

# Modellierung primärer multisensorischer Mechanismen der räumlichen Wahrnehmung

## D I S S E R T A T I O N

zur Erlangung des akademischen Grades  
Doktoringenieur (Dr.-Ing.)

vorgelegt der Fakultät für Informatik und Automatisierung  
der Technischen Universität Ilmenau

von Dipl.-Inf. Carsten Schauer

Gutachter: 1. Prof. Dr. Horst Michael Groß  
2. Prof. Dr. Ralf Möller  
3. Prof. Dr. Hanspeter Mallot

Tag der Einreichung: 10.01.2006

Tag der wissenschaftlichen Aussprache: 12.10.2006



# Zusammenfassung

In der vorliegenden Arbeit werden visuelle, auditive und multimodale Formen der räumlichen Wahrnehmung und deren Relevanz für den Entwurf technischer Systeme erörtert. Der dabei vertretene wissenschaftliche Ansatz hat interdisziplinären Charakter und berücksichtigt im Umfeld der Neuroinformatik und Robotik methodische Aspekte der Neurobiologie, Wahrnehmungspsychologie und Informatik gleichermaßen. Im Ergebnis sind einerseits neue und weitergehende Interpretationen der Befunde über die natürliche Wahrnehmung möglich. Andererseits werden Defizite bestehender Simulationsmodelle und technischer Anwendungen benannt und überwunden.

Den Ausgangspunkt der Untersuchungen bildet in Kapitel 1 die Diskussion und kritische Wertung etablierter Aufmerksamkeitsmodelle der Wahrnehmung, in denen frühe multisensorische Hirnfunktionen weitgehend unbeachtet bleiben. Als Grundgedanke der folgenden Untersuchungen wird die These formuliert, dass eine konzeptionelle Trennung zwischen primärer Aufmerksamkeit und höheren kognitiven Leistungen sowohl die Einordnung von sensorischen Merkmalen und neurologischen Mechanismen als auch die Modellierung und Simulation erleichtert. In den Kapiteln 2 und 3 werden zunächst die primären räumlichen Kodierungen der zentralen Hörbahn und des visuellen Systems vorgestellt und die Spezifika von projizierten und berechneten sensorischen Topographien beschrieben. Die anschließende Modellierung von auditorisch-visuellen Integrationsmechanismen in Kapitel 4 dient ausdrücklich nicht der Klassifikation oder dem Tracking von Objekten sondern einer frühen räumlichen Steuerung der Aufmerksamkeit, die im biologischen Vorbild unbewusst und auf subkortikalem Niveau stattfindet. Nach einer Erörterung der wenigen bekannten Modellkonzepte werden zwei eigene multisensorische Simulationssysteme auf Basis künstlicher neuronaler Netze und probabilistischer Methoden entwickelt. Kapitel 5 widmet sich der systematischen experimentellen Untersuchung und Optimierung der Modelle und zeigt, wie unbewusste Wahrnehmungsleistungen und deren Simulation unter Bezugnahme auf qualitative und quantitative Befunde über multisensorische Effekte im Mittelhirn evaluiert werden können. Die Diskussion des Modellverhaltens in realen audio-visuellen Szenarien soll unterstreichen, dass die frühe Steuerung der Aufmerksamkeit noch vor der Objekterkennung einen wichtigen Beitrag zur räumliche Orientierung leistet.

# Abstract

The presented work concerns visual, aural, and multimodal aspects of spatial perception as well as their relevance to the design of artificial systems. The scientific approach chosen here, has an interdisciplinary character combining the perspectives of neurobiology, psychology, and computer science. As a result, new insights and interpretations of neurological findings are achieved and deficits of known models and applications are named and negotiated.

