for computer science. This theme will be dealt with in greater depth in the next section.

1.2 The Goals and the Methods

1.2.1 Goals and Research Question

Apart from the gain in understanding of an interesting brain structure, studying multisensory integration (MSI) in the dSC can help solve problems in more classical areas of computer science. Biomimetic approaches have had applications in many areas, like materials science, mechanical sciences, and sensor technology (Vincent 2009). They were particularly successful in robotics (Vepa 2009). In hindsight, this is not very surprising since this discipline is concerned with creating whole organisms which are to perform in their ecological niche. Nature has been doing this for far longer and on a far greater scale than humans have, and it is to be expected that it has developed a trick or two that we have not yet discovered. The design and implementation of neurorobotic robotic systems is therefore an active field of research with considerable progress made in recent years (Wermter et al. 2005).

In a way, artificial intelligence (AI) is inherently biomimetic as its goal is to create intelligent machines.[2] Intelligence can be defined independent of living things, but it would be hard to deny that at least the inspiration comes from the observation of humans' and other animals' behavior.[3] Still, the choice of methods in realizing intelligent behavior in machines varies considerably from highly artificial to strongly bio-inspired (and there have been recurring trends in one direction or the other in the history of AI). Both approaches, the deductive and the inductive, have had considerable success. Some of the reasons why researchers in AI have turned to biomimetic approaches are that artificial systems are outperformed by natural systems in many tasks in terms of

[2] We deliberately do not formally define AI here.

[3] In fact, the father of both computer science and AI, Alan Turing, designed his machines to mimic human computation (Turing 1937). And he evoked those biomimetic computers as an example of possibly intelligent machines when he introduced his famous 'Imitation Game' or 'Turing Test' (Turing 1950).

speed, flexibility, robustness, generality, adaptivity, and energy efficiency (Adams et al. 2013; Wermter et al. 2005).

What we can hope to gain from a better understanding of the SC—given the tasks the dSC is involved with[4]—are direct and indirect contributions to machine learning in general, sensorimotor learning, and even to social robotics and artificial consciousness (Merker 2007; O'Regan 1992).

Therefore, the questions we will discuss in this thesis are the following: a) how is the way in which the dSC solves its task implemented in its biology, and b) how can we use some of the mechanisms in its implementation for similar tasks in AI and specifically in robotics?

1.2.2 The Methods

How a given scientific question can be approached depends on the question itself and on the nature of the field. In our case, it is the field in particular, and the perspective of the question, which warrants a closer look at the methods used to approach it. Computational neuroscience is a relatively new discipline and the computer scientist's specific interest in the field is somewhat unconventional both for computational neuroscience and the computer scientist. We will therefore use this section to discuss the methods we choose to study the dSC and approach an answer to the question posed above. In addition to the overall, unified view of the methods chosen in this thesis, a discussion of the differences in methodology between the two branches of science to which we contribute, and therefore in the differences in methodology of the respective parts of this thesis, can be found in Appendix A.

Modeling

There is some debate about the merit of models in science. To us, it seems that one of the main reasons for this debate are not deep, epistemological concerns, but a confused understanding, or rather lack of consensus, of what a model is. Suppes (1960) has dealt with this problem and found that different branches of science have different, sometimes incompatible, definitions of a model. Especially the relationship between models and theories is a very confused issue.

By Tarski's (1953) definition, a theory is (roughly) a collection of mathematical statements, and a model is a substitution of all the variables in these statements that makes them true statements. This use is common in much of physics, and in certain branches of computer science. Another understanding of the word 'model,' exemplified by Humphreys's (2007) definition, is a collection of mathematical statements (thus, a theory in Tarski's sense) together with a mapping of the entities in those statements to physical entities. These definitions are already incompatible, but it is shades between these extremes which make a scientific discussion about the usefulness of models really confusing. Suppes (1960) found that in the social and behavioral sciences, 'model' tends

[4] See Section 1.1.2.

to mean a quantitative version of a qualitative theory, and thus a specification or, in Tarski's sense, a partial substitution of variables.

Finally, there is a fourth meaning of the word 'model,' which can be defined as: a theory which abstracts from much of the detail and focuses on describing certain salient features of that phenomenology. Suppes (1960) states that this definition is used primarily in those branches of physics which do not yet largely explain the phenomenology of their domain of discourse. To us, this use seems to be the one most common in computational neuroscience as well. As we will see, it suits the requirements of this young science. Unless otherwise noted, this will therefore be what we mean when we use the word 'model' in this thesis.

With this definition, we can discuss what the use of a model is in general, and what makes a model a good model. Popper's (2002) principled and strict proposal for a scientific method is to come up with a theory and then test it against all available empirical data. If there is empirical data contradicting the theory, then the theory is to be rejected. If there is no more empirical data contradicting the theory, experiments are to be designed which could potentially disprove the theory if it is wrong. If no possible experiment could disprove the theory, that is, if it does not make predictions about more than the empirical data already known to us, then it is to be rejected as non-scientific. A good theory, according to Popper, is one which explains a lot of known phenomenology and makes many predictions which can practically be tested.

Following this proposal, every model in the sense defined above would have to be rejected immediately. After all, these models deliberately neglect parts of the empirical data and focus on explaining others. As valuable as Popper's proposal is, it is not fully adequate for such a young branch of science as computational neuroscience, where there are no theories which explain most of the data.

We argue that it should be the goal of modeling to come up with individual mechanisms and combinations of mechanisms which explain important aspects of our subjects of study. All things being equal, a model describing strictly more phenomenology than another should be preferred. Describing more phenomenology in the case of computational neuroscience can mean: 1) it describes the hardware in more and more realistic detail, 2) it explains more of the important system-level phenomena (e.g. patterns of activation), 3) it covers greater variety in observed input-output patterns (more complex stimuli, more complex behavior). Successful models will still be false scientific theories, but combinations of mechanisms which have shown to explain large parts of the phenomenology will serve as building blocks for even more powerful models. Over time, this should lead to a state like in currently more mature branches of science, in which all theories account for most of the empirical data and directed experiments are designed to distinguish between variants. At that point, our models will be theories and Popper's approach will fully become practical.

Levels of Modeling

So far, we have only decided to develop mathematical descriptions which abstract from detail and aim to describe as large a part of the empirical data as possible. However, there is not only one level at which an information processing system can be modeled. Marr proposes three levels at which a complex information processing system needs to be described to be understood (Marr 1983, pp. 24–25).

The first is what he calls the computational theory. It is a formalization of the goal of the computation implemented by the system, a justification of that goal, and a mathematical description of the 'logic of the strategy by which [the computation] can be carried out.' The second level deals with the implementation of the computation in terms of representations of input and output and the algorithm transforming one into the other. Last, there is the explanation of how the physical structure of the system realizes these computations.

All three levels are necessary to describe, explain, and understand the whole range of the phenomenology of the system. Computational theories, for example, typically fail to explain epiphenomenal aspects, like energy consumption and often timing in the case of biological systems (Jones and Love 2011). Mechanistic descriptions on the other hand can produce phenomenology, and thus help us verify that the model components are complete, but they are not the right level at which to understand complex behavior—simplifying abstraction is key for that (Krasne et al. 2011; Rosenblueth and Wiener 1945). And without an explanatory middle layer linking them, they are both weak scientific theories in that they only describe the status quo and neither makes any prediction about what happens when there are changes in the realm of the other one (in the task or in the physical composition of the system, respectively).

For our case, we will identify the level of the computational theory of dSC behavior with psychometric modeling of those sensorimotor behaviors which integrate multisensory information for object localization. We will identify the mechanistic level with the level of low-level neurobiological detail. As we will see in Sections 2.1 and 3, there have been considerable results at both of these levels.

However, less work has been done on connecting the top-most with the bottom-most layer. That middle layer, which is concerned with representations and algorithms, and therefore with the stock and trade of computer science, will be the focus of this work.

Testing Models

We explained above that theories, and models, are rated by the scope of phenomenology they cover. What the scope of a theory is is determined by applying its formulae and solving them for concrete cases—by generating predictions, which are compared to empirical data. These thought experiments are also called 'simulations' and, as argued compellingly by Humphreys (2007), can be done on biological or artificial computing hardware.

We have argued that theory in computational neuroscience is still in the process of catching up with the body of empirical data to be explained. In principle, it is therefore perfectly valid and common practice in the field to test new models against existing data only. For examples of studies which have followed that approach, see Chapter 3.

On the other hand, a model's strength can be boosted considerably by generating predictions and testing them against *new data*. This is what we do in experiments, which are of primary concern in some of the older branches of science, where unexplained phenomena are rare. In computational neuroscience, this practice is especially important in two cases: either to distinguish between two models which have the same predictive power on existing data, or to strengthen evidence for critical assumptions of a model.

Experiments with humans and other animals are one way to generate new data against which to test hypotheses. Another is to use physical models[5] of the real thing, which behave similar in aspects important to the point to be made with the experiment. Rosenblueth and Wiener (1945) have argued that using a physical model instead of the original subject of study can be the right thing to do if either that physical model is better understood (and therefore the data gained from the experiment is easier to interpret) or the loss in external validity due to using a model is outweighed considerably by the increased feasibility of the experiments.

Neurorobotic experiments are one case of this practice and a good example of Rosenblueth and Wiener's two criteria. Suppose we have a model of some part of an animal's brain. Then, a biorobotic experiment replaces all parts of the animal—motor, sensory, and cognitive—which are not part of the model with artificial substitutes. Thus, the robot can interact with a natural environment and generate plausible input to the model which is used to generate predictions about the behavior of the modeled brain region given equivalent input. To argue a convincing case with the results of such an experiment, it has to be clear about the analogy between the natural organism and the surrogate; about what is meant to be artificial and what is claimed to be similar (Datteri and Tamburrini 2007). For computational neuroscience, the benefits of biorobotic experiments is that they make effects observable and experiments practical, or even possible, which are not in real humans or animals (Bauer et al. 2012b; Brooks 1992; Rucci et al. 2007). For the computer scientist, such experiments also validate an algorithm's applicability to real-life problems.

Optimality

There is the notion that information processing in humans and other animals is statistically optimal in many ecologically relevant cases, in particular in many instances of MSI. We will allude to this notion in various places throughout this thesis and it is therefore important to discuss it here to prevent misunderstanding.

The idea that natural information processing is optimal comes in two flavors. The first, modest sense in which sensory processing is described as optimal stems from the rather innocuous observation that human and animal behavior often agrees with models of sensory processing in which the agent knows about the uncertainty in its sensory input and acts accordingly. Crucially, these models assume specific kinds of uncertainty and predict behavior which is statistically optimal given this kind of uncertainty. In the case of multisensory processing, models often assume Gaussian noise in unisensory estimation and predict that the agent will linearly combine unisensory estimates with factors optimally derived from the width of the Gaussian noise functions (Alais and

[5] Models in an entirely different sense, not to be confused with that used in the rest of this thesis.

Burr 2004; Ernst and Banks 2002; Hillis et al. 2004; Knill and Pouget 2004; Körding and Wolpert 2004; Landy et al. 2011).

The term 'optimal' in this sense is somewhat unfortunate as optimal performance without any qualification is virtually impossible to show of a real system with real input. This is probably the worst one can say about this notion of optimality, adding maybe that it is a purely computational way of looking at the phenomenon, which treats the actual processing as a blackbox, does not attempt to explain it, and does not account for the epiphenomenal aspects (like timing, energy consumption etc.) (Jones and Love 2011). Apart from that, it is simply a matter of agreement with empirical data whether it is correct to say that human sensory processing is 'optimal' in this sense, not a matter of methodology. Although slightly misleading, we will use the term 'optimal' in this sense when we discuss the performance of our system, because it is established. When we do, we will simply mean 'as if taking into consideration the uncertainty in the input, given a specific model of uncertainty'.

The radical view of optimal natural sensory processing is that human and animal performance truly is close to absolute statistical optimality wherever it has been important in our phylogeny. At first glance, this view appears theoretically supported by the idea that extant organisms are the products of millions of years of evolution in which selection pressure should have eradicated suboptimal information processing (Körding 2007). It seems backed up empirically by studies showing that human performance is comparable to certain Bayesian models of information processing which are provably optimal (see above).

On closer inspection, however, it becomes clear that human performance actually cannot be optimal. To perform strictly optimally, we would have to make use of *all information available to us, in the present and in the past, in the best possible way.* And that is impossible, for first, we do not have the storage capacity to keep all that information unabridged, and second, using information in *the best possible way* can require computations which are just intractable, even given the vast parallelism of the human brain[6] (Beck et al. 2012).

It is an apparent paradox that we can seem to find optimality almost wherever we look. This paradox is resolved when we realize that we implicitly factor in certain sources of suboptimality into the computational models to which we compare natural performance. The ventriloquist effect is a case in point: seeing a puppet move its lips in a 'conversation' with a seemingly motionless puppeteer, we get the impression that the puppet is the source of the words that are spoken, not the puppeteer (Chen and Vroomen 2013). If we did use all information optimally, this should not happen, because the fact that puppets do not speak is absolutely available to most of us. However, the maximum likelihood estimator (MLE) model used for this MSI task by Alais and Burr (2004) does not use that information.

Whether explicitly or implicitly: if we modify the hypothesis that natural cognition often is statistically optimal to say that it often is optimal *given its limitations*, we

[6] ...and that is saying nothing of the low fidelity of neurons as computational units and certain idiosyncrasies in our processing which we inherited from our ancestors to whom they were useful.

make it tautological since everything about natural cognition making it suboptimal is a limitation. Thus, the assumption of optimality in the radical sense is either wrong or not a scientific hypothesis.

It can inspire a method, though: it is true that evolution selects against suboptimal information processes (among other things) and we do often wonder at the effectiveness if not optimality of the results. We will therefore proceed in the spirit of Braitenberg's (1986) "law of uphill analysis and downhill invention", which postulates that it is harder to analyze a complex mechanism than to build something similar and see how it relates. Whenever we are unconstrained in our modeling by solid biological evidence, we will assume that Nature made the optimal choice and try to think what that would be, given the constraints that we know of. This has the benefits of often being right on the one hand and discovering something interesting when being wrong, on the other (Landy et al. 2011)... as long as we keep in mind that we are justified pragmatically, not epistemologically (see Jones and Love (2011) for a similar argument).

Transfer

Our strategy to look for optimal solutions, in the sense put forth in the last section, should be expected to produce results which do not only model, but may also have practical merit. Still, the methodology described so far primarily serves our goals in computational neuroscience. Ultimately, however, through our modeling work, we aim to provide not only an improved knowledge of the dSC, but also ideas for good solutions for problems in computer science.

In a way, computer science is agnostic to the significance of its problems in the physical world. It does not matter to a computer scientist for what purpose a list may be sorted, whether it is for typesetting the reference section in a book, like the one at the end of this thesis, or for deciding on the order in which a number of tasks are to be worked upon, depending on their relative importance. What matters is that there are better and worse ways to sort a list, and that algorithms for sorting can be described and evaluated quite independently of the ends to which lists are actually sorted (Skiena 2008, pp. 3–5). Similarly, solutions for loose coupling of software components with consistent state update (Gamma et al. 1994, pp. 293–304), universal function approximation from examples (Rumelhart et al. 1986), and relational structuring of data (Codd 1970) presumably all have been developed in a concrete context. Nevertheless, they have proven useful beyond that and can now be discussed without reference to any specific application. It is fortunate for us that solutions which were invented for one purpose can be abstracted from that problem, studied independently of it, and applied to a different problem, provided the two problems share similar mathematical formalizations. If this was not so, then, as computer scientists, all we could hope to achieve by studying cats' dSCs, for example, is to learn how they work, and maybe how to make one—not a desirable capability considering that there is no shortage of cat dSCs in the world.

We will analyze the models arising from our work, interpreting them from the point of view of information processing, and extract mechanisms which can be useful not only for the tasks of the dSC. In order to show that those mechanisms are useful for practical applications, we will apply those mechanisms—not the models themselves—to an actual problem that shares, as we wrote above, a similar mathematical characterization with the problem that is solved by the dSC. At the very least, this will provide a solution to a problem. Additionally, we argue, it will demonstrate the generality of the mechanisms we found when modeling the biological dSC.

1.3 Outlook to the End of the Journey—Novelty

In pursuit of our goals and following the methodology laid out in the last section, we will develop a novel model of the dSC which views it as a self-organizing network learning to perform statistical inference to compute a probability density function (PDF) for the position of a stimulus and represent it in a topographic probabilistic population code (PPC). Self-organization and probabilistic population codes (PPCs) have both been used to model the dSC, but not in combination. We will show that the model can account for the development of spatial organization of the dSC, the spatial principle and the principle of inverse effectiveness, at the neurophysiological level, as well as MLE-like MSI at the behavioral level. Extended by simplified cortical input, our model naturally produces enhancement of neural responses similar to the effects of spatial and feature attention, as well as task-dependent target selection. This it does without treating primary sensory and cortical input differently, thus supporting the view that attention may be better understood as an emergent phenomenon than as an inbuilt mechanism in natural cognition.

To machine learning and robotics, we contribute a novel neural learning algorithm, which we develop from our model, which learns a latent variable model (LVM) of its input and uses that to perform statistical inference. In contrast to other machine learning algorithms which learn LVMs, ours

- has a topographic interpretation,
- produces probabilistic output (estimates probabilities of all hypotheses, not just one most probable hypothesis),
- relies on very light assumptions on noise,
- supports highly non-linear LVMs,
- supports online learning.

Any one of the above items is shared with a number of other algorithms, but the combination, again, is novel.

To demonstrate the practical applicability of our novel algorithm, we will describe an adaptive system for binaural robotic sound-source localization (SSL) whose performance is comparable to the state of the art. That system has in its favor over established systems the ability to adapt online, as its physical form or the environment changes, and to integrate additional auditory or non-auditory information naturally.

1.4 Structure of this Thesis

The multidisciplinary nature of our topic and our strategy of first modeling a brain region and then transferring resultant knowledge to the application domain leads to a natural segmentation of our work.

First, we will take a closer look at our subject of study, the SC, and also at biological MSI, in Chapter 2. There, we will naturally focus on those aspects which are relevant in this thesis. However, additionally, we will also provide some background information, to give the reader who is not already familiar with these subjects some perspective.

Then, we will begin Part I, which is devoted to modeling the dSC and which represents the largest part of this thesis. We will start this modeling part by reviewing the literature on modeling the dSC, in Chapter 3. After that, in Chapter 4, we will introduce the artificial neural network (ANN) and an algorithm on which we will base our modeling. That algorithm will be used in Chapter 5 to model basic, bottom-up MSI in the dSC. In that section, we will demonstrate that our ANN model can reproduce several important phenomena occurring in the actual dSC, and we will report on a neurorobotic experiment in which we trained our model on real sensory data. The model will be extended from a purely bottom-up model to a model including cortical input in Chapter 6. There, we will demonstrate that effects of spatial and feature attention in MSI can arise from self-organized learning without cortical input requiring any different treatment from sensory input. Part I will close with an intermediate discussion of our efforts in modeling the dSC, in Chapter 7.

In Part II, our focus will change from modeling the dSC to extracting mechanisms from our models which may be useful in practice. As we will make the transition from explaining biology to using our knowledge for practical applications, we will argue that such applications do not lie in audio-visual localization. In accordance with the thoughts laid out in Section 1.2.2, we will therefore start by discussing the ANN algorithm we have used for modeling in Part I in more mathematical terms and compare it to other algorithms which do similar things, in Chapter 8. We will specifically address that algorithm's strengths and weaknesses. After that, in Chapter 9, we will introduce a version of our original algorithm which features optimizations that address some of those weaknesses. That new version of our algorithm will be applied in Chapter 10 to an actual task: robotic binaural SSL. Like Part I, Part II will close with an intermediate summary and discussion, in Chapter 11.

We will close the thesis in Part III with a global summary of our work and a discussion of the insight gained through it. The discussion will specifically highlight the interactions between our modeling and practical work, discuss questions that arise from those interactions, and point out directions for further research.

2 Multisensory Integration and the Superior Colliculus

The SC is by no means an understudied brain region. Our biological knowledge about it fills volumes. In fact, one of the publications most cited in the context of this brain region is a book, by Stein and Meredith, which was published as early as 1993. Since then, more data has accumulated steadily every year, and more papers, book chapters, and books have been written on the subject. The situation is similar with regard to the study of natural MSI, which is tightly related to that of the dSC. It can therefore not be the task of this section to provide anything approaching a comprehensive summary of what is known about the SC.

Instead, we will give an overview. In this overview, we will focus on knowledge which we perceive as paradigmatic for the community studying biological MSI and the SC, to provide the reader who is not part of that community with the necessary background. Where appropriate, we will add details which are not paradigmatic but merely important, if these details are relevant in some way for the discussion of our work to be described in later sections. Conversely, most of the information given here will be directly relevant in the context of this thesis. However, we do feel that a somewhat broader background is required to judge the adequacy of our treatment of the dSC in the rest of the thesis, knowing not only what is incorporated into our models, but also what is left out.

In the following, we will first look at MSI at the system- or behavioral level, and then focus on the neurobiology of the SC, in Sections 2.1 and 2.2, respectively. In particular, we will give the reader an overview over location and structure of the SC (Section 2.2.1), its role in guiding motor behavior (Section 2.2.2), the phenomenology of neural activity in the SC arising from uni- and multisensory stimulation (Section 2.2.3), the SC's spatial organization (Section 2.2.4), its involvement in attentional processes (Section 2.2.5), its internal connectivity and that to other brain regions (Section 2.2.6), and, finally, its development in the young animal (Section 2.2.7).

2.1 Multisensory Integration on the Behavioral Level

There is usually more than one way to perceive a behaviorally relevant object or event, and determine its qualities. Food can often be located by sight and smell, sometimes by hearing or even echolocation or electroception. Visual, haptic, and motoric cues can be used to estimate the size, shape, and material of an object.

MSI is the process of using cues from multiple sensory modalities to glean information about the world. There are three cases where that is advantageous: first, some kinds of information are only accessible by combining cues from different sources. An example is determining the material of an object as acrylic glass as opposed to glass or other synthetic materials, which cannot be done using vision or haptic information alone. Second, the same kind of information may be available in one modality sometimes, and sometimes in another. The location of an object comes to mind, which generally can be reckoned using vision or through auditory cues, but not always in both. Third, redundancy in two or more of one's senses can be exploited to reduce uncertainty. Vision and hearing can be used, for example, to get a better estimate of the position of an event than either of the modalities could provide individually, and vision and haptics can together improve an assessment of the size of an object. In this thesis, we will deal with the third and second kind of MSI, as those are the kinds in which the SC is involved.

A number of experiments on multisensory integration appear to indicate that we integrate by selecting that sensory source of information about a property of an object or an event which is most appropriate for that property. Information from all other modalities appears to be all but disregarded. The ventriloquism effects are cases in point: in the spatial ventriloquism effect, an auditory and a visual stimulus, separated spatially from each other, are perceived as one multimodal stimulus whose position appears to be that of the visual stimulus (Jack and Thurlow 1973). In contrast, the temporal ventriloquism effect occurs when auditory and visual stimuli are separated in time: in that case, they are often perceived as one multimodal stimulus which seems to occur at the actual point in time of the auditory stimulus (Aschersleben and Bertelson 2003; Bertelson and Aschersleben 2003; see also Chen and Vroomen 2013).

We wrote above that it *appears* that MSI selects the most reliable source of information and disregards the others, and that we *seem* to perceive an audio-visual stimulus at the location and point in time of the visual and auditory sub-stimulus, respectively. The computer scientist cringes at the idea of throwing away information from one source just because another source is more reliable. Bad information can always be used to make good information even better. And sure enough, on closer inspection, our brain seems to know that, too. More recent work has shown that we localize an audio-visual stimulus not at, but very close to the visual stimulus (Alais and Burr 2004; Battaglia et al. 2003). Thus, we actually integrate visual and other localizations.

The last decade saw a considerable amount of work which explored the precise way in which we integrate information from different modalities. Specifically, many studies compared human and animal performance to optimal strategies in the sense of Section 1.2.2: Given a model of a sensorimotor task which includes the sensory information available to an agent and the uncertainty of that information, it is often possible to derive the best way information can be integrated. It has been found that natural MSI often behaves similar to such optimal strategies not only in audio-visual localization, but in a range of other multisensory integration tasks as well (Alais and Burr 2004; Battaglia et al. 2003; Ernst and Banks 2002; Hillis et al. 2004; Körding and Wolpert 2004; Landy et al. 2011).[1] The ventriloquism effect and similar phenomena can be explained in this framework by the fact that one sensory modality is often much more reliable for a specific

[1] See Section 3.2.1.

task than others. An optimal integration strategy will thus produce results which are close to those which would have been produced by ignoring the less reliable modalities altogether.[2]

Especially higher animals learn to use their senses to their full extent only after birth, and they have to learn to use them in conjunction as well. It should come as no surprise that an animal like a cat, which is born blind and deaf, does not respond to combinations of visual and auditory stimuli either. However, studies show that development of MSI can continue for long after unisensory stimuli have become available. One way in which human children, for example, do not integrate the same as adults, is that their responses to multisensory stimuli are not as much shorter than those to unisensory stimuli (Neil et al. 2006). They also do not integrate visual and haptic cues consistent with the same model of optimal integration as adult performance (Gori et al. 2008).

Dependence on context is another feature differing in sensory processing between higher and lower animals. Multisensory integration is no exception in that it is not the same in all circumstances: the degree to which the perception of a stimulus in one modality is influenced by a stimulus in another modality is a function of stimulus features, but also of the cognitive state of the observer. Welch and Warren (1980) list historical factors, assumptions about the situation, and attention as important for integration. The 'compellingness' of a cross-modal stimulus, that is, its plausibility as a multimodal stimulus, bears strongly on whether or not the unisensory components are integrated (Jack and Thurlow 1973; Spence 2011; Warren et al. 1981).

2.2 Neurobiology of the Superior Colliculus

The SC is a multisensory brain region par excellence. It is responsible in some cases for the overt responses in MSI discussed in the previous section, and it is cited as a prototypical neural substrate for MSI even in studies dealing with forms of MSI it is *not* concerned with. We will in the following review the most salient and important aspects of SC neurobiology in general and its role in MSI in particular.

2.2.1 Location and Internal Structure

The tectum, a part of the vertebrate midbrain, consists of two symmetrical pairs of bumps (see Figure 2.1): the caudal pair are called the inferior colliculi (ICs), and they play only a supporting part in this thesis.[3] Conversely, the more rostral pair of bumps are the left and the right SC, or optic tecta (OTs), in non-mammals (Stein and Stanford 2013).

Each SC comprises seven layers (Chalupa and Rhoades 1977; May 2006; Sparks and Hartwich-Young 1989; Stein and Stanford 2013, see Figure 2.2). The top three layers are together referred to as the superficial superior colliculus (sSC). Some ambiguity exists with respect to the four layers below that: they are variously either all grouped into

[2] See Section 3.2.1 for a more detailed discussion of optimal strategies.

[3] See Section 10.1.

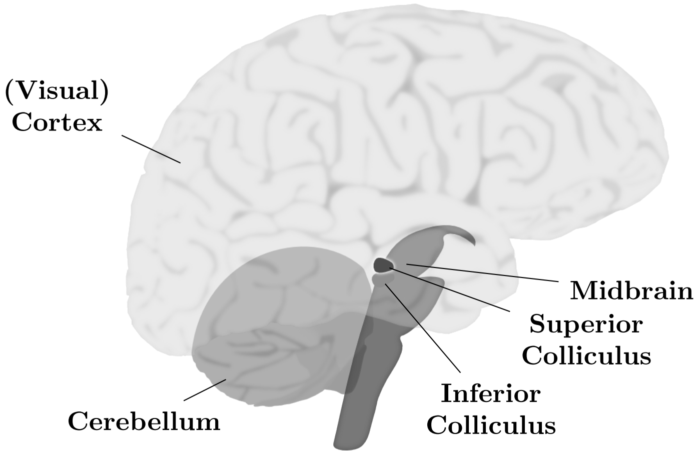


Figure 2.1: The SC's Location in the Human Brain

the deep superior colliculus (e.g. Sparks and Hartwich-Young 1989; Stein and Stanford 2013) or separated into intermediate superior colliculus (iSC) and dSC (e.g. Chalupa and Rhoades 1977; Middlebrooks and Knudsen 1984). In the following, we will adopt a pragmatic terminology. We will refer to the four deeper layers as the dSC, unless stated otherwise, and only distinguish the intermediate superior colliculus (iSC) where required (see Figure 2.2).

DSC and superficial superior colliculus (sSC) are mutually connected: it has been known for some time that sSC layers project to each other as well as to layers in the dSC (May 2006). The fact that the dSC projects to the sSC not only inhibitorily (Lee et al. 2007) but also excitatorily has been ascertained only recently (Ghitani et al. 2014).

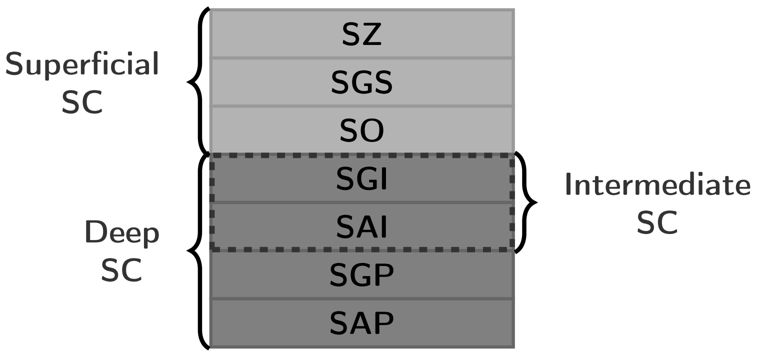
2.2.2 The Motor Function of the Superior Colliculus

It has been known for a long time that electrostimulation of the SC elicits saccades,[4] and that certain neurons in the dSC increase their activity before the onset of a saccade (Robinson 1972). Furthermore, deactivation or lesions of one SC (left or right) or its motor-related efferents will produce an inability to direct the eyes (and the rest of the body) towards stimuli in the contralateral visual field (Sprague and Meikle 1965). Thus, the SC is considered an important part of the oculomotor system.

However, the SC is not alone in this function. Like stimulating or deactivating the SC, stimulating or removing the frontal eye field (FEF) can elicit or lead to deficits in orienting responses (Schiller et al. 1980). While the effects of FEF ablation and some of the effects of SC ablation are temporary, removal of both FEF and SC leads to a drastic and permanent decrease of saccade frequency and range (Schiller et al. 1980). Thus, saccades seem to be elicited by more than one brain region.

Since the general ability to make saccades persists after SC removal, it is also clear that the actual motor signals must be generated somewhere else. Premotor neurons that

[4] (Eye) saccades are fast eye movements.



SZ:	stratum zonale	SAI:	stratum album intermediale
SGS:	stratum griseum superficiale	SGP:	stratum griseum profundum
SO:	stratum opticum	SAP:	stratum album profundum
SGI:	stratum griseum intermediale		

Figure 2.2: Layers and Sections of the SC.

generate the commands for eye movements then relayed to the ocular muscles by motor neurons are located in pons, medulla (horizontal movements), and the rostral midbrain (vertical movements) (Sparks 2002). DSC projections reach these neurons directly.

Another important brain region which is involved in generating eye movements is the cerebellum (Helmchen and Büttner 1995; Robinson et al. 1993). That brain region seems to be particularly important for adaptation of saccades, as lesions to the cerebellum can lead to impaired adaptivity of saccades (Straube et al. 2001; Xu-Wilson et al. 2009).

While the SC's role in generating eye-saccades is the most well-studied, it is also involved in a number of other motor behaviors. These include orienting behaviors of the head and body (Stein and Meredith 1993, p. 102; Gandhi and Katnani 2011). But they also include reaching and other forelimb-related actions (Reynolds and Day 2012; Song et al. 2011; Stuphorn et al. 2000), aversive behavior (Brandão et al. 1994), or even tongue-snapping in toads (Satou et al. 1985).

2.2.3 Neural Activity

The three layers of the sSC are almost exclusively visual (Chalupa and Rhoades 1977; Sparks 1986; Stein et al. 2014; but see Knudsen 1982). Neurons in the sSC have visual receptive fields (RFs), that is, angular ranges from which a visual stimulus evokes a response (Apter 1945; Cynader and Berman 1972). There are differences between species as to the selectivity of sSC neurons for size, motion, direction of motion etc. of stimuli (Cynader and Berman 1972). Another inter-species difference is whether or not RF sizes increase with eccentricity (e.g. hamster: Chalupa and Rhoades 1977; monkey: Cynader and Berman 1972).

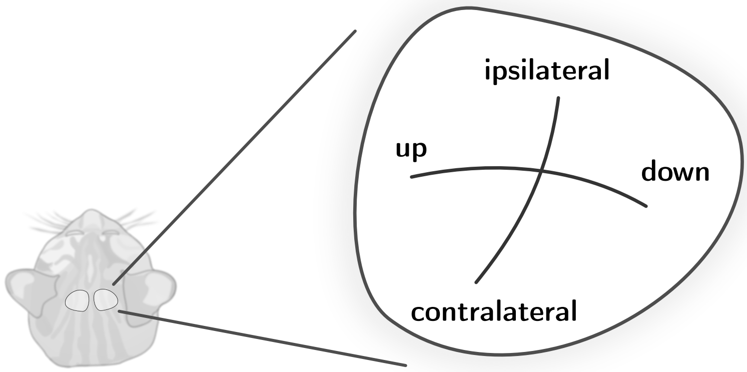
In contrast to the sSC, the dSC is multisensory. There are neurons in the dSC which respond to stimuli in the visual, auditory, somatosensory (Stein 2012a; Wickelgren 1971), and, in species that have them, the infrared, electroceptive, magnetic, and sonar sensory modalities (May 2006; Merker 2007). Many of these dSC neurons respond to stimuli in only one modality, but others are multisensory, and studies of the cat and monkey indicate that, in these species' dSCs, neurons can be found for every combination of visual, auditory, and somatosensory stimuli (Stein 2012a; Wallace and Stein 1996; Wickelgren 1971).

Like the neurons in sSC, dSC neurons show spatial selectivity and often prefer moving stimuli (Horn and Hill 1966; Wickelgren 1971). RFs for stimuli in different modalities tend to overlap in multisensory neurons (Krueger et al. 2009; Middlebrooks and Knudsen 1984; Wickelgren 1971). The size of visual RFs of dSC neurons is typically greater than that of sSC neurons (Stein and Stanford 2013; Stitt et al. 2013; Wickelgren 1971).

Neural responses in the dSC to spatially and temporally coincident cross-sensory stimuli can be much stronger than responses to unisensory stimuli (a phenomenon called 'enhancement'). In fact, they can be much greater than the sum of the responses to either stimulus alone ('superadditivity'). The increase in strength of a neural response to one stimulus due to another stimulus in a different modality tends to be greatest for weak stimuli and least for strong stimuli (the 'principle of inverse effectiveness', see e.g. Meredith and Stein 1986a; Stanford et al. 2005; Stein et al. 2014). If, on the other hand, two stimuli in the same or in different modalities are spatially or temporally separated, the responses to both are weaker than to either stimulus alone ('depression', see e.g. Kadunce et al. 1997). Enhancement and depression are together referred to as the 'spatial principle'. The spatial principle and the principle of inverse effectiveness are considered hallmarks of multisensory integration.

The enhancement and inverse effectiveness can also be observed in the temporal dynamics of responses to uni- and multisensory stimuli. Neurons which exhibit enhancement in the size of responses to multisensory stimuli compared to responses to unisensory stimuli also typically respond earlier to multisensory stimuli. Interestingly, response profiles may comprise superadditive, additive, and subadditive phases: they have superadditive phases even for cross-sensory stimuli whose unisensory components are strong, for which the aggregated response can be additive (Rowland and Stein 2013).

Neural responses in the dSC are affected by cortical functioning. For example, deactivating certain regions of anterior ectosylvian cortex (AES) or lateral suprasylvian cortex (LS) can completely eliminate responses of dSC neurons to stimuli in that modality to which the region in question is sensitive (Wallace and Stein 1994). Also, in contrast to sSC neurons, dSC neurons do react to color contrast—probably because color-related input is relayed to the dSC via a geniculo-cortical route (White et al. 2009). Thus, Wallace and Stein (1994) argue that some dSC neurons seem to receive sensory input only via cortex. Furthermore, superadditivity is contingent on a functioning association cortex: deactivating AES or the rostral part of the lateral suprasylvian cortex (rLS) in cats completely eliminates superadditivity (but not responses to input from different modalities) (Stein et al. 2014; Wallace and Stein 1994).



Each SC contains a two-dimensional sensorimotor map of external space. In that map, the rostrocaudal and mediolateral axes roughly correspond to azimuth (ipsi→contralateral) and elevation (up→down), respectively. The mapping is distorted, and size, shape, and location vary between species (Stein and Meredith 1993, p. 88). (Not to scale.)

Figure 2.3: Schematic Drawing of SC Orientation and Topology.

2.2.4 Spatial Organization

Those sensory modalities which are represented in the SC are spatial: part of the description of a stimulus perceived visually, acoustically, through somesthesia, or one of the other modalities is the stimulus' location or its direction from the perceiver. For most of these senses, the location of or direction to the stimulus is reflected in the identity, and in fact, the location, of the peripheral neurons being stimulated (Knudsen 1982). For example, a visual target stimulates neurons in different places in the retina depending on its location. The pressure-sensitive neurons stimulated by an object touching our skin are localized to the point of contact. The sensory maps—mappings from the location of the neurons being stimulated to the location of or direction to the stimulus in the outside world—which arise naturally in the peripheral neural populations in this way are preserved in many primary brain regions through topographic projections (see Kaas 1997, for a review).

The SC is such a brain region with a topographic representation of sensory space (see Figure 2.3). The retina projects onto the sSC in a topographic manner (Apter 1945; King 2013). RFs in deeper layers are in register across modalities (see above) and they roughly follow the spatial organization of the layers above them (Dager and Hubel 1975; King 2013). Sensory maps in the SC, and their registration across modalities, have been demonstrated in mice, cats, monkeys, guinea pigs, hamsters, barn owls, and iguanas (Sparks 1988).

As noted earlier, localized electrostimulation of dSC neurons can produce saccades.

The size and direction, of such a saccade, or, similarly, the direction of gaze after the saccade (see below), depends on the site stimulated. Typically, these saccades go into that general direction in which natural stimuli would lead to activation in the area that was electrically stimulated (Robinson 1972). In other words, sensory maps are in register not only with each other, but also with the motor map in the dSC (Sparks 1988; Stein et al. 2014). Thus, the position of a stimulus and the characteristics of an eye movement evoked by dSC activation are not encoded in the spiking behavior of a single neuron or neural assembly. Instead, we say that they are represented by a population code in the dSC: they can be read out of the activity of the entire population.

In the past, there has been some debate about the frame of reference of the various sensorimotor maps in the SC. In the mostly visual sSC,[5] the map of visual space is retinotopic (Stein and Stanford 2013). However, the sensorimotor coordinate systems in deeper layers might have been head- or body-centered. They are today thought to be largely retinotopic as well (Stein and Stanford 2013): one reason is that saccades typically bring a target into the center of the visual field with little to no error, regardless of previous eye position and the modality in which they are perceived (Hartline et al. 1995). Another reason is that moving eyes or pinnae, and thereby the map of visual space with respect to that of auditory space, shifts the auditory RFs of dSC temporarily (Hartline et al. 1995) or even permanently (Knudsen 1983, see Section 2.2.7) to keep their representation in the dSC in register. Finally, the gaze shifts induced by dSC electrostimulation are more consistent with the hypothesis that the motor map encodes target gaze positions in retinotopic coordinates than target gaze positions in head-centered coordinates or gaze displacement (Klier et al. 2001).

Therefore, and because the dSC is not only involved in orienting movements, but also in quite a number of other motor behaviors,[6], and spatial attention[7] dSC activity may not be best thought of as an encoding of a motor program for a saccade. Instead, dSC activity may code for the position of an object of interest in the visual field, which may, for example, become the target of a saccade. The actual motor commands are then the result of modulatory computations downstream of the SC, for example in pons and cerebellum (Gandhi and Katnani 2011; Klier et al. 2001; Sparks 2002).[8] Especially the latter is a good candidate considering the detrimental effects of cerebellar deactivation or lesions on saccades (e.g. Ritchie 1976; Robinson et al. 1993) and in particular on saccade adaptivity (e.g. Straube et al. 2001; Xu-Wilson et al. 2009).

2.2.5 The Superior Colliculus' Role in Attention

The SC is a phylogenetically old structure. It is often thought of as just one station in a rather automatic sensorimotor transformation circuit. The fact that it is involved in the cognitive phenomenon of attention may therefore be a little surprising.

[5] The sSC is exclusively visual in cats and apparently most other animals (Middlebrooks and Knudsen 1984), but strongly audio-visual in owls (Knudsen 1982).

[6] See Section 2.2.2.

[7] See Section 2.2.5.

[8] See Section 2.2.6.

The SC is related to attention in two ways: first, it seems to be involved in allocating spatial attention (Krauzlis et al. 2013). This is evident in the enhancement of responses of neurons in visual cortex following microstimulation of SC neurons with the same RFs as the cortical neurons (Born et al. 2012; Cavanaugh et al. 2006; Müller et al. 2005). Also, it seems impossible to plan a saccade to one location and focus spatial attention on another (Born et al. 2012; Deubel and Schneider 1996). Finally, deactivation of one SC can lead to hemi-neglect, which is similar to lack of attention to part of the visual field (Nummela and Krauzlis 2010; Sprague and Meikle 1965).

Second, SC activity is affected by attention. Spatial attention can enhance responses of neurons whose visual RFs overlap the attended region (Goldberg and Wurtz 1972; Ignashchenkova et al. 2004; Schneider 2011). Feature-based attention, that is, attention to color, shape, size etc. of a stimulus, enhances responses across the visual field in cortical neurons which are selective to these features (Born et al. 2012; Maunsell and Treue 2006). An functional magnetic resonance imaging (fMRI) study testing the effect of feature-based attention in SC and other subcortical structures produced inconclusive evidence: while the SC seemed to respond strongly to switching attention between different features (motion vs. color), but a general preference for either of them could not be shown (across subjects) (Schneider 2011). The differences between subjects and the fact that SC receives motion-related input directly, but color-related input only via cortex[9] makes the results difficult to interpret. Even if color-related attention has no effect on SC, attention to other features—possibly non-visual ones—might.[10]

A third line of evidence hinting at the SC's role in attention is the number of brain regions projecting to the SC which themselves have been implicated with attention. These include supplementary eye field (SEF), FEF, dorsolateral prefrontal cortex (DLPFC), and lateral intraparietal cortex (LIP) (Bon and Lucchetti 1997; Buschman and Miller 2007; Kastner and Ungerleider 2000) as well as AES, in cats, which has been linked to selective attention (Dehner et al. 2004).

These facts inspire the 'premotor theory of attention,' the theory that states that orienting the body and directing spatial attention share implementing neural circuitry (Rizzolatti et al. 1987), in particular, the dSC (Sprague and Meikle 1965). Both visual attention and sensory orienting direct cognitive resources at some objects at the expense of others (Born et al. 2012). The parallels between them, and possibly the shared neural hardware, thus may come at a surprise, but it seems natural, in hindsight. They are important and similar ways in which we 'probe' our visual world (O'Regan 1992).

2.2.6 Connectivity

It is fair to say that the SC is a hub within the brain. In fact, it is connected directly or indirectly to most parts of the brain (May 2006). Furthermore, there are considerable differences between the connectivity, and, even more so, in our knowledge of the connectivity between the SC and other brain regions. In this thesis, we are only interested

[9] See Section 2.2.6.

[10] See Chapter 6.

in the general pattern of input and output. Even more than in the other parts of this section, we will therefore be brief in our description of SC connectivity and provide only a very short overview.

SC afferents. SSC receives projections directly from the retina and from visual cortex (e.g. May 2006; Pollack and Hickey 1979; Schiller et al. 1979). Most of the retinotectal connections are contralateral. However, in primates and some other species, significant projections from the nasal part of the retina are ipsilateral (May 2006; Pettigrew 1986). At least in cats and primates, these projections are almost exclusively via the color-insensitive Y- and W-like ganglion cells (Fukuda and Stone 1974; Schiller et al. 1979; White et al. 2009).

Primary auditory input comes mainly from the IC (more precisely from the external nucleus of the inferior colliculus (ICx)). The sensory connections carrying this input reach the intermediate and deep layers of the SC (DeBello and Knudsen 2004; Edwards et al. 1979; Knudsen and Knudsen 1983; May 2006). DSC receives tactile localization-related inputs from the trigeminal nucleus (Stein et al. 2014).

Additional incoming connections from subcortical regions originate in the parabigeminal nucleus (nucleus isthmii in non-mammals), pretectum, dorsal and ventral lateral geniculate, and cerebellum (May 2006).

A number of cortical regions provide input to the SC as well. Among them are parts of visual cortex (White et al. 2009), auditory cortex (Bajo et al. 2007), primary somatosensory cortex (Triplett et al. 2012; Wise and Jones 1977), and motor cortex (Day 2014; Fries 1985; see also Fries 1984). Many non-primary cortical areas project to SC, often localized to only superficial or only deeper layers (May 2006), including the attention-related areas mentioned above, SEF (Huerta and Kaas 1990), FEF (Fries 1984), DLPFC (Selemon and Goldman-Rakic 1988), LIP (Colby and Goldberg 1999), and AES (Wallace et al. 1993; see also May 2006).

SC efferents. The deeper layers of the SC project strongly to the brainstem; spinal cord, especially to those regions involved in moving eyes, ears, head and limbs; to sensory and motor areas of thalamus; and cerebellum (Stein and Stanford 2013). Both deeper and superficial layers of the SC send projections to different regions in IC, especially ICx (Hyde and Knudsen 2000), creating a loop from SC to IC and back.

The lateral geniculate nucleus (LGN) receives topographic input from sSC, and the same regions in LGN which receive input from sSC project to visual cortex (Harting et al. 1991). Another indirect route from sSC to cortex may go through the pulvinar to the middle temporal visual area (MT) (Berman and Wurtz 2010).

Additional outgoing connections to subcortical regions terminate in the parabigeminal nucleus/nucleus isthmii (see above), pretectum, midbrain reticular formation, regions in thalamus, and hypothalamus (Benevento and Fallon 1975; May 2006)

For in-depth studies and a review, see Benevento and Fallon (1975) and Fries (1984, 1985), and May (2006).

2.2.7 Development

The maturational state of the SC in newborn animals depends on the species. For example, the response properties of sSC neurons in mice do not develop much after birth (Wang et al. 2010), and newborn macaque monkey's dSCs contain multisensory neurons (Wallace and Stein 2001). In contrast, cats' sSC neurons develop considerably, and their dSC neurons do not even start responding to auditory stimuli until days, to visual stimuli until weeks after birth (Kao et al. 1994; Stein and Stanford 2013).

At least a rough retinotopic organization of visual RFs in sSC seems to exist in many species. Chemical markers guiding neural growth are a probable mechanism for their early development. Later refinement of the retinotopic maps in the SC depends on sensory and in particular cross-sensory experience (Drescher et al. 1997; Fraser 1992; Stein and Stanford 2013). Whether the same is true for topographic maps in other modalities, in the SC, is yet unknown, but it is likely (Stein et al. 2014).

One way in which response characteristics of SC neurons change after birth is a shrinking of their RFs (Stein et al. 2014). Another important aspect is the way they integrate multisensory input. While neurons in the adult dSC often respond superadditively to stimulation from multiple senses, those in young animals' dSCs do not, even if they are responsive to more than one modality (Stein and Stanford 2013).

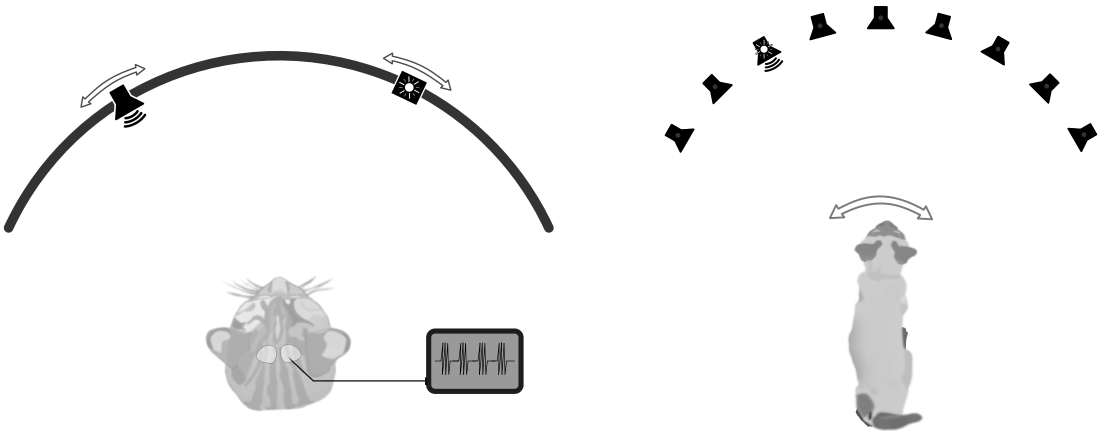
From experiments in which animals were reared under abnormal sensory conditions, we know of striking demonstrations of plasticity in the SC. The map of auditory space which develops in the OT of owls which are raised either with optic prisms fitted to their eyes or one ear partially occluded is shifted drastically, compared to that in normally raised owls. This shift largely compensates for the translation of the perceived visual world compared to the auditory world, introduced by prisms or ear plugs (Bergan et al. 2005; Knudsen 1983; Knudsen and Brainard 1995).

Similarly, raising cats in an environment in which visual stimuli always have a constant spatial offset from auditory stimuli also leads to a compensatory shift of the representation of auditory space in the SC (Wallace and Stein 2007). In contrast, raising them with only visual and auditory, but no systematic cross-modal stimuli at all, prevents the development of the spatial principle of MSI.[11,12] The mere exposure to concomitant audio-visual stimuli—with or without a spatial offset—is enough for SC neurons to learn multisensory integration, though tuned to the specific spatio-temporal conditions of the cats' environment (Wallace and Stein 2007; Xu et al. 2012). No behavioral relevance of these stimuli is needed for that (Xu et al. 2012).

A final, extreme example of the flexibility of map registry in the OT is reported by Law and Constantine-Paton (1981), who implanted tadpoles with additional eye primordia. These tadpoles grew up to be three-eyed frogs in which the retinae of the additional eyes projected to one of the OTs. The maps of visual space were relatively normal, in these frogs, and the OTs developed ocular dominance stripes: alternating stripes which were either innervated by one eye or the other. Such dominance stripes are common throughout the central nervous system in animals in which some parts of the retina

[11] See Section 2.2.3.

[12] Unfortunately, the authors do not report on map register in these cats.



(a) Neurobiological Experiments on SC.

(b) Behavioral Experiments on MSI.

> **Left:** An animal's head is fixed in the experimental apparatus and it is presented moving visual and auditory stimuli. At the same time, single- or multi-unit recordings are made of SC activity.
> **Right:** An animal is trained to respond to uni- and multisensory stimuli. The animal's behavior is monitored and compared to the results of the neurobiological experiments.

Figure 2.4: Paradigmatic Experiments on SC and MSI.

project contralaterally and others project ipsilaterally, but they are very unusual in frogs.

The simulation and robotic studies of SC models described in the later sections of this thesis can be seen as parallels to the neurobiological and behavioral studies conducted on cats, owls, and other animals, which gave us much of the knowledge summarized above, and which are depicted prototypically in Figure 2.4.

2.3 Focus and Convention in this Thesis

As stated above, the goal of this section was to provide the reader who is not a member of the research community studying the SC with background information in the form of paradigmatic knowledge. We therefore included details which are important for a general understanding of the biology of the SC and its role in MSI, but which are not central to the research presented in the following sections. For clarity, we will briefly state on what we will focus in our modeling in Part I of this thesis and thus implicitly also in Part II.

To the computer scientist, the SC is most interesting as an information processor. Its capability to integrate information from different sensory sources, taking into account their reliabilities, is especially interesting. Furthermore, its learning to integrate information and the reliability of each sensory source, apparently even without feedback, is a fascinating feature. Lastly, the influence of situational parameters—in the form of cognitive content—on its functioning has been observed, but the underlying processes have not been worked out sufficiently, considering the practical importance of the phenomenon.[13] Interest in these areas certainly overlaps in the scientific fields relevant in this thesis, computational neuroscience and computer science, and we will make them the focus of our research.

As perhaps the most striking feature of the SC, spatial organization and sensory map registry will play an important role in our efforts. Our models will produce and exploit spatial organization and it is one hypothesis of our work that it is an important strategy for implementing MSI. This hypothesis will be tested in Part II of this thesis.

To make our task feasible, we will abstract from certain details. One is the fact that there are two SCs in the midbrain. We will treat the two SCs as one structure which forms a continuous map of the sensory world. Another simplification will concern the internal structure of the SC: we will not distinguish between the different layers of the SC and, in fact, concern ourselves mostly with the multisensory dSC.

The timing aspects of MSI and of neural responses in the SC will not be considered in this thesis. They are important, and shorter response latencies may well be one of the greatest benefits of audio-visual integration under normal sensory conditions.[14] We will still exclude time from our considerations as it would overly complicate our work.

The representation of the connectivity of the SC will be greatly simplified in our modeling. We will not model more than one sensory population per sensory modality or kind of cortical input. This is because we want to extract the principles of MSI in the SC independently of the details of the respective neural codes. Since we are interested mainly in information processing, generating motor responses from SC activations will also not be dealt with. Instead, we will assume that the motor-related projection targets of the SC alone translate these activations into appropriate motor commands. All of these are not uncommon simplifications and most of the studies reviewed in Section 3.2.2 make them all.

Certain details of SC neurobiology will be important as markers of our models' adequacy in Part I. These include the spatial principle and the principle of inverse effectiveness, the effects of attention on neural activity in the SC, and spatial organization. The latter will thus play a double role as both an effect we want to reproduce and an effect which, when reproduced, we believe facilitates learning of MSI.

[13] see Chapter 6.

[14] See Section 12.2.

Part I

Modeling Multisensory Integration in the Superior Colliculus

We have stated in Section 1.2.1 that the goal of this thesis is to learn how the dSC integrates multisensory information for the double purpose of understanding more about this important brain region and discovering techniques which can also be applied in AI and especially in robotics. This part of the present thesis will focus on the first part: understanding what the dSC does, and how, by modeling it. Later, in Part II, we will take a pragmatic look at the insights gleaned here.

Natural MSI is an area of research which is as intriguing as it is broad. MSI is intriguing, because it is such a wide-spread phenomenon, being present in virtually all organisms possessing multiple means of sensation (Stein and Meredith 1993) and because it is so constitutive for our perception of the world, as becomes especially apparent on the rare occasion that it fails, like in the ventriloquism and McGurk effects (Chen and Vroomen 2013; McGurk and MacDonald 1976), among other reasons. MSI is a broad area of research, because it can be studied from so many different aspects: MSI can be studied in different species; involving different sensory modalities and different stimuli; in the time or the spatial domain; on the physical, behavioral, or neural level; how it develops ontogenetically and phylogenetically; how it can be modeled and understood physiologically, mathematically, or algorithmically; or from the point of view of bottom-up or top-down processes.

One reason why *we* are interested in MSI in the dSC is that we want to learn about how the dSC accomplishes its task, on the algorithmic level.[15] For the purposes of this thesis, we will take the view that that task is to use the neural responses of its afferents to uni- or cross-sensory stimuli to localize these stimuli and initiate motor responses accordingly (Stein et al. 2014). This view is taken wherever maps of sensory space in the SC and their retinotopicity or mototopicity are discussed (e.g. Kaas 1997; King 2013; Sparks 1988). It is a particularly interesting view from the perspective of computer science, as integrating information from different sources is an important part of many information processing tasks. While it could be argued that the SC does far more—or less—we will work on this assumption.[16]

We have decided, in Section 1.2.2, that we want to model the dSC at the algorithmic level, which means that we are interested in understanding how the dSC performs its task in terms of representations and operations. That is another choice we make in deciding which aspects of the dSC to study. Finally, MSI is at least partially learned, and there is evidence suggesting that learning of MSI is at least partially unsupervised (see Section 2.2.7). This is an interesting feature of the dSC, and we want to include that in our model: first, because it is important from a biological point of view, but also because what we can learn about it may have useful technical applications. After all, like animals, artificial systems and especially robots generally have many more opportunities to learn from unlabeled than from labeled data.

Also like animals, artificial systems have to cope with uncertainty. Recent developments have led researchers to believe that humans and other animals use information about uncertainty in sensory processing (e.g. Alais and Burr 2004; Landy et al. 2011;

[15] See Section 1.2.2.

[16] See Section 7.1.2 for a justification.

Ma et al. 2006, see also Section 1.2.2). How uncertainty is handled in sensory processing is an important question both from the biological and engineering perspective. We will therefore develop a model which explicitly explains how the dSC might learn to handle uncertainty.

In contrast, we will leave aside evolution of MSI, MSI in the time domain, MSI in any other brain region than the dSC, and a large part of the neurobiological detail of dSC neurons and their connectivity where that goes beyond the algorithmic view of the dSC, past the interface to and deep into the neural implementation. Fascinating aspects of natural MSI as these are, they are beyond the scope of this thesis.

In the rest of this modeling part, we will first review comparable models of MSI in general, and models of the SC in particular, in light of the choices laid out above. We will then go on to introduce the ANN which will serve as the core of our model. Then, we will report on simulations which show that our algorithm can be trained on biologically plausible input and that it reproduces important aspects of natural MSI in the dSC on the behavioral and neurophysiological level. We will focus on modeling and leave technical applications to Part II.

3 Previous Models

In this chapter, we will review previous work on modeling the SC. The SC has been modeled before and many of the models proposed by other researchers are related to ours in one way or another. The history of SC modeling is long and models have been proposed to account for many different aspects of SC phenomenology. We therefore will not aim for completeness in this review, but instead first attempt to offer an overview of the various kinds of SC models and other models of basic natural MSI and then focus on those which are comparable or even related to our model. Specifically, we will point out where we believe the literature is lacking so far to motivate our own approach to be laid out in the next sections. Some of the models presented here are not models of the SC explicitly, but of one of its functions.

3.1 Models of the Superior Colliculus' Oculomotor Function

It has long been recognized that the SC is an important brain structure in guiding eye movements.[1] One class of SC models has therefore explored the possibilities of how the sensorimotor processing in and around the SC could lead to the behavior and neural dynamics observed during saccades. Models in this class engage such problems as mapping from spatial sensory codes to temporal motor codes, intrinsic SC connectivity leading to the observed dynamics, and different placements within feedback loops within the oculomotor system. Girard and Berthoz (2005) have a review of SC models focusing on such models. Especially early models of the SC concentrated on its oculomotor function, but models in that area are still being proposed (e.g. Saeb et al. 2011; Tabareau et al. 2007). Since the focus of this thesis is on sensory processing, and in particular on multisensory integration, we will not review models of saccade generation in detail.

3.2 Models of Multisensory Integration

The dSC integrates multisensory stimuli. Together with MSI on the organism level as well as in other brain regions, this aspect has attracted considerable attention since the late 1980s (e.g. Meredith and Stein 1986b; Sparks 1986; Stein and Meredith 1993) and especially in recent years (e.g. Spence and Driver 2004; Stein 2012b; Trommershäuser et al. 2011). Models of MSI and in particular dSC models with a focus on the integration

[1] See Section 2.2.2.

of stimuli by now form a large and diverse corpus (see Ursino et al. 2014, for a review with a more general perspective).

3.2.1 Behavioral, System-Level Models

MSI has been studied at the purely behavioral level, and mathematical models ('computational' models in Marr's (1983) sense) have been developed. One particularly successful model has been a simple MLE model of MSI. This model predicts that MSI attempts to minimize uncertainty about world properties like the location of a stimulus or the size of an object by integrating estimates from different modalities. Assuming that the error of each of these estimates is normally distributed with known variability, the optimal strategy (the one that minimizes uncertainty) is to linearly combine the estimates with weights determined by the variability of the noise distribution (Ghahramani et al. 1997): suppose one has m estimates $(x_i)_{i=1}^m$ of some world property X from m different modalities. If the noise in each of these modalities is normally distributed around the true value x and if the variability of the noise distribution in modality i is σ_i^2, then the combined estimate x_o which minimizes overall variability is given by

$$x_o = \frac{\sum_{i=1}^m \sigma_i^{-2} x_i}{\sum_{i=1}^m \sigma_i^{-2}}. \tag{3.1}$$

The behavioral model based on Equation 3.1 has been successfully used to describe many instances of natural MSI (Landy et al. 2011), perhaps most prominently integration of auditory and visual cues for localization (Alais and Burr 2004).

Although the linear MLE model has been very successful at explaining a wide range of experimental results, it also has failed in some instances. Two other behavioral models of MSI which may account for human behavior in some instances are model selection and probability matching (Wozny et al. 2010). Model selection, in MSI, is the strategy of maintaining two or more models of the situation and generating the response to a cross-modal stimulus from that model under which the current stimulus is most likely. Thus, one could perceive an auditory and a visual stimulus and decide that they either have a common source and should thus be integrated (e.g. using an MLE model) or that it is more likely that they are independent from one another and need to be localized individually. This strategy has also been termed 'causal inference' (Fetsch et al. 2013; Körding et al. 2007; Shams and Beierholm 2010) and variants of mathematical formulations have been proposed (Roach et al. 2006; Sato et al. 2007).

Probability matching is the strategy of choosing one alternative out of a set of possible responses to a (single or cross-modal) stimulus with a probability proportional to the probability of that alternative being the correct response. This is a suboptimal strategy if the goal is to minimize the (mean squared) error of the estimate but it leads to a form of exploration of actions and consequences and can thus promote learning (Wozny et al. 2010).

All behavioral models discussed so far have been purely static models in that they do not explain how an agent might acquire the knowledge required to instantiate any of the strategies. Weisswange et al. (2011) offer an explanation on the basis of reward-mediated learning. In their model, a simple ANN learns to integrate a simulated auditory and a simulated visual localization of a stimulus so as to optimize the stochastic reward given in each round. Weisswange et al. demonstrate that this can lead to behavior consistent with either the linear MLE or the causal inference model, as long as stimulus reliabilities are kept constant, and that it quickly adapts when those reliabilities change. We count the Weisswange et al. (2011) model as a behavioral model because the biological plausibility of the ANN which implements it (and indeed the quality of the ANN's input) is neither discussed by the authors nor at all obvious.

The behavioral models discussed above are important in the context of this thesis: they describe natural MSI computationally. In accord with our goals set forth in Section 1.2.2, we will use this computational description as a target to work towards from the algorithmic perspective. Especially the linear MLE model and model selection will be important in Sections 5 and 6, respectively.

3.2.2 Neuroscientific Models

Neuroscientific models of MSI can be classified roughly along two axes: one axis is the distinction between neuron-level and network-level models. The other axis separates interpretative, theoretical models from models which focus on integrating neurophysiological detail and reproducing phenomenology.

Neurophysiological Single-Neuron Models.

One of the simplest single-neuron dSC models is the first of four models discussed by Rowland et al. (2011). In that model, sensory and cortical input is simply summed and passed through a sigmoid squashing function. That squashing function leads to inverse effectiveness:[2] The sum of weak inputs generally falls into the super-linear part of the sigmoid and thus produces a superadditive response.

In their 2003 model, Anastasio and Patton show how a two-stage learning procedure based on the self-organizing map (SOM) algorithm (see below) can produce cortical modulation of enhancement and depression in a network in which cortical and sensory input are combined multiplicatively. Interestingly, the competitive aspect of SOM learning is not used to learn to distinguish stimuli from different directions but stimuli with different combinations of sensory modalities. The topology of the SOM is not taken to correspond to the sensorimotor map of the dSC[3] and neurons' responses are studied only individually which is why, for our purposes, the Anastasio and Patton (2003) model fits in better with the single-neuron than with the network models. The Anastasio and Patton (2003) model is the second model discussed by Rowland et al. (2011).

The third model from Rowland et al. (2011) is the one the authors previously advanced in 2007. In that model, axons from (visual and auditory) sensory neurons and

[2] See Section 2.2.3.

[3] See Section 2.2.4.

from unisensory regions in association cortex connect to different compartments of target multisensory dSC neurons. Superadditivity arises in that model because axons from different unisensory regions in association cortex connect to the same compartment and their presynaptic activities interact in non-linear ways. Since axons from visual and auditory sensory neurons connect to different compartments, they interact only additively. The result is that deactivating association cortex extinguishes superadditivity.[4]

Finally, Rowland and Stein (2013) present a quite different single-neuron model, which is based on the leaky-integrate-and-fire neuron. Their model neuron receives simulated noisy sensory and cortical input, which is excitatory, and, with a certain delay, global inhibitory input. The timing of excitatory and inhibitory signals, together with the computations of the leaky-integrate-and-fire neuron model allow the model to reproduce the temporal dynamics of actual multisensory dSC neurons: at the beginning of the response to a multisensory stimulus, the activity of the model neuron is superadditive. With time, that superadditivity is reduced and the response becomes additive and even sub-additive. The model parameters can be fine-tuned to reproduce with some accuracy the time course of responses in real dSC neurons.

The relevance in neurophysiological single-neuron models for our work is that of a benchmark and a source of constraints. The models presented here are specifically designed to reproduce phenomenology and to be consistent with current neuroanatomical knowledge. While computation-level modeling cannot be expected to surpass models like these in biological plausibility, it should be the aim at least to approach their reproduction of natural response patterns. At the same time, these models map out the computational possibilities which may be presumed by more interpretative models.

Theoretical Single-Neuron Models.

Both the models proposed by Anastasio et al. (2000) and Patton and Anastasio (2003) as well as that proposed by Colonius and Diederich (2004) suggest that each dSC neuron's firing should be interpreted as encoding the probability of a stimulus being in the neuron's RF given sensory input. Anastasio et al. (2000) show that this hypothesis can explain multisensory enhancement: input from more than one modality decreases the uncertainty about whether a stimulus is in a neuron's RF. Thus, given multisensory information, a neuron in whose RF is a cross-sensory stimulus will tend to be more certain of that fact and respond with greater activity.

Wickelgren (1971) and later Wallace and Stein (1996) have found that dSC neurons in the same animal show very diverse patterns of responsiveness to input from different modalities.[5] Patton et al. (2002) as well as Colonius and Diederich (2004) also explain this finding in light of above hypothesis: according to Patton et al. (2002), a neuron might receive enough input from one modality to decide whether a stimulus is in its RF with sufficient certainty. It may thus 'preserve' specificity to one modality by ignoring input from other modalities. Colonius and Diederich (2004) go a bit further in theorizing

[4] See Section 2.2.3.

[5] See Section 2.2.3.

that a unisensory neuron cannot be confused by input from another sensory modality and is thus more effective at detecting unisensory stimuli.

In our own modeling, we will make a similar assumption about the meaning, as it were, of dSC neurons' activities. However, in contrast to the approaches described above, we will be interested in how the necessary knowledge can be acquired and how the topological organization seen in the dSC can arise.

Neurophysiological Network Models.

A relatively early model of audio-visual map alignment and realignment in the dSC was presented by Rucci et al. (1997). Their model focuses on the changes in the barn owl OT brought about by fitting the owl with optic prisms.[6] In addition to the OT, it comprises modules of the retina, central nucleus of the inferior colliculus (ICc), ICx, a motor plant and a value system. That value system is driven by motor error and it modulates learning of the weights in neural connections between ICc and ICx as well as those between OT and motor system. Thus, the OT module is not adaptive in Rucci et al.'s (1997) model. Adaptive aspects of their model are only a) the translation from the tonotopic representation in ICc to the retinotopic representation in ICx and b) generation of motor commands from OT activity.

Their assumption of value-dependent learning in ICx allows Rucci et al. (1997) to explain the formation of a retinotopic map without visual input. While such value-dependent learning is still a possibility, it has more recently been found that the ICx (in owls) receives topographic projections from the sSC and those projections may be involved in map formation in the ICx (Gutfreund and King 2012; Hyde and Knudsen 2000).[7]

The dSC model proposed by Ursino et al. (2009) is a simple network model. It consists of three populations of neurons: two sensory and one modeling the dSC. The neurons in these populations are modeled using differential equations. The hand-crafted connectivity between the sensory populations and the dSC population as well as the connectivity within the dSC are such that the network produces multisensory enhancement, inverse effectiveness, superadditivity, and cross- and within-modality depression.

Cuppini et al. (2012) present a highly detailed network model of the dSC. That model extends its close relatives, the last of the four models discussed by Rowland et al. (2011, see above) as well as the model proposed by Cuppini et al. (2010, 2011), and adds Hebbian learning mechanisms. A major difference to the Ursino et al. (2009) model is the inclusion of neural input from cortical areas as well as explicit modeling of populations of inhibitory interneurons. Because of these additions, dSC neurons in this model seize to integrate cross-sensory stimuli when cortical inputs are deactivated.[8] The model thus manages to reproduce many features of dSC neurophysiology including several experimentally observed aspects of the developmental time course of MSI in the dSC. However, the authors do not offer a functional interpretation of the—somewhat

[6] See Section 2.2.7.

[7] See Section 2.2.6.

[8] See Section 2.2.3.

speculative—specifics of their model and, in fact, do not analyze their network from a behavioral point of view.

Theoretical Network Models.

The facts that information is encoded in population codes in many parts of the brain (including the dSC, see Figure 2.2.4), that sensory information is noisy and thus sensory processing is statistical inference, and that some ANNs naturally perform statistical processing lead to the idea that population codes might not only encode the brain's best guesses at the properties being encoded, but PDFs over those properties (Barber et al. 2003). Population codes which encode PDFs are referred to as PPCs (Pouget et al. 2003).

There have been a number of suggested models for the neural implementation of computation on PPCs. Some, like the ones proposed by Beck et al. (2008), Cuijpers and Erlhagen (2008), and Ma et al. (2006) assume that neural input is already encoded in a PPC. Other studies, by Barber et al. (2003) and Jazayeri and Movshon (2006), have proposed ways in which neural networks can approximate and then encode PDFs from sensory input. Yet these models do not cover learning but rely on hard-wired connectivity and parameterization. Zhou et al. (2011) and Bauer et al. (2012a) did present unsupervised learning algorithms which produced spatial organization and neural populations computing PDF. However, both assumed (biologically implausible) Gaussian noise in neural input activities and the computations required of their neurons were not easily implementable in biological neurons.

The approach taken by Ma et al. (2006) and Beck et al. (2008) makes use of the fact that neural noise tends to be what they call 'Poisson-like', meaning that the variability of a neural response to a stimulus is proportional to its magnitude (Butts and Goldman 2006; Tolhurst et al. 1983; Vogels et al. 1989). Given that, two population codes can be optimally integrated by simply adding them. This contrasts with observations of the response of multisensory neurons in the dSC and elsewhere, whose responses can be strongly super-additive. Fetsch et al. (2013) reconciled the theory with the biology, arguing that population codes may be normalized, thereby performing optimal integration while keeping each neuron's response within its dynamic range. Ohshiro et al. (2011) showed that such a normalization model can reproduce a host of different neurophysiologically observed phenomena. Like Ma et al.'s (2006) and Beck et al.'s (2008) model, Ohshiro et al.'s (2011) model only produces an integrated PPC if its inputs are already PPCs and it does not learn.

The model we will present in the later sections is an unsupervised learning model, it reproduces neurophysiological phenomenology, it approximately computes a PPC from non-PPCs, and it makes only very light assumptions on neural noise and tuning functions. The ANN our model is based on borrows heavily from the SOM algorithm, which is why we discuss SOMs and SOM-based models of MSI separately in the following.

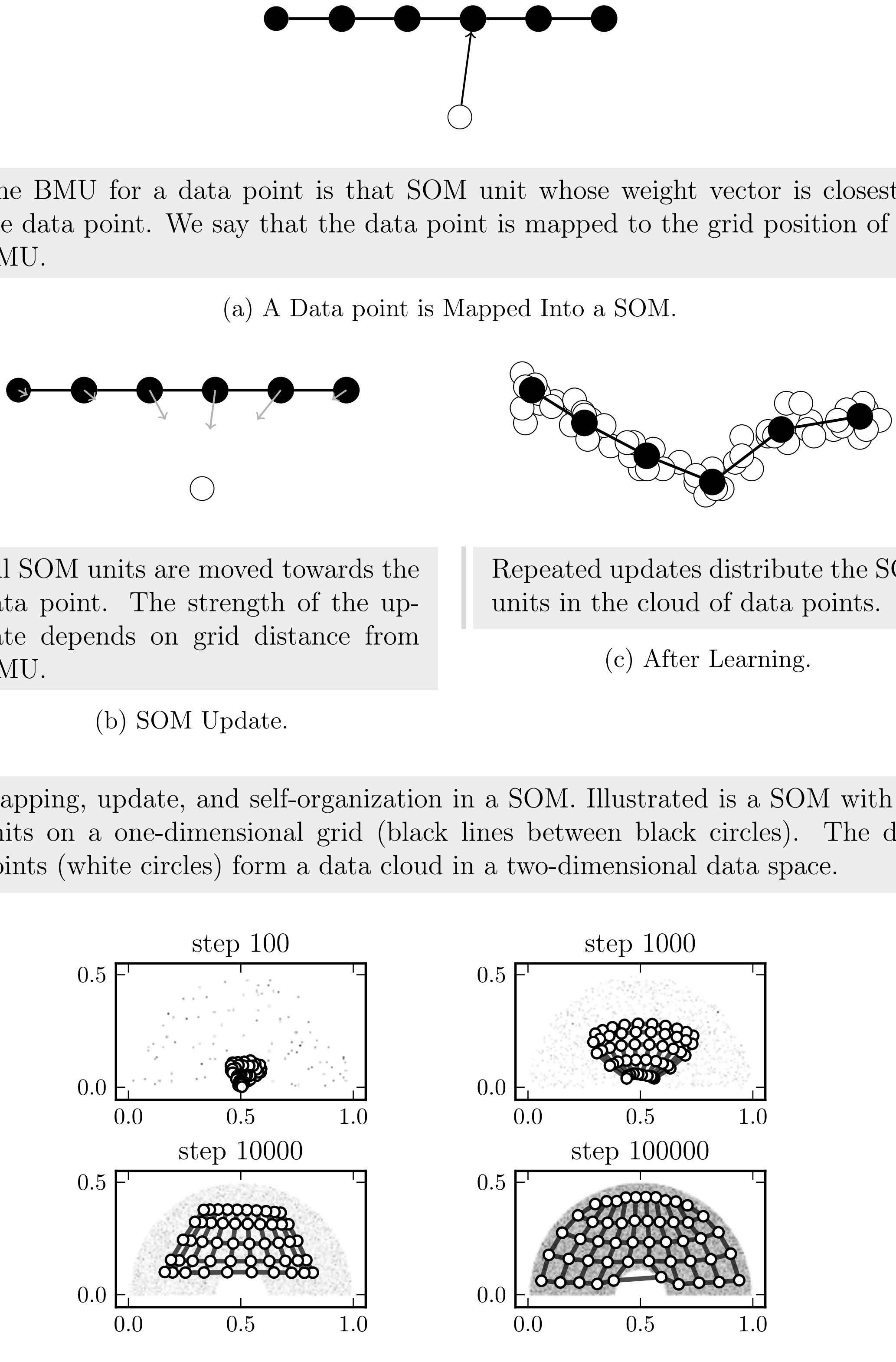
The Self-Organizing Map (SOM) and SOM-Based Models

The Kohonen Map, Kohonen Feature Map, or self-organizing map (SOM) algorithm is an abstract unsupervised ANN learning algorithm which has been shown to produce topology-preserving mappings from data space into network topology (Kohonen 2001, pp. 84–107). A SOM is a network in which a population of $n_{\mathbf{o}}$ output neurons, or 'units,' $\mathbf{O} = (\mathbf{o}_k)_{k=1}^{n_{\mathbf{o}}}$ is placed on a grid. In the most common form of the SOM algorithm, each output neuron $\mathbf{o}$ is fully connected to the population of input neurons $\mathbf{I} = (\mathbf{i}_k)_{k=1}^{n_{\mathbf{i}}}$ via a vector of 'weights,' $\mathbf{W_o} = (w_{\mathbf{o},k})_{k=1}^{n_{\mathbf{i}}}$ also called the neuron's 'prototype.' A data point, or input activation, $\mathbf{A_i} = (\mathbf{a}_{\mathbf{i},m})_{m=1}^{n_{\mathbf{i}}}$ is mapped into the grid by selecting that so-called best-matching unit (BMU) whose weight vector is closest to the data point in Euclidean space (see Figure 3.1a).

Learning often starts with a randomly initialized network. In single-data point[9] learning, each step consists of choosing one data point, finding the BMU $\mathbf{o}_B$, and updating $\mathbf{o}_B$ and its neighborhood. A neuron $\mathbf{o}$ is updated by moving its weight vector $W_{\mathbf{o}}$ closer to the data point in Euclidean space. How much closer a unit's weight vector is moved towards the data point, i.e. the strength of the update, depends on the unit's distance from the BMU in the grid: the closer the unit to the BMU, the greater the update strength (see Figure 3.1b). By starting with a large neighborhood width and strong overall updates and decreasing them as learning progresses, the SOM tends to produce an organization in which similar input vectors are mapped to units which are close to each other (see Listing SOM and Figure 3.1c).

Successful SOM learning leads to a network of units whose weight vectors are representative of the data vectors. As a consequence, the SOM algorithm can be seen as learning which values across the input dimensions typically go together. This makes the SOM a natural choice for modeling MSI: if trained on data in which the input dimensions pertain to different sensory modalities, a SOM should learn to associate those input values with each other which are the result of the same cross-modal stimuli. SOMs have therefore been previously used to model map registration in MSI. Pavlou and Casey (2010) presented a simple model of the dSC and its afferents consisting of four neural populations (two sensory, one cortical, one the actual dSC), each of which implements a SOM-like network. Their model reproduces topographic mapping and some of the effects of cortical input to the dSC, but it does not consider sensory noise or compare it behaviorally to natural MSI. An even more abstract model of the dSC and its afferents which does not consider cortical input but is also based on a network of SOM-like neural populations was proposed by Casey et al. (2012) (see below). The models presented by Zhou et al. (2011) and Bauer et al. (2012a) are SOM-based models and they are evaluated for their performance, but not for their response properties. Anastasio and Patton (2003) used SOMs to model MSI albeit not focusing on the spatial aspect (see above).

[9] By single-data point learning, we mean updating the network with one data point at a time. This is sometimes called 'stochastic gradient descent' in the context of machine learning, but that is not a good fit for SOM learning. It is also sometimes called 'online learning,' but that term can imply learning from live sensor data, and particularly using every data point only once. We want to avoid this connotation here to be able to distinguish between online and offline learning later on.



> The BMU for a data point is that SOM unit whose weight vector is closest to the data point. We say that the data point is mapped to the grid position of the BMU.

(a) A Data point is Mapped Into a SOM.

> All SOM units are moved towards the data point. The strength of the update depends on grid distance from BMU.

(b) SOM Update.

> Repeated updates distribute the SOM units in the cloud of data points.

(c) After Learning.

> Mapping, update, and self-organization in a SOM. Illustrated is a SOM with six units on a one-dimensional grid (black lines between black circles). The data points (white circles) form a data cloud in a two-dimensional data space.

> Self-organization distributes the SOM units' prototypes (circles) in the cloud of data points (dots).

(d) A Two-Dimensional SOM Unfolding.

Figure 3.1: Illustration of the SOM Algorithm

The SOM Algorithm.

```
function MAP($\mathbf{A_i} = (\mathbf{a}_{\mathbf{i},k})_{k=1}^{n_\mathbf{i}}$)
    return $\arg\min_{\mathbf{o}\in\mathbf{O}} (|W_\mathbf{o} - \mathbf{A_i}|)$
end function

procedure UPDATE($\mathbf{A_i} = (\mathbf{a}_{\mathbf{i},k})_{k=1}^{n_\mathbf{i}}, t$)
    $\mathbf{o}_B \leftarrow$ MAP($\mathbf{A_i}$)
    for $n_\mathbf{o} \in \mathbf{O}$ do
        $d \leftarrow \mathrm{d}(\mathbf{o}, \mathbf{o}_B)$
        $s \leftarrow \mathrm{h}(d, t)\alpha(t)$
        $W_\mathbf{o} \leftarrow (1 - s)W_\mathbf{o} + s\mathbf{A_i}$
    end for
end procedure

for $t \leftarrow 1 \rightarrow n_a$ do
    UPDATE($\mathbf{A}_{\mathbf{i},t}, t$)
end for
```

Given:

D :	Data set. $D = \left(\mathbf{A}_{\mathbf{i},m} = (\mathbf{a}_{\mathbf{i},k})_{k=1}^{n_\mathbf{i}}\right)_{m=1}^{n_a}$
$\mathbf{O}$:	Population of output neurons. $\mathbf{O} = (\mathbf{o})_{k=1}^{n_\mathbf{o}}$
$\mathrm{d}(\cdot,\cdot)$:	metric on $\mathbf{O}$
$\alpha(t)$:	learning rate, decreasing with t e.g. $\alpha(t) = 0.1 * (n_a - t)/n_a$
$\mathrm{h}(d, t)$:	neighborhood interaction, decreasing with d, t e.g. $\mathrm{h}(d, t) = \exp\left(-d^2/\sigma_t^2\right)$, σ_t^2 decreasing with t

The model proposed by Cuppini et al. (2012, see above) can be seen as implementing the SOM algorithm in neural detail.

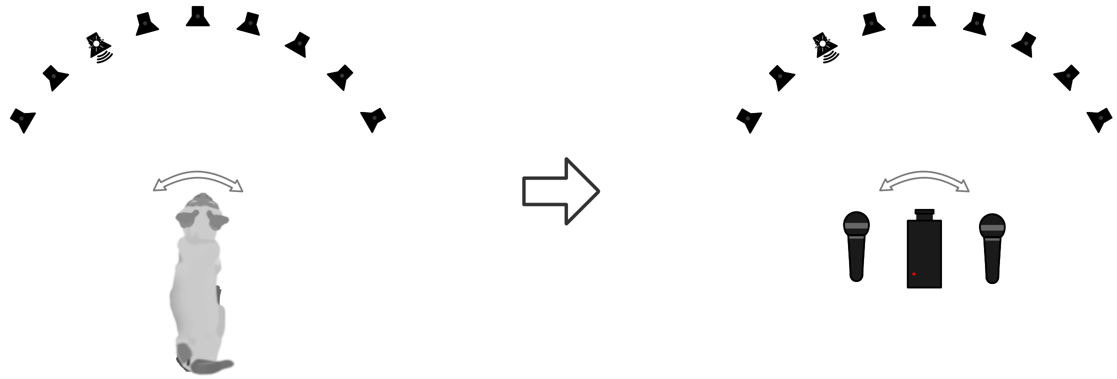
3.3 Neurorobotic Studies of the Superior Colliculus

Casey et al. (2012) implemented a robotic system on top of their model of learning multisensory integration.[10] With its focus on real-time operation on actual sensory input, Casey et al.'s (2012) model is relatively abstract and simple and was not shown to reproduce many of the patterns seen in neural responses in the dSC. They did show that sensory register could be learned from simple saliency-like visual input and population-coded, interaural level difference (ILD)-based auditory SSL.

Ravulakollu et al. (Ravulakollu et al. 2009, 2012) extracted visual cues (change between two images) and interaural time difference (ITD) between two microphones and integrated the result using radial basis function networks to integrate auditory and visual stimuli. Besides good localization performance, the authors demonstrated cross-sensory depression in responses to incongruent audio-visual stimuli. However, auditory stimuli were not presented in a biologically plausible neural code, their network did not reproduce, for example, topographic organization, and the system's behavior was not compared to that found in biological systems.

Rucci et al. (1999, 2000) used their model of natural multisensory integration to implement a robotic system which uses vision and sound to localize a cross-modal stimulus. Through what they term 'value-dependent learning,' their robotic system learns to use

[10] See Section 3.2.2.



> By replacing the cat or owl in classical neurorobotic experiments by a camera and a set of microphones, Rucci et al. (1999, 2000) were able to demonstrate the effectiveness of their model as a basis for an adaptive audio-visual localization system.

Figure 3.2: Transformation from Stein and Meredith (1993)-like to Neurorobotic Experiments.

visual and auditory (interaural time difference) cues to localize uni- and multisensory stimuli and orient towards them. Rucci et al.'s approach is very similar to ours in that they implement an ANN model of parts of the midbrain and other subcortical areas in a robot and test them on real audio-visual stimuli. One difference between their work and ours is that they assume reward-based learning, while ours is based on self-organization. Of course, the two can be combined and fruitfully applied together. More importantly, we also study the result of incongruent audio-visual input, and compare both the resulting behavior and the neural activations in response to these stimuli to what has been found in biological experiments. Finally, we extend our modeling to cortico-collicular processing.

3.4 Intermediate Summary: The State of dSC and MSI Modeling

Various types of models of the dSC and of dSC-type MSI have been proposed. Apart from the motor function, which has been at the focus of some models but which will not play a role in this thesis, MSI at the behavioral level, MSI at the network level, and MSI at the single-neuron level have been subject to modeling in the past. Models at all levels of Marr's (1983) hierarchy[11] have been devised at each level of granularity. These models have been either static, or they have encompassed adaptation mechanisms, thus modeling the dSC in the developing organism. Particularly models based on self-organization have

[11] See Section 1.2.2.

been successful, shedding light on how development in the dSC without task-relevant feedback can be explained.

Many aspects, and especially combinations of aspects, of the dSC's neurobiology and the biology of MSI are still in need of an explanation. We choose the set of aspects that we want to focus on in this thesis in concordance with our goals and choices laid out in Section 1.2.1: we will strive to connect theoretical and mechanistic levels of modeling algorithmically, and thereby explain how the phenomenology is a result of the implementation of the function of the dSC. We will particularly focus on experience-dependent development, that is, development without feedback, in principle. The result will be a model which describes how a theoretically motivated ANN algorithm learns to produce both the overt behavior and the neurophysiology of MSI in the SC. This set of constraints has not been modeled before and, to us, it is one of the most interesting unmodeled sets of constraints in dSC modeling.

4 A Network for Sensory Integration

When we motivated the relevance of the SC for the computer scientist, in Section 1.1.3, we hinted at a problem faced by the individual dSC neuron. Every neuron in the dSC is connected to a great number of other neurons in- and outside the SC. It receives action potentials from various neurons in sensory or cognitive brain regions via synapses on its dendrites or soma, and it sends action potentials along its axon to other neurons in motor, sensory, or cognitive brain areas.[1] Its purpose is to transform the patterns of postsynaptic potentials (PPs) to action potentials so as to fulfill its task as well as possible. It is clear that not all PPs can be treated equally. How to react to a given PP depends on the presynaptic neuron. If it is a retinal neuron, it may carry information about the light intensity of a stimulus and its meaning is further determined by where in the retina is the RF of that presynaptic neuron. The situation is similar for an auditory presynaptic neuron except that it responds to some feature of an auditory target. However, whatever information the input activity from an auditory neuron carries, that information is generally much less reliable than information from a visual neuron. In fact, even within a modality, different presynaptic neurons' activities have different meanings and different reliabilities in the context of the postsynaptic dSC neuron's task.

When an SC-possessing animal is born, some sensory capabilities may or may not be present already. However, in many species, sensory processing develops drastically after birth, both in function and in neurophysiology. In dSC neurons of altricial animals like cats, no responses to visual or auditory stimuli can be measured right after birth (Stein and Stanford 2013; Stein et al. 2014). Later in life, neural responses to sensory input become highly differentiated, as do overt reactions. Thus, it can be assumed that neurons adapt their responses mediated by experience, which means that they learn the significance of activity at their input synapses.

The SC is a highly preserved brain structure and it is present and functionally equivalent between animals with vastly different sensory worlds. That, together with the fact that even sensory neurons within the same modality may have very different response properties, and the fact that a dSC neuron has no way of knowing in which brain region an axon connecting to it originated, suggest that dSC neurons might not differentiate between modalities at all, but rather learn the activity patterns of each input connection individually. In this chapter, we will operate under this hypothesis. We will therefore develop an ANN model which shows how a network of neurons might self-organize and learn to extract and integrate information in its population-coded input without any inbuilt knowledge of the response properties of its individual input neurons. In line with

[1] See Section 2.2.6.

the above considerations, we will at this point not discuss multisensory input in this section. We will also proceed purely analytically, leaving simulations, experiments, and implications of this model in the case of bi-modal and cortical input to Chapters 5 and 6, respectively, and algorithmic properties to Part II.

The discussion of our algorithm in this section will start with a mathematical formulation of the problem of localizing a stimulus using population-coded information. It will then introduce basic mechanisms of models and algorithms we build upon. The algorithm itself will be explained in two parts: first, we will explain the architecture and the mathematical operations of our network and how these operations integrate input from multiple sensory modalities. This first step will assume that the network can be trained effectively. In the second step, we will offer a procedure for training the network with sensory input.

4.1 The Problem

The neural responses used by the SC to localize stimuli (as per our assumption formulated in Part I) are stochastic. This means that the same stimulus presented twice will not evoke the same neural responses every time (Seung and Sompolinsky 1993; Tolhurst et al. 1983; Vogels et al. 1989). Thus, the SC's task can be described as a statistical inference problem:[2] Suppose there are projections from $n_{\mathbf{i}}$ sensory neurons $\mathbf{I} = (\mathbf{i}_k)_{k=1}^{n_{\mathbf{i}}}$ to the dSC. Then their activities $\mathbf{A}_i = (\mathbf{a}_k)_{k=1}^{n_{\mathbf{i}}}$ can be thought of as observable variables from which the location L of the stimulus (a latent variable) is to be inferred. Bayes' theorem gives the general solution to this problem:

$$P(L \mid \mathbf{A}_i) = \frac{P(\mathbf{A}_i \mid L)}{P(\mathbf{A}_i)} P(L) \tag{4.1}$$

In words, this means that the probability of any possible location l (a realization of L), given the input activity, is proportional to how likely it is that the input neurons respond with that activity if the stimulus is at l, multiplied by how probable l is in general.

4.2 Probabilistic Population Codes

We know that humans perform multisensory localization as if they correctly applied this formula or a more specific version of it[3]. Behavior like that does not necessarily show that humans do in fact perform statistical processing. It would be explained, however, if they did. Neuroscientists have therefore looked for correlates of statistical processing, and they have found them (Gold and Shadlen 2001; Platt and Glimcher 1999; Yang and Shadlen 2007).

[2] It *is* a statistical inference problem. Whether or not what happens in the SC is actually an implementation of efficient statistical inference in the narrow sense is a different matter. However, we will suggest that it might well be that.

[3] See Section 2.1.

This is where our normative method comes into play:[4] We understand the task as one of manipulating probabilities. Neurons sometimes seem to encode probabilities. The brain often encodes quantities in population codes, and activity in the dSC seems to population-code for the location of stimuli[5]. Given these constraints, it would seem natural to implement the system as one in which PDFs are encoded in population codes and which manipulates PDFs by manipulating population codes.

Population codes which encode PDFs are referred to as probabilistic population codes (PPCs) (Pouget et al. 2003). Various neural models have been proposed to implement statistical computation.[6] However these either assume that neural input is already encoded in a PPC, they do not learn (unsupervised); they make unrealistic assumptions about the tuning functions, computational capabilities, or noise properties of neural input; or they do not operate on or produce population codes.

4.3 Network Structure and Computations

Perhaps the best way to describe our algorithm is to start with the desired state of our network and explain how the network can perform its tasks when successfully trained. Let us therefore first discuss the way a trained network could compute a PPC from sensory input. We will start by taking a closer look at the responsibilities of a single neuron.

In a PPC encoding the values of some sensorimotor variable V, each neuron $\mathbf{o}$ has a preferred value $v_{\mathbf{o}}$ of V and encodes in its activity the estimated probability of V being $v_{\mathbf{o}}$, given the input. For the dSC, the sensorimotor variable being encoded is the location L of a stimulus. Thus, each neuron $\mathbf{o}$ needs to compute the probability $p(L = l_{\mathbf{o}})$ of its preferred location $l_{\mathbf{o}}$ being the actual location of the stimulus. In effect, this means solving Equation 4.1 for the value $l_{\mathbf{o}}$:

$$p(L = l_{\mathbf{o}} \mid \mathbf{A}_i) = \frac{p(\mathbf{A}_i \mid L = l_{\mathbf{o}})}{p(\mathbf{A}_i)} p(L = l_{\mathbf{o}}) \tag{4.2}$$

This equation allows an arbitrarily complex statistical relationship between the value of L and the input activity $\mathbf{A}_i$. Solving it for a specific value could require much more involved calculations than could be reasonably assumed to be computable by a single neuron or a small sub-population of neurons. We therefore make the assumption that noise in individual neurons is uncorrelated.[7] This assumption greatly simplifies the calculations required of our neurons:

$$p(L = l_{\mathbf{o}} \mid \mathbf{A}_i) = \frac{\prod_{k=1}^{n_i} p(\mathbf{a}_{i,k} \mid L = l_{\mathbf{o}})}{p(\mathbf{A}_i)} p(L = l_{\mathbf{o}}) \tag{4.3}$$

[4] See Section 1.2.2.

[5] See Figure 2.2.4.

[6] See Section 3.2.2.

[7] See Section 8.3.

Next, let us assume that the variable L is uniformly distributed. If we then neglect the prior from Equation 4.3, we arrive at:

$$p(L = l_{\mathbf{o}} \mid \mathbf{A}_i) \propto \frac{\prod_{k=1}^{n_i} p(\mathbf{a}_{i,k} \mid L = l_{\mathbf{o}})}{p(\mathbf{A}_i)} \tag{4.4}$$

We can also drop the normalizing factor $p(\mathbf{A}_i)$, which is independent of $l_{\mathbf{o}}$, and get

$$p(L = l_{\mathbf{o}} \mid \mathbf{A}_i) \propto \prod_{k=1}^{n_i} p(\mathbf{a}_{i,k} \mid L = l_{\mathbf{o}}). \tag{4.5}$$

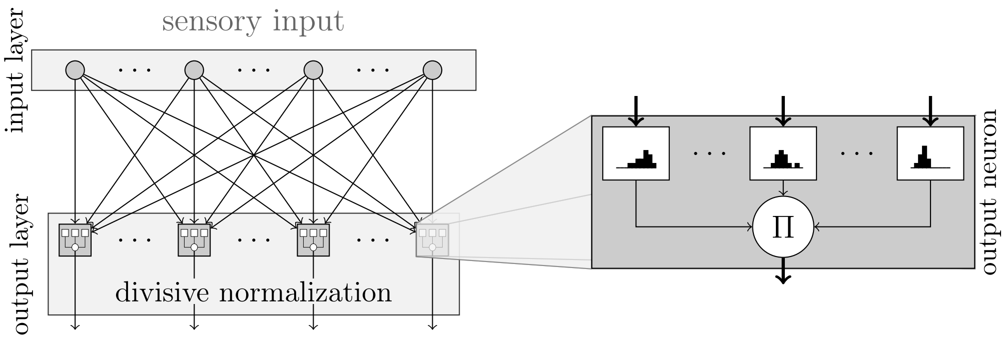
Given independent noise and equal probabilities for all values of L, the probability of any value l of L, given a population activity $\mathbf{A}$, is proportional to the product of the likelihoods of each individual neuron activity, given that L has the value l. Thus, if an output neuron could compute the likelihood of each of its input neurons' activities under the assumption that the actual value l of L is its preferred value of L, then, by multiplying all these likelihoods, it could compute the likelihood of the entire input under that assumption, which is proportional to the probability $p(L = l_{\mathbf{o}} \mid \mathbf{A}_i)$. If each of the output neurons performed this calculation and encoded the result in its activity, then we would have a PPC for L.

Now, the input neurons $\{\mathbf{i}_k\}$, whose activities are $\{\mathbf{a}_{i,k}\}$, can be neurons in any of the sensory and brain regions projecting to the dSC. Those neurons might react in very different ways to stimuli at a specific location. Therefore, it is difficult to specify a single family of functions for the individual likelihoods

$$g_{\mathbf{o},k}(\mathbf{a}_{i,k}) = p(\mathbf{a}_{i,k} \mid L = l_{\mathbf{o}}).$$

It would therefore make sense if dSC neurons $\mathbf{o}$ did not respond according to a simple transfer function but would adapt to input statistics as necessary. Relatively simple models can already explain how neural subpopulations may compute any continuous function of their input (Auer et al. 2008; Hornik et al. 1989) This would explain how certain neurons in the brain might compute and represent in their activities decision variables (Yang and Shadlen 2007). We will assume this capability in our model neurons and focus on the global learning algorithm.

To represent the computation, we use here a histogram-based mechanism which allows our model neurons to compute probabilities under the assumptions described above. It is important to note that this mechanism, useful as it is, does not come with any claim of biological realism of its implementation. Biological realism is claimed only with respect to the mechanism's capabilities, that is, with respect to the capability to gradually adapt to compute the functions $g_{\mathbf{o},k}$ as needed and as described in the next two sections. We will still describe the computations in detail as they are ultimately what determines the simulated learning behavior and neural responses which, in turn, are part of our model's predictions.



> Each neuron **o** in the integrative output layer maintains one histogram for the activity of each input neuron **i** in the input layer. Neurons **o** use histograms to estimate the likelihood of the input under the hypothesis that the stimulus is at their respective preferred locations $l_{\mathbf{o}}$. Output neurons have no information about input neurons' tuning functions.

Figure 4.1: The Structure of our Network.

4.3.1 Structure of Neurons and Network.

Figure 4.1 shows the structure of our neurons and the network. Every input neuron $\mathbf{i} \in \{\mathbf{i}_q \mid 1 \leq q \leq n_i\}$ projects to every output neuron $\mathbf{o} \in \{\mathbf{o}_r \mid 1 \leq r \leq n_o\}$. Every output neuron $\mathbf{o}$ has a preferred value $l_{\mathbf{o}}$ of L and maintains one histogram per input neuron $\mathbf{i}$.

For our purposes, a histogram for a bounded real random variable X is represented by a list $(f_i)_{i=0}^{n_b}$ of $(n_b + 1)$ frequencies, a bin size s, and a minimum value m. Then the bin $f_{\lfloor (x-m)/s \rfloor}$ contains the frequency of values x' of X in the same bin as x, i.e. values x' such that $\lfloor (x - m)/s \rfloor = \lfloor (x' - m)/s \rfloor$. If the histogram is sufficiently representative and the bin size is small enough, then the probability $p(X = x)$ of any value x can be approximately read out:

$$p(X = x) \simeq \frac{f_{\lfloor (x-m)/s \rfloor}}{\sum_{k=1}^{n_b} f_k}. \tag{4.6}$$

In the following, we will assume appropriate values for s, m and n_b and use $\text{bin}(x)$ for the bin $f_{\lfloor (x-m)/s \rfloor}$ and $\text{dens}(x)$ for the right-hand side of Equation 4.6.

Intuitively, the histogram maintained by output neuron $\mathbf{o}$ for input neuron $\mathbf{i}$ is to keep track of $\mathbf{i}$'s activity for those cases in the past when L was $l_{\mathbf{o}}$. Let $\text{dens}_{r,q}(\mathbf{a}_{i,q})$ be the dens function defined above for the histogram kept by output neuron $\mathbf{o}_r$ for the activity of input neuron $\mathbf{i}_q$. Then $\mathbf{o}_r$ can approximate the likelihood $p(\mathbf{A}_i \mid L = l_{\mathbf{o}_r})$ of the input given that the location L of the stimulus is its preferred location $l_{\mathbf{o}_r}$ by:

$$p(\mathbf{A}_i \mid L = l_{\mathbf{o}_r}) \simeq \prod_{q=1}^{n_i} \text{dens}_{r,q}(\mathbf{a}_{i,q}). \tag{4.7}$$

If we let the activity $\mathbf{a}_{o,r}$ of each output neuron $\mathbf{o}_r$ be proportional to the right hand side of Equation 4.7, then the resulting population code would be a PPC for the location of the stimulus. However, the gain of the population response would greatly vary with the overall likelihood of the input. Since biological neurons have a limited dynamic range and since their precision is greatest around the middle of that range, we let our network perform divisive normalization as suggested by Fetsch et al. (2013). Thus, the activity $\mathbf{a}_{o,r}$ of each output neuron $\mathbf{o}_r$ is

$$\mathbf{a}_{o,r} = \frac{\prod_{q=1}^{n_i} \mathrm{dens}_{r,q}(\mathbf{a}_{i,q})}{\sum_{r'=1}^{n_o} \prod_{q=1}^{n_i} \mathrm{dens}_{r',q}(\mathbf{a}_{i,q})}. \tag{4.8}$$

Note that normalization effectively re-introduces the factor

$$\frac{p(L = l_{\mathbf{o}})}{p(\mathbf{A}_i)}$$

that was dropped in Equations 4.4 and 4.5.

Under the assumptions made above and given that the histograms are representative, the output population response $\mathbf{A}_o = (\mathbf{a}_{o,r})_{r=1}^{n_o}$ is a PPC for an approximate, discretized PDF over the possible values of L. Next, we will discuss how the histograms used to approximate likelihood functions for input activities can be learned.

4.3.2 Learning Histograms and Dividing Responsibilities.

Suppose every output neuron $\mathbf{o}$ already had a preferred location $l_{\mathbf{o}}$ and suppose we know the location l of the stimulus for each data point. Then learning histograms would be easy: for every data point $\mathbf{A}_i = (\mathbf{a}_{i,q})_{q=1}^{n_i}$ for which the stimulus location was l, we could select that output neuron $\mathbf{o}_r$ whose preferred value was closest to l. We could then, for each of $\mathbf{o}_r$'s input neurons $\mathbf{i}_q$, update $\mathbf{o}_r$'s histogram for $\mathbf{i}_q$'s activity:

$$\mathrm{bin}_{r,q}(\mathbf{a}_{i,q}) \leftarrow \mathrm{bin}_{r,q}(\mathbf{a}_{i,q}) + 1. \tag{4.9}$$

Unfortunately, it is precisely the dSC's task to infer the true position l. We can therefore not assume that our learning algorithm is given this piece of information. Also, unless we assume that some knowledge of the relationship between neural input and stimulus position is already present in the network, there cannot be any coherent topology of preferred values. The learning algorithm must therefore accomplish two things: first, it will have to organize the network so that it implements a mapping from neural input into the topology of the network which maps inputs derived from stimuli at nearby positions to points in the network which are also close to each other. Second, it will have to update the neurons' histograms such that they reflect the statistics of the input.

We borrow the basic idea for our algorithm from the SOM algorithm. Through competitive learning and neighborhood interaction, the SOM algorithm updates the neurons' weights so that it produces a mapping from data topology to network topology.[8] Since

[8] See Section 3.2.2.

our network does not have weights per se to update, what we update instead are the histograms. Thus, the supervised learning rule above (Equation 4.9) can be changed into an unsupervised learning rule:

$$\mathrm{bin}_{r,q}(\mathbf{a}_{i,q}) \leftarrow \mathrm{bin}_{r,q}(\mathbf{a}_{i,q}) + \alpha_t \, \mathrm{h}_t(\mathrm{d}(\mathbf{o}_r, \mathbf{o}_B)), \tag{4.10}$$

where α_t is the global learning rate, $\mathrm{d}(\mathbf{o}, \mathbf{o}')$ is the distance between neurons in the network's grid, and $\mathrm{h}_t(d)$ is a function governing the neighborhood interaction. A typical choice of neighborhood interaction function for regular SOMs is a Gaussian function

$$\mathrm{h}_t(d) = \exp\left(-\frac{d^2}{\sigma_t^2}\right),$$

where the neighborhood interaction width σ_t depends on the learning step t. This choice works for our case as well.

Both neighborhood interaction function h_t and learning rate α_t depend on the learning step t: the neighborhood interaction parameter σ_t in h_t decreases over time such that neurons around the BMU are updated less and less strongly, while α_t increases sublinearly with t. The latter fact may be confusing for those familiar with the regular SOM algorithm: there α_t usually decreases. Since the histogram bins in our network are never normalized, the effect of adding to them on the PDF represented by the histograms decreases over time. Therefore, we increase the update strength α_t to ensure that later updates still have some impact. Since α_t increases sublinearly with the learning step, that impact still decreases, though not as strongly as with a constant α_t.