In chapter one, the discussion starts with a review on established models of attention, which largely disregard early neural mechanisms. In the following investigations and experiments, the basic idea can be expressed as a conceptual differentiation between early spatial attention and higher cognitive functions. All neural mechanisms that are modelled within the scope of this work, can be regarded as primary and object-independent sensory processing. In chapter two and three the visual and binaural spatial representations of the brain and the specific concept of the computational topography in the central auditory system are discussed. Given the restriction of early neural processes, the aim of the actual multisensory integration, as it is described in chapter four, is not object classification or tracking but primary spatial attention. Without task- or object-related requirements all specifications of the model are derived from findings about certain multisensory structures of the midbrain. In chapter five emphasis is placed on a novel method of evaluation and parameter optimization based on biologically inspired specifications and real-world experiments. The importance of early perceptual processes to orienting behaviour and the consequences to technical applications are discussed.

# Danksagung

Von meinen ersten studentischen Implementierungsversuchen an einfachen Schallortungsalgorithmen bis zur Fertigstellung dieser thematisch recht weit gefächerten Arbeit war es ein langer, nicht immer geradliniger aber oft sehr interessanter Weg. Die Zahl derer, die mich mit Ermutigungen, konstruktiver Kritik und Freundschaft auf diesem Weg begleitet haben, ist glücklicherweise zu groß für eine vollständige Aufzählung.

Mein betreuender Hochschullehrer, Prof. Dr. Horst Michael Groß, hätte es sich vor sieben Jahren sicher nicht träumen lassen, welche vielen Hirnareale, kortiko-tektales Verknüpfungen und sonstigen neurobiologischen Befunde ich ihm im Laufe der Zeit aufzählen würde. Trotzdem reservierte er bei allem Engagement für die kognitive Robotik eine Nische für meine Thematik am unteren Ende des sensorischen Abstraktionsniveaus. Auch gilt mein Dank allen Mitarbeitern, die zum wissenschaftlichen, technischen und organisatorischen Wohl unseres Neuroinformatik-Fachgebietes beitrugen, meinen Kollegen aus dem DFG Graduiertenkolleg 164 und dem Projekt CarDiKon sowie meinen fleißigen Studenten. In Ilmenau konnte ich vieles lernen, was in den Lehrbüchern oder mit den Suchmaschinen des Internet nicht zu finden ist: die kritische Neugier von Peter Paschke, die konsequente Gewissenhaftigkeit im Forschungs- und Lehrbetrieb von Dr. Klaus Debes oder die Hilfsbereitschaft und das enzyklopädiehafte Wissen meiner ehemaligen Kommilitonen Thomas Vesper und Dr. Volker Stephan sind wertvolles Rüstzeug für eine wissenschaftliche Arbeitsweise.

Nicht unerwähnt bleiben sollen die Mitstreiter, die mir beim letzten Schliff der vorliegenden Dissertationsschrift halfen. Ein großes Dankeschön an Steffen Müller für die vielen Diskussionsrunden zum Bayesfilter, an Dr. Kathrin Baldauf für ihren prüfenden Blick auf die neurologischen Textpassagen und an Dr. Hans Joachim Böhme für den allgemeinen, fachlichen Beistand und die beachtliche Menge an Kommata, die sich segensreich in meinen  $\LaTeX$  Dateien verteilten.

Ein noch größeres Dankeschön an Dani, für ihre Geduld mit mir, für den leckeren Kuchen und für die gestiftete Motivation, die mich gut durch die vergangenen zwei Jahre gebracht haben.