Again, since the modeling results in the following sections, i.e. our model's predictions, are ultimately determined by specifics of our algorithm, we provide a condensed, algorithmic summary of our network in Listing HistoSOM.

4.3.3 Determining the Mapping

Being an unsupervised learning algorithm, the novel ANN algorithm described above does not learn what it is told, but something that is intrinsic in the data. In order to use the trained network or interpret its output on new data, we need to determine what it has learned. In a regular SOM, it is possible to either inspect the prototypes of the SOM units and interpret those, or analyze the mapping of all or some of the data points and label each cluster of data points mapped to some unit or neighborhood of units. This process is often done manually, especially where the goal of applying the SOM algorithm is visual analysis.

In our case, the prototypes are very informative of the statistical relationships in the data. However, they are highly multidimensional histograms which can be difficult to interpret for the human observer. Another (solvable) complication is divisive normalization, which makes each neuron's response dependent on all other neurons' responses. This is not a problem from the modeling perspective: to our knowledge, there is no instance in the brain which analyzes the connection strengths between neurons to determine the meaning of their activity. In fact, 'meaning' of neural activity is notoriously

The HistoSOM Algorithm. The Basis of Our Modeling.

```
procedure UPDATE($t$)
    $(\mathbf{a}_{\mathbf{i},k})_{k=1}^{n_\mathbf{i}} \leftarrow \mathbf{A}_{\mathbf{i},t}$
    $(\mathbf{a}_{\mathbf{o},l})_{l=1}^{n_\mathbf{o}} \leftarrow$ RESPONSE($(\mathbf{a}_{\mathbf{i},k})_{k=1}^{n_\mathbf{i}}$)
    $\sigma_t \leftarrow (\alpha_> - \alpha_<) * \epsilon^{t/t_\delta} + \alpha_<$
    $\alpha_t \leftarrow \sqrt{t}$
    $b \leftarrow \arg\max_l(\mathbf{a}_{o,l})$
    for $l \leftarrow 1 \rightarrow n_o$ do
        $h \leftarrow \exp\left(-(b-l)^2/\sigma_t^2\right)$
        for $k \leftarrow 1 \rightarrow n_i$ do
            $m \leftarrow \mathbf{a}_{\mathbf{i},k}$
            $\mathrm{bin}_{l,k,m} \leftarrow \mathrm{bin}_{l,k,m} + h\alpha_t$
        end for
    end for
end procedure

// Initialization
for $l \leftarrow 1 \rightarrow n_\mathbf{o}$ do
    for $k \leftarrow 1 \rightarrow n_\mathbf{i}$ do
        for $m \leftarrow 1 \rightarrow \max(\mathbf{a}_{\mathbf{i},k,t})$ do
            $\mathrm{bin}_{l,k,m} \leftarrow \epsilon'$
        end for
    end for
end for

// Training
for $t \leftarrow 1 \rightarrow n_a$ do
    UPDATE($t$)
end for
```

```
function DENS($l, k, \mathbf{a}_{\mathbf{i},k}$)
    $m \leftarrow \mathbf{a}_{\mathbf{i},k}$
    return $\mathrm{bin}_{l,k,m} \,/\, \sum_{m'=1}^{n_\mathbf{o}} \mathrm{bin}_{l,k,m'}$
end function

function RESPONSE($\mathbf{A}_\mathbf{i} = (\mathbf{a}_{\mathbf{i},k})_{k=1}^{n_\mathbf{i}}$)
    for $l \leftarrow 1 \rightarrow n_\mathbf{o}$ do
        $r_l \leftarrow \prod_{k=1}^{n_i}$ DENS($l, k, \mathbf{a}_{i,k}$)
    end for
    $s \leftarrow \sum_{l=1}^{n_\mathbf{o}} r_l$
    for $l \leftarrow 1 \rightarrow n_o$ do
        $r_l \leftarrow r_l/s$
    end for
    return $(r_l)_{l=1}^{n_o}$
end function
```

Given:

n_i :	number of input neurons
n_o :	number of output neurons
n_d :	number of training steps
D :	Data set. $D = \left(\mathbf{A}_{\mathbf{i},t} = (\mathbf{a}_{\mathbf{i},k,t})_{k=1}^{n_\mathbf{i}}\right)_{t=1}^{n_d}$
$\alpha_>, \alpha_<$:	max, min update strength
ϵ, ϵ' :	small constants

difficult to define,[9] a problem that we fortunately do not have to solve here. Instead, we operate under the assumption that the dSC generates an activity that *implicitly* encodes the location of a stimulus, and that other brain areas, like the motor areas responsive for eye movements, *implicitly* use that encoding to fulfill their tasks without the need to ever *explicitly* decode it.

To make statements about the behavior of our network we still need to know what it has learned, after the training phase. One way to do that is to generate or select from a corpus, in what we call the mapping phase, more of the kind of data that was used for training, and observe which neuron $\mathbf{o}$ data points derived from which locations l are mapped. For any particular neuron $\mathbf{o}$, we can then define $\mathbf{o}$'s preferred value $l_{\mathbf{o}}$ of L as the mean of all values l for which $\mathbf{o}$ was BMU in the mapping phase. That mean will be (close to) the expected value of all l of which data points mapped to $\mathbf{o}$ are derived.[10] Note that this slightly changes the notion of a preferred value. In biological research, the preferred value of a biological neuron is that value for which its average activity is greatest. For us, the preferred value of a model neuron is the average of those values for which its activity tends to be the greatest of all model neurons in the network. We need this re-interpretation to relate simulated neurophysiology to behavior as we will do in the next sections. Without formal proof, we argue that these two things are usually the same in our network, after successful training, largely due to divisive normalization.

4.4 Intermediate Summary: A Network for Modeling

The ANN learning algorithm presented in this chapter was strongly motivated by the task we assume the dSC fulfills and by a mathematical and algorithmic approach to solving that task. Important features of this algorithm are unsupervised learning, more specifically topological self-organization; population coding; learning of statistical inference; minimal assumptions about the tuning functions of input neurons; minimal assumptions about the noise in the input; and, related to that, no assumptions about the differential origin of input. In the following, we will develop models of the dSC based on our ANN algorithm and aim to reproduce phenomena that occur in the biology of

[9] Does the activity in a certain visual neuron represent a feature of the visual display (Krüger et al. 2012)? Is it part of a modal representation of a concept (Barsalou et al. 2003)? Does it not represent anything (Wilson and Golonka 2013)? See Section 7.1.2.

[10] Any set of values of size n_s for which a given neuron $\mathbf{o}$ is the BMU is a sample from a random variable with finite variance σ_M (assuming L has a finite range), and it follows from the central limit theorem that its mean will approach the expected value μ of that random variable as $n_s \to \infty$. In particular, the mean of sample of size n_s will be approximately normally distributed with variance σ_M^2/n_s around μ (Billingsley 1995, p. 257).

The formal argument becomes a bit more complicated if the values l for which data points are generated are not randomly drawn but systematically chosen evenly spaced across all possible values of L, as in the experiments reported below. However, assuming L follows a uniform distribution, the end result stays the same. The benefit is that the number of data points used to determine the preferred value of each neuron is more evenly distributed across neurons.

In the simulations reported in the following sections, we chose very large n_s to ensure that our estimate was adequate.

the dSC. We will thus strive to corroborate the hypothesis that the task we ascribed to the dSC in Part I and the features outlined above indeed characterize learning and functioning of the dSC. These features will therefore be important both in the context of our modeling work and as potential features of algorithms in practical applications, to be investigated in Part II.

5 Bottom-Up Multisensory Integration

In this chapter, we will use the ANN algorithm introduced in Chapter 4 to model natural MSI in the dSC. In Section 5.1, we will describe the architecture of our model and the kind of simulated input on which we will train our model and study its behavior. We will demonstrate, in Section 5.2, that our model reproduces several important neurophysiological and behavioral phenomena of natural MSI in the dSC. After that, in Section 5.3, we will describe a neurorobotic experiment in which our network is trained and tested on real sensory input. Finally, in Section 5.4, we will summarize and discuss the results of our experiments and their significance for our modeling specifically of bottom-up MSI. The model will be extended in the next chapter, and the whole model will be discussed in depth in Chapter 7.

5.1 The Model: Multisensory Stimuli

Figure 5.1 shows the structure of our model. At the core of the model is the network described in the last section. Each output neuron receives input from all sensory input neurons. The population of input neurons is logically partitioned into 'visual' (V) and 'auditory' (A) neurons. It is an important feature of our model that this partitioning is opaque to the output neurons.

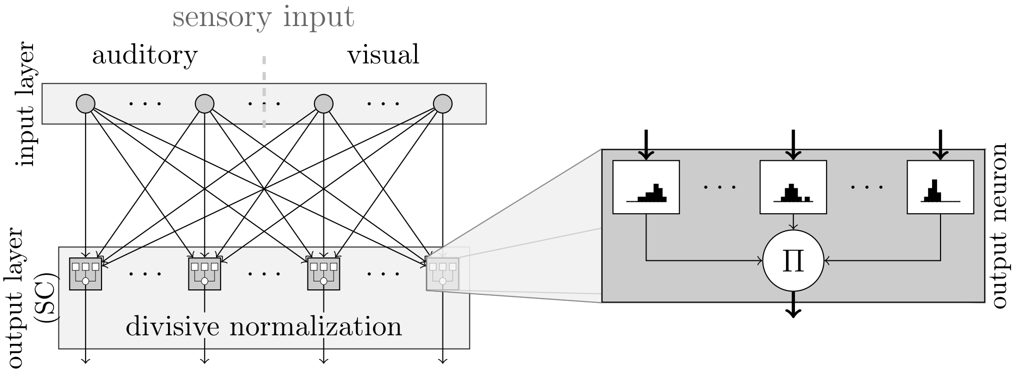
The population response of the input population is caused by a cross-sensory stimulus whose location is $l \in [0, 1]$. Each one of the $n_{\mathbf{i},m}$ visual or auditory input neurons $\mathbf{i}_{m,k}$, $m \in \{V, A\}$, $1 \leq k \leq n_{\mathbf{i},m}$ has a preferred location $l_{\mathbf{i}_{m,k}}$ of the stimulus. Input neurons' preferred locations are distributed evenly across the range of possible stimuli:

$$l_{m,k} = \frac{k-1}{n_{\mathbf{i},m} - 1}.$$

The activity $\mathbf{a}_{\mathbf{i}_{m,k}}$ of neuron $\mathbf{i}_{m,k}$ in response to a stimulus at location l is governed by a Poisson-noisy Gaussian tuning function:

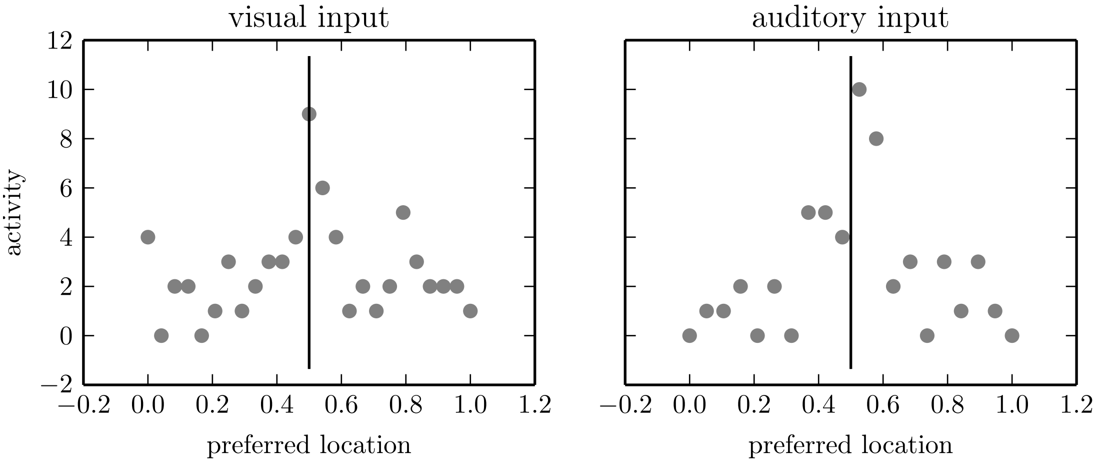
$$\mathbf{a}_{\mathbf{i}_{m,k}} \sim \text{Pois}\left(g_m \times \exp\left(-\frac{(l_{\mathbf{i}_{m,k}} - l)^2}{\sigma_m^2}\right) + \nu_s\right), \tag{5.1}$$

where g_m and σ_m are the modality-specific gain and width of the tuning functions. The sensory background noise parameter ν_s is there mainly for consistency with the modeling in Chapter 6. See Figure 5.2 for example input.



The input population is divided into visual and auditory neurons. Tuning functions in the two sub-populations are different; output neurons have no information on the modality any given input neuron **i** belongs to.

Figure 5.1: Our Model of Multisensory Integration.



An audio-visual stimulus elicits activity in visual and auditory input neurons in our model. Differently parameterized tuning functions and different population sizes lead to different amounts of information on the stimulus location in the two input populations. **Gray dots:** activity of the neurons. **Vertical line:** position of the simulated stimulus.

Figure 5.2: Example Sensory Input.

Simulation Steps		Tuning Functions			Population Sizes	
			vis.	aud.		
training	200000	gain ($g_{v/a}$)	8	7	vis. input ($n_{\mathbf{i},v}$)	25
mapping	50000	width ($\sigma_{v/a}$)	0.005	0.01	aud. input ($n_{\mathbf{i},a}$):	20
incongruent	100000	basel. noise ($\nu_{v/a}$)	2	2	output ($n_{\mathbf{o}}$):	500
varied intensity	2000					

Table 5.1: Parameter Values in Simulation of Multisensory Integration.

The particular shape of the above tuning functions and the kind of noise is not important in the context of our model. However, Gaussian tuning functions are a simple choice and realistic in that they have a central peak, and fall off with distance from the center. Tuning curves or tuning functions of neurons in Retina, sSC, striate cortex, and other regions in the brain show these characteristics and have been modeled using Gaussian functions or linear combinations of Gaussian functions (Butts and Goldman 2006; Cynader and Berman 1972; Hawken and Parker 1987; Jones and Palmer 1987). Poisson-like noise has the property that the variance is proportional to the mean, which is true of the variability of actual neural responses (Butts and Goldman 2006; Tolhurst et al. 1983; Vogels et al. 1989). What is important is that the parameters of the tuning functions are different between the two sub-populations. The amount of information about the true stimulus location L contained in each of the sub-population responses grows with the number of neurons and the gain and shrinks with increasing width of the tuning functions. For the simulations reported in the following, we chose parameters which reflect the fact that visual information is usually much more reliable for localization than is auditory information (Alais and Burr 2004). We trained a network of $n_{\mathbf{o}} = 500$ output neurons over 200000 training steps, generating a new data point at a random location $l \in [0, 1]$ for each step. Both values were chosen to be high enough to avoid artifacts like sampling error (too few neurons) or incomplete training (number of training steps). Smaller values easily yielded qualitatively similar results to the ones reported in the following. See Table 5.1 for the parameter settings used in the simulations.

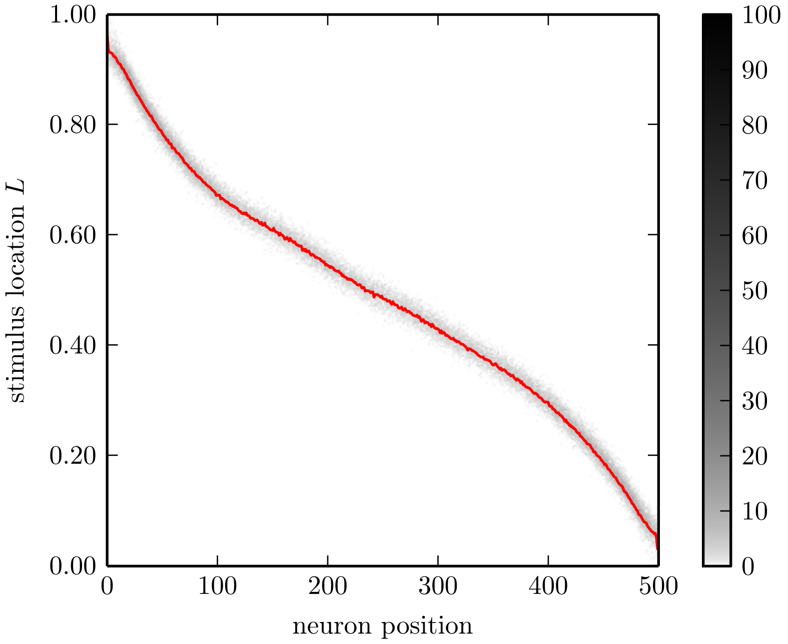
5.2 Multisensory Integration in our Model

After training had finished, we ran experiments to study the likeness of its neural responses and behavior to those seen in nature.

5.2.1 Neurophysiology

Topographic Mapping.

First, we determined the preferred location $l_{\mathbf{o}}$ of each neuron $\mathbf{o}$ in our network. To do that, we generated another 50000 data points at locations evenly spaced across the



> Gray levels indicate the number of data points from a particular location l mapped to a given neuron $\mathbf{o}$ in the mapping phase.

Figure 5.3: Mapping of Locations to Neurons.

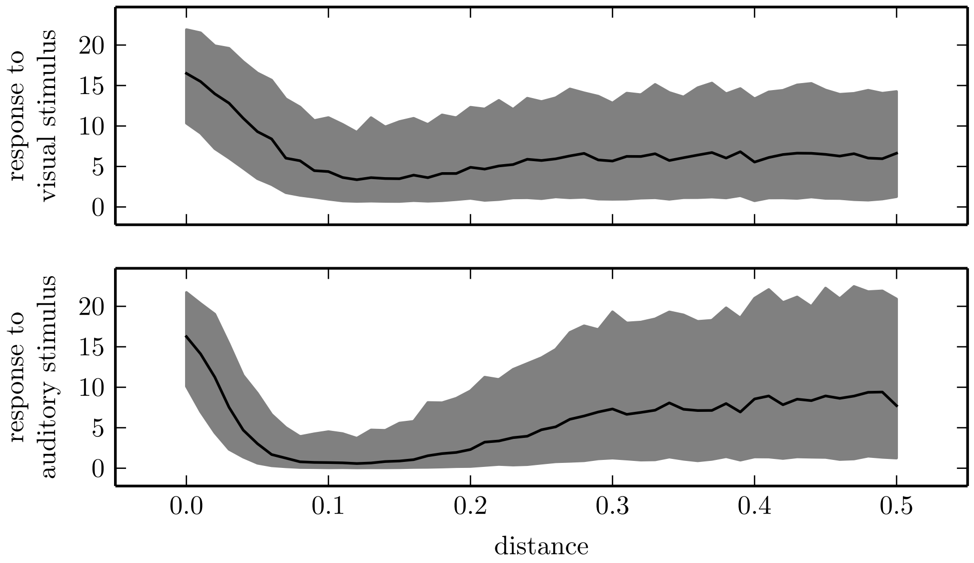
interval $[0, 1]$ and determined for each neuron $\mathbf{o}$ the mean of all locations l for which it was the BMU.[1] With $100 = 50000/n_{\mathbf{o}}$ locations per neuron on average, we were quite certain that the preferred location $l_{\mathbf{o}}$ of each neuron $\mathbf{o}$ was close to that mean. Figure 5.3 shows where data points for each location were mapped: we can see that each neuron was BMU for data points from a narrow range of values for l and that similar values of l are mapped to neurons which are close to each other in the network.

Depression.

The response to a stimulus in one sensory modality can be diminished by another stimulus in another modality if that other stimulus is not in the same place as the first.[2] To see whether our network reproduces this effect, we simulated incongruent input, that is, input in which visual neurons respond to a stimulus in one location and auditory neurons respond to a stimulus in a different location. 100000 data points were generated for which the locations l_V and l_A of the visual and auditory sub-stimulus were chosen randomly and independently. Figure 5.4 shows how the average response of the neuron $\mathbf{o}$ whose preferred location $l_{\mathbf{o}}$ was closest to l_V changed depending on the absolute distance

[1] See Section 4.3.3.

[2] See Section 2.2.3.



> Strength of the response of the output neuron whose preferred location was closest to the visual stimulus component depending on distance from auditory stimulus and vice versa. **Line:** mean response. **Gray area:** mean deviation for stimuli above, below mean.

Figure 5.4: Cross-Sensory Depression.

between the two component stimuli, and vice versa.

Enhancement and Inverse Effectiveness.

The response to a stimulus in one modality can be greatly enhanced by a stimulus in another modality if the two stimuli are close together in time and space. The size of this enhancement depends on the strength of the two component stimuli. Enhancement is strong for weak stimuli and weak for strong stimuli. This circumstance is referred to as the principle of inverse effectiveness.[2]

We presented another 2000 data points to the network to test whether our model also reproduces this principle. This time, the locations of the visual and auditory substimuli were identical and we varied the strength of the stimuli (g_V, g_A in Equation 5.1). We then computed the enhancement of the response to the visual stimulus by the auditory stimulus for each combination of visual and auditory stimulus strength g'_V and g'_A and vice versa: let $r_{g'_V,g'_A}$ be the average response of the neuron whose preferred location is closest to the stimulus. Then, we define the enhancement of the response to the visual stimulus by the auditory stimulus and vice versa as:

$$E^{g'_V}_{g'_A} = \frac{r_{g'_V,g'_A}}{r_{g'_V,0}} \qquad\qquad E^{g'_A}_{g'_V} = \frac{r_{g'_V,g'_A}}{r_{0,g'_V}}. \tag{5.2}$$

The graphs in Figure 5.5 show $E^{g'_V}_{g'_A}$ and $E^{g'_A}_{g'_V}$ for the different combinations of strengths

g'_A and g'_V. From those graphs, it is clearly visible that weak visual stimuli are strongly enhanced and strong visual stimuli are only weakly enhanced by an auditory stimulus.

5.2.2 Behavior

Next, we tested the similarity between MSI in human and animal behavior and in our model. Alais and Burr (2004) showed that the way humans integrate vision and hearing in localizing cross-sensory stimuli is well modeled by a Bayesian MLE model:[3] assuming that the errors in visual and auditory localization follow Gaussian distributions around the true stimulus positions with standard deviations of σ_V and σ_A, respectively, the optimal way of combining their estimates is by weighting them linearly with factors

$$w_V = \frac{\sigma_A^2}{\sigma_V^2 + \sigma_A^2} \quad \text{and} \quad w_A = \frac{\sigma_V^2}{\sigma_V^2 + \sigma_A^2}, \tag{5.3}$$

respectively. The model predicts that audio-visual localizations are distributed normally around a mean of

$$l_{VA} = w_A l_A + w_V l_V$$

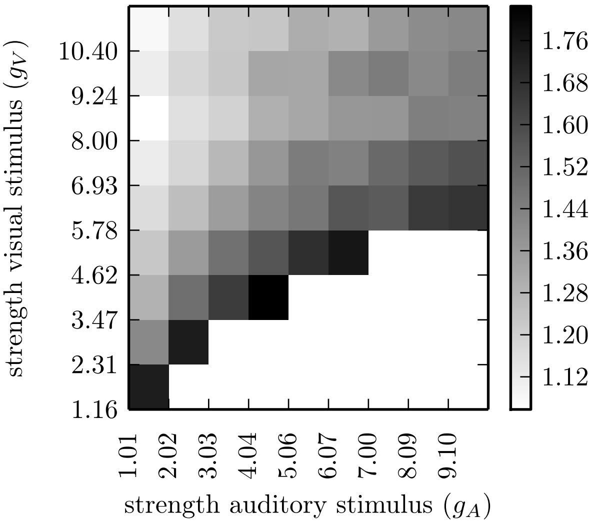
(where l_V and l_A are the possibly mutually offset locations of the auditory and visual sub-stimuli), with a standard deviation σ_{VA} given by

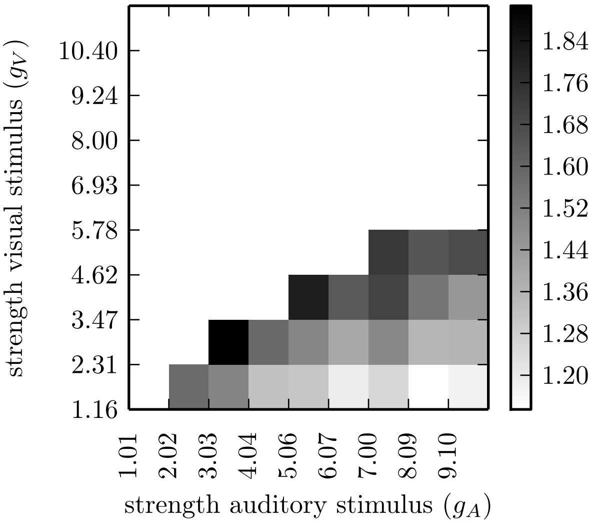
$$\frac{1}{\sigma_{VA}^2} = \frac{1}{\sigma_V^2} + \frac{1}{\sigma_A^2}. \tag{5.4}$$

Alais and Burr (2004) found that the distribution of localizations by human participants was well-described by this model.

To determine whether our model also seems to linearly combine unisensory localizations, and whether the factors in that linear combination are determined by the uncertainty in unisensory localization, as predicted by the MSI model, we first had to determine the distribution of errors in unisensory localization. We generated 100000 data points in which the visual component was zero and the same number of data points in which the auditory component was zero. We then computed the mean squared errors (MSEs) σ_V^2 and σ_A^2 the network made in localizing the unisensory stimuli in either condition. From these, we derived the standard deviation σ_{VA}^2 of the normal distribution of localizations that would be theoretically expected from optimal linear integration of auditory and visual localization as well as the weighting factors w_A and w_V (see Equation 5.3 and Equation 5.4). Figure 5.6 shows that the errors in unisensory localization are well modeled by normal distributions with the calculated MSEs. More importantly, the actual distribution of audio-visual localizations relative to the auditory and the visual stimulus location is very close to the one predicted by the linear MLE model.

[3] see Sections 2.1 and 3.2.1.



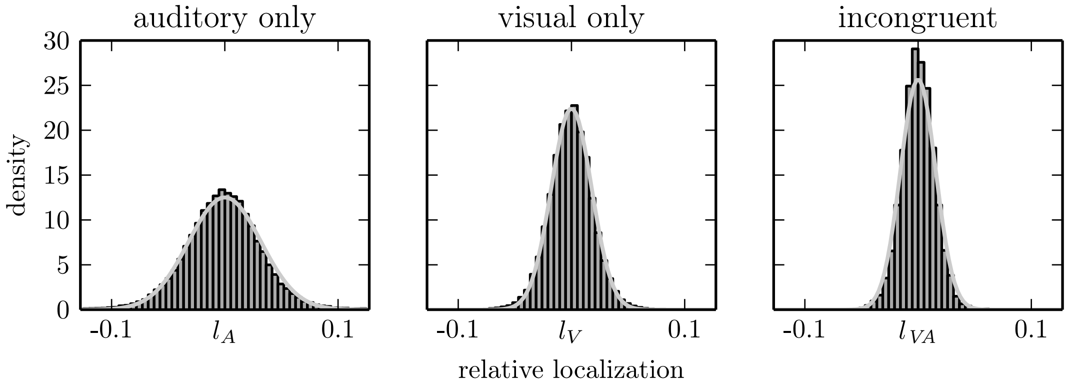


The response to a weak visual stimulus is enhanced more strongly by an auditory stimulus than that of a strong visual stimulus (and vice versa). Gray value: enhancement

$$E_{g'_A}^{g'_V} = \frac{r_{g'_V,g'_A}}{r_{g'_V,0}} \quad \text{and} \quad E_{g'_V}^{g'_A} = \frac{r_{g'_V,g'_A}}{r_{0,g'_V}}.$$

White areas: other modality stronger.

Figure 5.5: Inverse Effectiveness.

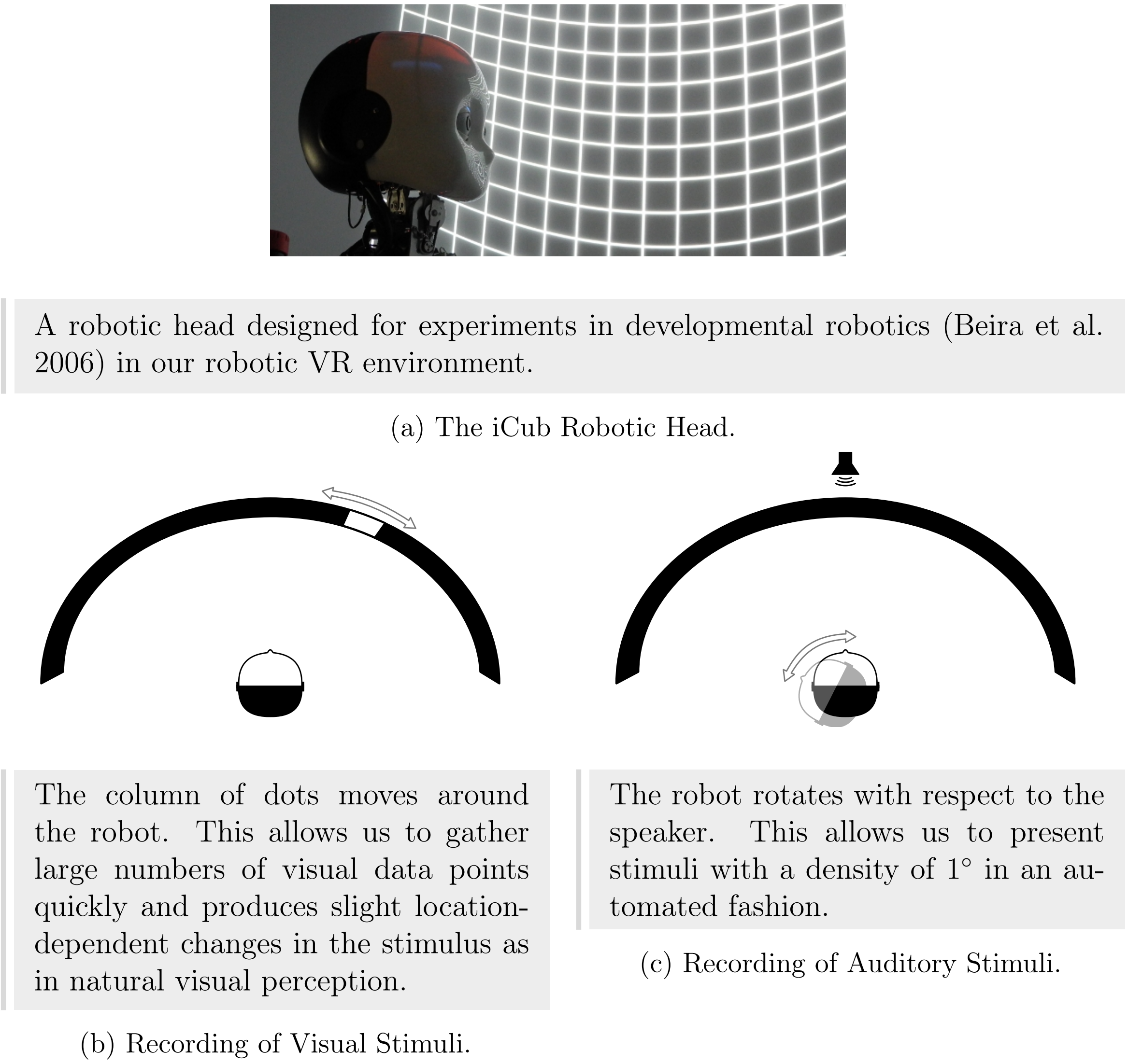


Left, Middle: normalized histograms of unisensory localizations relative to the locations l_V and l_A of the visual and auditory stimuli, respectively. Gray curves show fitted normal distributions. **Right:** normalized histogram of audio-visual localizations relative to the mean l_{VA} predicted by the MLE model. Gray curve shows predicted normal distribution.

Figure 5.6: Distribution of Localizations Relative to Stimuli in Uni-Sensory and Incongruent, Multi-Sensory Localization..

5.3 Neurorobotic Audio-Visual Localization

In their study, Alais and Burr had their participants localize visual, auditory, and combined audio-visual stimuli, manipulating the extent of the blobs of light serving as the visual localization targets, and thus changing the reliability of visual localization. When presenting their subjects with incongruent cross-sensory stimuli, that is, stimuli with a certain spatial offset between auditory and visual component, they found that their subjects behaved as if using Equation 5.3 to linearly combine estimates of stimulus positions from the two sensory modalities. In order to show that our algorithm can train a network to handle real sensory data and reproduce natural behavior found in humans, we conducted a similar experiment with a robot. The robot, an iCub robotic head (Beira et al. 2006), was placed in our robotic virtual reality environment (Bauer et al. 2012b). This environment allows us to present visual and auditory stimuli at arbitrary locations along a half-circle around the robot (see Figure 5.7). In the following, we will describe the combined, spatially congruent auditory and visual stimuli we presented to the robot and how we pre-processed the sensory data to generate input for training our network. We will report on the system's performance when presented with auditory, visual, congruent audio-visual, and incongruent audio-visual data, and how that behavior compares to the one observed by Alais and Burr (2004).



A robotic head designed for experiments in developmental robotics (Beira et al. 2006) in our robotic VR environment.

(a) The iCub Robotic Head.

The column of dots moves around the robot. This allows us to gather large numbers of visual data points quickly and produces slight location-dependent changes in the stimulus as in natural visual perception.

(b) Recording of Visual Stimuli.

The robot rotates with respect to the speaker. This allows us to present stimuli with a density of 1° in an automated fashion.

(c) Recording of Auditory Stimuli.

Figure 5.7: Experimental Setup.

5.3.1 Stimuli and Training.

Sensory signals reaching the integrative dSC are never raw. They are first shaped by the embodiment of and preprocessing in the sensory organs themselves, like filtering and compression in retina and optic nerve (Marr 1983; Stone 2012), or semi-mechanical frequency decomposition in the cochlea (Slaney 1993; Yates et al. 1992). The signals are further changed before reaching the dSC by later stages of processing, like the visual sSC and the IC, which integrates various auditory spatial cues (Schnupp et al. 2010, p. 205–209) and which is one major source of auditory information of the SC.[4]

Any one model of visual and auditory input to the SC would be complex and species-specific, and if we had had to commit to one specific model for generating input to our network from raw stimuli in our experiment, the validity of any claims derived from the experiment's results would be limited by the validity of that model. Fortunately, a strength of our model is that it explains how the dSC can learn to integrate information without built-in knowledge of the relationship between its input and the location of a stimulus. Therefore, our focus in choosing appropriate pre-processing of our stimuli was not so much on generating perfectly realistic SC input. Instead, we aimed to generate input which roughly preserved important properties of actual SC input (details below), which was experimentally and computationally feasible to produce, and, in the case of visual input, whose reliability for localization was easily modulated.

Visual Stimuli

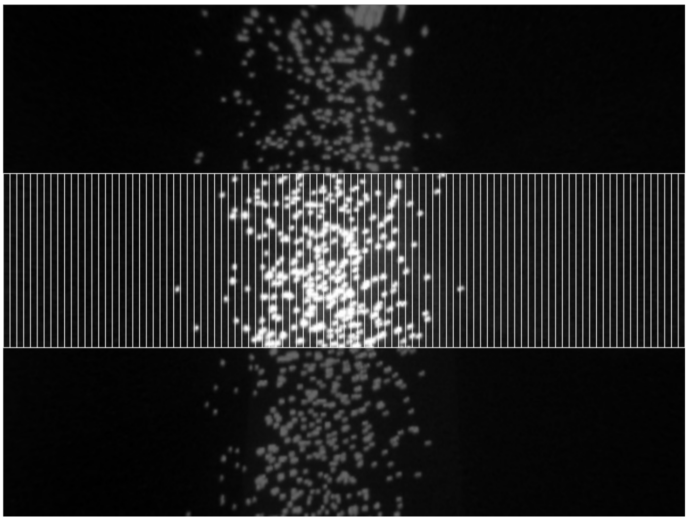
To emulate the important features of visual SC input in Alais and Burr (2004)'s experiment, we chose stimuli whose dominant property was their location and whose reliability could be manipulated to introduce significant uncertainty about that location.

Therefore, to generate a visual stimulus at angle α, we projected a column of dots against the screen which were randomly distributed around that angle α following a normal distribution. Visual input activations for training and testing our network were generated from the images then taken by the robot as described in Figure 5.8. Reliability of the visual stimulus and computational complexity was tuned by empirically choosing appropriate numbers for the number of visual input neurons, the factor by which the mean intensities were scaled, and the height of the strips cut from the image. We chose 320 for the number of visual input neurons, 0.1 for the scaling factor, and, for training, 140 pixels for the height of the strips.

Auditory Stimuli

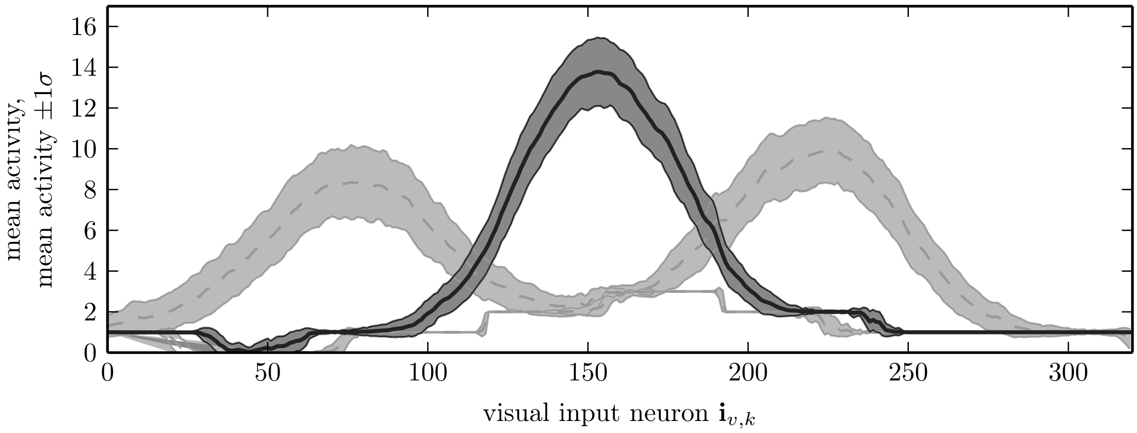
Binaural recordings for auditory localization were made by playing white noise on a speaker at a distance of about 1.30 m and rotating the robot with respect to the speaker. Auditory input activity was generated from these recordings using the IC model from Dávila-Chacón et al. (2012). Since projections from the IC to the intermediate and deep layers of the SC are a major source of auditory input to the SC (see above),

[4] See Section 2.2.6.



Column of dots projected on the screen right in front of the robot. A horizontal strip (area between horizontal white lines) of the camera image is horizontally subsampled. The mean intensity of pixels in each vertical column of the subsampled strip is computed, scaled by 0.1, and reduced to the nearest integer (see Figure 5.8b).

(a) Example Visual Stimulus.

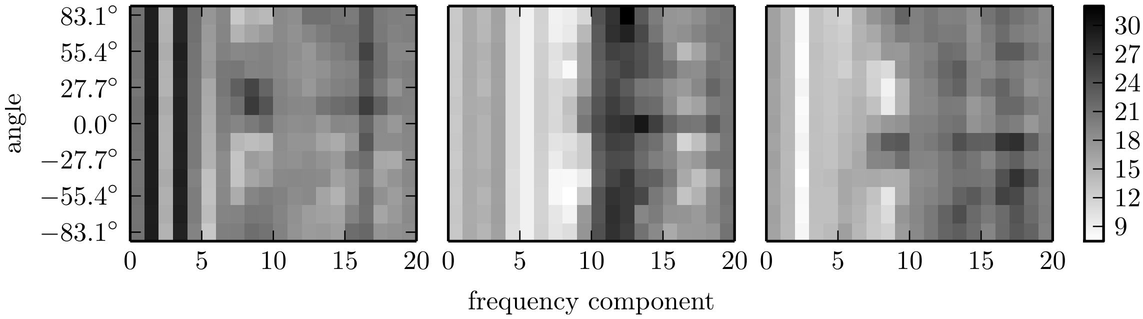


Visual input activity. **Solid line:** mean over all data points at $\alpha = 0°$. **Dashed lines:** mean over all data points at $\alpha = +10°$, $\alpha = -10°$, resp. **Gray Backgrounds:** mean ± one standard deviation.

(b) Example Visual Input.

One example of raw visual input and distributions of processed visual input for three different angles.

Figure 5.8: Visual Input.



> Mean auditory input for angles −15°, 0°, and 15°.
> IC activations are population codes of shape 20 × 13. Each column of neurons analyzes a different frequency, each row prefers a different angle. See Dávila-Chacón et al. (2012) for details. Note that there is some regularity in the input, but it is fairly complex, relatively high-dimensional, and noisy (even when it is averaged as seen).

Figure 5.9: Example Auditory Input Data.

the auditory input activity generated by the model described by Dávila-Chacón et al. (2012) is a reasonable approximation to the actual auditory input activity to the SC. For computational efficiency, auditory input was normalized by dividing each input activity $\mathbf{a}_{a,k,t}$ in auditory data point $\mathbf{a}_{a,t} = \mathbf{a}_{a,0,t}, \mathbf{a}_{a,1,t}, \dots, \mathbf{a}_{a,n_a,t}$ by the mean input $\langle \mathbf{a}_{a,k,m} \rangle_m$ over all data points $\mathbf{a}_{a,m}$ (see Figure 5.9).

Training

In total, we collected 12300 visual data points (300 per angle) and 19660 auditory data points (476–480 per angle) at 41 angles between −20° and 20°.

20 % of both visual and auditory data points were set aside for testing, leaving 9840 visual and 15728 auditory data points for training. The rest was used to train the network. In each training and testing step, we randomly chose one whole angle α between −20° and 20° and randomly selected a visual and an auditory input activity for that angle, generated as described above. The input to our network was then the concatenation of these two activations into one vector of length

$$n_v + n_a = 320 + 260 = 580 = n_i.$$

It is important to note at this point that the separation of the input population into visual and auditory sub-populations and the spatial relationship of neurons within the sub-populations are opaque to the output neurons: an output neuron $\mathbf{o}$ simply receives connections from a (logical) population of input neurons and makes no assumptions about the modality and the point in space from which the information conveyed by these input connections originates.

We trained the network over 40000 training steps. With 9840 visual and 15728 auditory data points belonging to only 41 classes, we did not expect or observe overfitting. After training, we determined the mapping learned by the network: we first determined the BMUs for another 6000 data points, randomly selected from the training set. Then, we computed for all data points with the same BMU $\mathbf{o}$ the mean over the angles at which the constituent visual and auditory input activations had been recorded. That mean angle was then chosen as the preferred value $l_{\mathbf{o}}$ of the neuron $\mathbf{o}$.

Then, to compute the accuracy of our network we presented it with 3000 data points from the test set. We took the preferred value $l_{\mathbf{o}_B}$ of the BMU $\mathbf{o}_B$ for each data point as the network's estimate of the true angle α at which the visual and auditory component of that data point were recorded (cross-sensory, congruent condition). To test the similarity between our network's behavior and that of human beings, we also tested its performance with visual and auditory input alone (unisensory conditions).

Finally, to emulate incongruent multisensory input as in Alais and Burr's [2004] experiment, we presented the network with data points from one visual activation from an angle α_v and an auditory data point with an angle α_a, where α_a was $\alpha_a = \alpha_v + 5^\circ$ (cross-sensory, incongruent conditions).

5.3.2 Results.

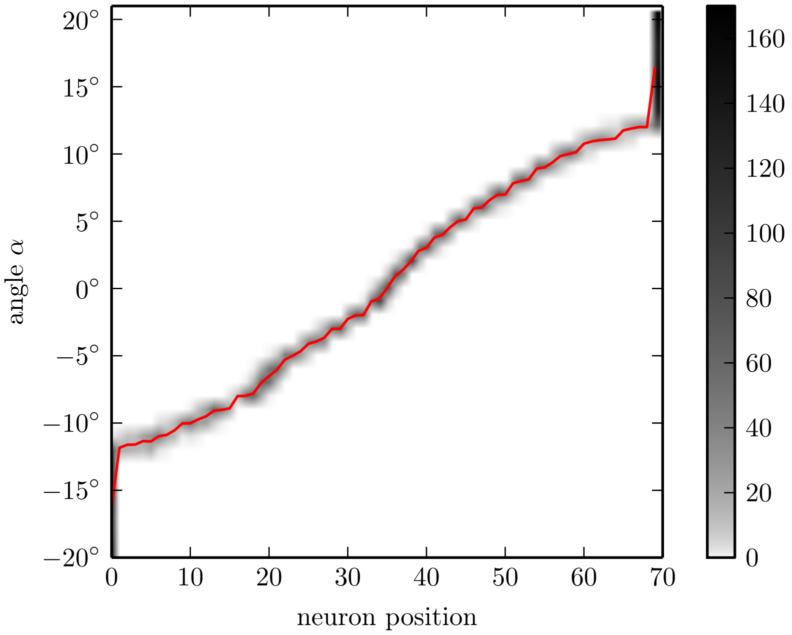
Figure 5.10 shows the spatial organization of the network after training. Stimuli around the center of the visual field are clearly mapped according to their topology. The first and the last neuron in the population attract stimuli from a range of stimuli at one end of the spectrum, each. This is due, in part, to stimuli in the periphery actually being harder to distinguish: visual stimuli are spread over more pixels as they move towards the edge of the visual field. The effect is similar, although the reasons are more complex, in auditory localization (Middlebrooks and Green 1991). Another reason for many stimuli at the side being mapped to the two outermost neurons is the border effect observable in other SOM-like algorithms (Kohonen 2001, 2013).

Figure 5.11 shows the responses of neurons to visual-only, cross-sensory congruent and cross-sensory incongruent stimuli whose visual component was located at their preferred angle. We see that the responses to cross-sensory congruent stimuli were stronger than the responses to visual-only stimuli for many neurons. We also see that the response to incongruent stimuli were strongly suppressed compared to responses to visual-only stimuli.

In the unisensory and cross-sensory, congruent conditions, we found that the MSE of the network given visual, auditory, and multisensory input was $\sigma_v^2 = 4.68^\circ$, $\sigma_a^2 = 4.82^\circ$, $\sigma_{ms}^2 = 2.46^\circ$, respectively. Given the unisensory MSEs σ_v and σ_a, the MLE model predicts a multisensory MSE of:

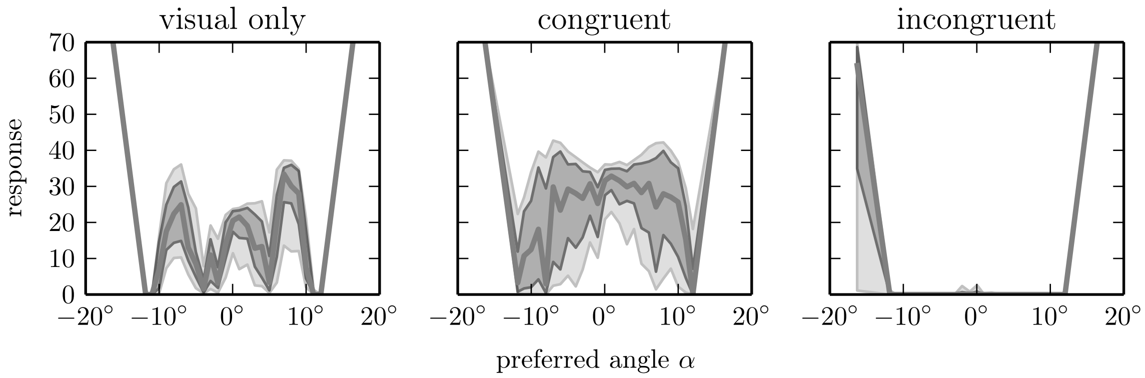
$$\hat{\sigma}_{ms}^2 = \frac{1}{\frac{1}{\sigma_v^2} + \frac{1}{\sigma_a^2}} \simeq 2.38^\circ.$$

The difference can be explained by the relatively large errors of the network in the



> Gray levels indicate the number of data points from a particular location l to a given neuron $\mathbf{o}$ in the mapping phase.

Figure 5.10: Mapping of Locations to Neurons in the Neurorobotic Experiment.



> **X-axis:** preferred angle α of each neuron. **Graphs:** median (thick gray line), upper, lower quartile (dark gray area), and 9th, 91st percentile (light gray area).

Figure 5.11: Responses to Visual-Only, Congruent, and Incongruent Stimuli.

unisensory conditions caused by the border effect, which is stronger for stimuli with greater ambiguity (unisensory) than for relatively reliable stimuli (cross-sensory).

In the cross-sensory, incongruent condition, the stimulus was located on average $\overline{\delta} \simeq 2.13°$ right of the visual stimulus. The predicted offset for this condition is $\delta_{MLE} \simeq 2.46°$.

5.4 Intermediate Discussion: Bottom-Up Multisensory Integration

The algorithm and network proposed in Chapter 4 are simple, yet we have shown that they can learn to localize stimuli from noisy multisensory data and that integration of multisensory stimuli then is comparable to MLE, which has also been shown for natural multisensory integration. In simulations and in a neurorobotic experiment, we have demonstrated the network's ability to reproduce formation of topographic multisensory maps, the spatial principle, and the principle of inverse effectiveness which are important aspects of multisensory integration in the dSC (Stein and Meredith 1993).

The network learns to approximately compute a PDF over the latent variable behind its input, and represent it in a PPC.[5] Since the activity of a neuron therefore is the result of computing a PDF, the neurophysiological effects produced by the network are caused by statistical inference in multisensory neurons, as previously theorized by Anastasio et al. (2000).

Simulations in which auditory and visual positions were far apart showed that the network does not integrate such cross-sensory stimuli. Instead, the network roughly selects one of the unisensory stimuli, usually the strongest. The reason for this is, in the simulations described above, the baseline noise. To understand how this leads to the observed behavior, consider two output neurons $\mathbf{o}_v$ and $\mathbf{o}_i$ whose preferred values are l_V and l_i, the intermediate location of the visual stimulus and a location between that of the visual and the auditory stimulus l_A. Each output neuron computes the likelihood of the activity of sensory input neurons under the hypothesis that the true location of the stimulus is its preferred value. With increasing distance between the component stimuli l_V and l_A, that likelihood decreases for neuron $\mathbf{o}_i$ until it becomes less than for neuron $\mathbf{o}_V$. The greater the background noise, the more likely becomes $\mathbf{o}_v$'s interpretation of auditory activity as spurious. Therefore, the distance between component stimuli at which the network stops integrating decreases with increasing background noise.[6]

The behavior thus explained is consistent with what we see in human observers: a visual stimulus and an auditory stimulus are perceived as one cross-sensory stimulus at one location in space if the actual distance between them is below a certain threshold. Otherwise, they are perceived as two unisensory stimuli at different locations (Jack and Thurlow 1973; Roach et al. 2006).

[5] See Section 8.1 for a discussion of this claim.

[6] But see Bauer et al. (2014) for a similar simulation without background noise in which the same effect emerges due to SOM-style learning and limited numerical precision.

A prediction following from this analysis is that the strength of sensory noise during development determines the susceptibility to the ventriloquism effect. Xu et al. (2012) found that dSC neurons in cats which are raised in extremely low-noise conditions[7] develop low tolerance to spatial or temporal incongruency, integrating cross-sensory stimuli only within a small spatio-temporal window. Whether this effect varies systematically with ecological noise and what its effects are at the behavioral level remains to be seen.

[7] Conditions in which all audio-visual stimuli were congruent and highly discernable.

6 Top-Down Modulation and Emergent Attention

In the last section, we have shown that this model reproduces important aspects of natural MSI, namely the spatial principle, the principle of inverse effectiveness, and so-called optimal multisensory integration (Alais and Burr 2004; King 2013; Meredith and Stein 1986a; Stein and Stanford 2008). Like other models of the dSC (or comparable MSI) (Beck et al. 2008; Deneve et al. 2001; Fetsch et al. 2013; Ohshiro et al. 2011; Ursino et al. 2009), ours has so far been purely stimulus-driven. The models due to Anastasio and Patton (2003), Martin et al. (2009), Pavlou and Casey (2010), Rowland et al. (2007), and Cuppini et al. (2012) do include cortical regions projecting to the SC. However, both Anastasio and Patton (2003) and Martin et al. (2009) have modeled *only* the effect of cortical input on multisensory enhancement in the dSC, leaving aside the topographic organization which is characteristic of dSC neurons' RFs (King 2013; Sparks 1988; Wallace and Stein 1996). The models put forward by Rowland et al. (2007) and Cuppini et al. (2012), while modeling the effect of cortical input on multisensory integration in the dSC, focus on replicating biology and refrain from interpreting the meaning of cortical input, network connectivity, and neural computations, functionally. They are pure bottom-up models in the sense of Krasne et al. (2011).

Our model on the other hand was specifically developed with functionality and mathematical interpretation in mind: in our model of the dSC, a self-organizing network learns the statistics of its input which it uses to infer the location of a stimulus from noisy, population-coded input. Its output approximates a population-coded PDF over that location. In this section, we extend that model from a stimulus-driven model to one which also considers top-down input: specifically, we test the idea that effects of spatial and feature-based attention are based on very similar mechanisms (Maunsell and Treue 2006). In fact, we model attentional input as just another source of input, indistinguishable to dSC neurons from sensory input. We show that statistical self-organization, the basic mechanism of our ANN model, produces effects very similar to those of natural spatial and feature-based attention observed *in vivo*. It also naturally produces specialization to different stimulus combinations in dSC neurons (Stein 2012a; Wickelgren 1971), a feature which has been interpreted in mathematical terms by Colonius and Diederich (2004) but whose development has not been modeled, to our knowledge.

While extending the modeling, we preserve the network and algorithm. The goal is to test the hypothesis that the effects of attention can be explained (in part) by the same mechanisms used to model learning of multisensory integration. The only aspect we therefore change in our model is the nature of the input, which now is not only stimulus-driven, but also reflects higher cognitive processes.

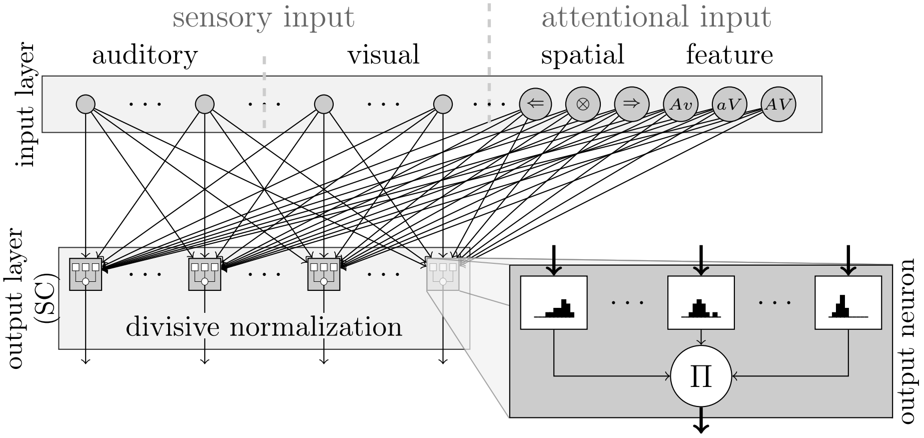


Figure 6.1: Our Model of Attention in Multisensory Integration

6.1 Extended Model: Sensory and Attentional Input

The structure of our network is shown in Figure 6.1: Again, all input neurons are modeled as part of one conceptual input layer regardless of their actual origin.

Sensory Input. The network was trained on simulated input consisting of ‘sensory’ and ‘attentional’ components (see Figure 6.1). The sensory component was in itself separated into ‘visual’ and ‘auditory’ parts. Stimuli were defined by their location $L \in [0, 1]$ and their stimulus class $C \in \{\mathrm{V}, \mathrm{A}, \mathrm{AV}\}$. The class determined the strength of the individual components. Stimuli of class V (‘visual’) or AV (‘audio-visual’) had strong visual components. Stimuli of class A (‘auditory’) or AV had strong auditory components. Concrete realizations l and c of the stochastic variables L and C were selected randomly and uniformly distributed in every step during training.

All sensory input neurons responded to a simulated stimulus at location l according to Poisson-noisy Gaussian tuning functions: each one of the $n_{\mathbf{i}} = 25$ auditory and visual input neurons $\mathbf{i}_{m,k}, m \in \{V, A\}, k \in [1..n_{\mathbf{i}}]$ had a preferred location

$$l_{\mathbf{i}_{m,k}} = \frac{k-1}{n_{\mathbf{i}}-1}.$$

Each input neuron’s Gaussian tuning function was centered around this preferred location. The activity $\mathbf{a}_{\mathbf{i}_{m,k}}$ of $\mathbf{i}_{m,k}$ in response to a stimulus of class c at location l was then determined by the stochastic function

$$\mathbf{a}_{\mathbf{i}_{m,k}} \sim \mathrm{Pois}\left(s(m,c) \times g_m \times \exp\left(-\frac{(l_{\mathbf{i}_{m,k}} - l)^2}{\sigma_m^2}\right) + \nu_s\right), \tag{6.1}$$

where

$$s(m,c) = \begin{cases} 1 & \text{if } m = \mathrm{V} \wedge c \in \{\mathrm{V}, \mathrm{AV}\} \\ 1 & \text{if } m = \mathrm{A} \wedge c \in \{\mathrm{A}, \mathrm{AV}\} \\ 0.5 & \text{otherwise.} \end{cases} \tag{6.2}$$

Again, g_m and σ_m are the modality-specific gain and width of the tuning functions, ν_s is the sensory background noise parameter, and again, the particular shape of the above tuning functions and the kind of noise is not important in the context of our model, but our choice is biologically plausible and in good keeping with modeling practice.[1] The differences between Equation 6.1 and Equation 5.1 from the previous section are the additional gain factor $s(m, c)$, which depends on the stimulus class. See Section 6.2 for a discussion of the effects of different choices of values for ν_s and other parameters.

Attentional Input. Apart from sensory input neurons, our model includes two types of input neurons from higher-level cognitive brain regions. The first type of what we will call 'attentional' input neurons encode information about the general region in which the stimulus is. These three neurons code for stimuli which are on the left ($\Leftarrow$), in the middle ($\otimes$), or on the right ($\Rightarrow$) of the simulated visual field. Another three neurons code for the type of stimulus. One neuron each codes for stimuli which are highly visible (aV), highly audible (Av), or both (AV).

We will call the former type 'spatial' input neurons and the latter type 'feature' input neurons. The intuition behind these additional input neurons is that we often have an expectation of which kind of stimuli we will be presented with. Often, we will expect a stimulus on the left or the right side of our visual field, and we will expect something that is very loud, or bright, or both. The encoding and activation of this knowledge (mostly in cortical areas) is represented in our model in the strongly simplified form of neurons whose activity is either 1 or 0 depending on whether the location or type of the expected stimulus is the one preferred by the respective conceptual input neuron.

Like sensory input, attentional input is modeled as stochastic, modeling non-determinism of ecological conditions, cognitive processes, and neural responses. More specifically, the activity of our attentional input neurons in every trial is modeled as a Bernoulli process whose parameter p depends on the location and class of the stimulus. The (deterministic) activation $\hat{\mathbf{a}}_{\Leftarrow}$, $\hat{\mathbf{a}}_{\otimes}$, and $\hat{\mathbf{a}}_{\Rightarrow}$ of the spatial input neurons $\mathbf{i}_{\Leftarrow}$, $\mathbf{i}_{\otimes}$, and $\mathbf{i}_{\Rightarrow}$, respectively, is modeled by the three functions:

$$\hat{\mathbf{a}}_{\Leftarrow} = \frac{\upsilon}{1 + \exp((l - 0.1) * 40)} + \nu_c, \tag{6.3}$$

$$\hat{\mathbf{a}}_{\otimes} = \exp\left(\frac{-(l - .5)^2}{0.05}\right) * \upsilon + \nu_c, \tag{6.4}$$

$$\hat{\mathbf{a}}_{\Rightarrow} = \frac{\upsilon}{1 + \exp(-(l - 0.9) * 40)} + \nu_c, \tag{6.5}$$

where $\nu_c = 0.05$ and $\upsilon = 0.9$ are noise parameters. These seemingly complex functions are in fact just two sigmoidal functions which have large values to the left and to the right of the interval $[0, 1]$, respectively, and a Gaussian function centered around 0.5 (see Figure 6.2). The activation of feature input neurons was simply $\hat{\mathbf{a}}_c = 1 - \nu_c$ whenever

[1] See Figure 5.1.

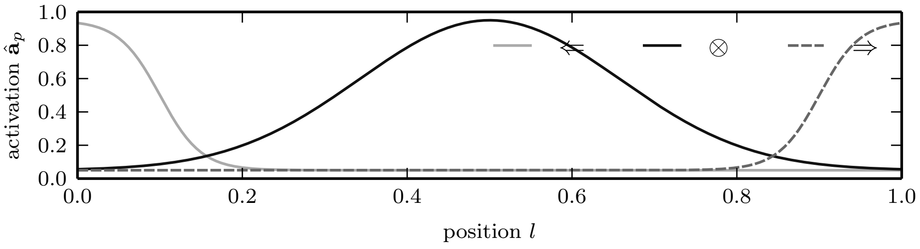


Figure 6.2: Activation $\hat{\mathbf{a}}_p$ of Attentional Input Neuron $\mathbf{i}_p$ for $p \in \{\Leftarrow, \otimes, \Rightarrow\}$
.

the actual stimulus class was c for $c \in \{aV, Av, AV\}$, and $\hat{\mathbf{a}}_c = \nu_c$ otherwise. Activity of each attentional input neuron was then stochastically computed from the activation:

$$\mathbf{a}_p \sim \mathrm{Bern}(\hat{\mathbf{a}}_p), \text{ for } p \in \{\Leftarrow, \otimes, \Rightarrow, aV, Av, AV\}.$$

The SC receives descending projections from various areas in the cortex.[2] Some of those play a role in attention, like FEF, DLPFC, and LIP.[3] In cats, AES plays an especially important role: its deactivation eliminates neurophysiological MSI (Wallace and Stein 1994) and drastically alters audio-visual orientation behavior (Wilkinson et al. 1996). It has been implicated with selective attention (Dehner et al. 2004; Foxe 2012), due to its effect on neural responses in the SC. Since orienting behavior is linked to attention (Kustov and Robinson 1996; Ignashchenkova et al. 2004, more recently), this implication is potentiated by the behavioral findings of Wallace and Stein (1994). In our model, 'attentional' input may relate, for example, to FEF, for more spatial input (Bruce et al. 1985), or AES, in cats, for more feature-related input (Dehner et al. 2004).

6.1.1 Training

We trained a network of $n_{\mathbf{o}} = 500$ output neurons extensively for 300000 training steps (as in Section 4.3.2). The one distinct feature of our parameter setting was the minimum neighborhood width (of $0.001 < \frac{1}{n_{\mathbf{o}}}$) which we chose deliberately small. With a small neighborhood width, neurons which are close to each other are permitted to learn to respond to different stimuli. Given that training sets up a roughly topography-preserving mapping from data space into the grid while the neighborhood interaction is still large, we expected that neurons which were close to each other would learn to respond to different special cases of similar input. Specifically, we expected that they would self-organize to have similar preferred locations but different stimulus classes.

[2] See Section 2.2.6.
[3] See Section 2.2.5.

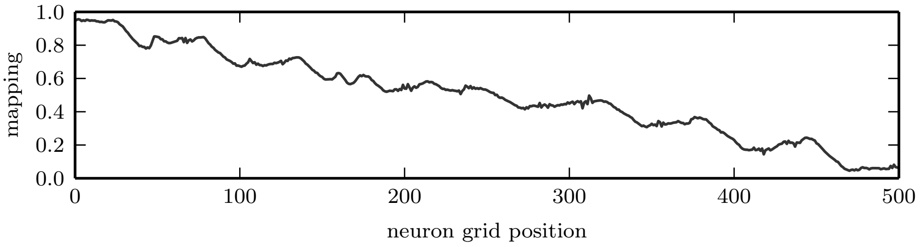
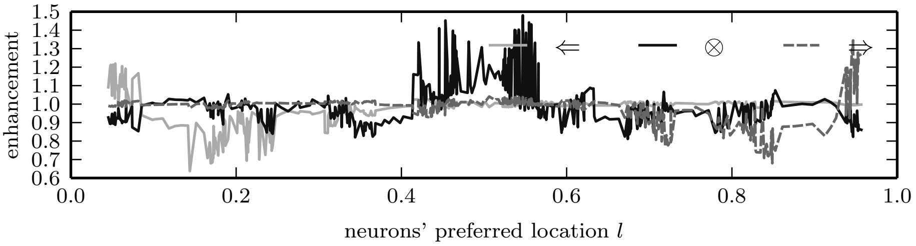


Figure 6.3: Mapping of Neurons to Stimulus Positions.



> Average activation of neurons given cortical input coding for spatial classes ⇐ (▒), ⊗ (■), ⇒ (▓) divided by average activation given zero attentional input.

Figure 6.4: Effect of Spatial Attention on Neural Responses.

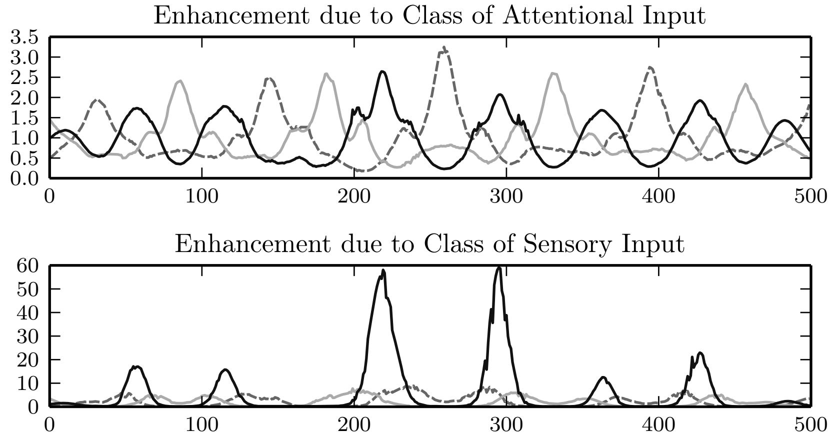
6.2 The Effect of Cortical Input

Mapping. To determine the preferred location of each neuron, we simulated input at 50000 positions, evenly spaced across the interval $[0, 1]$. At each location, we generated input for each stimulus class, and determined for each neuron $\mathbf{o}$ the mean of all locations l for which it was the BMU.[4] See Figure 6.3 for the resultant mapping from neurons to locations.

Enhancement. Endogenous spatial attention enhances the activity of SC neurons whose receptive fields overlap the attended region.[5] To demonstrate similar behavior in our network, we divided the mean activity of each neuron for trials in which 'attentional' input signaled a stimulus of spatial class $\Leftarrow$ by the mean activity of that same neuron with zero attentional input (and the same for $\otimes$ and $\Rightarrow$, compare Equation 5.2 in the previous section). Figure 6.4 shows that activating the neurons coding for $\Leftarrow$, $\otimes$, and $\Rightarrow$ clearly enhanced mean activity in those neurons whose preferred values were in the respective region.

[4] See Section 4.3.3.

[5] See Section 2.2.5.



Enhancement due to Class of Attentional Input

Enhancement due to Class of Sensory Input

Top: Average activation of neurons given cortical input coding for spatial classes aV (■), Av (■), AV (■) divided by average activation given zero cortical input. **Bottom:** Average activation of neurons given Va, vA, VA *sensory* input divided by average activation given va input.

Figure 6.5: Feature Selectivity.

In contrast to spatial attention, feature-based attention enhances activity of neurons selective to the features attended to across the visual field (Born et al. 2012; Maunsell and Treue 2006). We tested whether this was also true for our network by simulating multisensory input at 100 regular positions between 0 and 1. For each of these positions, we generated 100 sensory input and corresponding spatial activations which we combined once with feature activations coding for each of the stimulus classes $c \in \{aV, Av, AV\}$ and for no stimulus class (av), respectively. From the network's output activation, we computed enhancement for each of the stimulus classes: for each output neuron $\mathbf{o}$, we selected those cases where the difference between the actual stimulus location l and $\mathbf{o}$'s empirical preferred value $l_{\mathbf{o}}$ was within ±0.01. For each stimulus class, we divided the neuron's mean activity in cases where the cortical activity coded for that class by the mean activity in cases where the cortical activity did not code for any stimulus class. See top graph of Figure 6.5 for plots of enhancement for each of the stimulus classes. We can see that neurons specialized in *attentional* input coding for different stimulus classes.

Stimulus Selectivity. To test whether neurons also specialized in different types of *sensory* input, and whether they generally specialized in the same kind of attentional and sensory input, we evaluated for each neuron the enhancement of activity due to Va, vA, VA *sensory* input to va sensory input. Specifically, we divided, for each neuron, the mean given Va, vA, and VA sensory input by the mean activity given av input (when the stimulus was close to their preferred stimulus position, using the input and output activities generated to compute selectivity for *attentional* input, see above). The bottom graph in Figure 6.5 shows the result: we see, again, that neurons specialized in different kinds of input—this time, in different kinds of *sensory* input. A comparison of the two graphs in Figure 6.5 also suggests that the same neurons were generally selective for sensory input from one combination of modalities and for attentional input coding for such a stimulus. Especially, neurons selective for AV stimuli were also selective for the corresponding attentional input. Note also that some neurons' responses were depressed by attentional activation coding for their non-preferred stimulus combination (values < 1).

Since the same is a bit hard to see for Va and vA stimuli, in Figure 6.5, the relationship between responsiveness t each combination of modalities and attentional enhancement is plotted in Figure 6.6. What the figures show is that neurons which responded strongly to Va stimuli also tended to have their response enhanced by attentional input coding for Va input. More strikingly, their response was depressed by attentional vA input.

Localization. Having tested the effect of attention on the network's activity, we next tested how this effect was reflected in decisions made using the network's responses. To do that, we simulated input in which the visual and auditory component had different locations l_v and l_a, respectively. Both components were strong unisensory components ($c = AV$), but each sensory component was combined once with attentional input coding for each of the stimulus classes and for no stimulus class. Using the empirical mapping of neurons to positions, we then derived a localization of the incongruent input.

Figure 6.7 shows the distribution of relative localizations made by the network depending on the stimulus class represented by the feature-encoding input neurons. The individual graphs show histograms of the localization l_n of incongruent stimuli relative to the location of visual and auditory sub-stimuli l_v and l_a, depending on the absolute distance $|l_v - l_a|$ and cortical input. We see that attentional input coding for the stimulus class influences localization of incongruent audio-visual stimuli: at larger distances, visibly more stimuli were localized close to the auditory sub-stimulus if attentional content coded for a vA stimulus than in other conditions. Also, already at lower inter-stimulus distances, the mean of localizations in that condition is closer to the auditory stimulus. With attentional input coding for a Va stimulus, less stimuli were localized close to the auditory stimulus at large distances, and on average localizations were shifted towards the visual stimulus, compared to the other conditions.

Finally, to test whether spatial attention affected localization, we simulated incongruent audio-visual stimuli paired with spatial attention: In 10000 steps, we simulated a visual stimulus in the left third of the interval $[0, 1]$ and an auditory stimulus in the right third. We then combined the sensory input with attentional input coding for each combination of each of the spatial classes $\Leftarrow$, and $\Rightarrow$, and each of the stimulus classes *Va*, *vA*, *VA*, and *va*. After that, visual and auditory stimulus positions were switched, in every step, and combined with attentional input as above, giving us a total of 80000 input activations. We found that the network localized the combined stimuli on average at a position of 3.97×10^{-1}, relative to the interval $[l_v, l_a]$, as above, when spatial atten-

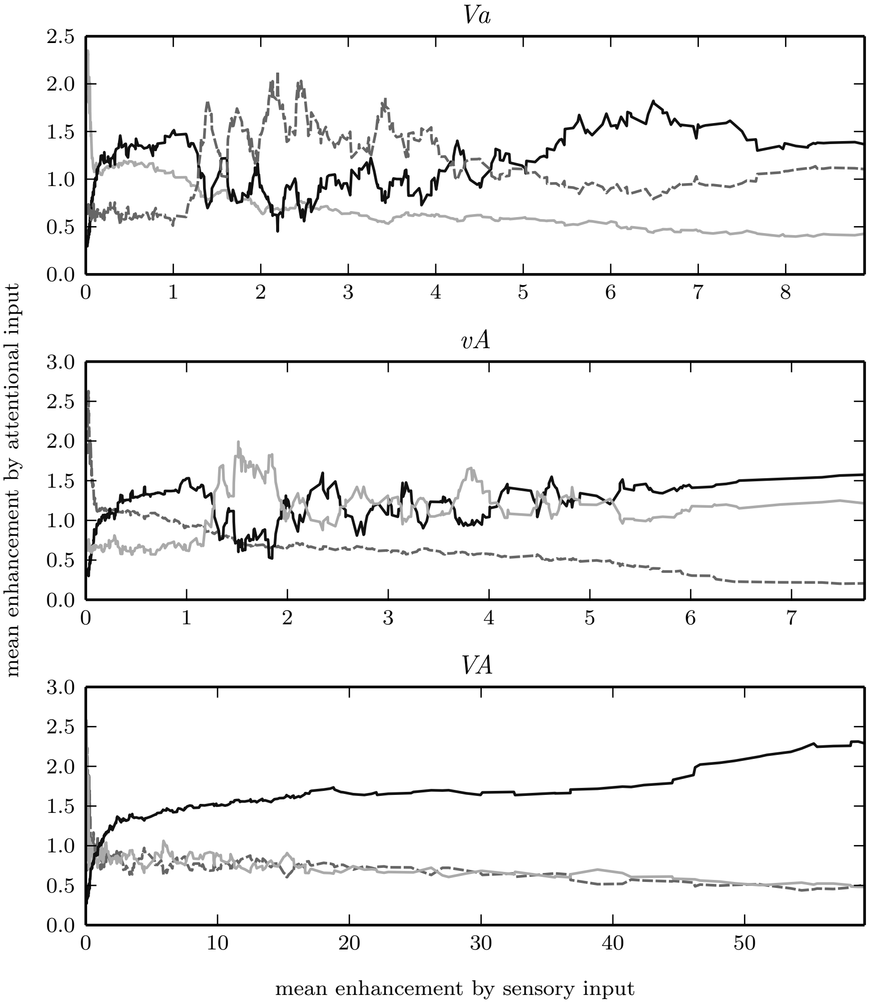
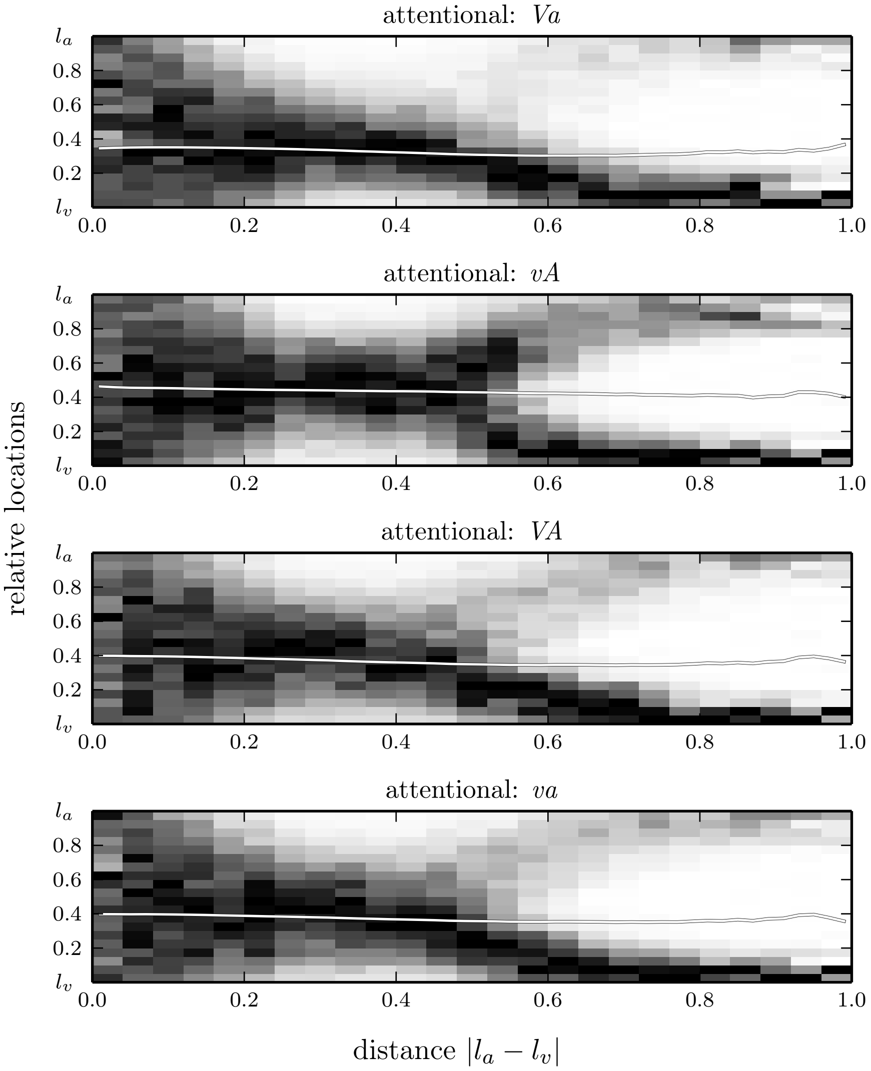


Figure 6.6: Effect of Feature-Based Attention Related to Sensory Selectivity



Gray scale: Frequency of relative localizations between visual (l_v) and auditory (l_a) sub-stimulus, depending on distance between l_v and l_a, given different attentional input. The values in each of the columns were normalized by dividing them by the maximum value in that column to improve legibility.
White lines: Mean relative localization.

Figure 6.7: Integration vs. Decision by Relative Stimulus Distance.

tion was on the side of the visual stimulus and 4.61×10^{-1} when it was on the side of the auditory stimulus. This means that spatial attention had a sizable effect on localization.

Parameters. All effects discussed in the next section were qualitatively robust under broad ranges of parameter settings. However, we did observe interesting quantitative effects due to tuning function parameters, which determined the information available for localization: information increased with lower background noise ν_s and greater gains g_a, g_v (Equation 6.1).

We ran experiments in which either the relative size of the sensory gains g_a, g_v was manipulated (Table 6.1a), they were jointly scaled, (Table 6.1b), or the baseline noise parameter ν_s was manipulated (Table 6.1c). For each experiment, we then computed the mean localizations given incongruent sensory and varying attentional input relative to the interval $[l_v, l_a]$, as above (columns μ_{Va}, μ_{vA}, μ_{VA}, μ_{va} in Table 6.1). We also fitted two models to the distributions of relative localizations at different absolute distances $|l_a - l_v|$: one model was a simple Gaussian model, while the second was a mixture of two Gaussians whose respective modes were at the location of the visual stimulus, l_v, and the auditory stimulus, l_a. Thus, the first was an integration model, while the other was a stimulus selection model. We then used Akaike's information criterion (AIC) (Akaike 1974; deLeeuw 1992) to determine the least distance $|l_a - l_v|$ at which the stimulus selection model described the distribution of localizations better than the integration model (columns a_{Va}, a_{vA}, a_{VA}, a_{va} in subtables of Table 6.1).

Unsurprisingly, more information in the visual or less in the auditory modality (larger gain g_v, lower gain g_a) caused localizations to generally move towards the visual stimulus in incongruent conditions (see Table 6.1a). More interestingly, the amount of sensory information was reflected in the maximum distance at which stimuli were integrated: what we can see in Tables 6.1b and 6.1c is a tendency for the mean of localizations to move towards the visual stimulus in *Va* conditions and towards the auditory stimulus in *vA* conditions with *less* sensory information (columns μ_{Va}, μ_{vA} in Tables 6.1b and 6.1c). Also, the network tends to stop integrating and start selecting one of the sub-stimuli earlier with less sensory information (strong background noise ν_s, low sensory gains g_v, g_a) than with more sensory information (smaller values in columns a_{Va}, a_{vA}, a_{VA}, a_{va}).

Unfortunately, as we can see, it is hard to make out a consistent pattern in the relationship between the amount of sensory information, attentional input, and integration versus stimulus selection. While there are appreciable differences between the columns a_{Va}, a_{vA}, a_{VA}, and a_{va} of Tables 6.1b and 6.1c, these differences do not coherently point into one direction. To be able to make a statement about the effect of sensory information on that of attentional input on integration and stimulus selection, many more simulations would be necessary. Additionally, a statistic different from the one used here, the minimum distance between l_v and l_a at which AIC favors the stimulus selection model, may be more appropriate for our purposes. Since the focus of, here, is more on qualitative effects of attention than on quantitative differences with varying parameter settings, we leave these aspects for future work.

g_v, g_a	a_{Va}	a_{vA}	a_{VA}	a_{va}	μ_{Va}	μ_{vA}	μ_{VA}	μ_{va}
8.0,5.0	0.629	0.591	0.619	0.611	0.241	0.280	0.252	0.250
8.0,7.0	**0.663**	**0.653**	**0.649**	**0.643**	**0.364**	**0.458**	**0.405**	**0.407**
10.0,7.0	0.685	0.665	0.651	0.663	0.281	0.331	0.301	0.303

(a) Alternative Relative Sensory Gains g_v, g_a

g_v, g_a	a_{Va}	a_{vA}	a_{VA}	a_{va}	μ_{Va}	μ_{vA}	μ_{VA}	μ_{va}
3.0,2.6	0.591	0.649	0.617	0.599	0.341	0.532	0.412	0.423
4.0,3.5	0.597	0.629	0.617	0.587	0.346	0.492	0.417	0.412
5.0,4.4	0.589	0.655	0.681	0.635	0.349	0.482	0.412	0.412
6.0,5.2	0.667	0.683	0.619	0.635	0.358	0.469	0.407	0.408
7.0,6.1	0.697	0.667	0.621	0.647	0.366	0.448	0.410	0.406
8.0,7.0	**0.663**	**0.653**	**0.649**	**0.643**	**0.364**	**0.458**	**0.405**	**0.407**
10.0,8.8	0.705	0.709	0.655	0.689	0.383	0.447	0.413	0.414
12.0,10.5	0.745	0.747	0.733	0.741	0.367	0.432	0.417	0.407
14.0,12.2	0.727	0.737	0.733	0.733	0.380	0.451	0.424	0.420
16.0,14.0	0.647	0.659	0.675	0.663	0.380	0.466	0.421	0.421
18.0,15.8	0.747	0.745	0.703	0.735	0.381	0.452	0.406	0.410

(b) Scaled Sensory Gains g_v, g_a

ν_s	a_{Va}	a_{vA}	a_{VA}	a_{va}	μ_{Va}	μ_{vA}	μ_{VA}	μ_{va}
0.5	0.810	0.802	0.818	0.812	0.398	0.458	0.417	0.423
1.0	0.705	0.725	0.721	0.715	0.387	0.446	0.406	0.410
1.5	0.717	0.743	0.689	0.717	0.381	0.440	0.407	0.406
2.0	0.745	0.721	0.687	0.713	0.374	0.446	0.416	0.410
2.5	0.655	0.663	0.661	0.655	0.370	0.450	0.407	0.406
3.0	**0.663**	**0.653**	**0.649**	**0.643**	**0.364**	**0.458**	**0.405**	**0.407**
4.0	0.661	0.685	0.629	0.641	0.373	0.458	0.408	0.411
5.0	0.619	0.645	0.639	0.631	0.371	0.450	0.409	0.409
6.0	0.621	0.665	0.619	0.615	0.356	0.474	0.405	0.411
7.0	0.623	0.667	0.621	0.639	0.352	0.464	0.405	0.404
8.0	0.607	0.625	0.625	0.613	0.356	0.475	0.408	0.412

(c) Alternative Baseline Noise Levels ν_s

Changing baseline noise levels and sensory gains affected the maximum distance at which stimuli were integrated and how strongly localization was influenced by attentional input. $a_c, c \in \{Va, vA, VA, va\}$: the least distance at which Akaike's information criterion was in favor of a stimulus selection model given attentional input of class c. $\mu_c, c \in \{Va, vA, VA, va\}$: mean of all relative localizations given c (analogous to y-axes in Figure 6.7). **Bold rows:** same parameters as in the rest of the paper.

Table 6.1: Comparison of Alternative Parameter Settings.

6.3 Intermediate Discussion: Attention

Figures 6.5 and 6.6 show clearly that some neurons reacted much more strongly to attentional and sensory input related to one stimulus class than others. Neurons whose activity was strongly enhanced by AV-class stimulus were different from those whose enhancement for Av-class stimuli was strong, and vice versa. This enhancement was reflected in the decision made by the network: attentional input coding for an Av stimulus led to substantially more localizations close to the auditory sub-stimulus than cortical input coding for any other stimulus class. This can be seen in Figure 6.7, where the upper 'arm' of the distribution at greater inter-stimulus distances has visibly more weight for attentional vA input, and in the mean of localizations, which is closer to the visual stimulus at all distances (see also columns $\mu_{Va}, \mu_{vA}, \mu_{va}$ in Table 6.1.

We relate these effects to those of feature-based attention: attention focused on the visual features of an object enhances the activity of neurons across the visual field in whose receptive fields is a stimulus with the attended features if they are sensitive to those features. On the behavioral side, attending to certain stimulus features will increase detection of objects with these features (Andersen et al. 2009; Born et al. 2012; Maunsell and Treue 2006). Similarly, activating the cortical content coding for a stimulus with high auditory and low visual salience enhanced the activity of specific neurons in our network and it increased the likelihood for the network to choose the location of the auditory sub-stimulus over that of the visual sub-stimulus.

Moreover, like in the experiments by Warren et al. (1981), Jack and Thurlow (1973), and Vatakis and Spence (2007), semantic content changed the extent to which stimuli in different modalities were integrated: whenever the 'unity-assumption' was strong (the belief that there is one, cross-sensory stimulus, see Vatakis and Spence (2007) and Warren et al. (1981)), i.e. when cognitive content coded for a highly audible and visible stimulus, stimuli had a higher likelihood of being integrated than when it was weak, i.e. when cognitive content coded for just a strong auditory or just a strong visual stimulus. This is reflected in the mean relative localizations under the different conditions, visualized in Figure 6.7, and in the respective columns in Table 6.1.

Mechanistically, the effects described above emerge because competitive learning leads to specialization among neurons such that different neurons react to different stimuli. Each neuron specializes in stimuli from a specific position in simulated space, and, to varying extent, to a specific stimulus combination. SOM-style self-organization tries to embed the topology of data space into the network's grid. Since data space is two-dimensional (stimulus position and stimulus type) but the grid only has one dimension, this cannot succeed completely. One of the dimensions—generally the one describing less variance in the data—would have been ignored by the network if we had kept the neighborhood size during learning above a certain threshold. Intentionally decreasing the neighborhood size to a very small number allowed the network to have some non-monotonicity in the mapping (see Figure 6.3), as it were, an effect similar to what Kohonen (1995, p. 87 *f*) calls 'zebra stripes.'

Miikkulainen et al. (2005, p. 62 *f*) call this effect 'folding' and they showed how it can produce structures resembling ocular dominance stripes or stripes of neurons selective for different stimulus orientations in the visual cortex. Ocular dominance stripes are also present naturally in the SCs of monkeys (Pollack and Hickey 1979) and they have been shown to arise in the tecta of tadpoles when they are implanted with a third eye (Law and Constantine-Paton 1981). In our context, multiple neurons came to code for the same location, but combined with a different stimulus class.

Specialization of neurons not only in stimuli from some direction but also of a certain stimulus class implements an important feature of natural multisensory integration. Wallace and Stein (1996) have found that not all dSC neurons react to stimuli in all or even more than one sensory modality. This has been modeled computationally by Colonius and Diederich (2004) who make a normative argument for why there are unisensory neurons in the dSC. That argument goes along the lines that a neuron which uses only evidence from one sensory modality to decide whether a stimulus is in its receptive field is not affected by noise in any other modality. Our model produces such a specialization, as can be seen in Figures 6.5 and 6.6, and it makes this argument more specific: according to our account, a mixture of unisensory and multisensory neurons effectively evaluates hypotheses about stimulus combinations and stimulus locations. It then chooses that stimulus combination and location which is most consistent with the evidence. In this context, cognitive content (attention) can either be seen as additional evidence or, equivalently, as a prior over stimulus locations and combinations.

Together, these findings show that attentional input to the dSC needs no different wiring from that of sensory input to have the neurophysiological and behavioral effects seen in experiments, which is the main result of this paper. Of course, this does not preclude the possibility that goal-directed learning may play a role. Weber and Triesch (2009a) have shown how essentially unsupervised learning can be extended and combined with mechanisms from reinforcement learning to emphasize learning of goal-relevant over goal-irrelevant features. Similarly, if there is a goal-directed feedback signal to the dSC, that feedback signal could modulate the unsupervised training process. What we show here is that in our model neither feedback is needed to produce the neurophysiological and behavioral effects shown here, nor do projections from different sources of input need to be treated differently in the overall architecture or in how they are treated by integrative dSC neurons.

Our model fits in well with the view of Desimone and Duncan (1995) that attention may be not so much a mechanism in itself, but a phenomenon emerging from of competition between stimuli for representation. Competition is one of the main features of SOM-like algorithms and it is additionally implemented by the mechanism of divisive normalization employed by our algorithm. As in the biased competition model of visual attention, attentional input to our model biases competition towards one combination of substimuli. More recently, and similarly, Krauzlis et al. (2014) suggested that the effects of attention really arise as by-products of processes serving the need for effective information processing. Krauzlis et al. (2014) argue that, for example, the SC is involved in regulating spatial attention behaviorally, but neural activity related to selective attention in visual cortex remains after collicular deactivation. Furthermore, even animals without a well-developed neocortex or even SC show signs of selective attention. Since no single brain region or circuit seems necessary for an organism to exhibit behavioral effects of attention, Krauzlis et al. (2014) argue that attention and its known neural correlates emerge simply because effective biological (and artificial) information processing requires state estimation. The estimated state at any point then modulates action and perception. We would add that zooming into loci which seemingly evolved to implement state estimation, like the SC, may show that, there again, attention is not an inbuilt mechanism but an emergent effect resulting from neurons using all available information to accomplish their function in the best possible way.

7 Modeling Summary and Discussion

In this Part, we have presented a model of learning of MSI in the dSC. The model assumes a few basic mechanisms and employs them to explain how the immature dSC may learn to integrate noisy stimuli from multiple sensory modalities. We have been able to demonstrate that our model can produce phenomenology in the simulated neural activities which is analogous to important phenomenology in natural MSI: with sensory input only, it reproduces topographic mapping and the spatial principle as well as the principle of inverse effectiveness. Extended by cortical and multi-class sensory input, the neural activity simulated by the model shows enhancement effects which we relate to effects of spatial and feature-based attention. More specifically, neurons in the model specialize in different classes of input (visually, acoustically, and visuo-acoustically salient stimuli) and their responses are enhanced by cortical input coding for their respective stimulus class.

On the behavioral side, we have seen that the decisions made by the model show similar patterns to those made by humans. Without any bias, auditory and visual stimuli are integrated, up to a certain spatial disparity, and the cross-sensory stimulus is localized between the actual position of the two component stimuli. Where, between the component stimuli, they are localized depends on the reliability of the two modalities: our network integrates those stimuli it does integrate as if using a linear combination of localizations of the two component stimuli, in which the weights applied to either component are determined by the reliability of the respective modality, consistent with an MLE model of MSI. This same model has been shown to describe the behavior of human participants well in audio-visual and other cross-sensory integration tasks.

The decision of whether or not to integrate cross-sensory stimuli depends on cognitive content. The network learns to use input from simulated cortical neurons which encode in their activity the type of stimulus. Cross-sensory stimuli with incongruent visual and auditory components have a higher chance of being integrated when cognitive content signals a cross-sensory stimulus than when it signals a largely visual or largely auditory stimulus—in our model and in humans.

What is particularly interesting about our model is that our model neurons do not have access to any information about the origin of their inputs in sensory or cortical areas. Rather, the network learns topographic mapping and localization in an unsupervised fashion which treats each of the inputs in the same way. Unsupervised learning is important because learning of MSI in the dSC seems to be at least partially unsupervised. And that makes sense for this brain region because there is much more sensory experience to learn from than opportunities to derive feedback from the environment. Homogeneous treatment of all input neurons is an important feature because even if dSC neurons had some way of distinguishing input neurons with different origins, the significance of their firing would be too different across neurons from the same origin to use that information for integration.

While our model incorporates only unsupervised learning as a mechanism for setting up topographic mapping, this is not inconsistent with other mechanisms which may be at play in the dSC. In the biological dSC, prenatal processes produce rough topographic retinotectal connectivity (e.g. Drescher et al. 1997; Fraser 1992; discussed in Stein and Stanford 2013), with large differences between species (Stein and Stanford 2013; Wallace and Stein 2001). It is likely that projections from other sensory areas also possess a rough topographic organization at birth (Stein and Stanford 2013). In the context of our modeling, this would correspond to a pre-initialization of the network's topology. Such pre-initialization is in fact recommended, where possible, for SOM learning (Kohonen 2013) to speed up convergence of learning. Similarly, a vertebrate which is born with roughly topographic retinotectal projections which mature with experience would benefit from fast learning without sacrificing the flexibility to adapt to the environment (and its own body). Such an optimization would present a significant evolutionary benefit.

Furthermore, it has been shown that adaptation in the (adult) dSC can be influenced by active involvement with multisensory stimuli (Bergan et al. 2005). Such results point towards reward-mediated learning or at least activity-mediated plasticity in the dSC. This, too, is not reflected in our model but it could easily be included and discussed in terms of SOM learning: unsupervised learning algorithms generally learn to distinguish between different kinds of input. However, tasks might make discrimination between some kinds of input more important than others and thus may call for stronger weighting of some input dimensions or features than others (Saeb et al. 2009; Weber and Triesch 2009a). Thus, including value-based aspects in an unsupervised algorithm such as the SOM algorithm can help emphasize the detection of those differences that are important. In the dSC, these may be predator- or pray-related cues which might be given priority over cues likely to be derived from features of the environment. Both mechanisms, pre-initialization and reward- or activity-mediated influences of learning, would be too good an idea for evolution *not* to develop them. What we have shown in this chapter is that self-organized statistical learning alone can already account for a great deal of phenomenology and may thus be an important mechanism at work in the dSC.

One prediction of our model is the existence of neurons whose activity is *depressed* by a strong stimulus in their non-preferred modality, even if that stimulus is in their receptive field. To our knowledge, this effect has not been observed experimentally. We see a number of possible reasons for this. First, depression has not been studied as extensively as enhancement. Second, it is difficult to precisely determine the best stimulus location of a neuron and thus to tell with any confidence whether the (perceived) location of an auditory stimulus is exactly at the same location as the visual stimulus. This is especially true for a neuron which does not respond strongly to an auditory stimulus to begin with. Third, it might be that ecologically sensory noise is so great in relation to sensory information that depression vanishes or at least becomes hard to detect (see Section 6.2). In this case, depression due to congruent stimuli in a non-preferred modality would be more likely to develop under unusually noiseless conditions. Finally, it may just be that the neural implementation does not permit this kind of depression. If sensory noise is typically high compared to sensory information, then our simulations show that depression would be weak and therefore its behavioral benefits could become negligible. In that case, it may be economical to completely prune connections to input neurons from the non-preferred modality, thereby eliminating the small amount of depression that would be present otherwise. Targeted neuroscientific inquiry into this issue could give valuable insight here.

Finally, the probabilistic origins of the network at the basis of our model suggest an elegant functional interpretation of all the effects reproduced by the model. According to this interpretation, self-organization produces a PPC in which each neuron represents in its activity the probability of a specific hypothesis about the position and the quality of a stimulus in terms of modality combination.[1]

7.1 Methodology

7.1.1 Level of Modeling

Our model is not a model on the level of biological implementation. As we have mentioned, histograms as a method of approximating likelihood functions are not biologically plausible and we do not specify any particular methods for implementing divisive normalization and winner selection. We see our model on the algorithmic level, in Marr's (1983) sense, akin for example to that due to Ohshiro et al. (2011): it is situated between computational theories of multisensory integration, like the MLE model due to Alais and Burr (2004), and theories of hardware implementation like for example the one due to Cuppini et al. (2012). In this position, it connects the interaction of the parts of the hardware implementation to the behavior described by the mathematical theory and thus allows for predictions about what happens to the hardware if there are changes on the behavioral level, and vice versa.

That said, we believe that the operations used by our algorithm are simple enough to be implemented by a network of biological neurons. Divisive normalization has explained many aspects of neural responses in sensory processing, and possible mechanisms have been identified (Carandini and Heeger 2011). A literal implementation of histograms by neurons which count the times they observe a particular input activation at each of their synapses is implausible. However, neurons apparently encoding probabilistic variables, computed from sensory input, have been found (Yang and Shadlen 2007) and Soltani and Wang (2010) have shown how neurons may learn to compute such variables from input using biologically plausible mechanisms. Extensions of Soltani and Wang's work are a promising direction for implementing the statistical computations necessary for our model to reach further into Marr's (1983) domain of hardware implementations. Such extensions are the subject of future work. In the meantime, our assumption, backed by results such as those from Yang and Shadlen (2007), is that neurons or neural

[1] see Sections 4.2 and 8.1

microcircuits *can* gradually learn to compute likelihood functions from input. We use histograms to model this ability and propose that it may play a role together with self-organization in learning multisensory integration.

If neural microcircuits possess this ability to compute likelihoods, and if self-organization similar to the account in this study occurs, then that could explain how the PPC presupposed by studies such as those due to Beck et al. (2008), Cuijpers and Erlhagen (2008), and Fetsch et al. (2013) could come about. Like the models proposed by Ohshiro et al. (2011) and Fetsch et al. (2013), our model uses divisive normalization and the neurophysiological effects it reproduces depend in part on that. Our model differs, however, in that input neurons whose activity is excitatory for some integrative output neurons can be directly inhibitive for others and that excitation and inhibition develop through learning of input statistics. In the divisive normalization models cited above, such an inhibition may arise indirectly through normalization, but direct inhibition is not considered. The presence of direct differential inhibition is thus a prediction by which to test our model against other models incorporating divisive normalization.

7.1.2 Representations

In the introduction of this chapter, we stated that it is our view that the dSC's task is to localize uni- and cross-sensory stimuli and generate motor responses. Throughout our experiments, we have worked on and evaluated the behavior of our model in light of this assumption. Specifically, we have presented a model which computes the position of a stimulus from neural input and represents that position in a probabilistic population code (PPC).

Some researchers have pointed out that it is not the primary task of biological cognitive systems to represent anything in the real world. Rather, it is the task of a brain to do its part in generating behavior that is adequate to the situation. Sometimes, such behavior is best generated through extensive transformation of sensory information through the complex dynamics of neural systems. At other times, the best response derives directly from the way the world interacts with the body and only little neural processing is necessary, if any at all. Thus, goes the argument, we should not be too fixated on trying to explain all neural phenomena in terms of amodal representations of world properties and actions, and information processes acting upon those representations. Instead, we should interpret the whole system of task, resources, and dynamics of the world, the body, and the brain in understanding cognition in a very wide sense (Clark 1999; Engel et al. 2013; Wilson and Golonka 2013).

The question here is: where does this leave approaches like ours which are built on the idea that the brain learns to distinguish states of aspects of the world which it represents in topographic population codes? One might take the extreme position that a description of the dynamics of body, brain, and environment can replace a description of human cognition in terms of representations and computational processes—what is called the Replacement Theory (of embodiment) (Wilson and Golonka 2013).

While we are sympathetic to the ideas of the embodiment movement and acknowledge embodiment as an important aspect of cognition in principle, we think the Replacement Theory goes a bit far. The reason we cling to our—possibly old-fashioned and computation-laden—way of seeing things is that it offers elegant, though by no means unique, explanations of parts of the phenomenology. A very short summary of this Chapter might be the following:

> We used an ANN algorithm to show how self-organized learning of world statistics can lead to topographic PPCs coding for world properties like the position of a stimulus. We showed how such an ANN algorithm can produce both 'optimal' multisensory integration on the behavioral level and neurophysiological phenomena observed in multisensory integration *in-vivo*.

Thus, representations and information processing acting upon them are invoked to interpret cognition and to connect behavior to neurophysiology. Since all we model happens in brains which are parts of bodies in a physical world, there is no doubt that an equivalent explanation could be phrased in terms of brain-body-world dynamics. Our explanation has in its favor the virtues of being short, intuitive, compositional, and explanatory, virtues arguably important for scientific theories (Carandini and Heeger 2011; Humphreys 2007; Machamer et al. 2000; Simon 1992; Sun 2009). Embodied theories of dSC behavior may be possible which are also short, intuitive, compositional, and explanatory, but it is unlikely that they are all of these things at the same time as explaining the same aspects of dSC functioning. These considerations once more highlight the need for theories of the same subject matter at different levels of granularity, and with different metaphysical assumptions and foci (Carandini and Heeger 2011; Machamer et al. 2000; Marr 1983).

We will finish this section with one last thought to motivate the use of information-processing machinery for modeling the dSC. An approach in Embodied Cognition is to analyze what is the task to be solved by a given organism in a given situation and what resources the organism has and how those resources might be used to solve that task, and then to test whether the organism really uses its resources in the way one hypothesizes that it does (Wilson and Golonka 2013). The SC is a brain structure which is preserved in all vertebrates, big and small, aquatic, terrestrial, and aerial, primitive and highly developed. In all those animals, its functioning is similar. At the same time, it is involved in a diverse array of behaviors, dependent of the necessities of the worlds the different organisms inhabit.[2] One could analyze the task and resource sets of vertebrates from the first possessor of an SC down to the modern species in their ecological niches and explain why the dynamics their specific SCs are involved in use the available resources to solve their respective tasks. Or one could say that the locations of objects are important determinants for solving many problems and a brain region dedicated to localizing objects is therefore advantageous for almost every organism. This account would almost certainly fail to describe the state of affairs in every SC-possessing species in one aspect or another.[3] But it would describe and explain the big picture at a manageable level of abstraction, as is the goal of a scientific theory.

[2] See Section 2.2.

[3] To encompass those aspects, it could be extended by more embodied notions like reward-mediated learning, as mentioned in the general discussion above.

Part II

All Things Practical: From SC-Modeling to the Real World

In our modeling work described in Part I of this thesis, our focus was strictly on reproducing biological phenomena in response to the first part of our research question: how does the dSC do what it does?[4] In the following chapters, we will concentrate on the second part of our research question, that is, how an understanding of an implementation of the dSC's tasks in the dSC can help solve problems in robotics.

Specifically, we will study the applicability of the network developed in Chapter 4 for practical purposes. True to our modeling strategy,[5] which assumes that mechanisms in information processing in nature are effective and which lets itself be guided by that assumption, we did motivate our model with functionality in mind.[6] However, once we conceived our algorithm, we neither analyzed its mathematical properties, nor compared it to other algorithms, nor evaluated its performance on real data of any type, because our only concern was its adequacy as a model of the dSC.

In contrast, in Part II, it is our goal to study the practical applicability of our algorithm. For this purpose, we need to know what it is it does, how well it does it, and when one might choose this algorithm over another. The first question, what it really is our algorithm does, will be the subject of Chapter 8: there, we will discuss our work and compare it to other approaches on a purely algorithmic level. We will describe the functionality of the algorithm proposed in Chapter 4 in more formal terms than previously and then compare it to other algorithms that are similar. In our analysis and comparison, we will pay special attention to weaknesses of our algorithm as it stands. Some of these weaknesses will be addressed in an application-oriented version of our algorithm which we will introduce in Chapter 9. Thus, we will open up solutions for one problem found in a biological information processing system, the dSC, to other problems in computer science.[7]

Considering that the dSC was assumed, in Part I, to localize audio-visual stimuli, it might seem natural to apply our new algorithm, which is based on a model of the dSC, to this problem as well. However, a closer look reveals that audio-visual localization is probably not a good application for our algorithm. Simply speaking, this is because visual localization is very precise and pragmatic approaches to localization may use auditory localization only as a backup in case visual localization fails. Integration of the two, in the narrow sense, does not improve localization under realistic conditions. This is different for *detection* (e.g. Mühling et al. 2012), and *disambiguation/target selection* (e.g. Kushal et al. 2006; Li et al. 2012; Sanchez-Riera et al. 2012; Voges et al. 2006; Yan et al. 2013), which can indeed profit from the combination of visual and auditory information. In principle, our algorithm may be used for these tasks. However, there already exist specialized algorithms for detection and disambiguation, whose outstanding performance our localization algorithm is not likely to approach.[8] To determine the practicality of

[4] See Section 1.2.1.