# Inhaltsverzeichnis

<b>1</b>	<b>Gegenstand und Motivation</b>	<b>1</b>
1.1	Sehen, räumliches Hören und Aufgabenstellung . . . . .	1
1.2	Aufmerksamkeit und Objektbegriff . . . . .	4
1.2.1	Hierarchische und parallele Wahrnehmungskonzepte . . . . .	4
1.2.2	Aufmerksamkeitsmodelle . . . . .	10
1.3	Frühe Aufmerksamkeit als Ziel der Modellierung . . . . .	14
1.3.1	Kritik an den Paradigmen der Objekterkennung . . . . .	14
1.3.2	Methodik, Randbedingungen und Lösungsansätze . . . . .	17
1.3.3	Orientierung am biologischen Vorbild . . . . .	21
<b>2</b>	<b>Primäre Mechanismen beim räumlichen Hören</b>	<b>23</b>
2.1	Spezifik auditorischer Rauminformationen . . . . .	23
2.2	Modellentwurf am Vorbild der Hörbahn . . . . .	27
2.2.1	Periphere Schallverarbeitung . . . . .	28
2.2.2	Innenohr und Hörnerv . . . . .	30
2.2.3	Auditorische Pfade in Hirnstamm und Mittelhirn . . . . .	36
2.2.4	Spezifische Kodierungen im Nucleus cochlearis . . . . .	39
2.2.5	Extraktion räumlicher Informationen im Hirnstamm . . . . .	42
2.2.6	Auditorische Karten im Inferior Colliculus . . . . .	52
2.3	Vergleich und Wertung der Modelle . . . . .	63
2.3.1	Optionen der Modellierung . . . . .	63
2.3.2	Experimenteller Ansatz . . . . .	66
2.3.3	Ergebnisse und Wertung . . . . .	69
<b>3</b>	<b>Primäre visuelle Verarbeitung</b>	<b>75</b>
3.1	Auge, Retina und Sehnerv . . . . .	75
3.2	Retinotopie Organisation . . . . .	77
3.3	Subkortikale Aspekte der Aufmerksamkeit . . . . .	81
3.4	Modellierung . . . . .	88

3.4.1	Bewegung als universelles Merkmal . . . . .	88
3.4.2	Abstraktion im Simulationsmodell . . . . .	89
<b>4</b>	<b>Frühe auditorisch–visuelle Integration</b>	<b>95</b>
4.1	Klassifizierung multisensorischer Effekte . . . . .	95
4.2	Multimodale Integration im Superior Colliculus . . . . .	98
4.2.1	Neuroanatomie des Superior Colliculus . . . . .	100
4.2.2	Sensorische und motorische Karten . . . . .	103
4.2.3	Multisensorische Integration auf Ebene des Neurons . . . . .	106
4.3	Modellierung . . . . .	111
4.3.1	Allgemeines Modellkonzept . . . . .	111
4.3.2	Dynamisches Neuronales Feld nach Amari . . . . .	113
4.3.3	Probabilistische Modelle . . . . .	118
4.3.4	Gegenüberstellung der Ansätze . . . . .	130
<b>5</b>	<b>Experimentelle Untersuchungen</b>	<b>135</b>
5.1	Motivation des experimentellen Ansatzes . . . . .	135
5.2	Szenarien, Experimente und Benchmarks . . . . .	138
5.2.1	Datenbank für audio–visuelle Szenarien . . . . .	138
5.2.2	Kriterien zur Evaluierung . . . . .	141
5.2.3	Benchmarkbasierte Optimierung . . . . .	145
5.2.4	Interpretation der Simulationsergebnisse . . . . .	149
5.3	Hardware Realtime–Demonstrator . . . . .	154
<b>6</b>	<b>Diskussion und Ausblick</b>	<b>159</b>
<b>A</b>	<b>Neuronale Modelle</b>	<b>167</b>
<b>B</b>	<b>Probabilistische Modelle</b>	<b>171</b>
B.1	Informationstheoretischer Delay-Schätzer . . . . .	171
B.2	Sensorisches Bayesfilter . . . . .	175
<b>C</b>	<b>Aufnahme-Setup und Datenbank</b>	<b>179</b>
<b>D</b>	<b>Multisensorische Benchmarks</b>	<b>183</b>
	<b>Glossar</b>	<b>189</b>
	<b>Literaturverzeichnis</b>	<b>195</b>