[5] See Section 1.2.2.

[6] See Section 4.3.2.

[7] See Section 1.2.2.

[8] Appendix B has a more in-depth discussion of practical audio-visual localization and a treatment of the question of why the dSC does seem to integrate vision and hearing for localization when that does not appear to be useful.

our algorithm, we therefore apply it to a task that it fits more naturally: in Chapter 10, we will describe the use of our algorithm in a practical application, binaural SSL in humanoid robots.

8 A Pragmatic View of Our Algorithm

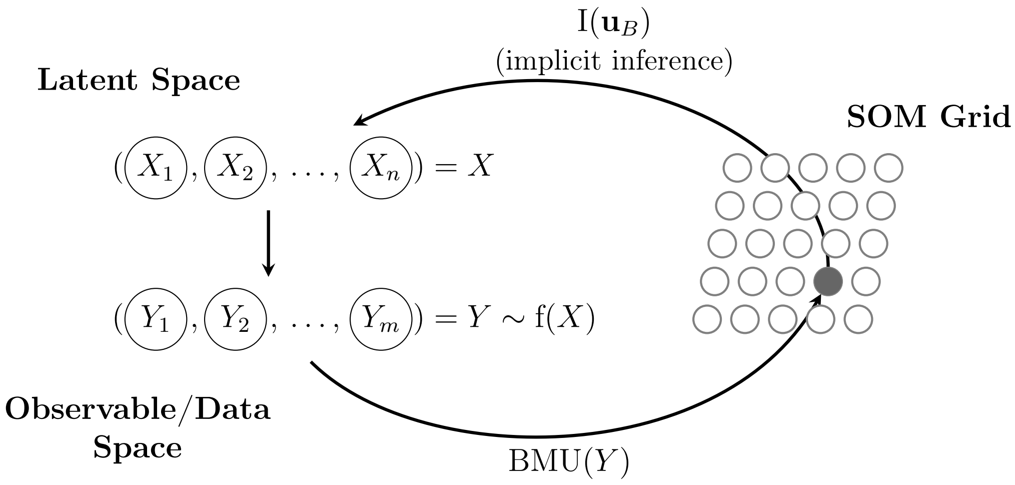
In Chapter 3, we reviewed the State of the Art (SOTA) in modeling the dSC and natural MSI in order to put our work in modeling into perspective. As we have just argued, an analogous discussion of the SOTA in practical MSI or in audio-visual localization is not meaningful. Instead, this chapter will be dedicated to a discussion of our algorithm at a functional level and of its status within the field of algorithms which are comparable at that functional level.

We will start our discussion in Section 8.1 by describing the relationship of SOMs—and thus of the algorithm we developed for modeling—to LVMs, which affords us a more functional view than the view taken so far, which was marked by a focus on phenomenological aspects. In Section 8.2, we will then compare our algorithm to others which are similar in that they share this relationship to LVMs. Section 8.3, will discuss weaknesses of our algorithm and ways in which these weaknesses can be addressed. Finally, we will summarize the status of our algorithm as a practical machine learning algorithm, in Section 8.4.

Here, as well as in the rest of Part II, we will strive to use notation consistent with that originally used to introduce our ANN algorithm. The terminology used here, on the other hand, will be a more application-oriented one, to mark the shift in perspective from modeling to developing useful machine learning algorithms. We will, for example, refer to 'units' instead of 'output neurons' and to 'data points' instead of 'input activations,' indicating that the entities under discussion are not to be seen as representations of anything present in natural neural information processing.

8.1 SOMs and Latent Variable Models

In Section 3.2.2 we described the SOM algorithm as an unsupervised learning algorithm which learns, when successful, which values in the dimensions of the data points in a high-dimensional data set typically co-occur. Of course, for this to make sense, there must be typical co-occurrence between the values of the dimensions of those data points. That means that there must be a causal structure reflected in the data. Such a causal structure can be described in terms of LVMs: according to that view, a potentially multidimensional latent variable is expressed stochastically in a number of observables, the dimensions of the data points. An LVM is a model of how the latent variables are expressed in the data points. Given a data point, an LVM can be used to perform probabilistic inference on the latent variable.


$\mathrm{I}(\mathbf{u}_B)$
(implicit inference)
Latent Space
SOM Grid
$(X_1, X_2, \ldots, X_n) = X$
$(Y_1, Y_2, \ldots, Y_m) = Y \sim \mathrm{f}(X)$
Observable/Data Space
BMU(Y)

> The probability distribution of a multidimensional observable variable is a function of the value of a multidimensional latent variable. Successful SOM learning leads to a network whose units' prototypes are distributed in data space such that mapping data points to the SOM units with the closest prototypes maps data points arising from similar values of the latent variable to the same SOM unit. Thus, mapping a data point into a trained SOM's grid is an approximation of implicit inference from observables to latent variables.

Figure 8.1: SOMs as LVMs.

In some sense, a successfully trained SOM can loosely be thought of as an LVM for the process that generated the data (Yin 2007): assume, for each of the network's N-dimensional prototypes $p = (p_i)_{i=1}^{N}$, that the values $p_1, p_2, \ldots, p_N$ really do usually go together, and that each data point has one unit whose prototype is much closer to it than any other unit's. Then, each unit can be interpreted as a placeholder for all values of the latent variable which are typically expressed in the space of observables by the values of the unit's prototype. If, for every prototype, there is only one value of the latent variable, or a few, similar ones, that would usually be expressed in the observable space by that prototype, then mapping a data point to a SOM unit is equivalent to mapping it to the value of the latent variable represented by that SOM unit (see Figure 8.1).

So far, our discussion is as true for the SOM algorithm as for the standard K-Means algorithm.[1] The essential difference between K-Means and the SOM algorithm is the latter's neighborhood updating mechanism: a SOM comes with a topological relationship between its nodes. While only one unit is updated with every presentation of a data point in K-Means, SOM learning in principle updates all units with the update strength decreasing with distance from the BMU with respect to the topology.[2] Thus, units

[1] For brevity, we will just refer to the standard K-Means algorithm as 'K-Means', although that term really refers to the problem, not the algorithm solving it.

[2] See Section 3.2.2.

close to each other in a SOM's topology tend to have prototypes close to each other in Euclidean space, after learning. If the image of the latent variable from which the data points are generated is a topological space, and if that topology is reflected in the Euclidean distances between data points, then the mapping from the units of a trained SOM to the values of the latent variable they represent is often at least roughly and locally homeomorphic.

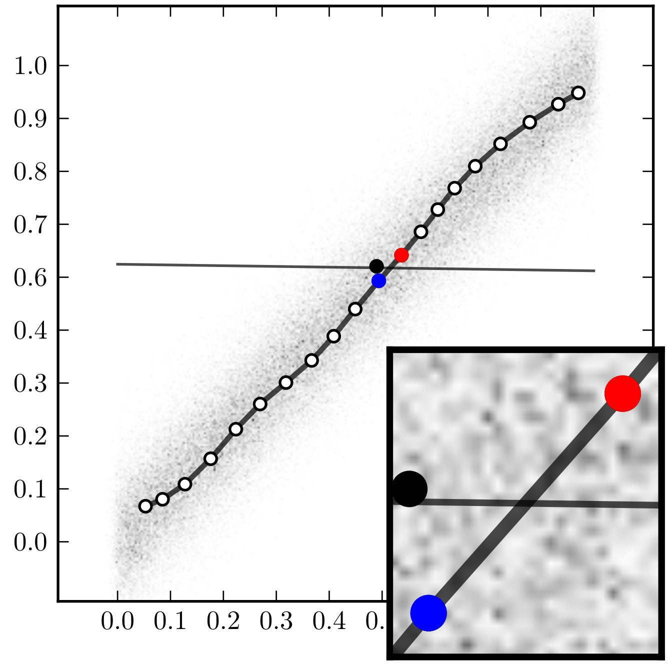
In our modeling, the data points, that is, the observables, are neural activations related to sensory input. The latent variable giving rise to the observables is the position of a stimulus. Therefore, a SOM trained on the data points may learn to map sensory activations into the network's grid such that the topology of the grid reflects the topology between different stimulus positions in space.

We said earlier that a trained SOM can be thought of as implementing an LVM *loosely*. There are at least three reasons for this reservation. The simplest one is that SOM learning is rarely perfect. Limited training data, asymptotic convergence, border effects, overfitting, and other effects generally lead to an imperfect representation of the expression of latent variables in observable space by the SOM units' prototypes. The other, more serious, reasons why a SOM does not implement an LVM are due to the noise present in most interesting data sets:

Assume a data point is to be mapped into a trained SOM. The algorithm normally selects that SOM unit whose prototype has the least Euclidean distance to the data point. That SOM unit, however, is not generally guaranteed to represent the most likely value of the latent variable expressed by the data point: depending on the noise which displaces data points from the optimal representations of the latent variable in observable space, a different unit might be the better choice. This case is exemplified in Figure 8.2: a one-dimensional, discrete latent variable is expressed by two-dimensional data points. The SOM units' prototypes are assumed to perfectly represent the optimal expressions of the values of the latent variable. Both dimensions are corrupted by independent Gaussian noise, but noise in one dimension is much stronger than in the other. Therefore, in looking for the unit most likely representing the value of the latent variable expressed by a data point, distance between units and the data point in the low-noise dimension is more significant than in the high-noise dimension. The basic SOM algorithm does not take the strength of noise into account and therefore does not perform optimally in mapping data points into the network's topology.

Another problem is related to this one: a SOM may be used simply to infer latent variables of data points. However, the result may also be used in further processing. In that case, it can be desirable not only to know what is the most likely value of the latent variable, but also how likely other values are. Since the Euclidean distance between data points and prototypes is only a crude proxy of probability, the value of the population response of a SOM (the distances between the data point and all the prototypes) may be limited, depending on the application.

In contrast, the algorithm proposed in Chapter 4 aims to learn precisely the noise characteristics of the input. It uses these noise characteristics to map data points into the network's grid and thus, in the language discussed above, to probabilistically infer the most probable value of the latent variable, that is, in our setting, the position of the



A regular one-dimensional SOM, trained on two-dimensional data. Data points were generated from a one-dimensional, uniformly distributed latent variable to which Gaussian noise was added. Noise was greater in the dimension shown as vertical than in the one shown as horizontal.

In the example, the SOM maps a data point (black circle) to the unit whose weight vector is closest (blue circle). However, another unit's weight vector is more likely closer to the true value of the latent variable, given the noise distribution. In fact, the red unit is the better match than the blue unit for every point above the horizontal gray line.

Figure 8.2: Suboptimal SOM Mapping.

stimulus in space. The response of each unit of our network has a well-defined meaning: it is the approximate probability of that unit being the representative of the value of the latent variable. Thus, the population response of our network is an approximate probability density function (PDF) (or rather a probability mass function (PMF) since it is discretized) over the values of the latent variable.[3]

A strength of our algorithm compared to the standard SOM algorithm is its resilience to differences in the range of values between the input dimensions. SOM learning and mapping performance is obstructed by drastically different ranges of values in input dimensions. The reasons for this are the same as for the problems with different noise characteristics discussed above (in fact, different ranges can be seen as a special case of different noise characteristics). The standard approach to solving such problems is to normalize the data such that all data points fall into the same range (usually into the interval $(0, 1)$) (Kohonen 2013). Kangas et al. (1990) have proposed a SOM variant which automatically learns the necessary scaling factor per dimension. Our algorithm neither requires normalization (in principle), nor does it encompass an explicit scaling mechanism. The easiest way to see that is to observe that it does not assume input values to be metric. The standard SOM algorithm would consider the data point $d = (2, 100)$ to be closer to one unit's prototype $u_1 = (1, 100)$ than to another unit's prototype $u_2 = (2, 200)$. Given only the two prototypes, u_1 would be the BMU. In our algorithm, units do not have prototypes and if there is anything like a distance between a unit and a data point it is the probability of that data point given the unit's histograms.

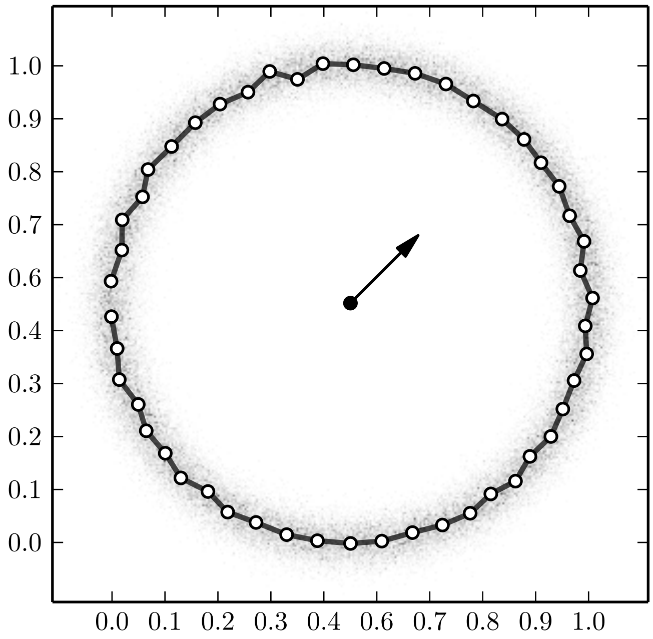
That said, it can still be good to apply a rough scaling of input dimensions to ensure the distribution of the values is sampled well by the histograms and for the efficiency of implementations which otherwise require too much memory for sparsely used histogram bins.

8.2 Relationship to Other Algorithms for Unsupervised Learning of LVMs

Two of the simplest techniques for learning LVMs are factor analysis (FA) and the related principal component analysis (PCA). Both assume the data is generated from a latent variable whose dimensionality is lower than that of the data space. FA and PCA learn LVM: given an n_λ-dimensional latent variable $\Lambda = (\lambda_k)_{k=1}^{n_\lambda}$, they assume that the process that generates the n_d-dimensional data uses Λ to linearly combine the n_d-dimensional factors resp. component vectors $(F_k)_{k=1}^{n_\lambda}$:

$$d = \sum_{k=1}^{n_\lambda} \lambda_k F_k + \varepsilon,$$

[3] This is true analytically under the condition that learning succeeds—for simulations supporting that claim empirically, see (Bauer et al. 2014).



A one-dimensional SOM (connected, black-and-white circles) approximates the two-dimensional data (dots) quite well and may serve as a good approximation to an LVM. A FA with one factor, on the other hand, represents the data by a mean and a component (black circle, arrow), which does not account for the fact that the data is confined to a narrow, circular band around the mean.

Figure 8.3: SOMs Can Learn Non-Linear LVMs—PCA and FA cannot.

where ε is a noise term. The assumptions, the interpretation, and the algorithm of PCA and FA differ, but for the purposes of this comparison, they are very similar.[4] Both are unsupervised, and both generate *linear* LVMs. Thus, in contrast to the SOM and SOM-like algorithms like ours, their utility is limited for non-linear relationships between latent variables and data (see Figure 8.3). Non-linear extensions of PCA exist (see Kruger et al. 2008, for a review), however, these come with significantly greater baggage in computational complexity, choice of parameters and/or assumptions on the data. Finally, PCA and FA are methods which are applied once, on a given data set. SOM-like algorithms like ours can be trained incrementally and are thus very suitable for learning in changing environments.

Another specific method which learns a non-linear, topographic LVM of its input is generative topographic mapping (GTM) (Bishop et al. 1998). This feature is inherited by both GTM and our algorithm from the original SOM algorithm (Yin 2007). What our algorithm shares with GTM and not with the original SOM algorithm is its prob-

[4] PCA is not as explicitly and strictly a method for learning LVMs as FA, but it is an approximation in much the same way the SOM is.

abilistic interpretation. Like GTM, our algorithm computes the probabilities of *a set of* hypotheses for the value of the latent variable causing the input. To do that, both learn the noise statistics of the input to correctly integrate information from each of the data dimensions. One benefit of our algorithm over GTM is that its assumptions on noise distributions are weaker. GTM in its original formulation assumes Gaussian noise (Bishop et al. 1998). It can be modified to accommodate different noise distributions, but their shape is fixed for each input dimension before learning. Most importantly in the present context, although the histograms used by our units to approximate input likelihood functions are not biologically plausible, we argue that the underlying principle is (see above). Also, unlike GTM, our algorithm in its current form is not a batch learning algorithm and therefore allows for online learning, which is plausible for modeling natural learning and useful for applications in unknown and changing environments.

A weakness of the current algorithm compared to GTM and to previous SOM-based algorithms is that it potentially requires more data to train: since it makes no assumptions on the distribution of noise—not even that this distribution be uni-modal—it cannot learn from a data point in which one of the input dimensions has value $\mathbf{d} = v$ about the probability of data points where $\mathbf{d}' = v + \delta$ if $v + \delta$ is in a different histogram bin than v. This is different in the regular SOM and the variants due to Zhou et al. (2011) and Bauer et al. (2012a) as well as in GTM, which all assume continuity of the noise distribution. This weakness, of course, is related to the algorithm's strength of being able to learn any noise distribution in principle, as long as that noise distribution is well discretized by the bins of the histograms and as long as enough data is available for training.

8.3 Weaknesses of the Network Model from a Pragmatic Perspective, and Solutions

One weakness of our algorithm derives from our assumption of uncorrelated noise in each input dimension in designing our network. Unfortunately, noise will in fact be correlated in real-life applications, rendering our algorithm suboptimal. This is indeed a limitation. However, Zhang (2005) has shown that Naïve Bayes algorithms like the one proposed in Chapter 4 are near-optimal in many realistic situations even in the face of correlated noise.

Another limitation can be seen in the figures depicting the mapping learned by our networks (Figures 5.3, 5.10, 6.3) in the experiments of Part I: These figures show that more data points are mapped to units at the edges than to units at the center of the network topology. From the viewpoint of statistical inference, described in Section 8.1, this is a problem since it means that the values of the latent variable generating the data points mapped to one of the border units cannot be determined with the same precision as other data points. This problem is in principle independent of the density of data points in the data cloud, and it is common with regular SOMs as well (Kohonen 2013, 2001, pp. 138–142): during learning, each unit in the network is updated not only with data points for which it is the BMU, but also with data points for which its neighbors are BMUs. Every time this happens, the region in data space mapped to a specific unit moves a bit towards the data point with which it is updated. In the end, the region in data space for which each unit is responsible is a compromise of those data points for which it was the BMU itself, and for those data points for which its neighbors were BMUs. Since those units at the border of the network topology have neighbors only on one side (per dimension for which they are at the border of the topology), that compromise is biased towards the center of the data space.

There are a number of ways to remedy the border effect. Kohonen (2013, 2001, p. 141) suggests to update border units more strongly than other units for data points which are outside the range covered by all the units' weights. He also suggests that a training phase without neighborhood interaction (i.e. with a neighborhood interaction width of $\sigma = 0$, amounting to a degeneration of the SOM to the K-means algorithm) after normal training can eliminate the border effect. Another mechanism which could alleviate the border effect could be adding Conscience (DeSieno 1988), by which BMU selection is biased towards units which have not been BMU for many training steps. Finally, one could use a network topology without borders, like a spherical (actually: circular) or toroid topology (Kohonen 2013; Ritter 1999).

In the experiments reported in Part I, we did not use any of these techniques: we were more concerned with model simplicity than with good performance and therefore did not optimize. Also, at least borderless topologies are out of the question for modeling the dSC because the topology of the dSC is not borderless. In Section 9.2, we will develop an application-oriented version of our algorithm, geared towards effectiveness and computational efficiency on real-world input. It will therefore both use a borderless topology and move from single-data point to mini-batch learning.

The third weakness of our algorithm from an application perspective is the one mentioned at the end of the last section. By using histograms instead of any specific class of functions to approximate the likelihood function for each of the input dimensions given a particular value of the latent variable, our algorithm makes very few assumptions. However, that also means that the algorithm does not learn all it could learn from a given data point if it assumed *the correct* class of likelihood function, as for example a Poisson or normal distribution.[5] A way to address that problem can be to assume parametric PDFs and update the parameters of those instead of the bins of a histogram (Bauer et al. 2012a).

Finally, note that, in our original statement of HistoSOM, a single update of a network with $n_\mathbf{o}$ output neurons and $n_\mathbf{i}$ input neurons requires updating $n_\mathbf{o} \times n_\mathbf{i}$ histogram bins because of neighborhood interaction. This can make training large networks with many inputs quite costly. If neighborhood interaction did not have to occur for every single data point, but aggregated for the whole data at once, then many computations could be saved. This idea inspires a batch learning version of our network. In batch learning (in SOMs), a training step consists of mapping all data points to their respective best-matching network units at once, then aggregating the updates for the entire batch,

[5] See Figure 8.2.

	HistoSOM	SOM	K-Means	GTM	PCA/FA
topographic	✓	✓		✓	✓
nonlinear	✓	✓	✓	✓	
probabilistic	✓	✓	✓	✓	
online learning	✓	✓			
weak assumptions on noise	✓				

Table 8.1: Comparison of HistoSOM to Similar Algorithms.

and finally applying the aggregate update (Abe et al. 1999). If neighborhood interaction can be applied on the aggregate update and if that is less computationally intensive than applying it for every one of the data points, then batch learning can be much more efficient than single-data point learning. What is more, it also often leads to faster and more stable convergence (Kohonen 2013).

8.4 Intermediate Summary: A Pragmatic View on Our Algorithm

Our algorithm, like the SOM algorithm from which it is derived, can be seen as an unsupervised algorithm for learning an LVM for its data. Such an LVM-learning algorithm represents an analysis of a data set or source of data and it can be used to make statistical inferences about the presumed latent variables which generate the data. In contrast to simple PCA and FA, our algorithm is capable of learning non-linear LVMs. Like the SOM algorithm and in contrast to K-Means, the learned LVM comes with a topographic interpretation, meaning that two data points which activate close-by units are suggested to originate from similar values of the latent variable. Unlike the SOM algorithm, PCA and FA, our algorithm approximately computes probabilities for different hypotheses about the latent variable for a given input. It shares with GTM learning of non-linear LVMs and a strong probabilistic interpretation. However, its assumptions on the distribution of noise in the input are much weaker than those of GTM. Furthermore, in contrast to GTM, PCA, and FA, and like the SOM, our algorithm lends itself to online learning and thus for use in autonomous systems in changing environments (see Table 8.1 for an overview of this comparison).

Not having been developed originally for practical applications, our algorithm has a few drawbacks. First, it treats all input dimensions as conditionally independent given the latent variable. In applications with strong dependencies between noise in different dimensions, this can reduce performance. Second the algorithm is susceptible to the border effect: after training, the units at the borders of the network tend to represent a disproportionally large part of the data compared to units at the center of the network. Third, its weak assumptions on noise properties can render our algorithm somewhat inefficient in its use of training data. Finally, its statement as a single-data point learning algorithm makes it less computationally efficient and possibly less stable than necessary.

We have proposed ways to alleviate these weaknesses. As we will see in the next chapters, a few modifications are enough to make our algorithm practical and solve real-world problems.

9 An Application-Oriented, Circular-Topology, Mini-batch Version of Our Algorithm

The network and algorithm presented in Chapter 4 were designed to model learning of multisensory integration. We therefore did not consider certain optimizations which might improve computational performance, stability, or accuracy, but which would not have been motivated by biological fact or necessity. While we have shown that the network can learn to integrate information from input data near-optimally[1] in spite of the limitations discussed in Section 8.3, these limitations may become a problem in real-life applications.

The following sections will therefore be dedicated to an extension of our algorithm follows some of the strategies also discussed in Section 8.3 to improve its practical usefulness. Chapter 9 will present a variant of the algorithm originally proposed, which uses a circular instead of a linear network topology to address the border effects present in ours as well as in other SOM variants. That variant will also make use of batch learning to improve stability and performance, and allow for a more efficient implementation. The usefulness of the resultant algorithm will be demonstrated in a neurorobotic experiment, in Chapter 10.

9.1 Cyclic Topology and Mini-batch Learning

In our simulations in Sections 5.2 and 6, but more drastically in the neurorobotic experiment of Section 5.3, we have seen that our algorithm (HistoSOM) can suffer from border effects. These border effects are inherited from the original SOM algorithm and a number of solutions have been proposed.[2] The algorithm used in this section will implement a border-less topology to address the problem of border effects.

The second major change to our algorithm to make it more practical is making it a mini-batch learning algorithm. In general, batch learning algorithms do not update their state (network weights, statistics etc.) after each step, as single-data point learning does. Instead, they compute output given the current state on all or part of the input and then aggregate and apply the changes for all data points in the batch at once. As Kohonen (2013) notes, batch learning is usually both safer than single-data point learning and faster in terms of numbers of presentations of the training set. In our case,

[1] See Chapter 5.

[2] See Section 8.3.

as we will see, the batch version of the algorithm also lends itself to a much more efficient implementation.

The computational complexity of an update step in batch learning grows at least linearly with the size of the data set. On the other hand, no amount of data will let SOM training converge after the first batch update step, or even the second or third step, usually. Each step improves organization which in turn lets the SOM extract more information from the data in the next step, and use it for learning. It should therefore be intuitive that it is not always efficient to update a SOM-like network with all the data in every step, if the data set is very large. Mini-batch learning compromises between single-data point and batch learning by selecting, at every step, a random subset of the entire data set and updating the network with that mini-batch. We will therefore formulate our algorithm as a mini-batch learning algorithm. Note that setting the size of the mini-batch to the size of the training set makes the algorithm a mini-batch learning algorithm while setting the mini-batch size to one results in a single-data point algorithm.

The aim of our modifications is to produce an algorithm which can be used on real data, in practical applications. We will therefore report on a neurorobotic experiment in which our variant algorithm is used for learning binaural SSL in a humanoid robotic head. In Section 9.2, we will describe in detail the variant algorithm. We will then, in Section 10.1 introduce some of the necessary background of SSL in general and of biological SSL in particular. Section 10.2 will review previous approaches to binaural SSL and the state of the art. In Section 10.3 we will describe the experiment, including experimental set-up, auditory pre-processing, training, and results, before we conclude with a discussion of the new algorithm and the experimental results in Section 10.4.

9.2 The Algorithm

Let us start with the data structures of our new algorithm. Like HistoSOM, the new algorithm operates on a network of $n_{\mathbf{o}}$ units which are fully connected to the input of dimensionality $n_{\mathbf{i}}$. For each pair of a unit $\mathbf{o} \in \{\mathbf{o}_1, \dots, \mathbf{o}_{n_o}\}$ and an input dimension $\mathbf{i} \in \{\mathbf{i}_1, \dots, \mathbf{i}_{n_i}\}$, there is a histogram consisting of n_b bins $f_{\mathbf{o},\mathbf{i},k}$, for $1 \leq k \leq n_b$. Data points are assumed to be tuples of length $n_{\mathbf{i}}$ of non-negative integers less than n_b.

The best matching unit $\mathbf{o}_B$ for a data point $\mathbf{A} = (\mathbf{a}_m)_{m=1}^{n_{\mathbf{i}}}$ is that $\mathbf{o}_B$ such that

$$\mathbf{o}_B = \underset{\mathbf{o} \in \{\mathbf{o}_1, \dots, \mathbf{o}_{n_{\mathbf{o}}}\}}{\arg\max} \left(\prod_{m=1}^{n_{\mathbf{i}}} f_{\mathbf{o},\mathbf{i}_m,\mathbf{a}_m} \right).$$

Assume that the data points are the result of a random process with a latent variable L and the multidimensional observable variable $\mathbf{A}$ such that $\mathbf{a}_m, \mathbf{a}_n$ are conditionally independent given L for $1 \leq m, n \leq n_{\mathbf{i}}, m \neq n$. Assume further that every unit $\mathbf{o}$'s histograms correctly reflect the distribution of values in the respective input dimensions for some value $l_{\mathbf{o}}$ of the latent variable L. Then $l_{\mathbf{o}_B}$ is the most likely value of L given $\mathbf{A}$.

So far, not much is different between HistoSOM and the new algorithm, except for presentation. The main difference is in the way the network is updated. First of all, the network is not updated with a single data point but with an entire subset of the data set, making this algorithm a batch learning algorithm: in each update step, the algorithm randomly selects a sequence $D = (\mathbf{A}_d)_{d=1}^{n_d}$ of length n_d from the set of all data points. It then generates a sequence $M = (\mathbf{o}_d)_{d=1}^{n_d}$ such that $\mathbf{o}_d$ is the unit to which data point $\mathbf{A}_d$ is mapped given the current state of the network.

Update histograms are then compiled as follows: for unit $\mathbf{o}$ and input dimension $\mathbf{i}$, bin $f'_{\mathbf{o},\mathbf{i},k}$ is first set to the number of data points which were mapped to $\mathbf{o}$ and in which the value of the input dimension $\mathbf{i}$ was k for $1 \leq k \leq n_i$:

$$f'_{\mathbf{o},\mathbf{i},k} \leftarrow \left|\{d \mid (1 \leq d \leq n_d) \wedge (\mathbf{o} = \mathbf{o}_d) \wedge (\mathbf{A}_{d,\mathbf{i}} = k)\}\right|,$$

where $\mathbf{A}_{d,\mathbf{i}}$ is the $\mathbf{i}^{\text{th}}$ entry of data point $\mathbf{A}_d$.

Neighborhood interaction is implemented by linearly combining the update histogram of each unit with the update histograms of all other neurons with the factors depending on distance between units:

$$f''_{\mathbf{o},\mathbf{i},k} \leftarrow \sum_{r=1}^{n_\mathbf{o}} \mathrm{h}_t(\mathrm{d}(\mathbf{o},\mathbf{o}_r)) f'_{\mathbf{o}_r,\mathbf{i},k}.$$

As in the original algorithm, h_t is an unnormalized Gaussian function whose width parameter σ_t decreases exponentially from a maximum starting value σ_M and asymptotically approaches a minimum value σ_m with increasing update step t:

$$\sigma_t = \exp\left(-\frac{10t}{n_t}(\sigma_M - \sigma_m)\right) + \sigma_m, \tag{9.1}$$

where n_t is the number of training steps.[3] Using a cyclic (spherical, toroid) distance function $\mathrm{d}(\mathbf{o}_q, \mathbf{o}_r)$, for units $\mathbf{o}_q, \mathbf{o}_r$, the network becomes borderless, which is the second major difference between this and the original algorithm. In the following, we will, for simplicity, assume a one-dimensional topology and thus a circular distance function. As a matter of convention, we will define a distance of 1 as twice the maximal distance in the network. Thus, assuming numbered units $(\mathbf{o}_s)_{s=1}^{n_\mathbf{o}}$, we define:

$$\mathrm{d}(\mathbf{o}_q, \mathbf{o}_r) := \frac{\min(|q-r|, n_\mathbf{o} - |q-r|)}{n_\mathbf{o}}, \text{ for } 1 \leq q, r \leq n_\mathbf{o}. \tag{9.2}$$

Extensions to more than one dimension are possible.

Update histograms are normalized:

$$f'''_{\mathbf{o},\mathbf{i},k} \leftarrow \frac{f''_{\mathbf{o},\mathbf{i},k}}{\sum_{j=1}^{n_b} f''_{\mathbf{o},\mathbf{i},j}}$$

And, finally, the units' histograms are updated by combining them linearly with the update histograms. The factors in the linear combination depend on the global update strength α_t:

$$f_{\mathbf{o},\mathbf{i},k} \leftarrow (1-\alpha_t) f_{\mathbf{o},\mathbf{i},k} + \alpha_t f'''_{\mathbf{o},\mathbf{i},k}$$

[3] The constant 10 in Equation 9.1 was found empirically—it ensures that the neighborhood size goes to a minimum long before training is finished The actual value is not critical.

over the course of the training procedure, α_t decreases linearly from a maximum α_M to a minimum α_m:

$$\alpha_t = \alpha_M - \frac{t}{n_t}(\alpha_M - \alpha_m).$$

The above description is summarized in Listing CB-HistoSOM.

CB-HistoSOM has a number of parameters which need to be fixed in an application. We will explore the effects of most of the parameters empirically in our robotic experiments in Section 10.3. Here, we will briefly describe them, introducing notation to be used later, and providing theoretical implications.

9.2.1 Parameters of the Mini-Batch Algorithm

The number of training steps n_t. One possible termination criterion for an iterative training algorithm is the number of training steps n_t. Others are reaching a target performance of the algorithm on a validation set or convergence of performance, trained parameters, or output of the algorithm. The number of training steps is the simplest termination criterion and we choose to adopt it here. We choose it, first, to avoid complicating our algorithm, and, second, in order to be able to study the effects of under- and overtraining in our experiments.

The number of bins n_b in each histogram. When we introduced the first version of our algorithm for modeling, in Chapter 4, we did not worry about the number of bins and just silently assumed it was going to be large enough.[4] In an application context, this parameter does make a difference for two reasons: on the one hand, a large number of bins leads to large data structures which need to be kept in memory and evaluated and maintained by the algorithm. On the other hand, for small numbers of n_b to suffice, the actual values in the data on which the algorithm operates may need to be scaled and/or cropped. The choice of n_b therefore determines the resolution at which likelihood functions are discretized. Depending on how much training data is available, a finer or coarser discretization may be desirable.

Mini-batch size n_d. Generally, batch learning can lead to greater stability and it can speed up learning.[5] In contrast to batch learning, mini-batch learning does not update the algorithm on the whole data set, but on a subset of it. This is because the amount of knowledge that can be extracted from a batch of data can depend on the amount of knowledge already available. It thus does not always make sense to present a large data set to a training algorithm all at once.

In SOM learning, for example, initial clustering of the data is largely random, and does not follow the topology of the network. Only after the network has developed a rough organization does clustering converge to assigning close-by SOM units

[4] In fact, the implementation increases that number whenever necessary, which is an expensive operation.

[5] See Section 9.2.

The CB-HistoSOM Algorithm.

```
function NORMALIZE$((f_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b})$
    for $(l \leftarrow 1 \rightarrow n_\mathbf{o}), (m \leftarrow 1 \rightarrow n_\mathbf{i}), (n \leftarrow 1 \rightarrow n_b)$ do
        $f'_{l,m,n} \leftarrow f_{l,m,n} / \sum_{n'=1}^{n_b} f_{l,m,n'}$
    end for
    return $(f'_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b}$
end function

function BMU$((\mathbf{a}_n)_{n=1}^{n_b}, (f_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b})$
    return $\arg\max_{1 \leq l \leq n_o} (\prod_{m=1}^{n_\mathbf{i}} f_{l,m,\mathbf{a}_m})$
end function

function DIST$(l_1, l_2)$
    return $(\min(|l_1 - l_2|, n_\mathbf{o} - |l_1 - l_2|)/n_\mathbf{o}$
end function

function INTERACTION$((f'_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b}, \sigma_t)$
    for $(l \leftarrow 1 \rightarrow n_\mathbf{o}), (m \leftarrow 1 \rightarrow n_\mathbf{i}), (n \leftarrow 1 \rightarrow n_b)$ do
        $f''_{l,m,n} \leftarrow \sum_{l'=1}^{n_\mathbf{o}} \exp\left(-\frac{\textsc{dist}(l,l')^2}{2\sigma_t^2}\right) f'_{l',m,n}$
    end for
    return $(f''_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b}$
end function

function UPDTHIST( )
    $(f'_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b} \leftarrow 0$
    for $1 \rightarrow n_d$ do
        $(\mathbf{a}_{\mathbf{i},k})_{k=1}^{n_\mathbf{i}} \leftarrow$ random $\mathbf{A}_\mathbf{i} \in D$
        $m \leftarrow$ BMU$((\mathbf{a}_{\mathbf{i},k})_{k=1}^{n_\mathbf{i}})$
        for $(l \leftarrow 1 \rightarrow n_\mathbf{i}), (n \leftarrow 1 \rightarrow n_b)$ do
            $f'_{l,m,\mathbf{a}_n} \leftarrow f'_{l,m,\mathbf{a}_n} + 1$
        end for
    end for
    return $(f'_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b}$
end function

// Initialization
$H = (f_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b} \leftarrow$ RAND()
$H \leftarrow$ NORMALIZE$(H)$

// Training
for $t \leftarrow 1 \rightarrow n_t$ do
    $H' = (f'_{l,m,n})_{l=1,m=1,n=1}^{n_\mathbf{o},n_\mathbf{i},n_b} \leftarrow$ UPDTHIST()
    $\sigma_t \leftarrow \exp\left(-\frac{10t}{n_t}(\sigma_M - \sigma_m)\right) + \sigma_m$
    $H' \leftarrow$ INTERACTION$(H', \sigma_t)$
    $H' \leftarrow$ NORMALIZE$(H')$
    $\alpha_t \leftarrow \alpha_M - \frac{t}{n_t}(\alpha_M - \alpha_m)$
    $H \leftarrow (1 - \alpha_t)H + \alpha_t H'$
end for
```

Given:

$n_\mathbf{i}$:	number of input neurons
$n_\mathbf{o}$:	number of output neurons
n_b :	number of bins
n_d :	mini-batch size
n_t :	number of training steps
n_d :	number of data points
D :	Data set. $D = \left(\mathbf{A}_{\mathbf{i},t} = (\mathbf{a}_{\mathbf{i},k,t})_{k=1}^{n_\mathbf{i}}\right)_{t=1}^{n_d}$
α_M :	max update strength
α_m :	min update strength
σ_M :	max neighborhood size
σ_m :	min neighborhood size

to similar data points and thus updates units with the appropriate data points. A rough topology in a SOM can often be set up with only small amount of data. Thus, updating a SOM with an entire large data set in early stages of SOM training can be inefficient.

Training with very small mini-batches, on the other hand, reduces the benefits of stability and training speed of batch learning. The size of the mini-batches, n_d, therefore is a trade-off between computational efficiency and stability.

Neighborhood size. The neighborhood size shrinks exponentially in our algorithm, approaching a minimum neighborhood size as training proceeded. A large *initial* neighborhood size σ_M will update all neurons strongly with all data points. Depending on the topology of the subspace of the data cloud in data space, this can lead to a strong simplification of the mapping from data space into the network's grid. If subsequent training with smaller neighborhood sizes does not sufficiently adapt that initial mapping then this can adversely affect the performance of the algorithm. On the other hand, small initial neighborhood sizes may lead to incompatible local topologies of the mapping in the network. Later training steps may not be able to repair the global topology, again leading to suboptimal learning.

The asymptotic neighborhood size σ_m affects the amount of generalization of a SOM. Relatively large neighborhood sizes at the end of the training procedure will preserve previously learned local topology of the mapping at the cost of a potentially worse fit to the data set. Small neighborhood sizes will allow each unit to adapt optimally to the data points that are mapped to it in the end. This allows the network to learn certain inhomogeneities of the topology of the data cloud—which can mean a better fit or specialization of some units to unwanted noise dimensions.[6]

Note that the size of σ_M and σ_m is relative to the number of neurons in the network (see Equation 9.2).

Update strength α. The amount by which the network is updated in each training step decreases linearly from a maximal strength at the beginning to a minimum strength at the end. Weak update strengths usually protect learning algorithms from unlearning in one step what they have previously learned. On the other hand, they can also prevent the network from escaping bad local optima and they slow down learning.

Any small final value for the initial update strength α_m should generally be suitable for our algorithm—too small a value may render late stages in training ineffective, but it should not break the algorithm completely. The initial update strength α_M, however, is important for the reasons described above.

[6] As in the 'zebra-stripes' described in Section 6.3.

9.3 Intermediate Summary: An Application-Oriented Version of Our Algorithm

The algorithm presented in this chapter, CB-HistoSOM, is an adapted version of the algorithm we developed as a basis for our modeling work, HistoSOM, in Chapter 4. The changes we made were to improve stability of learning, to prevent the border effect, and to allow for a more efficient implementation. These changes are the introduction of a circular topology and batch or mini-batch learning. Our new algorithm preserves the topological interpretation of the network of units, competitive learning, self-organization, and histogram-based likelihood estimation for the hypotheses represented by the network's units.

Note that it also preserves the capability to learn online, that is, from live data: while we introduced mini-batch learning to improve computational performance and stability of the algorithm, there is nothing which prevents our algorithm from being used in, for example, an autonomous robotic system. Such a system could collect data points into mini-batches and update the network every time a sufficient number of data points have been collected. A good mini-batch size would then be a compromise between the time it takes the system to accrue enough data points for a mini-batch, and stability and performance properties of the system.

Circular topology and mini-batch learning had not been considered as optimizations in our original algorithm because they are not plausible in the context of dSC learning. As we will see in the next section, our new algorithm lends itself to solving an actual robotic task.

10 An Application: Robotic SSL

It is one of the goals of this thesis to learn from natural sensory processing something that we can use in robotics and generally in machine learning. To this end, we have analyzed the algorithm developed in Chapter 4 for modeling from a machine learning perspective, we have identified some of its weaknesses, and we have developed, in the last chapter, a practical version of our algorithm, which we hoped would address some of the weaknesses we have identified. This work would not be complete without an experiment that proves that the new algorithm can indeed be useful in a practical application.

As we have argued in the beginning of Part II, there is little merit in applying our algorithm to audio-visual localization: we could hardly expect to improve on or even reach the performance of the SOTA and anything less would not prove that our algorithm is a useful contribution. However, there are tasks in robotic sensory processing which still need improvement and which may profit from a bio-inspired approach. We will see in this chapter that robotic binaural SSL is one such task and we will see that our algorithm can be applied to this task successfully, approaching, and in some ways exceeding, the SOTA.

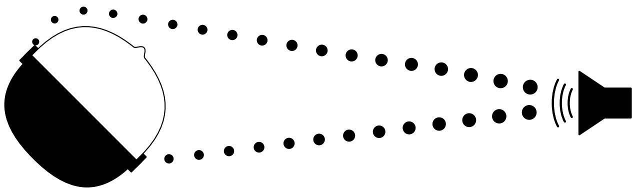
Before describing our robotic experiment, we will give a brief overview of the physical principles of SSL in Section 10.1. Since we will base our our system on biomimetic feature extraction, we will also, in that section, give a high-level introduction to the biological implementation of SSL. Section 10.2 will be dedicated to previous approaches to binaural SSL—biomimetic and otherwise—as a basis for an appraisal of our system. We will describe our experiment and report on results in Section 10.3 and finish this chapter with an evaluation and discussion of the approach and results in Section 10.4.

10.1 Physical and Biological Background of Sound-Source Localization

Auditory[1] SSL is an active field of research with obvious applications in various areas, including robotics and especially human-machine interaction (HMI). The basic principles are actually rather simple: among others, the two most important cues for SSL are the phase shift of a sound corresponding to the time difference of arrival (TDOA)/interaural time difference (ITD) between one ear/microphone and others and the difference in volume (ITD) (Middlebrooks 2015).

Both the interaural time difference and the interaural level difference between two ears or microphones arise from the difference in distance of the receptors from the sound

[1] Technically, a sound source may also be localized visually.



> One ear is farther from the sound source than the other. This difference in distance from the sound source introduces both ITD and ILD.

Figure 10.1: Interaural Time Difference and Interaural Level Difference.

source and, in the biological case, from the acoustic shadowing introduced by the head. That difference in distance and the shadowing effect are determined by the angle of incidence relative to the vertical plane separating the two receptors (the anteroposterior axis of the head): a sound whose source is left of that axis will travel further to the right than to the left receptor. That difference leads to a relative phase shift between the sounds perceived in the two receptors, and to a difference in intensity, as the intensity of a sound generally decreases with the square of the distance:

$$I(d) = \frac{P}{d^2},$$

where I is intensity, d is distance and P is the power of the sound.

A damping body between the receptors will cause the sound to travel even further around that body and introduce a greater ITD and interaural level difference (ILD) (see Figure 10.1). Humans and other animals make use of these two cues to localize sound sources (Yin 2002, p. 101; Schnupp et al. 2010, pp. 178–183; Middlebrooks 2015).

In vertebrates, two parts of the superior olivary complex (SOC), the medial superior olive (MSO) and the lateral superior olive (LSO), are known to extract those cues from neural auditory signals. A typical neuron in the MSO is selective to a specific pairing of an ITD and a frequency range. In contrast, neurons in the LSO are selective to combinations of *ILDs* and frequency ranges (Middlebrooks 2015; Yin 2002). Neurons in the ICc receive projections from MSO and LSO and integrate their inputs; they are selective to angles of incidence of sounds—and, again, frequency ranges (Knudsen and Konishi 1978b). Finally, in the ICx, a map of auditory space emerges (Knudsen and Konishi 1978a; Schnupp et al. 2010).

10.2 Previous Approaches to Binaural SSL and State of the Art

Basic binaural SSL can be realized by a relatively simple algorithm which maximizes the cross-correlation of the signals at the left and right sensor by time-shifting them with respect to each other. The time shift giving the greatest cross-correlation between the signals is an estimate of the ITD which can be translated to an angle given knowledge of the distance between the sensors (modulo a cone of confusion) (see Knapp and Carter 1976, for a somewhat more sophisticated view on cross-correlation for time-shift estimation).

A common, more involved approach to binaural SSL is to measure the head-related transfer function (HRTF) of the physical body of a system. The HRTF is a function describing the temporal and spectral modifications which a given sound will go through before reaching the sensor when emitted from a particular location relative to the system. It can be used to infer where a sound perceived at the two ears is coming from.

MacDonald (2008) proposed two algorithms for HRTF-based SSL. Suppose a signal was recorded at each of an array of microphones. Then, the first algorithm, called the inverse algorithm, tries to reconstruct the original signal by applying to the signal recorded by the microphone i the *inverse* of the head-related impulse response (HRIR)[2] $F^{(\theta,\phi,i)}$ associated with i and with horizontal and vertical candidate angles θ and ϕ, for each microphone. Given a perfect inverse function for $F^{(\theta,\phi,i)}$, this should result in the same signal for all microphones. The algorithm therefore chooses those θ and ϕ for which the Pearson correlation coefficient between the recovered signals is greatest. The second algorithm, termed cross-channel algorithm, attempts to compute the signal recorded by microphone i from the signal recorded by microphone j, for all i and j by applying the HRIR $F^{(\theta,\phi,i)}$ to the signal recorded by j. Again, similarity is assumed to be greatest for the correct candidate angles θ and ϕ and it is measured by the Pearson correlation coefficient. Neither method is limited to two microphones and they exploit spectral cues to resolve front/back ambiguities, thus making 360° localization possible. However, both algorithms suffer from high computational complexity, especially the first, and they require that HRIRs are known for all candidate angles and all microphones. MacDonald (2005, 2008) shows that the inverse algorithm can localize sounds reliably with a resolution of 5°.[3]

Wan and Liang (2013) have proposed a system which, in a first step, computes a rough angle estimate using cross-correlation, which is computationally cheaper than evaluating HRTFs for all possible angle to find the best match. In a second step, they narrow down the search space for classical HRTF-based localization using MacDonald's (2008) cross-channel algorithm around that rough estimate. In simulations, their system performs very well, with absolute localization errors between $\sim 0°$ and $\sim 2.5°$ mean absolute error, depending on simulated signal-to-noise ratio (SNR) and reverberation. Their method is faster and, at low SNRs, more accurate than the HRTF-based method alone. However, their results must be taken with a bit of caution as the input audio signals were simulated using the same HRTF that was used later in signal analysis.

Most recently, Talagala et al. (2014) measured the HRTF of a dummy head and body in a semi-anechoic chamber and used this HRTF to localize sounds. Their system reliably localizes sounds at a resolution of ±5°. A considerable amount of analysis was

[2] the time-domain representation of the HRTF

[3] The author reports a mean absolute error of down to 2.9°, depending on noise. This is especially impressive as front/back confusions occurred and are included in these absolute errors. However, test signals and horizontal candidate angles were only spaced in 5° steps around the system.

required to compute the HRTF from their recordings and the authors note that the system was effectively specialized to localizing sounds under the anechoic conditions of their experiments.

Bio-inspired approaches to binaural SSL have been very successful. Rucci et al. (1999) implemented as system for audio-visual localization based on a midbrain model comprising MSO, IC, (retina, motor neurons,) and SC.[4] That system is trained and tested on uni- and multisensory stimuli. The authors report a mean auditory-only localization error of $1.54^\circ \pm 1.01^\circ$ (presumably mean standard error) for angles in the range from -60° to 60° from the robot. Auditory stimuli were broadband noise bursts. The learning paradigm is a simplified version of reinforcement learning, termed 'value-dependent learning' by the authors.

Liu et al. (2010) achieved reliable localization of noise bursts and speech, at a resolution of first 30°, in later experiments of 15° over the full range from -90° to 90° (Dávila-Chacón et al. 2012). The system is built on top of a neural model of natural MSI, and it adapts to the physical body of the robot in a batch training procedure. The model proposed by Liu et al. (2010) is modular and we base our system on two of the modules of that model. We will therefore briefly provide the necessary detail of the structure of Liu et al.'s (2010) model here.

At the top level of Liu et al.'s (2010) model, there is a module corresponding to the ICc. It consists of a population of neurons each of which is sensitive to stimuli at a certain angle from the system and a certain frequency range. The ICc module integrates the neural activity of two other modules modeling the MSO and the LSO, respectively.

The part of Liu et al.'s (2010) model corresponding to the MSO is based on the highly influential qualitative model proposed by Jeffress (1948). In that model, some non-specified brain region which computes ITDs—presumably the MSO (Yin 2002, p. 114)—contains neurons which each receive signals from both ears. Those signals are frequency-specific and phase-locked to the sound arriving at the ears, and they arrive at the neurons in the ITD-computing brain region with different delays. The neurons act as incidence-detectors, that is, they respond vigorously to spikes of their afferents arriving at the same time. Frequency specificity, delays, and incidence detection make each neuron sensitive to a specific ITD in a specific frequency. Liu et al. (2010) implement the qualitative Jeffress (1948) model using a network of leaky integrate-and-fire neurons.

The LSO part of Liu et al.'s (2010) model consists of a map of neurons which each are selective to a specific ILD in a specific frequency channel. The biological detail of computing ILDs (in neurons) and generating spikes is not part of the model. In the system to be described in the following section, we use implementations of the two parts corresponding to LSO and MSO in Liu et al.'s (2010) model to preprocess sound.

Some of the methods described above are adaptive in that they involve training, in the broadest sense of the word, on real or simulated auditory data. Others, like the simple cross-correlation method, are purely analytic and thus require knowledge of the effect of the physical properties of an SSL system on sound. Of those methods which are adaptive, all are trained in a supervised fashion, that is, the system requires access

[4] See Section 3.3.

to both binaural recordings and the angle towards the sound source for each recording. Furthermore, many systems were tested only in simulations or in carefully designed noise-free physical experiments. In some of the studies which did study the effect of noise on performance, that noise was added artificially to the recordings. In many instances, systems were only tested on broadband noise bursts or a small number of natural auditory stimuli.