# Kapitel 1

## Gegenstand und Motivation

### 1.1 Sehen, räumliches Hören und Aufgabenstellung

Jede natürliche Wahrnehmung hat multisensorischen Charakter. So vielgestaltig die optischen, akustischen, chemischen und mechanischen Reize der Umwelt sind, so groß ist das Arsenal an Sinnesorganen, die sich im Zuge der Evolution entwickelt haben. Die Wichtung, mit der verschiedene sensorische Modalitäten zur Repräsentation der Umwelt beitragen, kann in verschiedenen Spezies sehr unterschiedlich sein – angepasst an den Lebensraum, aber auch an die Lebensweise. Tag- oder nachtaktive Tiere verfolgen ebenso spezifische Strategien zur Wahrnehmung wie sich die kognitiven Leistungen von Jäger und Beute unterscheiden. Prinzipiell von Vorteil ist es aber, Nutzen aus der Kombination von sich ergänzenden Sinneseindrücken zu ziehen. Im Tierreich ist die Motivation für eine multisensorische Repräsentation der Umwelt allgegenwärtig – ob bei der Nahrungssuche oder im Kontakt mit Artgenossen oder potentiellen Feinden. Die Aufmerksamkeit vieler Beutetiere gilt nicht nur Bewegungen, sondern auch Geräuschen und Gerüchen. Für ihre Jäger kann die Kombination von Sehvermögen, Gehör und Geruch ebenfalls hilfreich sein – allerdings nicht um einen Fluchtreflex zu initiieren, sondern zur gezielten räumlichen Orientierung. Dass den universellen multisensorischen Strategien, mit denen die „Was?“, „Wer?“ und „Wo?“ Probleme der natürlichen Wahrnehmung gelöst werden, zunehmend Bedeutung beigemessen wird, spiegelt sich in multimedialen technischen Anwendungen wie in den Forschungsschwerpunkten der Neurowissenschaften gleichermaßen wider.

Für uns Menschen besonders relevant und nachvollziehbar ist die multisensorische Verknüpfung von Sehen und räumlichem Hören. Intuitiv lassen sich die spezifischen räumlich-zeitlichen Konzepte dieser beiden, uns vertrauten Modalitäten beschreiben. Betrachtet man den zeitlichen Charakter der sensorischen Information, dann können

Geräusche als diskontinuierliche, separierte oder segmentierbare Zeitsignale verstanden werden. Verstummt eine Schallquelle, wird sie für unsere Wahrnehmung „unsichtbar“. Demgegenüber werden optische Reize, Intensitäten, Farben und Konturen, in aller Regel nicht plötzlich verschwinden. Selbst das zeitlich kodierte Merkmal Bewegung hat meist einen sehr viel kontinuierlicheren Charakter als die Erzeugung von Geräuschen. Auch bezüglich des räumlichen Aspektes der Wahrnehmung scheint das visuelle System im Vorteil. Viele räumliche Informationen können bereits monokular abgebildet werden, da die optischen Reize, der topographische Rezeptor (z.B. die Retina) und die visuellen Repräsentationen primär räumlich organisiert sind. Im Gegensatz dazu können Schallrichtungen nicht direkt wahrgenommen werden. Sie müssen vielmehr aus monauralen Signaleigenschaften wie der Klangfarbe oder binauralen Intensitätsunterschieden und Stereolaufzeiten abgeleitet werden. Natürlich erfolgte die aufwendige phylogenetische Entwicklung des auditorischen Systems nicht grundlos – abgesehen davon, dass die Mechanismen des räumlichen Hörens auch bei völliger Dunkelheit funktionieren, existieren eine Reihe weiterer auditorischer Leistungen, die das Sehen ergänzen. Während die visuelle Wahrnehmung gleichzeitig nur einen räumlich mehr oder weniger begrenzten Bereich unserer Umgebung erfasst, liefert das auditorische System eine komplette räumliche Abbildung und schließt das sonst unbemerkte Geschehen hinter oder über uns mit ein.