10.3 Robotic Experiment

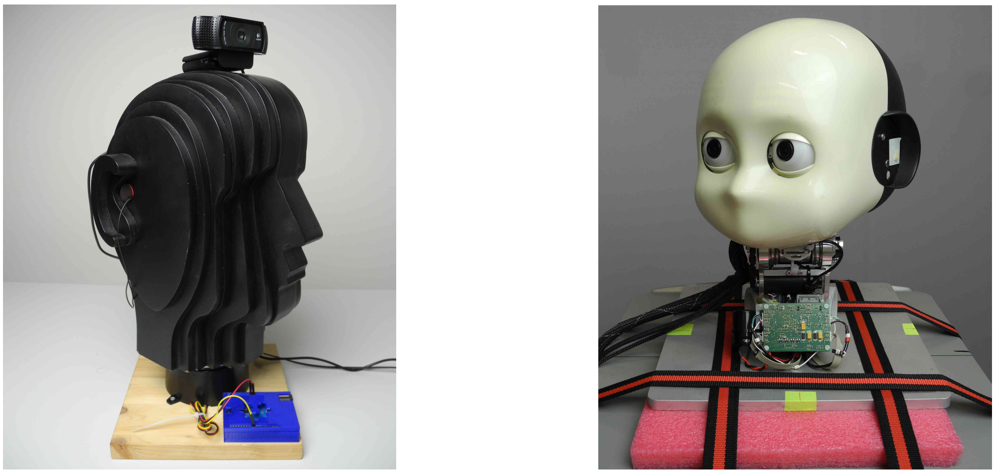
In what follows, we will describe a neurorobotic binaural SSL system based on the adaptive ANN algorithm presented in Section 9.2. That system differs from previous systems in that training is unsupervised. We will report on physical experiments in which we tested its performance. Those experiments involved two different robotic platforms, generating drastically different levels of ego-noise, and two different kinds of auditory stimuli: white noise and speech.

10.3.1 Experimental Procedure

We collected four data sets of binaural recordings: two robots were used in the experiments, the iCub robotic head (Beira et al. 2006) and a Kunstkopf mounted on a rotating plate. Each recorded, in two different conditions, white noise and speech samples (see Figure 10.2).[5] To collect sounds at different relative angles from the robots, we rotated the robots in 1° steps from left to right. At each of the positions, we collected ~ 500 quarter-second samples (see Table 10.1). The length of the samples was chosen to comply with earlier studies.

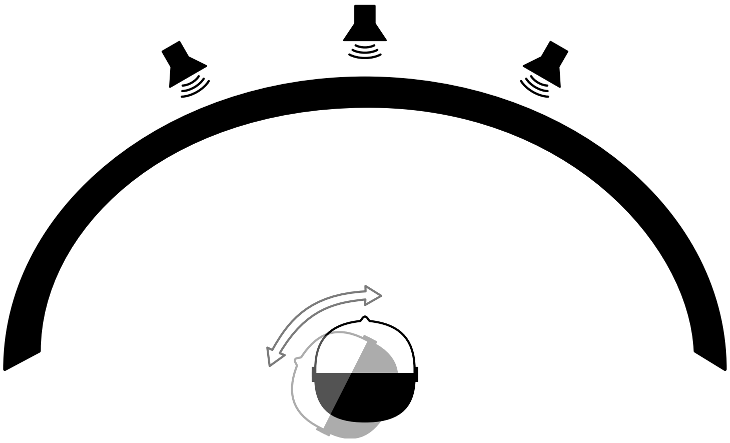
At each position, we recorded samples from three speakers, located at $-45°$, $0°$, and $45°$ from the center, for two reasons (see Figure 10.3). First, the iCub head can only rotate between $-55°$ and $55°$ from dead ahead. By combining recordings from all three speakers, we were in principle able to train and test on sounds from directions between $-100°$ and $100°$. Apart from this rather technical issue, recordings from only a single speaker may produce a data set in which an angle can not only be identified by true spatial cues, but also by artifacts: recordings which might normally have been hard to localize because of the frequency distribution in them (mainly some of the voice samples) could have, for example, contained easy-to-localize reverberations from the metal structure surrounding the robot. Alternatively, sounds from outside of our lab could have provided unfairly easy-to-localize cues. Also, in the case of the iCub, ego-noise might have been different depending on the head posture. Our network might have learned to exploit those cues instead of the actual spatial cues that would enable it to localize sounds outside the lab.

[5] At each angle, the same 28 sentences from the TIMIT core-test-set (Garofolo et al. 1993) were played, one after the other, and recorded by the robot.



Left: Dummy Head "John" by Soundman e.K., `http://www.soundman.de/en/dummy-head/` mounted on a rotating plate. **Right:** the iCub robotic head (Beira et al. 2006).

Figure 10.2: The Two Robots Used in Our SSL Study.



The speakers in our set-up are at angles −45°, 0°, and 45° from the robot. They are three out of an array of speakers behind a projection screen surrounding the robot (Bauer et al. 2012b). The robot rotates in 1° steps and records at each position one sound sample played on each of the speakers.

Figure 10.3: The Robot Rotates with Respect to the Three Speakers.

	iCub WN	iCub S	Kunstk. WN	Kunstk. S
training set	131688	131688	176850	176850
testing set	14632	14632	19650	19650

> The ratio of training set size to test set size was 90:10. Differences in data set sizes were due to the different recording methods. In particular, we were able to rotate the Kunstkopf around the full 180° whereas the iCub head could only move in the range from $-55°$ to $55°$. The results reported in Section 10.3.3 were robust and we are therefore confident that the differences did not affect the results. (WN: white noise, S: speech)

Table 10.1: Training and Test Set Sizes of SSL Data Sets.

10.3.2 Preprocessing and Training

As stated in Section 10.2, we used code implementing the MSO and LSO parts of the Liu et al. (2010) model to preprocess our recordings. All credit for these parts goes to the original authors. Both the MSO and LSO model were configured for 20 frequency components and 43 ITD/ILD steps. For each quarter-second recording, we therefore computed two 20×43 matrices which were flattened and concatenated into a 1720-dimensional data point. These values were the same as those chosen by Dávila-Chacón et al. (2012). The parameter values used by Dávila-Chacón et al. (2012) are in turn adapted to the iCub robot from the ones used by Liu et al. (2010) in their original study.

We normalized each dimension of the data points by subtracting the minimum value, dividing by the ninetieth percentile per dimension, and scaling by the number of bins in our network's histograms $n_b = 20.0$ and capping at n_b. Thus, for the k^{th} entry of the i^{th} data point $d_{i,k}$

$$d_{i,k} \leftarrow \min\left(n_b, n_b \frac{d_{i,k} - m_k}{p_k - m_k}\right),$$

where m_k was the minimum value in dimension k, $m_k = \min_i(d_i, k)$, and p_k was the ninetieth percentile in that data dimension. This normalization procedure promoted economic use of the data data structures of our algorithm: subtracting the minimum value ensured that no bins in our data structure were dedicated for small values that did not occur in the data. Dividing by the ninetieth percentile and multiplying by n_b ensured that the range between the minimum value and the ninetieth percentile per input dimension was sampled by our histograms. By capping at n_b, and thereby mapping all values above the ninetieth percentile to the largest bin, we avoided dedicating bins to outliers.

Each data set was partitioned into a training and a test set with set sizes as reported in Table 10.1 We trained a network of 200 neurons on each of the test sets. That number was chosen to allow for one neuron per angle plus a few neurons to counter

	iCub WN	iCub S	Kunstk. WN	Kunstk. S
mean abs. error:	1.53°	14.6°	1.06°	5.55°
mean abs. error (center)	1.0°	9.95°	0.72°	2.96°
accurate ($\lvert\varepsilon\rvert < 5.0°$)	96.71 %	50.49 %	99.0 %	85.82 %

> We computed the mean absolute error over the entire test set by averaging the mean absolute errors for stimuli at each angle, thereby correcting for different numbers of (randomly selected) test data points at different angles. We also computed the mean absolute error for stimuli at the center, that is, stimuli whose location was between −45° and 45°. We defined accuracy as the percentage of data points which were localized with an error of < 5° for comparison with other studies.

Table 10.2: Performance of our System.

slight inefficiencies due to imperfect self-organization.

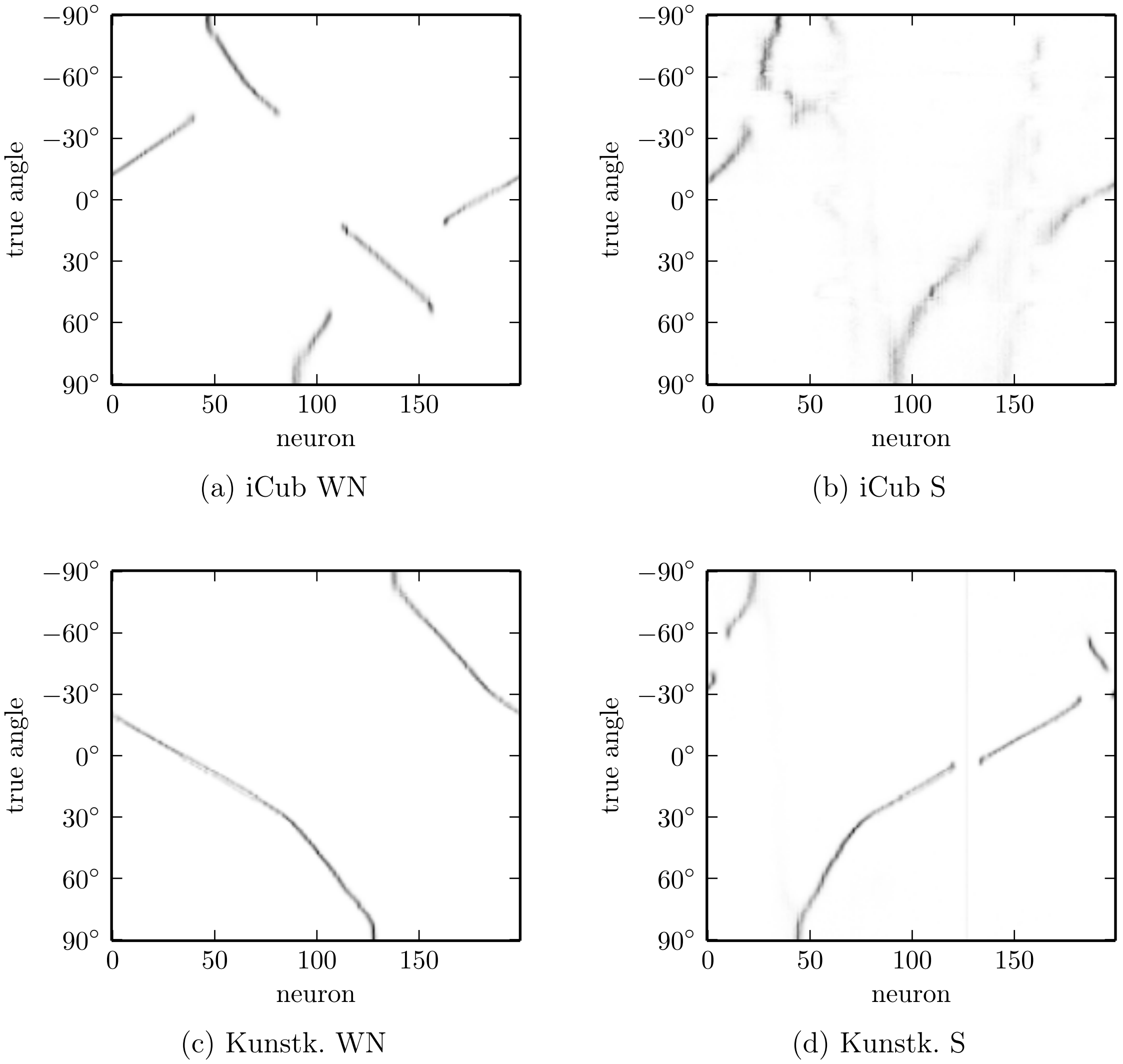
At each training step, we updated the network with a mini-batch of 500 data points randomly selected from the training set. Each network was trained over 200 training steps with the neighborhood size decreasing exponentially from 2.0 to 0.0125 and the update strength decreasing linearly from 0.04 to 0.001 (values empirically determined, see below).

After training, we determined the mapping that had been learned by computing the BMUs for 10000 random data points from the training set.[6] To test the network's performance, we mapped each data point from the test set into the network and chose the previously determined mapping of each of the data points as the network's estimate of the stimulus position. In order to make our results comparable to those of other studies which only tested SSL at coarser resolutions, we computed accuracy as the percentage of localizations with an absolute error of less than 5°. See Figure 10.4 for matching matrices and Figure 10.5 for confusion matrices for each of the data sets. See Figure 10.6 and Table 10.2 for localization errors.

The parameters reported above were chosen through cursory empirical experimentation. To see if our choice was a good one, we selected those parameters most likely to influence the results and systematically trained and tested our network with different parameter settings. Each parameter was changed in isolation, keeping all other parameters as previously described. Incidentally, all parameters of our algorithm are positive-valued. We assumed that small changes in parameters were likely to affect the algorithm's performance more strongly for small values than for large values. Therefore, we sampled the parameter space more finely for values smaller than the reference values described above than for values greater than those reference values. Parameter exploration was performed on the Kunstkopf/speech data set as that was the one with the greatest anticipated practical use.

The parameters which we explored and the figures summarizing the results are listed

[6] See Section 4.3.3.



(a) iCub WN

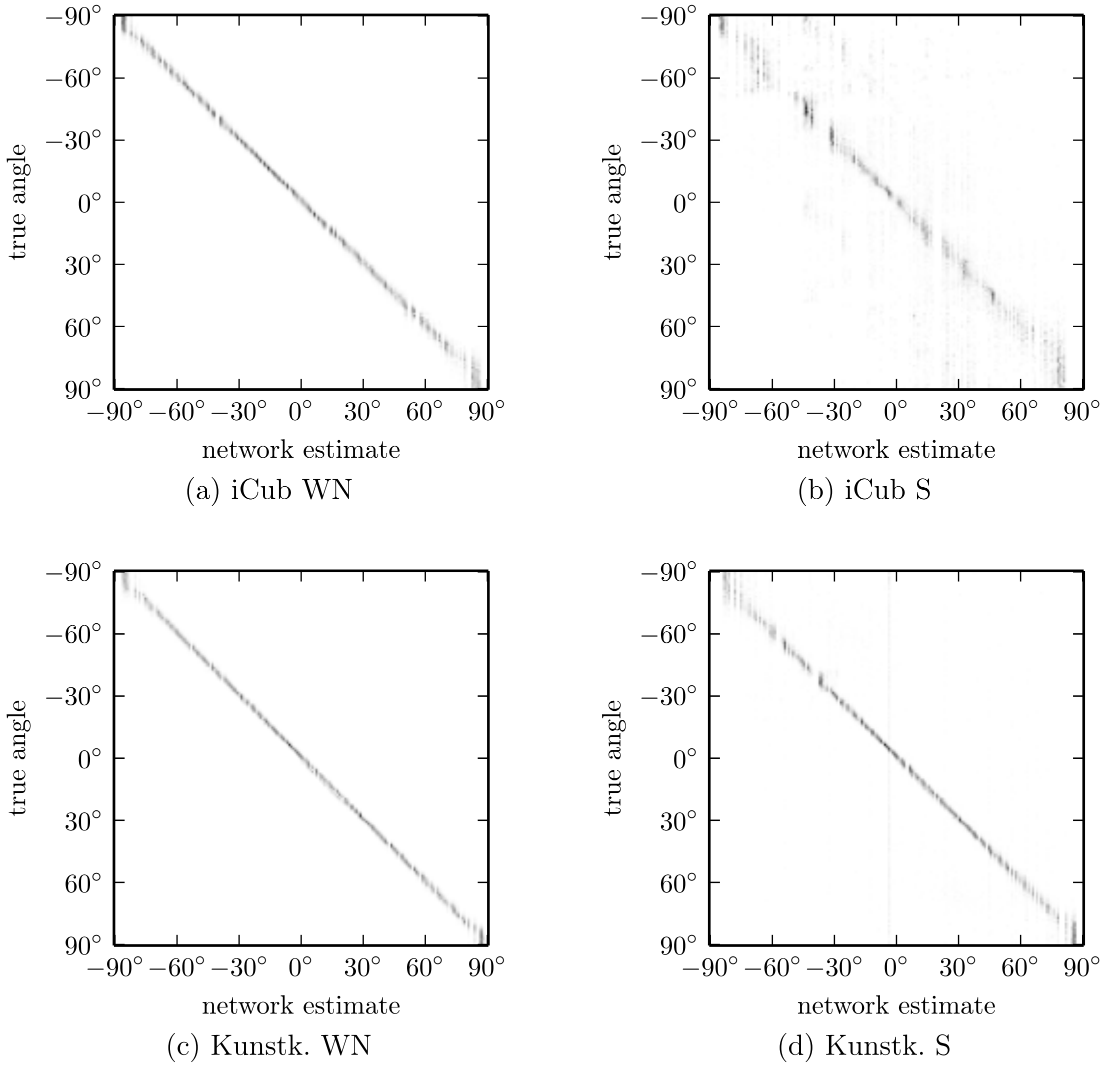
(b) iCub S

(c) Kunstk. WN

(d) Kunstk. S

> Frequency with which recordings from a true angle α were mapped to a given neuron. (WN: white noise, S: speech)

Figure 10.4: Matching Matrices for SSL Experiments.



(a) iCub WN

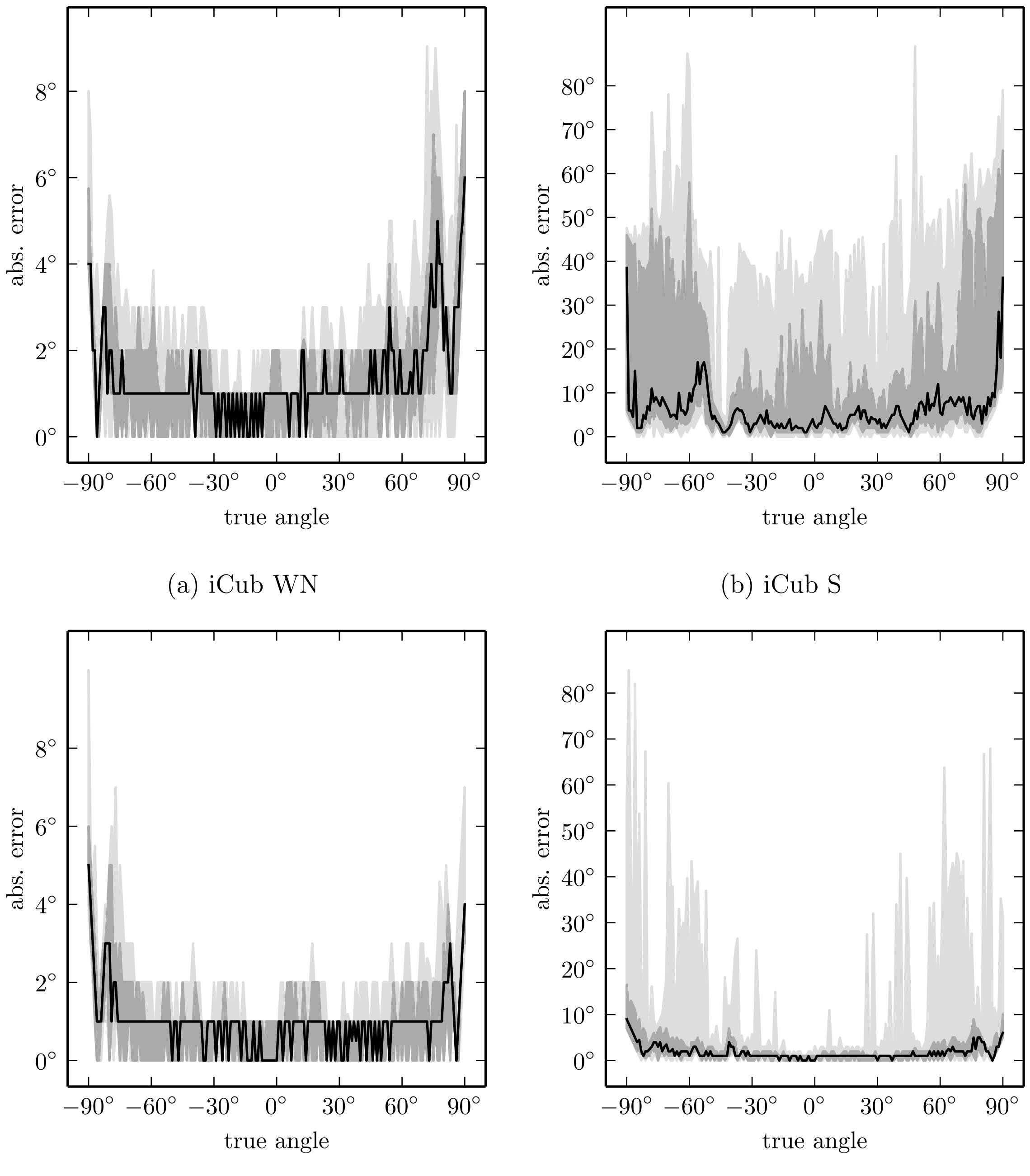
(b) iCub S

(c) Kunstk. WN

(d) Kunstk. S

> Frequency with which recordings from a true angle α were mapped to a given neuron. (WN: white noise, S: speech)

Figure 10.5: Confusion Matrices for SSL Experiments.



(a) iCub WN

(b) iCub S

(c) Kunstkopf WN

(d) Kunstkopf S

> Localization errors by true angle α, for each of the conditions.
> **Black Line**: median. **Dark gray area**: upper, lower quartile. **Light gray area**: 9th, 91st percentile. **WN**: white noise. **S**: speech

Figure 10.6: Absolute Errors at each Angle for the Four Experiments.

below:[7]

- **The number of training steps** n_t, summarized in Figure 10.7a.
- **The number of bins and input gain** n_b, summarized in Figure 10.7b.
- **The initial update strength** α_M, summarized in Figure 10.8a.
- **The mini-batch size** n_d, summarized in Figure 10.8b.
- **The asymptotic neighborhood size** σ_m, summarized in Figure 10.9a.
- **The initial neighborhood size** σ_M, summarized in Figure 10.9b.

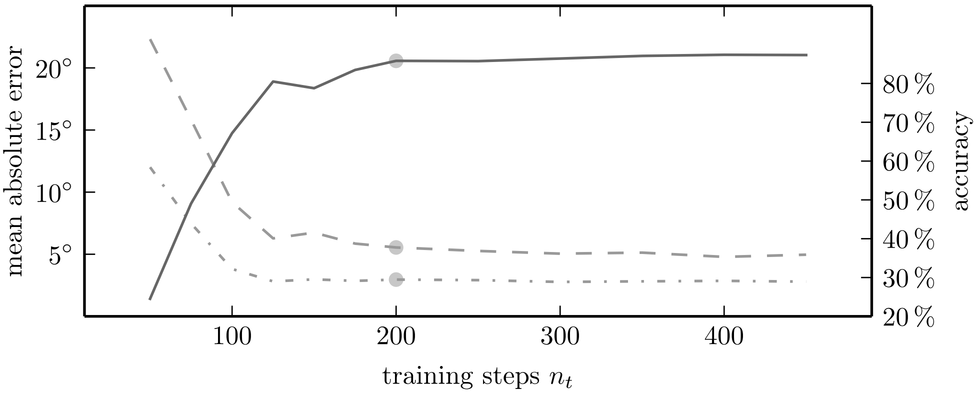
10.3.3 Results: State-of-the-Art Sound Source Localization

Figure 10.6 shows that localization of sounds from central angles was considerably better than for sounds from from peripheral angles. We therefore provide aggregate localization errors both for sounds from the entire range and from −45.0° and 45.0° azimuth ('central'). It is apparent from Figure 10.6 and Table 10.2 that our algorithm performs very well. For three of the data sets, the mean absolute localization error is below 3° for sounds from central angles, and it is below 6° over the entire range. For the third, the data set generated from recordings of speech made by the iCub head, the localization error is below 11° for sounds from central angles, and it is below 17° over the entire range. The mean absolute localization error on the best data set—the Kunstkopf/white noise data set—was around 1° everywhere.

Parameter exploration indicated that the algorithm was not highly sensitive to any of the parameters: we can see in Figures 10.7–10.9 that different values either did not have a strong effect at all, or they had large ranges which were close to optimal. The fact that the algorithm performs comparably with many of the parameter settings also points to its stability regarding the random presentation of data points. Note that the graphs seem to exhibit random behavior in some cases, but that the ranges in performance in these cases were small (e.g. for initial neighborhood width, Figure 10.9b). This indicates that the variability in performance was due to stochasticity in the training procedure rather than the influence of the respective parameters.

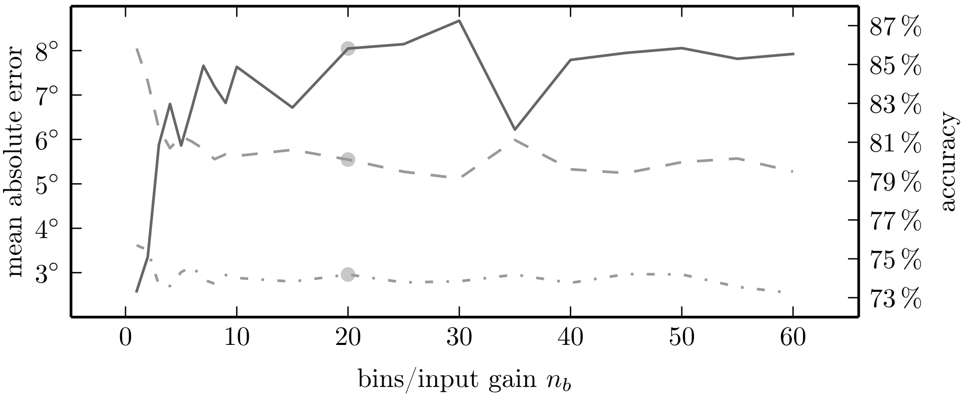
An interesting effect arose in our least successful experiment—SSL of speech using the iCub: the confusion matrix shown in Figure 10.6b indicates that a sizable proportion of the stimuli were localized about 45° or −45° from their true origin. This is especially true for stimuli presented in the interval between about −90° and −45°, but close inspection shows this effect for all angles. The corresponding matching matrix in Figure 10.6b displays a related effect: while most units were BMU only for stimuli from a small range of angles during mapping, units with numbers around 75 and 150 were BMU for stimuli from three different angles separated by about 45°.

[7] For notation and theoretical discussion, see Section 9.2.1.



The number of training steps affected the final performance, as is to be expected. After ∼ 200 training steps, improvement due to longer training stagnated. There was no detectable overfitting (probably due to the training set size).

(a) Performance by Training Steps.

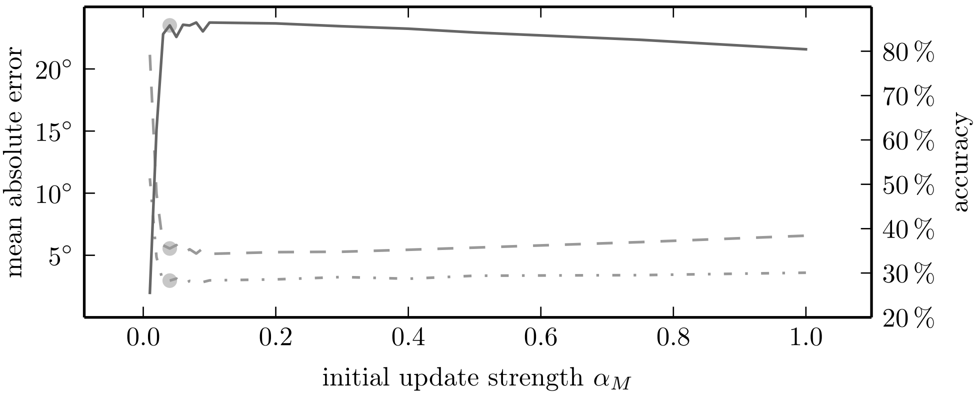


Above a certain threshold, the factor by which input data was scaled did not have a great impact on performance. The network's performance decreased considerably only when that factor was very small ($a < 5$).

(b) Performance by Number of Bins/Input Gain.

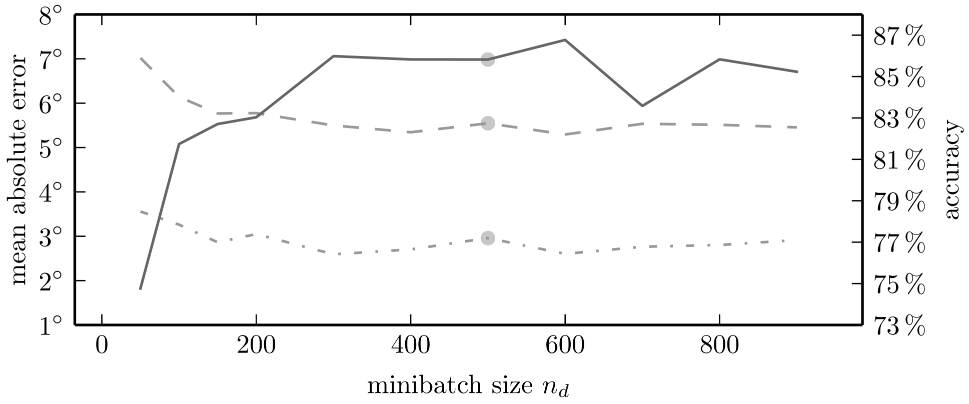
Dashed lines: absolute error on the whole angle range $[-90° .. 90°]$. **Dotted lines**: absolute error on the interval $[-45.0° .. 45.0°]$. **Solid lines**: percentage accurate localization (absolute error $|\varepsilon| < 5.0°$). **Circles**: settings as described in Section 10.3.2.

Figure 10.7: Performance After Training with Alternative Numbers of Training Steps, Input Gains.



The range of reasonable settings for the initial update strength was large. Values between 0.03 and 0.06 yielded the best results but performance degraded only slightly with greater values.

(a) Performance by Initial Update Strength.

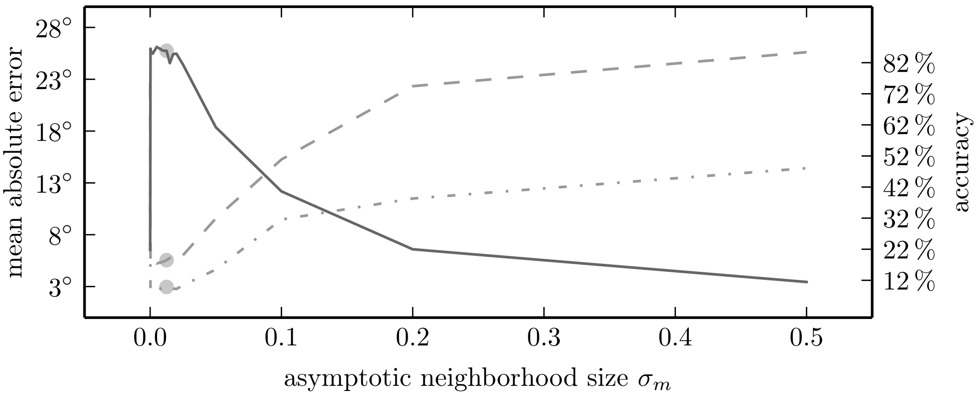


Performance was not strongly affected by the size of the mini-batch in each training step. Mini-batch sizes greater than 500 gave slightly better results but also increased the wall-clock time each training step took.

(b) Performance by Mini-batch Size.

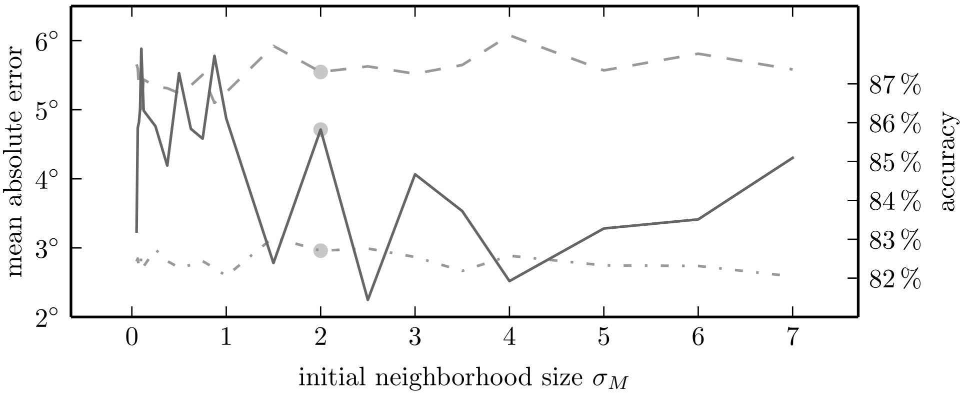
Dashed lines: absolute error on the whole angle range [−90° .. 90°]. **Dotted lines**: absolute error on the interval [−45.0° .. 45.0°]. **Solid lines**: percentage accurate localization (absolute error $|\varepsilon| < 5.0°$). **Circles**: settings as described in Section 10.3.2.

Figure 10.8: Performance After Training with Initial Update Strength, Mini-batch Size.



On this data set, a small asymptotic neighborhood size produced the best results. Performance dropped with increasing asymptotic neighborhood size and when the asymptotic neighborhood size was very small ($a \ll 0.001$).

(a) Performance by Asymptotic Neighborhood Size.



Initial neighborhood size did not strongly affect performance on this data set. Especially initial sizes greater than zero and smaller than the network's size produced good results reliably.

(b) Performance by Initial Neighborhood Size.

Dashed lines: absolute error on the whole angle range [−90° .. 90°]. **Dotted lines**: absolute error on the interval [−45.0° .. 45.0°]. **Solid lines**: percentage accurate localization (absolute error $|\varepsilon| < 5.0°$). **Circles**: settings as described in Section 10.3.2.

Figure 10.9: Performance After Training with Alternative Initial Update Strengths, Mini-batch Sizes.

10.4 Intermediate Discussion: Self-Organized Learning of Auditory Sound-Source Localization

In a series of robotic experiments, we have made binaural recordings of white noise and of speech from angles between $-90°$ and $90°$ using two different robots, the iCub head and a rotatable Kunstkopf. We created data sets for SSL from each of the sets of recordings by applying ANN models of the MSO and the LSO, which were previously proposed by Liu et al. (2010), and which extract ILD and ITD in different frequency ranges. Finally, we applied our new algorithm to each of these data sets.

Localization performance is generally good on three out of the four data sets. It is significantly worse on speech data recorded by the iCub. This is readily explained by the construction of that robot: the iCub head is essentially a plastic orb with facial features, with cameras, two microphones in simple pinnae, and a full-blown integrated computer inside. The head is mostly hollow, allowing some of the sound to travel through it, and its surface is smooth, thereby not shaping the sound traveling around it as strongly as a real head would. Much more importantly, though, the computer inside the head is cooled by small fans which are only centimeters from the microphones and separated from them only by a thin plastic wall. Therefore, all recordings made with the iCub suffer from the constant humming of the fans in the background at a sizable volume. White noise has the same energy in all frequencies (subject to the speakers' fidelity). Therefore, binaural recordings of white noise contain information usable for SSL even if they are corrupted by irrelevant background noise that is limited to certain frequency ranges, like that from the fan. Speech, on the other hand, is limited in the frequency ranges available, and therefore affected much more strongly. It thus comes as no surprise that performance on the white-noise data set is much better than on the speech data set.

Dávila-Chacón et al. (2012) showed that the system based on the Liu et al. (2010) model is capable of dealing with the noise in the Nao robot, which is similar in that respect to the iCub, and localize sound reliably (but not flawlessly) with a resolution of 15°. Later, Davila-Chacon et al. (2013) found that the same architecture required additional postprocessing steps to achieve acceptable localization performance on the iCub, which produces $\sim 60\,\mathrm{dB}$ of ego noise—considerably more than the $\sim 40\,\mathrm{dB}$ of ego noise in the Nao. They achieved good SSL on the iCub robot with a number of combinations of statistical and ANN methods, but again, the authors did not attempt to localize sounds at a resolution of less than 15° with any of the methods. Under those circumstances, the accuracy of our algorithm on the noisy data set seems acceptable.

In the previous section, we noted a pattern of errors in the iCub–Speech experiment. Some units apparently became selective to stimuli from three different angles each, separated by ±45°. This offset is also the angular distance between our three speakers. We conclude that these units did not, in fact, learn to localize the sound, but learned instead to determine the position of the robot head from spurious cues.[8]

It is clear that this is not the result we hoped for, and it shows a limitation of our

[8] See Section 10.3.1.

system. That limitation results from a general problem of unsupervised learning: the system successfully learns to distinguish data points, but the criteria by which they are distinguished cannot be controlled.

However, the failure of our system to learn to localize some of the stimuli in one of our settings points to a danger in all evaluations of adaptive SSL systems. Had we not used three different speakers, we would not have the confusion and all stimuli which in our experiment were localized according to the absolute position of the robot would have been counted as correct. Similarly, any experiment in which only the physical SSL system or only the sound source moves runs the risk that the system learns to use extraneous cues to distinguish experimental conditions. If relevant cues and irrelevant cues always agree—in this case, because either only the speaker or only the robot moves—then the system will never distinguish and use them according to their reliability. This can happen with supervised and unsupervised learning algorithms. From the studies cited in Section 10.2, only the one by Rucci et al. (2000) used multiple speakers *and* a moving robot.

In this part of the present thesis, we are interested in applications of our algorithm, not in modeling. Still, there is at least one interesting observation with regard to the relationship between our system and biological SSL. Comparing the subfigures of Figure 10.4, we see various levels of global continuity in the topographic mapping. Mapping of auditory stimuli in the SC is usually globally continuous (and in register with retinotopic mapping of visual stimuli). However, Knudsen (1988) has found discontinuities in the topology of the mapping of auditory stimuli in the optic tecta of dark-reared barn owls. Visual inspection of the matching matrices in Figure 10.4 and comparison to the performance in the different conditions shown in Table 10.2 reveals that continuity of the mapping goes along with accuracy. If self-organization plays a role in the development of an auditory space map as suggested by Gutfreund and King (2012), then the discontinuities seen in our topographic maps might be analogous to those seen in the optic tecta of dark-reared barn owls. Considering that these continuities seem to be correlated with the quality of the information available for localization, we get both an extended hypothesis on the role of vision in the formation of the auditory map and a prediction: the prediction is that discontinuities in maps of auditory space in dark-reared animals should depend on the quality of auditory localization cues: the less reliable the cues, the more discontinuities should be expected. The hypothesis is the one we have been making in the modeling part of the present thesis: according to that hypothesis, vision does not play a distinguished role at all, or rather, it plays a distinguished role only because of the quality of information it provides.

There are ways in which our algorithm could be improved to make it more practical, depending on circumstances. One of the most basic ways in which the algorithm could be improved is by using a more sophisticated stopping criterion than just reaching a maximum number of training steps. In supervised training, it is common to track performance on a validation set. In principle, this would be possible for our algorithm as well: one could, every so often, compute the current mapping of the network and evaluate performance on a validation set in much the same way as we have done in Section 10.3.2. One could then train until performance stagnates or starts deteriorating.

However, mapping and computing performance are computationally expensive with large networks and data sets. A much simpler alternative would be to stop training as soon as the clustering of training data points stops changing from one update step to another. The drawback of that method may be that the network may have learned a stable mapping for the data points, but not necessarily fully updated the statistics of the input in the units' histograms. Finally, one might track the change of the units' histograms—measured, for example, by Kullback–Leibler divergence between histograms before and after an update—and stop training once that change is below a certain threshold.

A more ambitious way to improve our algorithm would be to try to develop mechanisms by which the algorithm regulates its own parameters during learning. Especially mini-batch size, update strength, and neighborhood size are good candidates for adaptive parameters. Mechanisms for implementing adaptive parameters have been developed for regular SOMs and their relatives (e.g. Adams et al. 2013; Berglund 2010; Fritzke 1995). The probabilistic nature of our algorithm makes the search for good adaptive update regimes both more involved than for regular SOMs and possibly very effective as it may guide estimation of optimal parameters.

Adaptive parameters are especially appealing in situations where a system learns continuously in a changing world (Adams et al. 2013). These are also the situations in which an unsupervised system like ours has the greatest benefits: in the experiments described in this section, we had ground truth for every one of our data points. Therefore, a supervised learning method could have been used and might even have produced better results. However, an autonomous system in a real application may not always have access to ground truth. In fact, it is easy to imagine scenarios where a system almost never gets the chance to label its data points. In such situations, an algorithm like ours can at least learn to distinguish different states of the world. It thus drastically reduces the input dimensionality of decision problems,[9] making, for example, reinforcement learning a feasible method for learning correct behavioral responses.

[9] by implicit inference in an LVM, see Section 8.1.

11 Application Summary and Discussion

Part II of the present thesis was dedicated to the second part of the question posed in Section 1.2.1: how can we use some of the mechanisms which allow the dSC to do what it does in AI and especially in robotics?

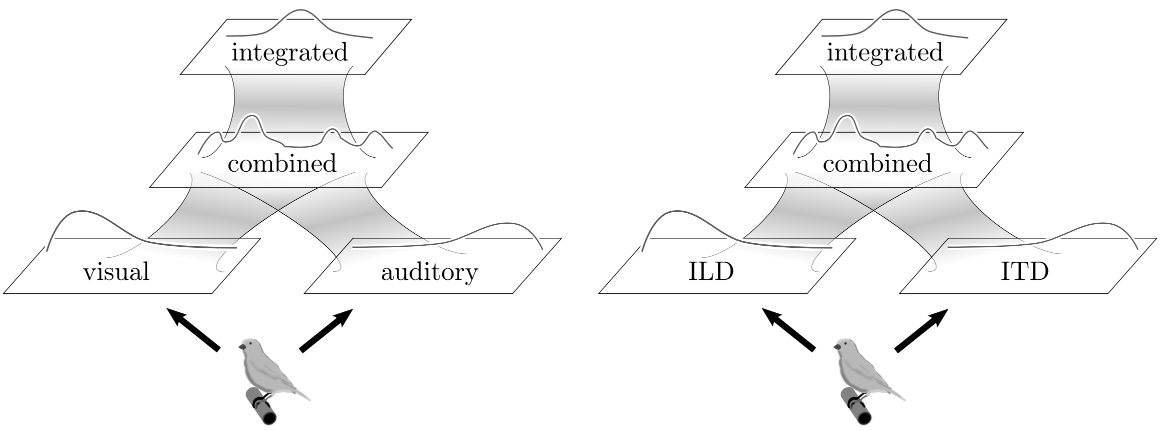
Our analysis and experiment provide one answer to this question. First, we have interpreted the ANN algorithm, which we proposed as a basis for modeling the dSC, as an algorithm which learns an LVM in an unsupervised fashion, and uses that LVM for statistical inference. Important characteristics of that algorithm—self-organization, local interaction, topographic mapping, probabilistic population-coding, learning of likelihood functions, and naïve Bayesian inference on high-dimensional data—were transferred directly to a new algorithm, CB-HistoSOM. In Chapter 4, these characteristics were motivated from biological fact and necessity, and demonstrated to lead to behavior similar to that seen in biology.

The new algorithm includes those characteristics and additionally caters to the specific settings of typical machine learning problems. In contrast to biological settings, we cannot assume an initialization which is close to adequate.[1] Instead, we often need to initialize our algorithms randomly. Another difference is the amount of time and data available for learning. It often takes days, weeks, or even months for young animals to develop certain basic sensorimotor skills and sometimes years to master adult-like performance.[2] In machine learning and robotics, we can usually only afford hours or maybe days to collect data and about the same time to train our algorithms. Also, the parallelism of processing in machines is still a far cry from that in natural information processing. And thus the number of neurons simulated, the granularity of data, and the speed of processing are still much lower.

On the other hand, in artificial information processing, we have the ability to store all our data points perfectly and present them to our algorithms repeatedly and synchronously. Also, we are free to simulate neurons in our algorithms in any topology we like, whereas biological neurons are bound to the topology of the real world, and non-Euclidean topologies in their connectivity come at metabolic and other costs. Thus, the algorithm we presented in Chapter 9 combined the above-mentioned mechanisms, which were extracted from our modeling of biological information processing, with batch learning and a circular topology of the simulated neural network to partially compensate for the shortcomings of *in-silico* processing.

[1] See Section 2.2.7.

[2] See Section 2.2.7.



(a) Audio-Visual Localization

(b) SSL

> To the computer scientist, audio-visual localization and SSL are very similar: in both cases, different properties of the stimulus give rise to different sets of features. These sets are combined into one set of features on which an algorithm may operate to generate one coherent response.

Figure 11.1: Similarity of AV localization and SSL.

In the spirit of abstracting solutions found for one problem and making them available to computationally similar problems,[3] we have shown that our algorithm, which was inspired by audio-visual localization in the dSC, could be used to solve a different problem, binaural SSL. This problem is indeed similar, at a computational level, to biological audio-visual localization (see Figure 11.1). In both cases, the input consists of high-dimensional data points in which there is a stochastic, highly non-linear relationship between the values of each of the input dimensions and the low-dimensional property about which we want to make inferences. In biological audio-visual localization, there is far more data available for learning than there is feedback. In binaural SSL, the stochasticity of the input makes large data sets favorable for learning. Thus, both problems favor solutions using unsupervised or partially unsupervised learning algorithms.

The transfer from neuroscientific modeling to the technical problem was very successful. In an experiment with two different robots, we showed that our solution is competitive with state-of-the-art specialized SSL systems in terms of localization accuracy. In contrast to most of these systems, our algorithm can in principle adapt online, it is unsupervised, and it is not limited to auditory input. Thus, it may be more applicable to the problem of SSL in autonomous robots, which require performance in unknown environments and with a changing body, and where additional contextual or cross-sensory information may be available for localization.

We have argued that the applications of our work cannot lie in robotic audio-visual

[3] See Section 1.2.2.

localization. Again, the reason is that, in contrast to auditory localization, visual localization comes as a by-product of detection in technical systems. When visual detection succeeds, then localization is near-perfect. Otherwise, it fails completely. Thus, auditory localization can serve as a backup for those cases where a stimulus cannot be detected, or when visual information is ambiguous. An integration in the sense of MSI in the dSC, however, is not useful for localization.

This raises two questions: a) Why does the dSC seem to integrate visual and auditory information and b) Can our work be used for audio-visual *detection* instead of localization? The former question concerns biology and will therefore deferred to Appendix B. The answer to the latter question is a speculative 'yes, in theory, but...'

Visual object detection, in Viola and Jones' (2001) system and in others, usually computes a set of statistics called 'features' for an image and then combines these features in a way specific to the system to decide whether the image depicts the target object or not. To detect an object in a large image, that image is segmented into smaller sub-images and each of these sub-images is classified as a depiction of the object or of something else. In principle, one could try and form a population code from features such as those used by classical computer vision (CV) algorithms and use that population code as visual input to our algorithm. Our algorithm could then learn to combine that visual input with auditory input.

All would hinge on the quality of these features, though. The features used by Viola and Jones' (2001) system are simple, general, and cheap to compute. The power of that system lies in the supervised learning algorithm which learns to combine them effectively. For our unsupervised algorithm, features would be necessary which are already specific to the kind of object to be detected. Otherwise, the visual input would not strongly depend on the position of the object and thus our algorithm could not learn to estimate that position.

Even given good multisensory features, it is far from clear that our algorithm would fare better than a purely visual standard CV algorithm: ours is a naïve Bayesian algorithm and as such cannot make use of complex causal structure in the data. This is in contrast to classical CV algorithms, which often implement sophisticated decision processes. An improvement could still be possible, but, in a specific application, it would likely be small and the cost may be high compared to that of better hardware or a more computationally intensive CV algorithm.

On the other hand, in the application we chose, binaural SSL, there is room and a need for improvement. Current algorithms are mostly supervised, and they lend themselves neither to integration of non-auditory information nor online learning. With our solution, a robot may be modified or brought into an environment with drastically new acoustic properties, and it could gradually adapt its learnt SSL without the help of a teacher. It could also naturally integrate supplementary information like expectations or additional acoustic features.

Thus, we have shown that the mechanisms incorporated in our algorithm are not only useful, apparently, in a biological application, but they can also be put to good use in an actual robotic task. These results are encouraging and they suggest that our algorithm may fare well also in other information processing applications, multi- or unisensory, or even amodal.

Part III

Closing

12 Conclusion

12.1 Thesis Summary

In the preceding pages, we have studied information processing in the deep superior colliculus (dSC) for the double purpose of learning about this brain region and extracting knowledge for applications in robotics and in artificial intelligence (AI) in general.

After a discussion of the goals we were going to pursue and the methods we were going to use, and after a short review of the biology of the superior colliculus (SC), we have first modeled the dSC. To this end, we have taken the view that it is the dSC's task to localize objects and events. Further, localization is encoded in a topographic population code and that population code is a probabilistic population code (PPC), meaning that the response of each neuron is an encoding of the estimate of the probability that the stimulus is in the neuron's best area. The dSC learns to localize objects and events and to encode its estimates in a PPC and it does so in a partially unsupervised fashion.

Out of this view, we have developed an artificial neural network (ANN) algorithm, which, as we have proceeded to show, replicates important neurophysiological and behavioral phenomena of biological audio-visual localization. Some of those effects were demonstrated in simulations and some, additionally, in a neurorobotic experiment.

We have then extended our model to include abstract cortical input. Cortical input was connected to the network in the same way as primary sensory input. Without, thus, having any information on the kind of input, our network learned to use that cortical input, and its simulated neural responses and its behavior exhibit effects which are comparable to those of spatial and feature attention.

Having thus developed a model of multisensory integration (MSI) in the dSC and having demonstrated its adequacy, we turned towards the practical use of our insights. First, we analyzed and interpreted the algorithm at the basis of our modeling and its mathematical properties. We then compared it to similar, established algorithms and techniques. We identified our algorithm's strengths and weaknesses, from a practical perspective, and developed a new algorithm which combined the basic principles of our biological model with optimizations for applications on standard computing hardware and in robotic applications.

Finally, we used our new algorithm to implement a robotic binaural sound-source localization (SSL) system and tested that system rigorously in a range of experiments with two different robots and two different kinds of input. We were able to show that this system, which is based on the information processing and learning principles we have found in the dSC, produces state-of-the-art SSL performance, while being more flexible than many of the more established SSL systems.

12.2 Discussion

As a consequence of the cross-disciplinarity of our research, our work has produced insight in two scientific areas: computational neuroscience and computer science and robotics.

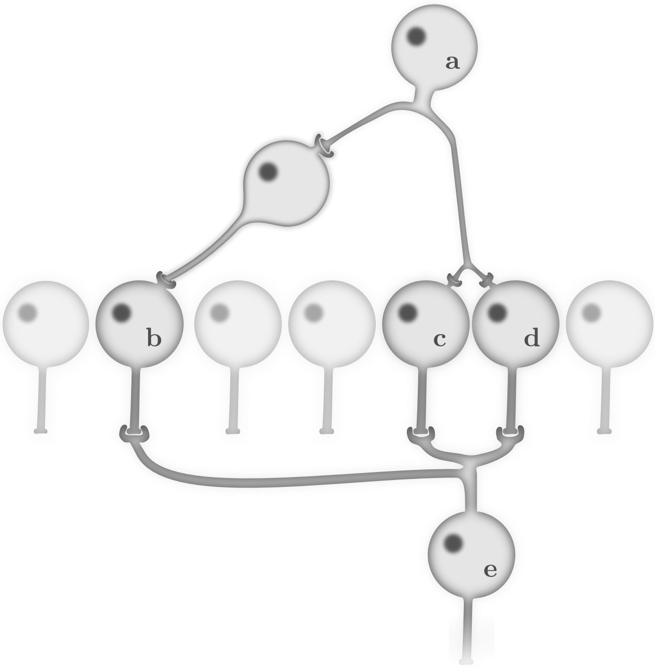
By assuming a function of the dSC and formulating an algorithm for the fulfillment of this function, we have implicitly proposed a hypothesis about and interpretation of its biology. Made explicit, that hypothesis states that the dSC is a brain region which learns, in a partially unsupervised fashion, to localize objects using input from various modalities. In particular, it learns to accurately take into account the tuning functions and noise characteristics of each of its sensory and non-sensory inputs in integrating them. The result is a topographic PPC for the position of the stimulus. Statistical learning and self-organizing map (SOM)-like self-organization are important principles in the process of learning MSI in our model of the dSC. By showing that important phenomena not only of bottom-up sensory, but also of top-down attentional processing can be explained by those principles and in light of our theory, we have demonstrated the adequacy of our theory.

We have advanced a novel unsupervised machine learning algorithm. That algorithm learns a non-linear latent variable model (LVM) for high-dimensional input and topographic mapping from the latent variable into a grid structure. In contrast to, for example, the original SOM algorithm and principal component analysis (PCA), its output does not only provide an estimate of the latent variable on the basis of the input, but also estimated likelihoods for other values of the latent variable. Its output can thus be used in later steps of statistical processing.