Eine weitere Besonderheit der auditorischen Wahrnehmung, die im Folgenden wiederholt eine Rolle spielt, ist weniger augenfällig und wurde bislang kaum diskutiert: Eine akustische Signalquelle, die einen lokalisierbaren Reiz bewirkt, ist meist gleichbedeutend mit einem Objekt, das unsere Aufmerksamkeit auf sich zieht. Das bloße Auftreten von Helligkeiten, Farben und Kontrasten kann dagegen für die Aufmerksamkeit eines Individuums vollkommen irrelevant sein. Erst die Kombination von spezifischen Ausprägungen dieser Merkmale führt zur Bildung von visuell interessanten Regionen, von Objekten, dann jedoch nicht mehr auf einem primären, sensorischen, sondern symbolischen Niveau. Einzig die Wahrnehmung von Bewegungen stellt offensichtlich ein allgemeines Kriterium zur unwillkürlichen visuellen Steuerung der Aufmerksamkeit dar. Angesichts dieser Gegenüberstellung erscheint es legitim, beim Sehen und Hören von komplementären Strategien zur räumlichen Wahrnehmung auszugehen und gewisse Vorteile eines auditorisch-visuellen Systems im Vergleich zur Verarbeitung unimodaler Sensorinformationen zu erwarten.

Ein häufig zitiertes Beispiel zur Veranschaulichung der komplexen auditorisch-visuellen Mechanismen ist der Bauchrednereffekt, bei dem ein Sprachsignal nicht seinem eigentlichen Ursprungsort, sondern der Richtung einer Handpuppe zugeord-

net wird, die sich visuell durch Gesten, Mund- und Kopfbewegungen auszeichnet [SM93, CC01, WRH<sup>+</sup>04]. Der Effekt sagt mehr über die Perzeption im Publikum als über die Kunstfertigkeit des Vorführenden aus und wirft eine ganze Reihe von Fragen und Problemen bei der konventionellen Objekterkennung auf. In einem unisensorischen Paradigma würde die genannte Situation zur Ausprägung zweier Objekthypothesen führen – schließlich werden im Ergebnis von Sehen und räumlichem Hören eindeutig unterscheidbare Richtungen erkannt. Um die vermeintlich separaten Objekte sequentiell zu analysieren, müsste nun eine fortlaufende Verlagerung der Aufmerksamkeit zwischen beiden Positionen stattfinden. Dabei tritt ein Widerspruch zwischen räumlicher Trennung und zeitlicher Korrelation der Reize zu Tage. Da wir selbst nicht etwa einen unsichtbaren Sprecher und eine stumme Puppe wahrnehmen, muss das Gehirn auf einem für uns unbewussten Niveau eine Lösung realisieren und die auditorische Repräsentation der Szene an die visuelle anpassen. Wie wäre dieses Problem in einem audio-visuellen Wahrnehmungsmodell zu beherrschen? Wie zuverlässig oder situationsabhängig ist die räumliche Orientierung durch Sehen und Hören und auf welchen Abstraktionsebenen oder anhand welcher sensorischer Informationen und Merkmale ist eine Integration der beiden Sinne hilfreich?

Etablierte Paradigmen der Aufmerksamkeitssteuerung und Objekterkennung werden aus der Sicht der Informatik und Ingenieurwissenschaft noch immer pragmatisch im Kontext isolierter Wahrnehmungsaufgaben interpretiert. Lokalisations- und Klassifikationsprobleme sollen möglichst exakt gelöst werden, und Leistungen des biologischen Vorbilds wie das scharfe foveale Sehen oder eine auf wenige Grad genaue Schallortung sind gleichermaßen Ansporn und Rechtfertigung für hoch spezialisierte Simulationsmodelle und technische Anwendungen. Die verbreitete Konzentration auf eine abstrakte und symbolische Verarbeitungsebene korrespondiert dabei meist pauschal mit primären oder assoziativen kortikalen Prozessen und lässt visuelle, auditorische und insbesondere multimodale Mechanismen auf einem frühen, subkortikalen Niveau außer Acht. Das Ziel dieser Arbeit ist es, Defizite und Risiken, die ein auf den Objektbegriff zentriertes Verständnis der Wahrnehmung birgt, zu benennen und Ansätze zur Modellierung der subsymbolischen und weitgehend unbewussten, sensorischen Verarbeitung vorzustellen.