We have shown that our biologically inspired machine learning algorithm can be used in a robotic application: in a binaural robotic SSL setting, it produces results which are comparable to those achieved with specialized systems. In contrast to these systems, learning is unsupervised in our algorithm, potentially online and continuous, and the information provided to the system can be naturally extended to additional, sensory or non-sensory information.

Apart from the use of an algorithm developed for modeling as a basis for practical applications, there are other interesting interactions between our modeling and technical work. First, both, the modeling and the practical ANN algorithm, use local interaction between neurons, and thus implement topographic mapping from data space into the network's grid. One might wonder what is the benefit of topographic mapping. After all, what is necessary to read out a population code is not the location of each neuron, but its best area or preferred location/value. Further, presuming there is a benefit, it is unclear that that benefit exists both in biological and technical settings.

In biology, one of the most important reasons for clustering neurons with similar response properties is probably the metabolic cost and time delays incurred by long neural connections (Kaas 1997; Koulakov and Chklovskii 2001): long, even polysynaptic connections are necessary to provide two spatially separated neurons with the same neural input. If, on the other hand, these neurons are close to each other, they can connect to the same axons carrying that input (see Figure 12.1, top). Thus, grouping neurons with



Top: Neurons **b**, **c**, and **d** all require input from neuron **a**. If they are clustered, like **c** and **d**, then shorter wiring is required than if they are scattered.
Bottom: Conversely, if projections from **b**, **c**, and **d** were clustered, then neuron **e**, which requires input from all of them, would have a higher probability of making correct connections, provided it is close to the common projection site of its input neurons.

Figure 12.1: Spatial Organization Minimizes Wire Length.

similar response properties, and thus likely similar dependencies on input connections, helps minimize wire length. While this explanation has recently been called into question for maps of higher-order stimulus features like spatial frequency or orientation, it is unproblematic for two-dimensional maps of visual space (Wilson and Bednar 2015).

We argue that topographic order also facilitates learning. Consider, for example, learning of motor behavior downstream of the dSC. The dSC is not only itself spatially organized; many of its projections are, as well (May 2006). A motor neuron in the brainstem saccade circuits whose activity drives saccades, say, to the right would have to make input connections to axons from dSC neurons coding for saccade targets on the right. If projections from the dSC are in topographic order, and if that motor neuron is close to the projection targets of the dSC neurons whose input it requires, the (non-random but) stochastic process with which the neuron makes and severs these connections has a higher chance to make appropriate connections than if the dSC and its projections were not topographically ordered (see Figure 12.1, bottom). The same goes, of course, for learning in the dSC and topographically ordered afferent projections to the dSC.

But where does that leave spatial organization in ANN? In typical implementations, the computational cost is the same for every connection, whether it is local or distant. And the way connections are strengthened and weakened does not need to be affected by the distance between neurons—that is a matter of programming. Thus, the reasons for topographic order mentioned so far do not hold for ANN as they do for biological neural networks.

In CB-HistoSOM, topographic order enters through the neighborhood interaction. Removing that mechanism or, equivalently, setting the neighborhood interaction width to zero, would have preserved competitive learning, but it would have eliminated learning of spatial ordering. Thus, learning might have succeeded in leading to specialization of neurons, but close-by locations would not have been mapped to close-by units after learning. However, since we determined the mapping of each unit empirically and independently, we would still have been able to read out a probability density function (PDF) from the network's activation.

There is at least one reason why SOM-like neighborhood interaction is beneficial for learning both in ANN and in natural neural networks. Consider an untrained network with a very large number of units. Without neighborhood interaction, every data point updates only one unit in each training step. Thus, some of the units will learn to respond strongly to the input early on, and the rest may stay in its initial state for a long time, or even forever. With neighborhood interaction, every training step affects many units at once. While, in every step, one neuron is updated most strongly, the rest gets a chance to learn from the input as well. This soon leads to a rough spatial organization. In turn, this leads to collateral updates updating units with data points for which they already have at least a weak preference. The same holds for natural neural networks, where learning episodes are used not only to train neurons which are exactly the best-suited for the given input, but which prefer input similar to the current input. Thus, data points/learning episodes are used more efficiently with spatial organization than without—given that the input data has an implicit spatial organization. This is both an insight for machine learning and for the interpretation of biological neural networks.

Another interesting implication from our practical work for biological neuroscience arises from the application of our algorithm for SSL. Like the SC, the external nucleus of the inferior colliculus (ICx) also features a map of space—only auditory, not retinotopic.[1] That order arises in part because of input from the SC, but it does not require visual input (Gutfreund and King 2012; Hyde and Knudsen 2000). Thus, one might think about modeling the ICx on the basis of our network algorithm, or a derivative of it. Less is known about the biology and development of the inferior colliculus (IC) and ICx than about the SC. Exploring the possibilities of modeling it using statistical self-organization may therefore provide valuable research hypotheses. As an in-depth literature study of the IC is beyond the scope of this thesis, we unfortunately cannot pursue this promising direction, here.

[1] See Section 10.1.

12.3 Future Work

The model of the dSC we presented in this thesis is a model at a relatively high level of abstraction. To explain phenomena of dSC biology in greater detail, it will need to be implemented in a more biologically plausible way. This means lifting some of the simplifications we made in Section 2.3 and Chapter 4.

First, it would be good to see how the approximation of likelihood functions could be implemented by biological mechanisms. One step in that direction could be implementing each neuron by an multilayer perceptron (MLP) and having that MLP learn a multi-dimensional PDF. The biological analogy would then be that each one of our units really represents a small cluster of neurons which cooperate to learn the statistics of the input.

Another direction towards greater biological detail is making our network recurrent. This would allow our algorithm to perform Bayesian belief update about the position of a stimulus. It could model the way individual action potentials from sensory areas are integrated as they arrive. Simple recurrency of this kind should not be very hard to implement, though some changes to our network and especially the learning algorithm would be necessary. However, sooner or later, one would want to implement integration and learning using a more biologically plausible spiking neuron model. Combining a leaky-integrate-and-fire-type model with the kind of self-organization and probabilistic information processing would require some very careful engineering.

A dynamic model would also allow for a more detailed implementation of the winner selection mechanism which is currently an explicit step in our algorithm. Appropriate lateral connectivity between neurons could be designed to replace the winner selection step by an emerging winner-take-all effect of the network. Such biologically plausible implementations of the SOM algorithm have been presented before (e.g. Panchev and Wermter 2001; Rumbell et al. 2014). However, though far from impossible, implementing this while preserving the probabilistic interpretation of the algorithm and network would not be straightforward.

In contrast, we expect learning motor behavior based on localization in our dSC model to be a relatively easy addition. It would probably suffice to implement a reward-mediated learning network, similar to that proposed by Weisswange et al. (2011), to get a reasonably realistic model of the saccade circuits in the brain stem and drive orienting movements in simulations or in a robot.[2] Moderate additional complexity would be required to include the influence of the cerebellum.

The most ambitious extension of our model would be closing the loop from cortical regions to the dSC and back. We have, so far, considered only the influence of cortical input to the dSC. This has been one theme in modeling the dSC. However, outside of modeling, the effect of SC activation on cortical function has received a fair amount of attention.[3] It would be fascinating to study the dynamics of reciprocal connections between cortical areas, which are presumably involved in associative, contextual, and

[2] See Section 2.2.2.

[3] See Section 2.2.5.

even semantic processing, and the dSC, which is usually thought of as a mostly sensorimotor brain region. A model that captures these dynamics may provide valuable insight into the interplay between the highly parallelized cognitive and the serialized motor and attentional processing and thus attach to such theories of consciousness as those of O'Regan and Noë (2001) and Merker (2007).

It is a natural consequence of our biomimetic approach that modeling tends to be ahead of applications for most of the time. One aspect of our modeling which has not yet seen use in our practical application is the inclusion of contextual information.[4] Thus, our work could be extended by higher-level processing which provides situational knowledge to our network, for example in an SSL setting. Visual or predictive cues heralding a sound source with specific acoustic properties to be localized would be one direction.

Finally, SSL is not a means in itself. It would be good to see our SSL system as a module in a larger robotic system performing a useful task.

12.4 Conclusion

In this thesis, we have taken a strongly multi-disciplinary approach to studying the dSC. That approach consisted of first modeling the dSC to understand how its function is implemented in its biology and then to extract principles of this implementation as a basis for advances in machine learning and robotics.

The first step, modeling, was characterized by an information processing view on the function of the dSC. This view produced a model which depicts the dSC as a self-organizing network of neurons which learns a topographic LVM of its input. The latent variable in the dSC, according to our model, is a combination of the position of a stimulus and the identity of the modalities which contributed to that stimulus. Our model includes projections from cortical regions. These projections are viewed as just another kind of input, indistinguishable to the network from sensory input. We have shown that the model we have advanced reproduces a number of phenomena known from basic MSI and from attentional processes. Reproduction of these phenomena corroborates the adequacy of our model as a model of the dSC.

Previous models of MSI and the dSC have featured self-organized learning of spatial order, cortical input, or a probabilistic interpretation of neural information processing. However, the model combining these and the range of phenomena explained by that model are a novel contribution to computational neuroscience.

In the second step, harvesting the knowledge gleaned from modeling the dSC, we have re-cast the principles embodied in our model into a practical machine learning algorithm. The resultant algorithm addresses some of the more serious computational problems of the original algorithm at the heart of our dSC model. It is an unsupervised, self-organized batch learning algorithm, based on the SOM, which makes only very weak assumptions on the kind of input and especially the noise in that input. The algorithm

[4] See Chapter 6.

learns to integrate information from each of its input dimensions, taking into account the stochastic properties of each dimension. Based on that algorithm, we have built a neurorobotic, binaural SSL system. We have shown that that system can perform state-of-the-art sound-source localization.

Self-organization is an old mechanism in machine learning, and other algorithms have been proposed which perform learning of LVMs and statistical inference. However, our algorithm is novel and a contribution to machine learning in that it makes very weak assumptions on the input and affords continuous learning. Our application in SSL is valuable in itself for various use cases. Being more flexible than traditional approaches, it may serve, out of the box or adapted to fulfill specialized demands, as a module in a robotic system.

Beyond those concrete contributions, we see value in this thesis as a case study of tightly connected research in computational neuroscience and computer science. The considerable advancements in both fields justify the approach taken, summarized at the beginning of this section, and suggest it as a model for future research. Especially the in-depth methodological considerations at the beginning of this thesis and in Appendix A may provide a template—to be revised and adapted—to scientists on a similar path.

Appendices

A The Merging of the Sciences: Methodology in Modeling and Machine Learning

This thesis contributes to two distinct areas of knowledge, machine learning and computational neuroscience, and therefore operates within different scientific communities with different goals, customs, and standards; different paradigms (Kuhn 2012, in particular Hacking's introduction). The differences in methodology between these areas are due to the differences in goals and problems. Consequently, we apply different methodology in this thesis where different scientific disciplines are concerned. This change of methodology from one part of the thesis to another may raise concern in readers, depending on their background. In the following, we will therefore explain, for those practices which are not the same in machine learning and computational neuroscience, why they are commonly accepted in their respective field and thus justify our use of them where we make claims belonging to either of the two fields.

A.1 The Role of Parameters in Studying Models and Algorithms

By modeling the dSC, we work to improve the understanding of that brain region. This effort is contained in the scientific discipline of computational neuroscience. Computational neuroscience tests computational instantiations of neuroscientific models by simulating computationally the implications of their models under certain conditions. For this to be possible, instantiations must be computationally feasible. Real neural systems have vast complexity and modeling them without significant simplifications is not generally feasible. To allow computationally feasible instantiations, a model must therefore allow simplifications and claim invariances between the effects observed in the simplified instantiation and the real system the model describes. Since simplifications usually drastically alter the quantitative behavior of a system, these invariances tend to be qualitative. Consider, for example, the dynamic properties of a model of some neural system which is instantiated for simulation's sake by a rate-coded, recurrent, discrete-time ANN comprising a few hundred units. Observing that the network settles into a stable state on a given type of input in a certain number of simulation steps, we will not usually make quantitative predictions about the time the original system will take to reach an equilibrium. Instead, we will typically predict that the system *will eventually* reach a stable state if its input shares certain qualitative similarity with the input we presented to our ANN.

The fact that predictions are typically qualitative does not mean that quantities do not matter entirely. Many parameters of an instantiation, like the number of units in an ANN or the number of training steps in a learning sequence, are important for the instantiation to produce the behavior that the model predicts to be similar in the real system. Often, these parameter settings cannot easily be translated into quantities in the system being modeled. Modelers therefore usually imply that the parameter choices they make do not represent quantitative predictions about the world: instead, the fact that a certain instantiation with a specific parameter setting exhibits a certain behavior proves that *there exist* an instantiation and parameter setting for which the instantiation shows this behavior. If the analogy between the qualities of the instantiation and the system is plausible and these qualities are plausibly identified by the model as responsible for the observed behavior in the instantiation, then such an existential proof for an instantiation and parameters strengthens the model's claim at explaining the behavior in the system.

Proving that there exists *another* parameter setting for the same instantiation which produces this behavior does not add much credibility to the model. This is unless it is shown for some parameter that all or almost all settings for this parameter produce results which are similar in ways important for the model's predictions. This is typically the case for seeds of pseudo-random number generators in stochastic instantiations. Otherwise, it is common to demonstrate a model instantiation's properties only for one parameter setting, or a few to show that the results were not accidental (see e.g. Krasne et al. 2011; Malsburg 1973; Spratling 2012; Weber and Triesch 2009a, for a somewhat arbitrary sample of models from different contexts).

All of this is very different in the second scientific field to which we contribute: in machine learning, and more generally in algorithm design, the goal is not to describe some aspect of the world, but to produce algorithms which solve a specific task as well as possible. To show that a new algorithm is useful, a designer or other analyst can study the algorithm's mathematical properties or apply it to data, which may be synthetic or from the actual problem domain. A new algorithm is good if it solves a problem more effectively or efficiently than previous algorithms.

Like computational instantiations of neuroscientific models, algorithms often have parameters. Having too many parameters can impede application of an algorithm if there is not a rule or at least a heuristic that tells the user of the algorithm which parameter settings will produce good results in a given situation. Therefore, a proponent of an algorithm will show that either there is one setting or range of settings for some parameter which will generally yield good results, or that there is a heuristic which will often identify such parameter settings. If this cannot be shown analytically, then it is shown empirically, which means by demonstrating good results for the parameter setting or many of the parameter settings in the range or identified by the heuristic.

The above discussion of the different methodologies in the two fields should make clear the differential roles of parameters: the computational neuroscientist claims that there is a parameter setting with which his or her simulation will generate certain results. It therefore suffices to find one such setting and demonstrate that it does. The computer scientist claims that either the algorithm's performance is not strongly affected by the choice of parameter setting, or that it is easy to choose a good setting for a given situation. It is therefore necessary to try out the algorithm with all or many of the suggested settings.

A.2 Quality of Input

The computational study of both neuroscientific models and algorithms in algorithm design consists of applying them to concrete data and suggesting that the results will be similar to other, qualitatively similar, data. In principle, the data to which algorithms and model instantiations are applied can be synthetic in both cases. However, the plausibility of the results then hinges on the plausibility of the similarity between the synthetic and the actual data. Since it can be difficult to argue convincingly that the synthetic data is similar in all the ways that matter to the data to which an algorithm will be applied, results from computational studies of algorithms in algorithm design are much more plausible if the data is 'real'.

The same argument applies to testing models in computational neuroscience. However, since models usually describe neural subsystems instead of entire nervous systems, it is often exceedingly hard to obtain real data: many brain regions receive incoming connections from a host of neurons elsewhere in the brain. It is difficult to measure the activity of just a few hundred neurons at the same time and know exactly where they connect in the target brain region. Measuring at the same time and over ontogenetic time spans the activity of only a significant proportion of all the neurons projecting to a given brain region under study is impossible with current technology. Measuring only some of the incoming activity and interpolating to the rest, on the other hand, again amounts to elaborated synthesis of simulation data.

A second problem of using real data in computational neuroscience is related to our argument about the necessary partial mismatch between the system being modeled and the computational instantiation of the model. Since the computational instantiation comes with gross simplifying abstractions, its response to the same data cannot usually be expected to be similar to that of the actual system. Since real input data is hard to come by in computational neuroscience and since it could not be expected to produce the same behavior in the computational instantiation as in the system being modeled, anyway, computational neuroscience usually settles for synthetic data which is, as we have explained above, qualitatively similar to real data in those ways which are asserted to be important by the model.

A.3 A Word on Neurorobotics

We have argued, in Section 1.2.2 and in Bauer et al. (2012b), that robotic experiments can be used to inspire and test neuroscientific models. It might appear at first glance that we have dismissed this argument in the previous sections. After all, implanting an instantiation of a model as a module of a robotic system is just another way of generating input data for the model. And as we have argued above, there are only two things regarding the input data that matter for the plausibility of a computational test of a neuroscientific model: the first thing is the plausibility of the similarity between the data in the simulation and in the real system. The second is the plausibility of the argument that those similarities in modeled and actual input are responsible for producing the behavior being explained. Thus, results of a neurorobotic experiment could only support a neuroscientific model if the input to the model's instantiation is plausibly argued to be similar to biological reality in all the ways that matter. However, although a robot's means of perception may use the same kinds of energy as biological sensory organs (light, sound etc.), the way they do that is usually very different, and therefore that argument is not trivial.

Take for example vision using a standard digital camera. The differences start the moment the light hits the sensor at the very latest. The digital image sensors (DISs) in digital cameras, which are analogous to the retinae in human eyes, are flat. Retinae, in contrast, follow the curvature of the inner eye. DISs take one full picture after another, whereas rods and cones transmit information continuously and asynchronously. DISs have uniform resolution while the retina is foveated (Weber and Triesch 2009b). Cells in the retina respond to complex temporal and spatial properties of stimuli (e.g. Barlow 1953; Barlow and Hill 1963; Enroth-Cugell and Robson 1966; Stone 2012, Chapters 2,3 for a review) and DISs do not. The list goes on for the comparison between retina and DIS, and it continues at every step along the path of visual processing. A neurorobotic experiment testing a model for the low-level responses of some higher-order part of the visual pathway would have to be very carefully designed to ensure that the final input to the model's instantiation within the robot is actually analogous to that in the living organism after all the processing stages. It may be easier to craft synthetic input and argue convincingly for its similarity.

We believe that the two arguments for and against neurorobotic sensory experiments are valid to different degrees in different cases. Too many layers of processing between a sensing device and a computational instantiation of a model of a brain region reduce the plausibility of low-level properties of the actual input being similar to that to the actual brain region. Thus, neurorobotic experiments may not be the right tool to test a model which explains what it explains on the basis of such low-level properties. While, however, minute features of neural activity patterns are unlikely to be reproduced precisely by long cascades of processing, some of the more global features may be preserved.

Take as an example speech processing. There are large and important differences in how sound waves are transformed into electrical signals in the human ear and in a microphone. Also, the pathway between the inner ear and those brain regions involved in semantic processing is long. Modeling that pathway on the level of individual neuron spiking is a difficult feat in itself. However, current neural models of language acquisition and understanding do not claim or rely on fine-grained similarity of their input to the brain regions they model (Heinrich et al. 2013; Hinoshita et al. 2011; Lawrence et al. 2000). Instead, their models' input is a coarse-grained neural representation of sequences of phonemes and graphemes, respectively. The implicit assumption is that the sequences of activations in their models bears analogy to sequences of activations of actual neurons in the auditory pathway in response to verbal utterances. The models' plausibility only hinges on that assumption and is therefore not affected by the undisputed difference between modeled and real neural input on a lower level. If that implicit assumption is correct, then many of the features of actual speech which make speech processing difficult are preserved in the model input—differences between speakers, uncertainty about units of speech being uttered, and ambiguity of natural language to name a few.

The arguments and examples in this section show that evidence from neurorobotic experiments can indeed strengthen a model's claim to explain biological reality. They also show, however, that this is not always the case: whenever the model explains real phenomena in terms of aspects of the input for which it is not clear that they are analogous in experiment and reality, such evidence is not stronger and possibly even weaker than that of pure, well-designed simulations.

A.4 Implications in the Context of This Thesis

The structure of this thesis reflects the two fields to which our research contributes. The two main parts, Part I and Part II, respectively deal with modeling learning of MSI in the dSC and with using the insights we have gained from modeling in a general machine learning and robotics context. Accordingly, the methodology applied in the two chapters is somewhat different:

In Part I, we are interested in how the dSC might learn to represent the location of cross-sensory stimuli in one unified population code. We derive a model of learning in the dSC and, to test that model, we perform simulations which show that, under certain conditions, our model's behavior is analogous in important ways to that of the real dSC. As we have argued in Section A.1, there would be nothing to be gained from testing other conditions. Suppose we did set our parameters differently and observed qualitatively the same behavior. Since there is no way to translate our learning rates, training steps, noise levels and so on quantitatively into actual biological conditions, it would not even be possible to tell whether those different conditions in the simulations correspond to significantly different conditions in reality. Suppose, on the other hand, that we did not observe the expected behavior. Then, we could only conclude that our model predicts different behavior of the dSC under different biological conditions, which is to be expected and has been shown experimentally for special cases (e.g. Bergan et al. 2005; Stein et al. 2014; Wallace and Stein 1994, 2007; Xu et al. 2012).

This is different in Part II. There, we want to evaluate the performance and robustness of our algorithms. We therefore test a range of parameter settings to find approximate optima for the tasks on which we try them and to demonstrate that their performance is stable over a wide range of parameters. In Part II, we also evaluate our algorithms on real data. We do this because testing our algorithms on real data strengthens our claim about their usefulness: binaural SSL in the iCub, for example, is in itself useful and our algorithm can be applied there. Furthermore, if our algorithm performs well in that setting, then it is plausible that it will also perform well in similar settings, like for example binaural SSL in different robots.

Real data for testing our modeling in Part I would be much harder to acquire and, as we have argued in Section A.2, any results would not be much more relevant for the evaluation of our claims than simulated data: even actual animal SCs do not exhibit the same phenomena under natural and altered sensory conditions or with full and only partial input (see above). Therefore, real data for testing our modeling would have to be real activations of neurons projecting to the dSC under natural conditions. At the level we are modeling the dSC, however, even real data would not be useful; Real neural activity is made up of individual spikes, whereas we assume rate-coded input. Real population codes comprise the activity of often millions of neurons whereas we model (and can model) only a few hundred. Real learning takes place over months, or often years, whereas we model it in only a few thousand training steps. Transforming real data into the kind of data we would need as input for our model would decrease its 'reality' to the level of simulated input, or possibly even below that.

The same argument goes for neurorobotic experiments, in principle. A robotic body and preprocessing in a robot's control computer does not normally produce data which is comparable to the neural input received by the dSC. Changing the robotic body and preprocessing such that they do generate biologically plausible data requires transformations so drastic and complex that it is not more plausible that the resultant input should be more realistic than simulated input. And again, if it were, it would not be the right kind of input for our model's instantiation and would have to be further transformed, with the contingent loss of plausibility. However, although there are no strong parallels to be expected between the actual neurophysiology of the dSC and the simulated activations of the neurons in our ANN algorithm in the machine learning experiments of Part II, these experiments demonstrate the effectiveness of the mechanisms we assume at work in the dSC and implement in our algorithm. Thus, these experiments yield evidence for the appropriateness of the extracted mechanisms for tasks which are similar to the dSC's task and therefore make it more plausible that the dSC employs them as well.

B Audio-Visual Integration in Practical Systems and in the dSC

In the introduction to Part II, we argued that our model of audio-visual localization was not likely a good basis for a technical audio-visual localization system, basically because visual localization is too accurate when it succeeds and useless otherwise. This is demonstrated in Figure B.1. We can see there that a standard computer vision (CV) algorithm has no difficulties localizing a face quite precisely and that adverse visual conditions do not affect its ability to *localize* a face as long as it *detects* it. Here, we want to expand on this argument. Specifically, we will treat the two questions of a) why there are practical studies which purportedly perform audio-visual localization and b) why the dSC appears to engage in audio-visual localization, as per our assumption in the introduction of Part I, when we believe that that is not useful in technical applications.

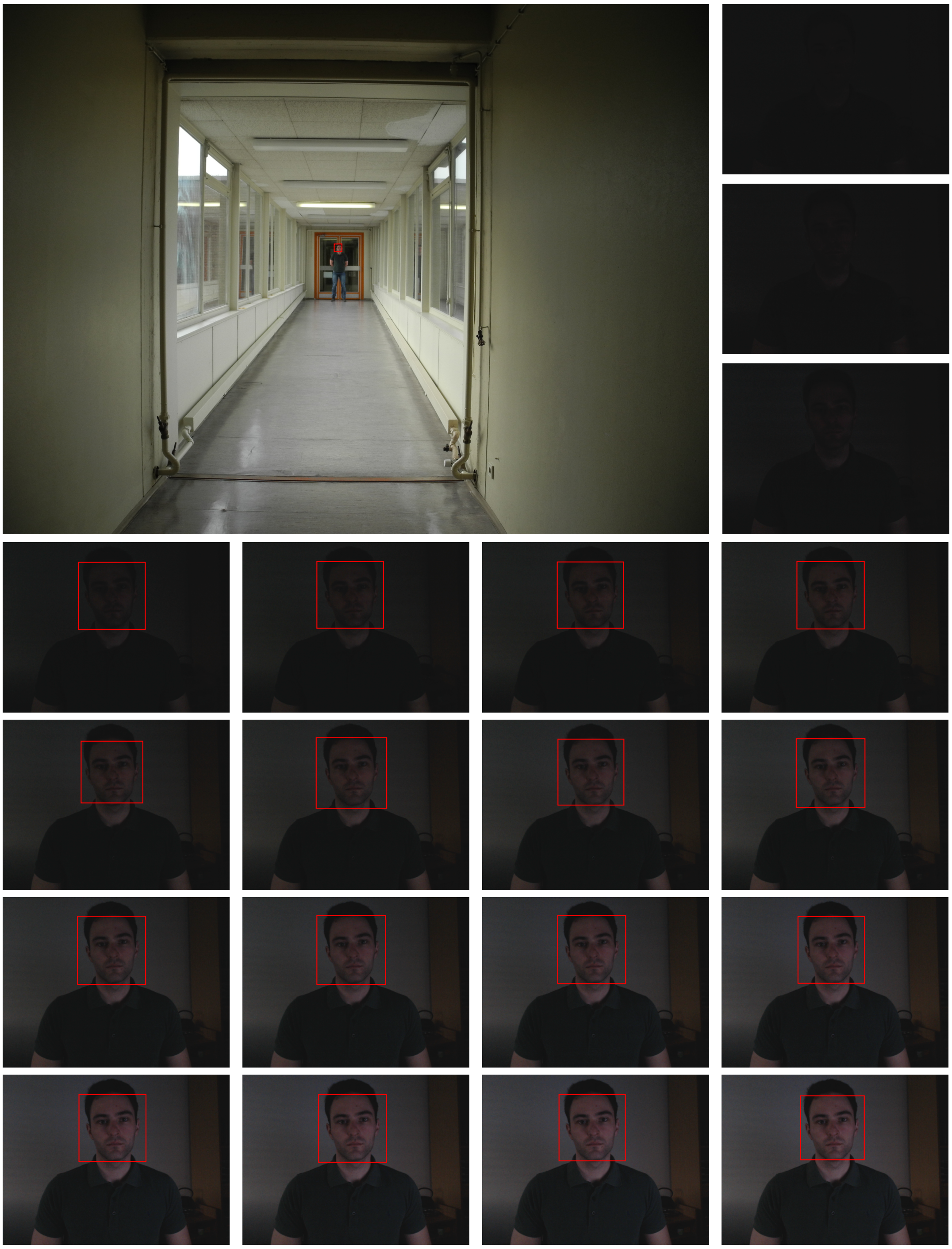
B.1 The Futility of Audio-Visual Integration for Practical Localization

To approach the first question, let us consider the studies of robotic audio-visual localization reviewed in Section 3.3. Revisiting them, one notes that most of these studies either used highly simplified visual stimuli,[1] or the systems being proposed did not perform very well.[2] In fact, our own experiment, described in Section 5.3, suffers from both deficiencies, from an application perspective. This is because it was not the goal of any of these studies to localize objects using vision and hearing. They were undertaken to test models of biological MSI.

However, other, practical studies which are ostensibly concerned with practical audio-visual localization do seem to perform well on realistic input (e.g. Kushal et al. 2006; Li et al. 2012; Sanchez-Riera et al. 2012; Voges et al. 2006; Yan et al. 2013). The key to answering the first question posed above, why these practical approaches seem to perform audio-visual localization, lies in the distinction between localization, detection, and disambiguation. To localize means to reckon the position of an object. To detect means to determine whether or not a target is present. To disambiguate means to decide which of a number of targets is the most relevant. All of these might in principle be done visually, acoustically, or by integrating visual and auditory signals.

[1] Ravulakollu et al. (2009, 2012) and Rucci et al. (1999, 2000)

[2] Casey et al. (2012), 'not very well' meaning: compared to much simpler solutions, see below.



Face detection and localization using Viola and Jones' (2001) algorithm. **top left:** A face is localized within ∼ 0.4° (camera aperture: ∼ 64°). **others:** The same scene, photographed at successively longer exposures, faces detected and marked. It is clear that uncertainty about the *location* of the face is not affected by exposure time in those images in which a face is detected. Only within a narrow range of low exposure times could the face potentially be detected but is not—room for improvement is scarce.

Figure B.1: Traditional Visual Localization.

We suggest that all the practical studies cited above really do not perform audio-visual integration to improve localization but to disambiguate.[3] This is apparent from the lack of any figures on the accuracy of audio-visual localization in these studies: Sanchez-Riera et al. (2012) report on correct detections instead; Li et al. (2012) do not present precise numbers, but state that audio-visual localization was generally as good as (not better than) visual localization; Yan et al. (2013) only use visual information to calibrate audio-motor maps and therefore do not report on the accuracy of visual or audio-visual localization; finally, Voges et al. (2006) and Kushal et al. (2006) simply do not report on accuracy of audio-visual localization at all. This would be peculiar if precise audio-visual localization was really the goal in these studies.

However, precise localization is not their goal: except for Yan et al. (2013), all of the studies cited above deal with the problem of active speaker detection. They all first localize speakers visually and then integrate with auditory information. Integration is either done simplistically, meaning a decision is made whether to use visual or auditory information, or some sort of probabilistic integration is performed. In the latter case, however, the mathematical procedures are only loosely based on a mathematical model of localization, which is presumably appropriate for the task the authors set themselves. Effectively, all of these lead to the same result: if visual detection and disambiguation is possible, then that is done and auditory localization has no or almost no effect. Auditory localization only has an effect in case visual detection fails or when it is needed to select one out of a number of valid visual targets. This fact is stated in so many words by some of the studies.

Thus, our argument stands: visual localization using standard CV methods is too precise to profit meaningfully from true integration with auditory localization. True audio-visual localization is therefore not done for practical purposes. Detection and disambiguation, on the other hand, can profit from auditory as well as visual information, and have been shown to do so in practical studies.

Our algorithm was developed for localization in situations where the position of a stimulus cannot be determined precisely from one modality (from one cue) alone. This is not the case in audio-visual localization where either visual localization dominates or it fails detectably and auditory localization can be used as a substitute. Therefore, we cannot expect it to generate a significant improvement over a simplistic if-then-else implementation of audio-visual localization.

Our Algorithm as an Algorithm for Detection or Disambiguation. The question is: would our algorithm perform detection or disambiguation? Quite simply, the answer is ‘probably.’ Given multidimensional target-related, spatial input with moderate noise, there is a good chance that our algorithm would learn to detect targets. On target-related audio-visual spatial input, one can expect our algorithm to learn to prefer combined audio-visual stimuli over visual-only stimuli of similar saliency, and thus to disambiguate.

The reason we have not tried this is that the results would likely not be impressive. Our algorithm is a Naïve Bayesian algorithm which evaluates all input dimensions in-

[3] With the exception of Yan et al. (2013), which has a different goal altogether.

dependently to determine the probability of the presence of a target. But visual signals contain complex interdependencies between the dimensions, which are handled by decision structures in Viola and Jones (2001)-like algorithms and by learned hierarchies of features in the even more powerful deep learning systems developed more recently (e.g. Krizhevsky et al. 2012).

B.2 Why Audio-Visual Localization in the dSC?

Now, if audio-visual integration does not improve localization, why does the dSC seem to to it? Generally speaking, there are three ways of approaching this question. We can argue that, in integrating visual and auditory localization, the dSC does something that is not practically useful. Obviously, this is not an attractive option. Alternatively, we could try and find differences between the conditions under which the dSC and technical systems operate. Or we might come to the conclusion that our assumptions are wrong and the dSC does not actually localize objects and events. We will examine both of the latter possibilities in turn.

One of the most glaring differences between biological and robotic visual processing is the granularity at which information is transmitted. In biological vision, a new visual target will stimulate light receptors in the retina which will then respond in a receptor-specific way, sending action potentials through retinal ganglion cells to visual brain areas like the superficial superior colliculus (sSC). Thus, initial information about the target reaches the sSC the moment the first action potentials arrive. As more and more action potentials arrive, information increases until near-certainty is achieved. By integrating visual input with input from other sources, the rate at which information arrives is increased and thus sufficient certainty for precise behavioral responses is reached earlier than with just visual information. Since fast localization of and orientation towards stimuli can be a matter of life and death for an animal in the wild, it is an evolutionary advantage to have the capability of audio-visual integration for orientation.

In contrast, to get visual information, a robotic system instructs its camera to take a photo, encode it, and route it through layers of transmission protocols until it reaches the main memory. For the purposes of a CV algorithm, all of the data is unavailable one moment and available in its entirety the next. There simply is no accumulation of certainty about the location of an object in this mode of operation, and integration of visual with auditory information can therefore not speed up the process.

It might be possible to construct a system in which visual and auditory information are processed as they come in. Such a system would have to be built from the ground up because available cameras are just not designed to transmit anything but a full picture, and the propagation of both visual and auditory information through a robotic system takes too long to make an arbitrarily small packaging feasible in any case. If such a system were to be built, however, then a dynamic version of our algorithm could be a candidate for implementing audio-visual localization. Our aim in Chapter 10 was to demonstrate the actual usefulness of our algorithm, and we therefore chose a more immediately feasible application.

Above, we wrote that the other viable way to attack the question of why the dSC seems to perform audio-visual localization is to deny that it performs localization. Most models of the dSC which state any function for it, including ours, seem to assume that the dSC does localize, but they might of course be wrong. One might argue in that direction, seeing as, in order to localize, there needs to be a target. At any given moment, there are hundreds or thousands of potential targets for localization around us. Yet, the dSC's activity most saliently leads to saccades to only one of them at a time. It might therefore be said that it really performs action selection, not localization.

However, a case may be made against action selection as well. When dSC activity affects spatial attention, there is no action. Also, to our knowledge, it has not been shown that dSC activity cannot at any time enhance visual processing in more than one location, so it is not clear that the dSC selects, either. In a similar vein, it does not really detect, because it lacks the complex processing hierarchy of the visual cortex to detect a familiar face or a sought-after object, for example. And it does not just use multiple sensory modalities for backup, in case vision fails, or for disambiguation, when multiple targets seem equally relevant. Otherwise there should be no ventriloquism effect at all, but only visual capture.[4]

Probably the only true thing that can be said about the function of the dSC is that it does what it does, and that is what it has evolved to do, which was whatever kept its possessor alive. However, that is not a very useful thing to say and a simplified account—or a multitude of such accounts—yields a much better handle for studying and understanding the dSC.[5]

B.3 Audio-Visual Localization in our Work

The overarching strategy of this thesis was to first model the dSC and then to extract mechanisms that can be used in practical applications, in that order. It follows that relevance in technical contexts was not our focus when we chose a function of the dSC to model, in the beginning of Part I. We might have looked out for a problem which both the dSC and a technical system could solve in a similar way. In that case, we would probably not have modeled audio-visual localization, and if we had, we would have prioritized addressing accumulation of evidence. However, as we argued in Section 1.2.2, computer science is not so much interested in solving concrete problems in the first instance, but in developing re-usable strategies for solving concrete problems. Although possibly somewhat surprising, our abstracting from the dSC's problem and applying the insight we have gained from modeling it to another problem is therefore perfectly in line with our goals and methodology.

[4] See Section 2.1.

[5] See Section 7.1.2.

C Publications Arising from this Thesis

Bauer, Johannes and Wermter, Stefan (2013a). “Learning Multi-Sensory Integration with Self-Organization and Statistics”. In: *Ninth International Workshop on Neural-Symbolic Learning and Reasoning (NeSy’13)*. Beijing, China, pp. 7–12.

– (2013b). “Self-Organized Neural Learning of Statistical Inference from High-Dimensional Data”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Beijing, China, pp. 1226–1232.

Bauer, Johannes, Weber, Cornelius, and Wermter, Stefan (2012a). “A SOM-based Model for Multi-sensory Integration in the Superior Colliculus”. In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*. Brisbane, Australia: IEEE, pp. 1–8.

Bauer, Johannes, Dávila-Chacón, Jorge, Strahl, Erik, and Wermter, Stefan (2012b). “Smoke and Mirrors — Virtual Realities for Sensor Fusion Experiments in Biomimetic Robotics”. In: *2012 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. Hamburg, Germany: IEEE, pp. 114–119.

Bauer, Johannes, Dávila-Chacón, Jorge, and Wermter, Stefan (2014). “Modeling development of natural multi-sensory integration using neural self-organisation and probabilistic population codes”. In: *Connection Science* 27.2, pp. 1–19.

Bauer, Johannes, Magg, Sven, and Wermter, Stefan (2015). “Attention modeled as information in learning multisensory integration”. In: *Neural Networks* 65, pp. 44–52.

D Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, den

Johannes Bauer

List of Abbreviations

AES anterior ectosylvian cortex. 22, 25, 26, 76

AI artificial intelligence. 8, 9, 33, 133, 139

AIC Akaike's information criterion. 82

ANN artificial neural network. 16, 34, 37, 40, 41, 44, 45, 47, 53, 55, 57, 73, 91, 97, 119, 130, 133, 139, 140, 142, 149, 150, 154

BMU best-matching unit. 41, 42, 53, 55, 60, 69, 77, 98, 101, 104, 122, 126

CV computer vision. 135, 155, 157, 158

DIS digital image sensor. 152

DLPFC dorsolateral prefrontal cortex. 25, 26, 76

dSC deep superior colliculus. 5, 6, 8, 9, 11, 14–17, 20–27, 29, 33–35, 37–41, 43–45, 47–50, 52, 55–57, 66, 71–73, 85, 87, 88, 90, 91, 95, 97, 104, 113, 133–135, 139–144, 149, 153–155, 158, 159

FA factor analysis. 101, 102, 105

FEF frontal eye field. 20, 25, 26, 76

fMRI functional magnetic resonance imaging. 25

GTM generative topographic mapping. 102, 103, 105

HMI human-machine interaction. 115

HRIR head-related impulse response. 117

HRTF head-related transfer function. 117, 118

IC inferior colliculus. 19, 26, 66, 68, 118, 142

ICc central nucleus of the inferior colliculus. 39, 116, 118

ICx external nucleus of the inferior colliculus. 26, 39, 116, 142

ILD interaural level difference. 43, 115, 116, 118, 121, 130

iSC intermediate superior colliculus. 20

ITD interaural time difference. 43, 115–118, 121, 130

LGN lateral geniculate nucleus. 26

LIP lateral intraparietal cortex. 25, 26, 76

LS lateral suprasylvian cortex. 22

LSO lateral superior olive. 116, 118, 121, 130

LVM latent variable model. 15, 97–99, 101, 102, 105, 132, 133, 140, 144, 145, 170

MLE maximum likelihood estimator. 13, 15, 36, 37, 62, 64, 69, 71, 87, 89

MLP multilayer perceptron. 143

MSE mean squared error. 62, 69

MSI multisensory integration. 8, 12, 13, 15–19, 27–29, 33–37, 39–41, 44, 45, 57, 62, 73, 76, 87, 97, 118, 135, 139, 140, 144, 153, 155, 169

MSO medial superior olive. 116, 118, 121, 130

MT middle temporal visual area. 26

OT optic tectum. 6, 19, 27, 39

PCA principal component analysis. 101, 102, 105, 140

PDF probability density function. 15, 40, 49, 52, 53, 71, 73, 101, 104, 142, 143

PMF probability mass function. 101

PP postsynaptic potential. 47

PPC probabilistic population code. 15, 40, 49, 50, 52, 71, 89–91, 139, 140

RF receptive field. 21–25, 27, 38, 47, 73

rLS rostral part of the lateral suprasylvian cortex. 22

SAI stratum album intermediale. 21

SAP stratum album profundum. 21

SC superior colliculus. 5–9, 16–21, 23–29, 33–35, 45, 47, 48, 66, 68, 73, 76, 77, 85, 86, 91, 118, 131, 139, 142, 143, 154, 169

SEF supplementary eye field. 25, 26

SGI stratum griseum intermediale. 21

SGP stratum griseum profundum. 21

SGS stratum griseum superficiale. 21

SNR signal-to-noise ratio. 117

SO stratum opticum. 21

SOC superior olivary complex. 116

SOM self-organizing map. 37, 40–43, 52, 53, 69, 71, 84, 85, 88, 97–105, 107, 108, 110, 112, 132, 140, 142–144, 169, 170

SOTA State of the Art. 97, 115

sSC superficial superior colliculus. 19–24, 26, 27, 39, 59, 66, 158

SSL sound-source localization. 15, 16, 43, 96, 108, 115–119, 121–124, 126, 130, 131, 134, 135, 139, 140, 142, 144, 145, 153, 154, 170, 171, 173

SZ stratum zonale. 21

TDOA time difference of arrival. 115

List of Figures

List of Tables

Bibliography

Abe, Takashi, Kanaya, Shigehiko, Kinouchi, Makoto, Kudo, Yoshihiro, Mori, Hirotada, Matsuda, Hideo, Del Carpio, Carlos, and Ikemura, Toshimichi (1999). "Gene Classification Method Based on Batch-Learning SOM". In: *Genome Informatics* 10, pp. 314–315.

Adams, Samantha V., Wennekers, Thomas, Denham, Sue, and Culverhouse, Phil F. (2013). "Adaptive training of cortical feature maps for a robot sensorimotor controller". In: *Neural Networks: Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society* 44, pp. 6–21.

Akaike, Hirotugu (1974). "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

Alais, David and Burr, David (2004). "The ventriloquist effect results from near-optimal bimodal integration". In: *Current Biology* 14.3, pp. 257–262.

Anastasio, Thomas J. and Patton, Paul E. (2003). "A Two-Stage Unsupervised Learning Algorithm Reproduces Multisensory Enhancement in a Neural Network Model of the Corticotectal System". In: *The Journal of Neuroscience* 23.17, pp. 6713–6727.

Anastasio, Thomas J., Patton, Paul, and Belkacem-Boussaid, Kamel (2000). "Using Bayes' Rule to Model Multisensory Enhancement in the Superior Colliculus". In: *Neural Computation* 12.5, pp. 1165–1187.

Andersen, Søren K., Müller, Matthias M., and Hillyard, Steven A. (2009). "Color-selective attention need not be mediated by spatial attention". In: *Journal of Vision* 9.6. Article 2.[†]

Apter, Julia T. (1945). "Projection of the Retina on Superior Colliculus of Cats". In: *Journal of Neurophysiology* 8.2, pp. 123–134.

Aschersleben, Gisa and Bertelson, Paul (2003). "Temporal ventriloquism: crossmodal interaction on the time dimension: 2. Evidence from sensorimotor synchronization". In: *International Journal of Psychophysiology* 50.1-2, pp. 157–163.

Auer, Peter, Burgsteiner, Harald, and Maass, Wolfgang (2008). "A learning rule for very simple universal approximators consisting of a single layer of perceptrons". In: *Neural Networks: Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society* 21.5, pp. 786–795.

Bajo, Victoria M., Nodal, Fernando R., Bizley, Jennifer K., Moore, David R., and King, Andrew J. (2007). "The Ferret Auditory Cortex: Descending Projections to the Inferior Colliculus". In: *Cerebral Cortex* 17.2, pp. 475–491.

Barber, Michael J., Clark, John W., and Anderson, Charles H. (2003). "Neural Representation of Probabilistic Information". In: *Neural Computation* 15.8, pp. 1843–1864.

[†] Articles independently paginated in this journal.

Barlow, Horace B. (1953). “Summation and inhibition in the frog’s retina”. In: *Journal of Physiology* 119.1, pp. 69–88.

Barlow, Horace B. and Hill, Richard M. (1963). “Evidence for a Physiological Explanation of the Waterfall Phenomenon and Figural After-effects”. In: *Nature* 200.4913, pp. 1345–1347.

Barsalou, Lawrence W., Simmons, W. Kyle, Barbey, Aron K., and Wilson, Christine D. (2003). “Grounding conceptual knowledge in modality-specific systems”. In: *Trends in Cognitive Sciences* 7.2, pp. 84–91.

Battaglia, Peter W., Jacobs, Robert A., and Aslin, Richard N. (2003). “Bayesian integration of visual and auditory signals for spatial localization”. In: *Journal of the Optical Society of America A* 20.7, pp. 1391–1397.

Bauer, Johannes and Wermter, Stefan (2013a). “Learning Multi-Sensory Integration with Self-Organization and Statistics”. In: *Ninth International Workshop on Neural-Symbolic Learning and Reasoning (NeSy’13)*. Beijing, China, pp. 7–12.

– (2013b). “Self-Organized Neural Learning of Statistical Inference from High-Dimensional Data”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Beijing, China, pp. 1226–1232.

Bauer, Johannes, Weber, Cornelius, and Wermter, Stefan (2012a). “A SOM-based Model for Multi-sensory Integration in the Superior Colliculus”. In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*. Brisbane, Australia: IEEE, pp. 1–8.

Bauer, Johannes, Dávila-Chacón, Jorge, Strahl, Erik, and Wermter, Stefan (2012b). “Smoke and Mirrors — Virtual Realities for Sensor Fusion Experiments in Biomimetic Robotics”. In: *2012 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. Hamburg, Germany: IEEE, pp. 114–119.

Bauer, Johannes, Dávila-Chacón, Jorge, and Wermter, Stefan (2014). “Modeling development of natural multi-sensory integration using neural self-organisation and probabilistic population codes”. In: *Connection Science* 27.2, pp. 1–19.

Bauer, Johannes, Magg, Sven, and Wermter, Stefan (2015). “Attention modeled as information in learning multisensory integration”. In: *Neural Networks* 65, pp. 44–52.

Beck, Jeffrey M., Ma, Wei J., Kiani, Roozbeh, Hanks, Tim, Churchland, Anne K., Roitman, Jamie, Shadlen, Michael N., Latham, Peter E., and Pouget, Alexandre (2008). “Probabilistic population codes for Bayesian decision making”. In: *Neuron* 60.6, pp. 1142–1152.

Beck, Jeffrey M., Ma, Wei J., Pitkow, Xaq, Latham, Peter E., and Pouget, Alexandre (2012). “Not noisy, just wrong: the role of suboptimal inference in behavioral variability”. In: *Neuron* 74.1, pp. 30–39.

Beira, Ricardo, Lopes, Manuel, Praça, Miguel, Santos-Victor, José, Bernardino, Alexandre, Metta, Giorgio, Becchi, Francesco, and Saltarén, Roque (2006). “Design of the robot-cub (iCub) head”. In: *IEEE International Conference on Robotics and Automation, 2006. ICRA 2006*. Orlando, FL, USA: IEEE, pp. 94–100.

Benevento, L. A. and Fallon, James H. (1975). “The ascending projections of the superior colliculus in the rhesus monkey (Macaca mulatta)”. In: *Journal of Comparative Neurology* 160.3, pp. 339–361.

Bergan, Joseph F., Ro, Peter, Ro, Daniel, and Knudsen, Eric I. (2005). “Hunting Increases Adaptive Auditory Map Plasticity in Adult Barn Owls”. In: *The Journal of Neuroscience* 25.42, pp. 9816–9820.

Berglund, Erik (2010). “Improved PLSOM algorithm”. In: *Applied Intelligence* 32.1, pp. 122–130.

Berman, Rebecca A. and Wurtz, Robert H. (2010). “Functional Identification of a Pulvinar Path from Superior Colliculus to MT”. In: *The Journal of Neuroscience* 30.18, pp. 6342–6354.

Bertelson, Paul and Aschersleben, Gisa (2003). “Temporal ventriloquism: crossmodal interaction on the time dimension: 1. Evidence from auditory-visual temporal order judgment”. In: *International Journal of Psychophysiology* 50.1-2, pp. 147–155.

Billingsley, Patrick (1995). *Probability and Measure*. Third Edition. New York, NY, USA: John Wiley & Sons.

Bishop, Christopher M., Svensén, Markus, and Williams, Christopher K. I. (1998). “GTM: The Generative Topographic Mapping”. In: *Neural Computation* 10.1, pp. 215–234.

Bon, Leopoldo and Lucchetti, Cristina (1997). “Attention-related neurons in the supplementary eye field of the macaque monkey”. In: *Experimental Brain Research* 113.1, pp. 180–185.

Born, Sabine, Ansorge, Ulrich, and Kerzel, Dirk (2012). “Feature-based effects in the coupling between attention and saccades”. In: *Journal of Vision* 12.11. a27 (w/o page numbers).

Braitenberg, Valentino (1986). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA, USA: MIT Press.

Brandão, Marcus L., Cardoso, Melo, Liana L., Motta, Vitor A., and Coimbra, Norberto C. (1994). “Neural Substrate of Defensive Behavior in the Midbrain Tectum”. In: *Neuroscience and Biobehavioral Reviews* 18.3, pp. 339–346.

Brang, David, Taich, Zack, Hillyard, Steven A., Grabowecky, Marcia, and Ramachandran, Vilayanur S. (2013). “Parietal connectivity mediates multisensory facilitation”. In: *NeuroImage* 78, pp. 396–401.

Brooks, Rodney A. (1992). “Artificial Life and Real Robots”. In: *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*. Cambridge, MA, USA: MIT Press, pp. 3–10.

Bruce, Charles J., Goldberg, Michael E., Bushnell, M. Catherine, and Stanton, Gregory B. (1985). “Primate Frontal Eye Fields. II. Physiological and Anatomical Correlates of Electrically Evoked Eye Movements”. In: *Journal of Neurophysiology* 54.3, pp. 714–734.

Buschman, Timothy J. and Miller, Earl K. (2007). “Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices”. In: *Science* 315.5820, pp. 1860–1862.

Butts, Daniel A. and Goldman, Mark S. (2006). “Tuning Curves, Neuronal Variability, and Sensory Coding”. In: *PLoS Biology* 4.4, e92+.

Carandini, Matteo and Heeger, David J. (2011). “Normalization as a canonical neural computation”. In: *Nature Reviews Neuroscience* 13.1, pp. 51–62.

Casey, Matthew C., Pavlou, Athanasios, and Timotheoue, Anthony (2012). “Audio-visual localization with hierarchical topographic maps: Modeling the superior colliculus”. In: *Neurocomputing* 15.4, pp. 783–810.

Cavanaugh, James, Alvarez, Bryan D., and Wurtz, Robert H. (2006). “Enhanced Performance with Brain Stimulation: Attentional Shift or Visual Cue?” In: *The Journal of Neuroscience* 26.44, pp. 11347–11358.

Chalupa, Leo M. and Rhoades, Robert W. (1977). “Responses of Visual, Somatosensory, and Auditory Neurones in the Golden Hamster’s Superior Colliculus”. In: *The Journal of Physiology* 270.3, pp. 595–626.

Chen, Lihan and Vroomen, Jean (2013). “Intersensory binding across space and time: A tutorial review”. In: *Attention, Perception, & Psychophysics* 75.5, pp. 790–811.

Clark, Andy (1999). “An embodied cognitive science?” In: *Trends in Cognitive Sciences* 3.9, pp. 345–351.

Codd, Edgar F. (1970). “A relational model of data for large shared data banks”. In: *Communications of the ACM* 13.6, pp. 377–387.

Colby, Carol L. and Goldberg, Michael E. (1999). “Space and Attention in Parietal Cortex”. In: *Annual Review of Neuroscience* 22.1, pp. 319–349.

Colonius, Hans and Diederich, Adele (2004). “Why aren’t all deep superior colliculus neurons multisensory? A Bayes’ ratio analysis”. In: *Cognitive, Affective & Behavioral Neuroscience* 4.3, pp. 344–353.

Cuijpers, Raymond H. and Erlhagen, Wolfram (2008). “Implementing Bayes’ Rule with Neural Fields”. In: *Proceedings of the 18th international conference on Artificial Neural Networks, Part II*. ICANN ’08. Prague, Czech Republic: Springer-Verlag, pp. 228–237.

Cuppini, Cristiano, Ursino, Mauro, Magosso, Elisa, Rowland, Benjamin A., and Stein, Barry E. (2010). “An emergent model of multisensory integration in superior colliculus neurons”. In: *Frontiers in integrative neuroscience* 4. Article 6.[†]

Cuppini, Cristiano, Stein, Barry E., Rowland, Benjamin A., Magosso, Elisa, and Ursino, Mauro (2011). “A computational study of multisensory maturation in the superior colliculus (SC)”. In: *Experimental Brain Research* 213.2, pp. 341–349.

Cuppini, Cristiano, Magosso, Elisa, Rowland, Benjamin A., Stein, Barry E., and Ursino, Mauro (2012). “Hebbian mechanisms help explain development of multisensory integration in the superior colliculus: a neural network model”. In: *Biological Cybernetics* 106.11-12, pp. 691–713.

Cynader, Max and Berman, Nancy (1972). “Receptive-field organization of monkey superior colliculus”. In: *Journal of Neurophysiology* 35.2, pp. 187–201.

Dager, Ursula C. and Hubel, David H. (1975). “Physiology of visual cells in mouse superior colliculus and correlation with somatosensory and auditory input”. In: *Nature* 253.5488, pp. 203–204.

Datteri, Edoardo and Tamburrini, Guglielmo (2007). “Biorobotic Experiments for the Discovery of Biological Mechanisms”. In: *Philosophy of Science* 74.3, pp. 409–430.

Dávila-Chacón, Jorge, Heinrich, Stefan, Liu, Jindong, and Wermter, Stefan (2012). “Biomimetic Binaural Sound Source Localisation with Ego-Noise Cancellation”. In:

[†] Articles independently paginated in this journal.

Artificial Neural Networks and Machine Learning – ICANN 2012. Ed. by Alessandro E. P. Villa, Włodzisław Duch, Péter Érdi, Francesco Masulli, and Günther Palm. Vol. 7552. Lecture Notes in Computer Science. Springer-Verlag, pp. 239–246.

Davila-Chacon, Jorge, Magg, Sven, Liu, Jindong, and Wermter, Stefan (2013). "Neural and statistical processing of spatial cues for sound source localisation". In: *The 2013 International Joint Conference on Neural Networks (IJCNN)*. Dallas, TX, USA: IEEE, pp. 1–8.

Day, Brian L. (2014). "Subcortical Visuomotor Control of Human Limb Movement". In: *Advances in Experimental Medicine and Biology* 826, pp. 55–68.

DeBello, William M. and Knudsen, Eric I. (2004). "Multiple Sites of Adaptive Plasticity in the Owl's Auditory Localization Pathway". In: *The Journal of Neuroscience* 24.31, pp. 6853–6861.

Dehner, Lisa R., Keniston, Leslie P., Clemo, Ruth R., and Meredith, Alex A. (2004). "Cross-modal Circuitry between auditory and somatosensory areas of the cat anterior ectosylvian sulcal cortex: A 'New' Inhibitory form of Multisensory Convergence". In: *Cerebral Cortex* 14.4, pp. 387–403.

deLeeuw, Jan (1992). "Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle". In: *Breakthroughs in Statistics*. Ed. by Samuel Kotz and Norman L. Johnson. Springer Series in Statistics. New York, NY, USA: Springer-Verlag, pp. 599–609.

Deneve, Sophie, Latham, Peter E., and Pouget, Alexandre (2001). "Efficient computation and cue integration with noisy population codes". In: *Nature Neuroscience* 4.8, pp. 826–831.

DeSieno, Duane (1988). "Adding a Conscience to Competitive Learning". In: *Neural Networks, 1988., IEEE International Conference on*. San Diego, CA, USA: IEEE, pp. 117–124.

Desimone, Robert and Duncan, John (1995). "Neural Mechanisms of Selective Visual Attention". In: *Annual Review of Neuroscience* 18.1, pp. 193–222.

Deubel, Heiner and Schneider, Werner X. (1996). "Saccade Target Selection and Object Recognition: Evidence for a Common Attentional Mechanism". In: *Vision Research* 36.12, pp. 1827–1837.

Drescher, Uwe, Bonhoeffer, Friedrich, and Müller, Bernhard K. (1997). "The Eph family in retinal axon guidance". In: *Current Opinion in Neurobiology* 7.1, pp. 75–80.

Edwards, Stephen B., Ginsburgh, Charles L., Henkel, Craig K., and Stein, Barry E. (1979). "Sources of Subcortical Projections to the Superior Colliculus in the Cat". In: *The Journal of Comparative Neurology* 184.2, pp. 309–329.

Engel, Andreas K., Maye, Alexander, Kurthen, Martin, and König, Peter (2013). "Where's the action? The pragmatic turn in cognitive science". In: *Trends in Cognitive Sciences* 17.5, pp. 202–209.

Enroth-Cugell, Christina and Robson, John G. (1966). "The Contrast Sensitivity of Retinal Ganglion Cells of the Cat". In: *The Journal of Physiology* 187.3, pp. 517–552.

Ernst, Marc O. and Banks, Martin S. (2002). "Humans integrate visual and haptic information in a statistically optimal fashion". In: *Nature* 415.6870, pp. 429–433.

Fetsch, Christopher R., DeAngelis, Gregory C., and Angelaki, Dora E. (2013). “Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons”. In: *Nature Reviews Neuroscience* 14.6, pp. 429–442.

Foxe, John J. (2012). “The Interface of Multisensory Processing and Selective Attention”. In: *The New Handbook of Multisensory Processing.* Ed. by Barry E. Stein. Cambridge, MA, USA: MIT Press, pp. 337–343.

Fraser, Scott E. (1992). “Patterning of retinotectal connections in the vertebrate visual system”. In: *Current Opinion in Neurobiology* 2.1, pp. 83–87.

Fries, Wolfgang (1984). “Cortical Projections to the Superior Colliculus in the Macaque Monkey: a Retrograde Study Using Horseradish Peroxidase”. In: *The Journal of Comparative Neurology* 230.1, pp. 55–76.

– (1985). “Inputs from Motor and Premotor Cortex to the Superior Colliculus of the Macaque Monkey”. In: *Behavioural Brain Research* 18.2, pp. 95–105.

Fritzke, Bernd (1995). “A Growing Neural Gas Network Learns Topologies”. In: *Advances in Neural Information Processing Systems 7.* Ed. by Gerald Tesauro, David S. Touretzky, and Todd K. Leen. Cambridge, MA, USA: MIT Press, pp. 625–632.

Fukuda, Yutaka and Stone, Jonathan (1974). “Retinal Distribution and Central Projections of Y-, X-, and W-cells of the cat’s retina”. In: *Journal of Neurophysiology* 37.4, pp. 749–772.

Gamma, Erich, Helm, Richard, Johnson, Ralph, and Vlissides, John (1994). *Design Patterns: Elements of Reusable Object-Oriented Software.* 1st ed. Professional computing series. Boston, MA, USA: Addison-Wesley.

Gandhi, Neeraj J. and Katnani, Husam A. (2011). “Motor Functions of the Superior Colliculus”. In: *Annual Review of Neuroscience* 34.1, pp. 205–231.

Garofolo, John S., Lamel, Lori F., Fisher, William M., Fiscus, Jonathan G., Pallett, David S., and Dahlgren, Nancy L. (1993). *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.*

Ghahramani, Zoubin, Wolpert, Daniel M., and Jordan, Michael I. (1997). “Computational Models of Sensorimotor Integration”. In: *Self-organization, Computational Maps, and Motor Control.* Ed. by Pietro G. Morasso and Vittorio Sanguineti. Vol. 119. Advances in Psychology. Elsevier, pp. 117–147.

Ghitani, Nima, Bayguinov, Peter O., Vokoun, Corinne R., McMahon, Shane, Jackson, and Basso, Michele A. (2014). “Excitatory Synaptic Feedback from the Motor Layer to the Sensory Layers of the Superior Colliculus”. In: *The Journal of Neuroscience* 34.20, pp. 6822–6833.

Girard, Benoît and Berthoz, Alain (2005). “From brainstem to cortex: Computational models of saccade generation circuitry”. In: *Progress in Neurobiology* 77.4, pp. 215–251.

Gold, Joshua I. and Shadlen, Michael N. (2001). “Neural computations that underlie decisions about sensory stimuli”. In: *Trends in Cognitive Sciences* 5.1, pp. 10–16.

Goldberg, Michael E. and Wurtz, Robert H. (1972). “Activity of Superior Colliculus in Behaving Monkey. I. Visual Receptive Fields of Single Neurons”. In: *Journal of Neurophysiology* 35.4, pp. 542–559.

Gori, Monica, Del Viva, Michela, Sandini, Giulio, and Burr, David C. (2008). “Young Children Do Not Integrate Visual and Haptic Form Information”. In: *Current Biology* 18.9, pp. 694–698.

Gutfreund, Yoram and King, Andrew J. (2012). “What Is the role of Vision in the Development of the Auditory Space Map?” In: *The New Handbook of Multisensory Processing*. Ed. by Barry E. Stein. Cambridge, MA, USA: MIT Press. Chap. 32, pp. 573–588.

Harting, John K., Huerta, Michael F., Hashikawa, T., and Lieshout, David P. van (1991). “Projection of the Mammalian Superior Colliculus upon the Dorsal Lateral Geniculate Nucleus: Organization of Tectogeniculate Pathways in Nineteen Species”. In: *The Journal of Comparative Neurology* 304.2, pp. 275–306.

Hartline, Peter H., Pandey Vimal, Ram L., King, Andrew J., Kurylo, Daniel D., and Northmore, David P. M. (1995). “Effects of eye position on auditory localization and neural representation of space in superior colliculus of cats”. In: *Experimental Brain Research* 104.3, pp. 402–408.

Hawken, Michael J. and Parker, Andrew J. (1987). “Spatial properties of neurons in the monkey striate cortex”. In: *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)* 231.1263, pp. 251–288.

Heinrich, Stefan, Weber, Cornelius, and Wermter, Stefan (2013). “Embodied Language Understanding with a Multiple Timescale Recurrent Neural Network”. In: *Artificial Neural Networks and Machine Learning – ICANN 2013*. Ed. by Valeri Mladenov, Petia Koprinkova-Hristova, Günther Palm, Alessandro E. P. Villa, Bruno Appollini, and Nikola Kasabov. Vol. 8131. Lecture Notes in Computer Science. Springer-Verlag, pp. 216–223.

Helmchen, Christoph and Büttner, Ulrich (1995). “Saccade-related Purkinje cell activity in the oculomotor vermis during spontaneous eye movements in light and darkness”. In: *Experimental Brain Research* 103.2, pp. 198–208.

Hillis, James M., Watt, Simon J., Landy, Michael S., and Banks, Martin S. (2004). “Slant from texture and disparity cues: Optimal cue combination”. In: *Journal of Vision* 4.12, pp. 967–992.

Hinoshita, Wataru, Arie, Hiroaki, Tani, Jun, Okuno, Hiroshi G., and Ogata, Tetsuya (2011). “Emergence of hierarchical structure mirroring linguistic composition in a recurrent neural network”. In: *Neural Networks: Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society* 24.4, pp. 311–320.

Horn, Gabriel and Hill, Richard M. (1966). “Responsiveness to Sensory Stimulation of Units in the Superior Colliculus and Subjacent Tectotegmental Regions of the Rabbit”. In: *Experimental Neurology* 14.2, pp. 199–223.

Hornik, Kurt, Stinchcombe, Maxwell, and White, Halbert (1989). “Multilayer Feedforward Networks are Universal Approximators”. In: *Neural Networks: Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society* 2.5, pp. 359–366.

Huerta, Michael F. and Kaas, Jon H. (1990). “Supplementary Eye Field as Defined by Intracortical Microstimulation: Connections in Macaques”. In: *The Journal of Comparative Neurology* 293.2, pp. 299–330.

Humphreys, Paul (2007). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method.* Oxford, UK: Oxford University Press.

Hyde, Peter S. and Knudsen, Eric I. (2000). “Topographic Projection from the Optic Tectum to the Auditory Space Map in the Inferior Colliculus of the Barn Owl”. In: *The Journal of Comparative Neurology* 421.2, pp. 146–160.

Ignashchenkova, Alla, Dicke, Peter W., Haarmeier, Thomas, and Thier, Peter (2004). “Neuron-specific contribution of the superior colliculus to overt and covert shifts of attention”. In: *Nature neuroscience* 7.1, pp. 56–64.

Jack, Charles E. and Thurlow, Willard R. (1973). “Effects of Degree of Visual Association and Angle of Displacement on the “Ventriloquism” Effect”. In: *Perceptual and Motor Skills* 37.3, pp. 967–979.

Jazayeri, Mehrdad and Movshon, Anthony A. (2006). “Optimal representation of sensory information by neural populations”. In: *Nature Neuroscience* 9.5, pp. 690–696.

Jeffress, Lloyd A. (1948). “A Place Theory of Sound Localization”. In: *Journal of Comparative and Physiological Psychology* 41.1, pp. 35–39.

Johnson, Mark H. (2005). “Subcortical Face Processing”. In: *Nature Reviews Neuroscience* 6.10, pp. 766–774.

Jones, Judson P. and Palmer, Larry A. (1987). “An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex”. In: *Journal of Neurophysiology* 58.6, pp. 1233–1258.

Jones, Matt and Love, Bradley C. (2011). “Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition”. In: *The Behavioral and brain sciences* 34.4, pp. 169–188.

Kaas, Jon H. (1997). “Topographic Maps are Fundamental to Sensory Processing”. In: *Brain Research Bulletin* 44.2, pp. 107–112.

Kadunce, Daniel C., Vaughan, J. William, Wallace, Mark T., Benedek, Gyorgy, and Stein, Barry E. (1997). “Mechanisms of Within- and Cross-Modality Suppression in the Superior Colliculus”. In: *Journal of Neurophysiology* 78.6, pp. 2834–2847.

Kangas, Jari A., Kohonen, Teuvo K., and Laaksonen, Jorma T. (1990). “Variants of self-organizing maps”. In: *IEEE Transactions on Neural Networks* 1.1, pp. 93–99.

Kao, Chang-Qing, McHaffie, John G., Meredith, M. Alex, and Stein, Barry E. (1994). “Functional development of a central visual map in cat”. In: *Journal of Neurophysiology* 72.1, pp. 266–272.

Kastner, Sabine and Ungerleider, Leslie G. (2000). “Mechanisms of visual attention in the human cortex”. In: *Annual Review of Neuroscience* 23.1, pp. 315–341.

King, Andrew J. (2013). “Multisensory Circuits”. In: *Neural Circuit Development and Function in the Brain.* Oxford, UK: Academic Press, pp. 61–73.

Klier, Eliana M., Wang, Hongying, and Crawford, J. Douglas (2001). “The superior colliculus encodes gaze commands in retinal coordinates”. In: *Nature Neuroscience* 4.6, pp. 627–632.

Knapp, Charles H. and Carter, G. Clifford (1976). “The Generalized Correlation Method for Estimation of Time Delay”. In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 24.4, pp. 320–327.

Knill, David C. and Pouget, Alexandre (2004). “The Bayesian brain: the role of uncertainty in neural coding and computation”. In: *Trends in neurosciences* 27.12, pp. 712–719.

Knudsen, Eric I. (1982). “Auditory and Visual Maps of Space in the Optic Tectum of the Owl”. In: *The Journal of Neuroscience* 2.9, pp. 1177–1194.

– (1983). “Early Auditory Experience Aligns the Auditory Map of Space in the Optic Tectum of the Barn Owl”. In: *Science* 222.4626, pp. 939–942.

– (1988). “Early blindness results in a degraded auditory map of space in the optic tectum of the barn owl”. In: *Proceedings of the National Academy of Sciences of the United States of America* 85.16, pp. 6211–6214.

Knudsen, Eric I. and Brainard, Michael S. (1995). “Creating a Unified Representation of Visual and Auditory Space in the Brain”. In: *Annual Review of Neuroscience* 18.1, pp. 19–43.

Knudsen, Eric I. and Knudsen, Phyllis F. (1983). “Space-Mapped auditory projections from the inferior colliculus to the optic tectum in the barn owl (Tyto alba)”. In: *Journal of Comparative Neurology* 218.2, pp. 187–196.

Knudsen, Eric I. and Konishi, Masakazu (1978a). “A Neural Map of Auditory Space in the Owl”. In: *Science* 200.4343, pp. 795–797.

Knudsen, Erik I. and Konishi, Masakazu (1978b). “Space and frequency are represented separately in auditory midbrain of the owl”. In: *Journal of Neurophysiology* 41.4, pp. 870–884.

Kohonen, Teuvo K. (1995). *Self-Organizing Maps.* Ed. by Thomas S. Huang, Teuvo Kohonen, and Manfred R. Schroeder. Springer series in information sciences. Berlin, Germany: Springer-Verlag.

– (2001). *Self-Organizing Maps.* Springer Series in Information Sciences. Berlin, Germany: Springer-Verlag.

– (2013). “Essentials of the Self-Organizing Map”. In: *Neural Networks: Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society* 37, pp. 52–65.

Körding, Konrad P. (2007). “Decision Theory: What “Should” the Nervous System Do?” In: *Science* 318.5850, pp. 606–610.

Körding, Konrad P. and Wolpert, Daniel M. (2004). “Bayesian integration in sensorimotor learning”. In: *Nature* 427.6971, pp. 244–247.

Körding, Konrad P., Beierholm, Ulrik, Ma, Wei Ji J., Quartz, Steven, Tenenbaum, Joshua B., and Shams, Ladan (2007). “Causal inference in multisensory perception”. In: *PloS one* 2.9, e943+.

Koulakov, Alexei A. and Chklovskii, Dmitri B. (2001). “Orientation Preference Patterns in Mammalian Visual Cortex”. In: *Neuron* 29.2, pp. 519–527.

Krasne, Franklin B., Fanselow, Michael S., and Zelikowsky, Moriel (2011). “Design of a Neurally Plausible Model of Fear Learning”. In: *Frontiers in Behavioural Neuroscience* 5. Article 41

Krauzlis, Richard J., Lovejoy, Lee P., and Zénon, Alexandre (2013). "Superior Colliculus and Visual Spatial Attention". In: *Annual Review of Neuroscience* 36.1, pp. 165–182.

Krauzlis, Richard J., Bollimunta, Anil, Arcizet, Fabrice, and Wang, Lupeng (2014). "Attention as an effect not a cause". In: *Trends in Cognitive Sciences* 18.9, pp. 457–464.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Lake Tahoe, NV, USA: Curran Associates, pp. 1097–1105.

Krueger, Juliane, Royal, David W., Fister, Matthew C., and Wallace, Mark T. (2009). "Spatial receptive field organization of multisensory neurons and its impact on multisensory interactions". In: *Hearing Research* 258.1-2, pp. 47–54.

Krüger, Norbert, Janssen, Peter, Kalkan, Sinan, Lappe, Markus, Leonardis, Aleš, Piater, Justus, Rodriguez-Sánchez, Antonio J., and Wiskott, Laurenz (2012). "Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1847–1871.

Kruger, Uwe, Zhang, Junping, and Xie, Lei (2008). "Developments and Applications of Nonlinear Principal Component Analysis – a Review". In: *Principal Manifolds for Data Visualization and Dimension Reduction*. Ed. by Alexander N. Gorban, Balázs Kégl, Donald C. Wunsch, and Andrei Y. Zinovyev. Vol. 58. Lecture Notes in Computational Science and Enginee. Springer-Verlag, pp. 1–43.

Kuhn, Thomas S. (2012). *The Structure of Scientific Revolutions: 50th Anniversary Edition*. Fourth Edition. Chicago, IL, USA: The University of Chicago Press.

Kushal, Akash, Rahurkar, Mandar, Li, Fei-Fei, Ponce, Jean, and Huang, Thomas (2006). "Audio-Visual Speaker Localization Using Graphical Models". In: *The 18th International Conference on Pattern Recognition*. Vol. 1. ICPR '06. Hong Kong: IEEE, pp. 291–294.

Kustov, Alexander A. and Robinson, David L. (1996). "Shared neural control of attentional shifts and eye movements". In: *Nature* 384.6604, pp. 74–77.

Landy, Michael S., Banks, Martin S., and Knill, David C. (2011). "Ideal-Observer Models of Cue Integration". In: *Sensory Cue Integration*. Ed. by Julia Trommershäuser, Konrad Körding, and Michael S. Landy. Oxford, UK: Oxford University Press, pp. 251–262.

Law, Margaret I. and Constantine-Paton, Martha (1981). "Anatomy and Physiology of Experimentally Produced Striped Tecta". In: *Journal of Neuroscience* 1.7, pp. 741–759.

Lawrence, Steve, Giles, C. Lee, and Fong, Sandiway (2000). "Natural Language Grammatical Inference with Recurrent Neural Networks". In: *IEEE Transactions on Knowlege and Data Engineering* 12.1, pp. 126–140.

Lee, Psyche H., Sooksawate, Thongchai, Yanagawa, Yuchio, Isa, Kaoru, Isa, Tadashi, and Hall, William C. (2007). "Identity of a pathway for saccadic suppression". In: *Proceedings of the National Academy of Sciences* 104.16, pp. 6824–6827.

Li, Zhao, Herfet, Thorsten, and Thormählen, Thorsten (2012). "Multiple Active Speaker Localization based on Audio-visual Fusion in two Stages". In: *2012 IEEE International*

Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). Hamburg, Germany, pp. 262–268.

Linzenbold, Walter and Himmelbach, Marc (2012). “Signals from the Deep: Reach-Related Activity in the Human Superior Colliculus”. In: *The Journal of Neuroscience* 32.40, pp. 13881–13888.

Liu, Jindong, Perez-Gonzalez, David, Rees, Adrian, Erwin, Harry, and Wermter, Stefan (2010). “A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation”. In: *Neurocomputing* 74.1-3, pp. 129–139.

Ma, Wei J., Beck, Jeffrey M., Latham, Peter E., and Pouget, Alexandre (2006). “Bayesian inference with probabilistic population codes”. In: *Nature Neuroscience* 9.11, pp. 1432–1438.

MacDonald, Justin A. (2005). “An Algorithm for the Accurate Localization of Sounds”. In: *New Directions for Improving Audio Effectiveness.* Neuilly-sur-Seine, France, pp. 28–1–28–10.

– (2008). “A localization algorithm based on head-related transfer functions”. In: *The Journal of the Acoustical Society of America* 123.6, pp. 4290–4296.

Machamer, Peter, Darden, Lindley, and Craver, Carl F. (2000). “Thinking about Mechanisms”. In: *Philosophy of Science* 67.1, pp. 1–25.

Maior, Rafael S., Hori, Etsuro, Uribe, Carlos E., Saletti, Patricia G., Ono, Taketoshi, Nishijo, Hisao, and Tomaz, Carlos (2012). “A role for the superior colliculus in the modulation of threat responsiveness in primates: toward the ontogenesis of the social brain”. In: *Reviews in the Neurosciences* 23.5-6, pp. 697–706.

Malsburg, Christoff von der (1973). “Self-organization of orientation sensitive cells in the striate cortex”. In: *Kybernetik* 14.2, pp. 85–100.

Marr, David (1983). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco, CA, USA: W.H. Freeman Publishers.

Martin, Jacob G., Meredith Alex, A., and Ahmad, Khurshid (2009). “Modeling multisensory enhancement with self-organizing maps”. In: *Frontiers in Computational Neuroscience* 3. Article 8.[†]

Maunsell, John H. and Treue, Stefan (2006). “Feature-based attention in visual cortex”. In: *Trends in Neurosciences* 29.6, pp. 317–322.

May, Paul J. (2006). “The mammalian superior colliculus: laminar structure and connections”. In: *Progress in Brain Research* 151, pp. 321–378.

McGurk, Harry and MacDonald, John (1976). “Hearing lips and seeing voices”. In: *Nature* 264.5588, pp. 746–748.

Meredith, Alex M. and Stein, Barry E. (1986a). “Spatial factors determine the activity of multisensory neurons in cat superior colliculus”. In: *Brain Research* 365.2, pp. 350–354.

Meredith, M. Alex and Stein, Barry E. (1986b). “Visual, Auditory, and Somatosensory Convergence on Cells in Superior Colliculus Results in Multisensory Integration”. In: *Journal of Neurophysiology* 56.3, pp. 640–662.

[†] Articles independently paginated in this journal.

Merker, Björn (2007). “Consciousness without a cerebral cortex: A challenge for neuroscience and medicine”. In: *The Behavioral and Brain Sciences* 30.1, pp. 63–81.

Middlebrooks, John C. (2015). “Sound localization”. In: *The Human Auditory System—Fundamental Organization and Clinical Disorders*. Ed. by Michael J. Aminoff, François Boller, and Dick F. Swaab. Vol. 129. Handbook of Clinical Neurology. Amsterdam, The Netherlands: Elsevier, pp. 99–116.

Middlebrooks, John C. and Green, David M. (1991). “Sound Localization by Human Listeners”. In: *Annual Review of Psychology* 42.1, pp. 135–159.

Middlebrooks, John C. and Knudsen, Eric I. (1984). “A neural code for auditory space in the cat’s superior colliculus”. In: *The Journal of Neuroscience* 4.10, pp. 2621–2634.

Miikkulainen, Risto, Bednar, James A., Choe, Yoonsuck, and Sirosh, Joseph (2005). *Computational Maps in the Visual Cortex*. 2005th ed. New York, USA: Springer-Verlag.

Mühling, Markus, Ewerth, Ralph, Zhou, Jun, and Freisleben, Bernd (2012). “Multimodal Video Concept Detection via Bag of Auditory Words and Multiple Kernel Learning”. In: *Advances in Multimedia Modeling*. Ed. by Klaus Schoeffmann, Bernard Merialdo, AlexanderG Hauptmann, Chong-Wah Ngo, Yiannis Andreopoulos, and Christian Breiteneder. Vol. 7131. Lecture Notes in Computer Science. Springer-Verlag, pp. 40–50.

Müller, James R., Philiastides, Marios G., and Newsome, William T. (2005). “Microstimulation of the superior colliculus focuses attention without moving the eyes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.3, pp. 524–529.

Nakano, Tamami, Higashida, Noriko, and Kitazawa, Shigeru (2013). “Facilitation of face recognition through the retino-tectal pathway”. In: *Neuropsychologia* 51.10, pp. 2043–2049.

Neil, Patricia A., Chee-Ruiter, Christine, Scheier, Christian, Lewkowicz, David J., and Shimojo, Shinsuke (2006). “Development of multisensory spatial integration and perception in humans”. In: *Developmental Science* 9.5, pp. 454–464.

Nummela, Samuel U. and Krauzlis, Richard J. (2010). “Inactivation of Primate Superior Colliculus Biases Target Choice for Smooth Pursuit, Saccades, and Button Press Responses”. In: *Journal of Neurophysiology* 104.3, pp. 1538–1548.

Ohshiro, Tomokazu, Angelaki, Dora E., and DeAngelis, Gregory C. (2011). “A normalization model of multisensory integration”. In: *Nature Neuroscience* 14.6, pp. 775–782.

O’Regan, J. Kevin (1992). “Solving the “Real” Mysteries of Visual Perception: The World as an Outside Memory”. In: *Canadian Journal of Psychology* 46.3, pp. 461–488.

O’Regan, J. Kevin and Noë, Alva (2001). “A sensorimotor account of vision and visual consciousness”. In: *Behavioral and Brain Sciences* 24.05, pp. 939–973.

Panchev, Christo and Wermter, Stefan (2001). “Hebbian Spike-Timing Dependent Self-Organization in Pulsed Neural Networks”. In: *World Conference on Neuroinformatics*. Vienna, Austria. Chap. Hebbian Spike-Timing Dependent Self-Organization in Pulsed Neural Networks, pp. 378–385.

Patton, Paul, Belkacem-Boussaid, Kamel, and Anastasio, Thomas J. (2002). “Multimodality in the superior colliculus: an information theoretic analysis”. In: *Cognitive Brain Research* 14.1, pp. 10–19.

Patton, Paul E. and Anastasio, Thomas J. (2003). “Modeling Cross-Modal Enhancement and Modality-Specific Suppression in Multisensory Neurons”. In: *Neural Computation* 15.4, pp. 783–810.

Pavlou, Athanasios and Casey, Matthew (2010). “Simulating the Effects of Cortical Feedback in the Superior Colliculus with Topographic Maps”. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. Barcelona, Spain: IEEE, pp. 1–8.

Pettigrew, John D. (1986). “Flying Primates? Megabats Have the Advanced Pathway from Eye to Midbrain”. In: *Science* 231.4743, pp. 1304–1306.

Platt, Michael L. and Glimcher, Paul W. (1999). “Neural correlates of decision variables in parietal cortex”. In: *Nature* 400.6741, pp. 233–238.

Pollack, Jay G. and Hickey, Terry L. (1979). “The Distribution of Retino-Collicular Axon Terminals in Rhesus Monkey”. In: *The Journal of Comparative Neurology* 185.4, pp. 587–602.

Popper, Karl R. (2002). *The Logic of Scientific Discovery*. 2nd ed. Routledge Classics. London, UK: Taylor & Francis.

Pouget, Alexandre, Dayan, Peter, and Zemel, Richard S. (2003). “Inference and Computation with Population Codes”. In: *Annual review of neuroscience* 26.1, pp. 381–410.

Ravulakollu, Kiran K., Knowles, Michael, Liu, Jindong, and Wermter, Stefan (2009). “Towards Computational Modelling of Neural Multimodal Integration Based on the Superior Colliculus Concept”. In: *Innovations in Neural Information Paradigms and Applications*. Ed. by Monica Bianchini, Marco Maggini, Franco Scarselli, and Lakhmi Jain. Vol. 247. Studies in Computational Intelligence. Berlin, Heidelberg, Germany: Springer-Verlag. Chap. 11, pp. 269–291.

Ravulakollu, Kiran K., Liu, Jindong, and Burn, Kevin (2012). “Stimuli Localization: An Integration Methodology Inspired by the Superior Colliculus for Audio and Visual Attention”. In: *Procedia Computer Science* 13, pp. 50–61.

Reynolds, Raymond F. and Day, Brian L. (2012). “Direct visuomotor mapping for fast visually-evoked arm movements”. In: *Neuropsychologia* 50.14, pp. 3169–3173.

Ritchie, Larry (1976). “Effects of Cerebellar Lesions on Saccadic Eye Movements”. In: *Journal of Neurophysiology* 39.6, pp. 1246–1256.

Ritter, Helge (1999). “Self-Organizing Maps on non-euclidean Spaces”. In: *Kohonen Maps*. Ed. by Merja Oja and Samuel Kaski. Amsterdam, The Netherlands: Elsevier, pp. 97–108.

Rizzolatti, Giacomo, Riggio, Lucia, Dascola, Isabella, and Umiltá, Carlo (1987). “Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention”. In: *Neuropsychologia* 25.1A, pp. 31–40.

Roach, Neil W., Heron, James, and McGraw, Paul V. (2006). “Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration”. In: *Proceedings of the Royal Society B: Biological Sciences* 273.1598, pp. 2159–2168.

Robinson, David A. (1972). "Eye movements evoked by collicular stimulation in the alert monkey". In: *Vision Research* 12.11, pp. 1795–1808.

Robinson, Farrel R., Straube, Andreas, and Fuchs, Albert F. (1993). "Role of the caudal fastigial nucleus in saccade generation. II. Effects of muscimol inactivation". In: *Journal of Neurophysiology* 70.5, pp. 1741–1758.

Rosenblueth, Arturo and Wiener, Norbert (1945). "The Role of Models in Science". In: *Philosophy of Science* 12.4, pp. 316–321.

Rowland, Benjamin A. and Stein, Barry E. (2013). "A model of the temporal dynamics of multisensory enhancement". In: *Neuroscience & Biobehavioral Reviews* 41.0, pp. 78–84.

Rowland, Benjamin A., Stanford, Terrence R., and Stein, Barry E. (2007). "A model of the neural mechanisms underlying multisensory integration in the superior colliculus". In: *Perception* 36.10, pp. 1431–1443.

Rowland, Benjamin A., Stein, Barry E., and Stanford, Terrence R. (2011). "Computational Models of Multisensory Integration in the Cat Superior Colliculus". In: *Sensory Cue Integration*. Ed. by Julia Trommershäuser, Konrad Körding, and Michael S. Landy. Oxford, UK: Oxford University Press, pp. 333–344.

Rucci, Michele, Tononi, Giulio, and Edelman, Gerald M. (1997). "Registration of Neural Maps through Value-Dependent Learning: Modeling the Alignment of Auditory and Visual Maps in the Barn Owl's Optic Tectum". In: *The Journal of Neuroscience* 17.1, pp. 334–352.

Rucci, Michele, Edelman, Gerald M., and Wray, Jonathan (1999). "Adaptation of orienting behavior: from the barn owl to a robotic system". In: *IEEE Transactions on Robotics and Automation* 15.1, pp. 96–110.

Rucci, Michele, Wray, Jonathan, and Edelman, Gerald M. (2000). "Robust localization of auditory and visual targets in a robotic barn owl". In: *Robotics and Autonomous Systems* 30.1-2, pp. 181–193.

Rucci, Michele, Bullock, Daniel, and Santini, Fabrizio (2007). "Integrating robotics and neuroscience: brains for robots, bodies for brains". In: *Advanced Robotics* 21.10, pp. 1115–1129.

Rumbell, Timothy, Denham, Susan L., and Wennekers, T. (2014). "A Spiking Self-Organizing Map Combining STDP, Oscillations, and Continuous Learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 25.5, pp. 894–907.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536.

Saeb, Sohrab, Weber, Cornelius, and Triesch, Jochen (2009). "Goal-directed learning of features and forward models". In: *Neural Networks: Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society* 22.5–6, pp. 586–592.

– (2011). "Learning the Optimal Control of Coordinated Eye and Head Movements". In: *PLoS Computational Biology* 7.11. Article e1002253+.[†]

[†] Articles independently paginated in this journal.

Sanchez-Riera, Jordi, Alameda-Pineda, Xavier, Wienke, Johannes, Deleforge, Antoine, Arias, Soraya, Cech, Jan, Wrede, Sebastian, and Horaud, Radu (2012). "Online Multimodal Speaker Detection for Humanoid Robots". In: *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids).* Osaka, Japan: IEEE, pp. 126–133.

Sato, Yoshiyuki, Toyoizumi, Taro, and Aihara, Kazuyuki (2007). "Bayesian Inference Explains Perception of Unity and Ventriloquism Aftereffect: Identification of Common Sources of Audiovisual Stimuli". In: *Neural Computation* 19.12, pp. 3335–3355.

Satou, Masahiko, Matsushima, Toshiya, Takeuchi, Hiroaki, and Ueda, Kazuo (1985). "Tongue-muscle-controlling motoneurons in the Japanese toad: topography, morphology and neuronal pathways from the 'snapping-evoking area' in the optic tectum". In: *Journal of Comparative Physiology A* 157.6, pp. 717–737.

Schiller, Peter H., Malpeli, Joe G., and Schein, Stan J. (1979). "Composition of Geniculostriate Input to Superior Colliculus of the Rhesus Monkey". In: *Journal of Neurophysiology* 42.4, pp. 1124–1133.

Schiller, Peter H., True, Sean D., and Conway, Janet L. (1980). "Deficits in Eye Movements Following Frontal Eye-Field and Superior Colliculus Ablations". In: *Journal of Neurophysiology* 44.6, pp. 1175–1189.

Schneider, Keith A. (2011). "Subcortical Mechanisms of Feature-Based Attention". In: *The Journal of Neuroscience* 31.23, pp. 8643–8653.

Schnupp, Jan, Nelken, Israel, and King, Andrew J. (2010). *Auditory Neuroscience: Making Sense of Sound.* 1st ed. Cambridge, MA, USA: MIT Press.

Selemon, Lynn D. and Goldman-Rakic, Patricia S. (1988). "Common cortical and subcortical targets of the dorsolateral prefrontal and posterior parietal cortices in the rhesus monkey: evidence for a distributed neural network subserving spatially guided behavior". In: *The Journal of Neuroscience* 8.11, pp. 4049–4068.

Seung, H. Sebastian and Sompolinsky, Haim (1993). "Simple models for reading neuronal population codes". In: *Proceedings of the National Academy of Sciences* 90.22, pp. 10749–10753.

Shams, Ladan and Beierholm, Ulrik R. (2010). "Causal inference in perception". In: *Trends in Cognitive Sciences* 14.9, pp. 425–432.

Simon, Herbert A. (1992). "What Is an "Explanation" of Behavior?" In: *Psychological Science* 3.3, pp. 150–161.

Skiena, Steven S. (2008). *The Algorithm Design Manual.* 2nd ed. London, UK: Springer-Verlag.

Slaney, Malcolm (1993). *An efficient implementation of the Patterson-Holdsworth auditory filter bank.* Tech. rep. Apple Computer, Perception Group.

Soltani, Alireza and Wang, Xiao-Jing (2010). "Synaptic computation underlying probabilistic inference". In: *Nature Neuroscience* 13.1, pp. 112–119.

Song, Joo-Hyun, Rafal, Robert D., and McPeek, Robert M. (2011). "Deficits in reach target selection during inactivation of the midbrain superior colliculus". In: *Proceedings of the National Academy of Sciences* 108.51, E1433–E1440.

Sparks, David L. (1986). “Translation of Sensory Signals into Commands for Control of Saccadic Eye Movements: Role of Primate Superior Colliculus”. In: *Physiological Reviews* 66.1, pp. 118–171.

– (1988). “Neural Cartography: Sensory and Motor Maps in the Superior Colliculus”. In: *Brain, Behavior and Evolution* 31.1, pp. 49–56.

– (2002). “The brainstem control of saccadic eye movements”. In: *Nat Rev Neurosci* 3.12, pp. 952–964.

Sparks, David L. and Hartwich-Young, Rosi (1989). “The deep layers of the superior colliculus”. In: *Neurobiology of Saccadic Eye Movements.* Ed. by Robert A. Wurtz and Michael E. Goldberg. Vol. 3. Reviews of Oculomotor Research. Amsterdam, The Netherlands: Elsevier. Chap. 5, pp. 213–255.

Spence, Charles (2011). “Crossmodal correspondences: A tutorial review”. In: *Attention, Perception, &amp; Psychophysics* 73.4, pp. 971–995.

Spence, Charles and Driver, Jon (2004). *Crossmodal Space and Crossmodal Attention.* 1st ed. Oxford, UK: Oxford University Press.

Sprague, James M. and Meikle, Thomas H. (1965). “The role of the superior colliculus in visually guided behavior”. In: *Experimental Neurology* 11.1, pp. 115–146.

Spratling, Michael W. (2012). “Predictive coding as a model of the V1 saliency map hypothesis”. In: *Neural Networks: Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society* 26, pp. 7–28.

Stanford, Terrence R., Quessy, Stephan, and Stein, Barry E. (2005). “Evaluating the Operations Underlying Multisensory Integration in the Cat Superior Colliculus”. In: *The Journal of Neuroscience* 25.28, pp. 6499–6508.

Stein, Barry E. (2012a). “Early Experience Affects the Development of Multisensory Integration in Single Neurons of the Superior Colliculus”. In: *The New Handbook of Multisensory Processing.* Ed. by Barry E. Stein. Cambridge, MA, USA: MIT Press. Chap. 33, pp. 589–606.

– ed. (2012b). *The New Handbook of Multisensory Processing.* Cambridge, MA, USA: MIT Press.

Stein, Barry E. and Meredith, M. Alex (1993). *The Merging Of The Senses.* 1st ed. Cognitive neuroscience series. Cambridge, MA, USA: MIT Press.

Stein, Barry E. and Stanford, Terrence R. (2008). “Multisensory integration: current issues from the perspective of the single neuron”. In: *Nature Reviews Neuroscience* 9.5, pp. 255–266.

– (2013). “Development of the Superior Colliculus/Optic Tectum”. In: *Neural Circuit Development and Function in the Brain.* Elsevier, pp. 41–59.

Stein, Barry E., Stanford, Terrence R., and Rowland, Benjamin A. (2014). “Development of multisensory integration from the perspective of the individual neuron”. In: *Nature Reviews Neuroscience* 15.8, pp. 520–535.

Stitt, Iain, Galindo-Leon, Edgar, Pieper, Florian, Engler, Gerhard, and Engel, Andreas K. (2013). “Laminar profile of visual response properties in ferret superior colliculus”. In: *Journal of Neurophysiology* 26, pp. 1333–1345.

Stone, James V. (2012). *Vision and Brain: How We Perceive the World.* 1st ed. Cambridge, MA, USA: MIT Press.

Straube, Andreas, Deubel, Heiner, Ditterich, Jochen, and Eggert, Thomas (2001). "Cerebellar lesions impair rapid saccade amplitude adaptation". In: *Neurology* 57.11, pp. 2105–2108.

Stuphorn, Veit, Bauswein, Erhard, and Hoffmann, Klaus-Peter (2000). "Neurons in the primate superior colliculus coding for arm movements in gaze-related coordinates". In: *Journal of Neurophysiology* 83.3, pp. 1283–1299.

Sun, Ron (2009). "Theoretical Status of Computational Cognitive Modeling". In: *Cognitive Systems Research* 10.2, pp. 124–140.

Suppes, Patrick (1960). "A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences". In: *Synthese.* Vol. 12. 2-3. Kluwer Academic Publishers, pp. 287–301.

Tabareau, Nicolas, Bennequin, Daniel, Berthoz, Alain, Slotine, Jean-Jacques J., and Girard, Benoît (2007). "Geometry of the superior colliculus mapping and efficient oculomotor computation". In: *Biological Cybernetics* 97.4, pp. 279–292.

Talagala, Dumidu S., Zhang, Wen, Abhayapala, Thushara D., and Kamineni, Abhilash (2014). "Binaural sound source localization using the frequency diversity of the head-related transfer function". In: *The Journal of the Acoustical Society of America* 135.3, pp. 1207–1217.

Tarski, Alfred (1953). "A General Method in Proofs of Undecidability". In: *Undecidable Theories.* Ed. by Alfred Tarski, Andrzej Mostowski, and Raphael M. Robinson. Amsterdam, The Netherlands: North-Holland Publishing Company. Chap. 1, pp. 1–30.

Tigges, Johannes (1970). "Retinal Projections to Subcortical Optic Nuclei in Diurnal and Nocturnal Squirrels". In: *Brain, Behavior and Evolution* 3.1, pp. 121–134.

Tolhurst, David J., Movshon, J. Anthony, and Dean, Andrew F. (1983). "The Statistical Reliability of Signals in Single Neurons in Cat and Monkey Visual Cortex". In: *Vision Research* 23.8, pp. 775–785.

Triplett, Jason W., Phan, An, Yamada, Jena, and Feldheim, David A. (2012). "Alignment of Multimodal Sensory Input in the Superior Colliculus through a Gradient-Matching Mechanism". In: *The Journal of Neuroscience* 32.15, pp. 5264–5271.

Trommershäuser, Julia, Körding, Konrad, and Landy, Michael S., eds. (2011). *Sensory Cue Integration.* Oxford, UK: Oxford University Press.

Turing, Alan M. (1937). "On Computable Numbers, with an Application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society* s2-42.1, pp. 230–265.

– (1950). "Computing Machinery and Intelligence". In: *Mind.* New Series 59.236, pp. 433–460.

Ursino, Mauro, Cuppini, Cristiano, Magosso, Elisa, Serino, Andrea, and Pellegrino, Giuseppe (2009). "Multisensory integration in the superior colliculus: a neural network model". In: *Journal of Computational Neuroscience* 26.1, pp. 55–73.

Ursino, Mauro, Cuppini, Cristiano, and Magosso, Elisa (2014). "Neurocomputational approaches to modelling multisensory integration in the brain: A review". In: *Neu-*

ral Networks: Official Journal of the International Neural Network Society, European Neural Network Society & Japanese Neural Network Society 15.4, pp. 783–810.

Vatakis, Argiro and Spence, Charles (2007). “Crossmodal binding: Evaluating the “unity assumption” using audiovisual speech stimuli”. In: *Perception & Psychophysics* 69.5, pp. 744–756.

Vepa, Ranjan (2009). *Biomimetic Robotics: Mechanisms and Control.* 1st ed. Cambridge, UK: Cambridge University Press.

Vincent, Julian F. V. (2009). “Biomimetics – a review”. In: *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of Engineering in Medicine* 223.8, pp. 919–939.

Viola, Paul and Jones, Michael (2001). “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001.* Vol. 1. Kauai, HI, USA: IEEE, pp. 511–518.

Vogels, Rufin, Spileers, Werner, and Orban, Guy A. (1989). “The response variability of striate cortical neurons in the behaving monkey”. In: *Experimental Brain Research* 77.2, pp. 432–436.

Voges, Christoph, Märgner, Volker, and Martin, Rainer (2006). “Algorithms for Audiovisual Speaker Localisation in Reverberant Acoustic Environments”. In: *The 3rd Workshop on Positioning, Navigation and Communication (WPNC’06).* Hannover, Germany: Shaker Verlag, pp. 75–80.

Wallace, Mark T. and Stein, Barry E. (1994). “Cross-Modal Synthesis in the Midbrain Depends on Input from Cortex”. In: *Journal of Neurophysiology* 71.1, pp. 429–432.

– (1996). “Sensory organization of the superior colliculus in cat and monkey”. In: *Extrageniculostriate Mechanisms Underlying Visually-Guided Orientation Behavior*. Ed. by Masao Norita, Takehiko Bando, and Barry E. Stein. Vol. 112. Progress in Brain Research. Elsevier, pp. 301–311.

– (2001). “Sensory and Multisensory Responses in the Newborn Monkey Superior Colliculus”. In: *The Journal of Neuroscience* 21.22, pp. 8886–8894.

– (2007). “Early experience determines how the senses will interact”. In: *Journal of Neurophysiology* 97.1, pp. 921–926.

Wallace, Mark T., Meredith, M. Alex, and Stein, Barry E. (1993). “Converging Influences from Visual, Auditory, and Somatosensory Cortices Onto Output Neurons of the Superior Colliculus”. In: *Journal of Neurophysiology* 69.6, pp. 1797–1809.

Wan, Xinwang and Liang, Juan (2013). “Robust and low complexity localization algorithm based on head-related impulse responses and interaural time difference”. In: *The Journal of the Acoustical Society of America* 133.1, EL40–EL46.

Wang, Lupeng, Sarnaik, Rashmi, Rangarajan, Krsna, Liu, Xiaorong, and Cang, Jianhua (2010). “Visual Receptive Field Properties of Neurons in the Superficial Superior Colliculus of the Mouse”. In: *The Journal of Neuroscience* 30.49, pp. 16573–16584.

Warren, David H., Welch, Robert B., and McCarthy, Timothy J. (1981). “The role of visual-auditory “compellingness” in the ventriloquism effect: Implications for transitivity among the spatial senses”. In: *Perception & Psychophysics* 30.6, pp. 557–564.

Weber, Cornelius and Triesch, Jochen (2009a). "Goal-Directed Feature Learning". In: *2009 International Joint Conference on Neural Networks*. Atlanta, Ga, USA: IEEE, pp. 3319–3326.

– (2009b). "Implementations and Implications of Foveated Vision". In: *Recent Patents on Computer Science* 2.1, pp. 75–85.

Weisswange, Thomas H., Rothkopf, Constantin A., Rodemann, Tobias, and Triesch, Jochen (2011). "Bayesian Cue Integration as a Developmental Outcome of Reward Mediated Learning". In: *PLoS ONE* 6.7, e21575+.

Welch, Robert B. and Warren, David H. (1980). "Immediate perceptual response to intersensory discrepancy". In: *Psychological Bulletin* 88.3, pp. 638–667.

Wermter, Stefan, Palm, Günther, Weber, Cornelius, and Elshaw, Mark (2005). "Towards Biomimetic Neural Learning for Intelligent Robots". In: *Biomimetic Neural Learning for Intelligent Robots*. Ed. by Stefan Wermter, Günther Palm, and Mark Elshaw. Vol. 3575. Lecture Notes in Computer Science. Berlin, Heidelberg, Germany: Springer-Verlag, pp. 1–18.

White, Brian J., Boehnke, Susan E., Marino, Robert A., Itti, Laurent, and Munoz, Douglas P. (2009). "Color-Related Signals in the Primate Superior Colliculus". In: *The Journal of Neuroscience* 29.39, pp. 12159–12166.

Wickelgren, Barbara G. (1971). "Superior Colliculus: Some Receptive Field Properties of Bimodally Responsive Cells". In: *Science* 173.3991, pp. 69–72.

Wilkinson, Lee K., Meredith, M. Alex, and Stein, Barry E. (1996). "The role of anterior ectosylvian cortex in cross-modality orientation and approach behavior". In: *Experimental Brain Research* 112.1, pp. 1–10.

Wilson, Andrew D. and Golonka, Sabrina (2013). "Embodied Cognition is Not What you Think it is". In: *Frontiers in Psychology* 4. Article 58.[†]

Wilson, Stuart P. and Bednar, James A. (2015). *What, if anything, are topological maps for?* in press.

Wise, Steven P. and Jones, Edward G. (1977). "Somatotopic and Columnar Organization in the Corticotectal Projection of the Rat Somatic Sensory Cortex". In: *Brain Research* 133.2, pp. 223–235.

Wozny, David R., Beierholm, Ulrik R., and Shams, Ladan (2010). "Probability Matching as a Computational Strategy Used in Perception". In: *PLoS Computational Biology* 6.8, e1000871+.

Xu, Jinghong, Yu, Liping, Rowland, Benjamin A., Stanford, Terrence R., and Stein, Barry E. (2012). "Incorporating Cross-Modal Statistics in the Development and Maintenance of Multisensory Integration". In: *The Journal of Neuroscience* 32.7, pp. 2287–2298.

Xu-Wilson, Minnan, Chen-Harris, Haiyin, Zee, David S., and Shadmehr, Reza (2009). "Cerebellar Contributions to Adaptive Control of Saccades in Humans". In: *The Journal of Neuroscience* 29.41, pp. 12930–12939.

[†] Articles independently paginated in this journal.

Yan, Rujiao, Rodemann, Tobias, and Wrede, Britta (2013). “Computational Audiovisual Scene Analysis in Online Adaptation of Audio-Motor Maps”. In: *IEEE Transactions on Autonomous Mental Development* 5.4, pp. 273–287.

Yang, Tianming and Shadlen, Michael N. (2007). “Probabilistic reasoning by neurons”. In: *Nature* 447.7148, pp. 1075–1080.

Yates, Graeme K., Johnstone, Biran M., Patuzzi, Robert B., and Robertson, Donald (1992). “Mechanical preprocessing in the mammalian cochlea”. In: *Trends in Neurosciences* 15.2, pp. 57–61.

Yin, Hujun (2007). “Learning Nonlinear Principal Manifolds by Self-Organising Maps”. In: *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer-Verlag. Chap. 3, pp. 68–95.

Yin, Tom C. T. (2002). “Neural Mechanisms of Encoding Binaural Localization Cues in the Auditory Brainstem”. In: *Integrative Functions in the Mammalian Auditory Pathway*. Ed. by Donata Oertel, Richard R. Fay, and Arthur N. Popper. Vol. 15. Springer Handbook of Auditory Research. Springer-Verlag, pp. 99–159.

Zhang, Harry (2005). “Exploring Conditions For The Optimality Of Naïve Bayes”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 19.2, pp. 183–198.

Zhou, Tao, Dudek, Piotr, and Shi, Bertram E. (2011). “Self-Organizing Neural Population Coding for improving robotic visuomotor coordination”. In: *The 2011 International Joint Conference on Neural Networks (IJCNN)*. San Jose, CA, USA: IEEE, pp. 1437–1444.