Die kognitive Illusion des Bauchrednereffekts verdeutlicht, dass Aufmerksamkeit, Lokalisation und Klassifikation bei der natürlichen Wahrnehmung eng zusammenwirken. Eine idealisierte Unterscheidung zwischen frühen Mechanismen der Aufmerksamkeit und symbolischer, objektspezifischer Verarbeitung darf daher nur unter dem Vorbehalt der massiven Parallelität und Rekurrenz neuronaler Prozesse

proklamiert werden. Beim Entwurf der auditorischen, visuellen und multisensorischen Modelle in den Kapiteln 2–4 werden konsequenterweise auch immer die Wechselwirkungen der frühen, sensorischen Mechanismen im Mittelhirn mit der kortikalen Ebene diskutiert. Die Unterscheidung zwischen allgemeinen oder objektspezifischen sensorischen Merkmalen und kognitiven Leistungen soll demnach kein Dogma separater Modellkomponenten darstellen, sondern eine Orientierungshilfe in der Vielzahl neurologischer Befunde und mathematisch–technischer Beschreibungsmittel bieten. Als thematischer roter Faden wird der Verzicht auf objekt– oder aufgabenspezifische Kriterien selbst bei der experimentellen Evaluierung der Simulationsmodelle im Kapitel 5 aufgegriffen – um Probleme zu behandeln, die sich gerade aus dem unbewussten Charakter einer subsymbolischen Verarbeitung ergeben.

## 1.2 Aufmerksamkeit und Objektbegriff

### 1.2.1 Hierarchische und parallele Wahrnehmungskonzepte

Die vorangestellten Überlegungen zum Zusammenwirken der sensorischen Systeme führten schnell zur intuitiven Verwendung der Begriffe Merkmal, Objekt und Aufmerksamkeit. Die Vermutung liegt nahe, dass diese in den Neurowissenschaften – Kognitionspsychologie und Neurobiologie einerseits sowie der Neuroinformatik andererseits – verschiedene Bedeutung haben können. Selbst innerhalb der genannten Disziplinen gibt es leider keine einheitlichen Definitionen, etwa für *die Aufmerksamkeit* oder *das Objekt*. Da die Terminologie im weiteren Verlauf jedoch wichtig zur Einordnung der Modelle und Ansätze ist, sollen Interpretationsmöglichkeiten und das Verständnis dieser zentralen Begriffe zu Beginn diskutiert werden.

Aufmerksamkeit beschreibt zunächst ganz allgemein einen selektiven Charakter der Wahrnehmung. Im einfachsten Fall kann allein die Intensität – ein besonders heller oder lauter Reiz – die Aufmerksamkeit lenken. Die Selektivität kann dabei räumlicher und zeitlicher Natur sein, d.h. die Lokalisiertheit, den Zeitpunkt oder die Veränderung eines Reizes bzw. die Bewegung seiner Ursache betreffen. Viele kognitive Selektionsprozesse basieren aber nicht nur auf dem Pegel der Reize, sondern auch auf den über die Amplitude transportierten sensorischen Merkmalen. Der Begriff der Aufmerksamkeit wird dann benutzt, um das Problem der visuellen Suche anhand von Farbe und Formeigenschaften oder, als auditorisches Pendant, die Fokussierung bestimmter Klänge und Geräusche beim sogenannten Cocktailparty Effekt, zu beschreiben. Of-

fensichtlich werden mit dem Phänomen der Aufmerksamkeit kognitive Leistungen in verschieden abstrakter Weise beschrieben – sowohl auf einer primären sensorischen als auch auf einer symbolischen Ebene. Aufgrund der Komplexität dieser Wahrnehmungsleistungen ist es für das Verständnis und die Modellbildung einfacher, Aufmerksamkeit und Objektbegriff in einem spezifischen Kontext, etwa bezüglich der neuronalen Architekturkonzepte oder kognitiver Orientierungs- oder Erkennungsaufgaben, zu erörtern. Daraus ergeben sich eine Reihe von Konzepten, mit denen jeweils bestimmte Aspekte und Teilprobleme der Wahrnehmung erklärt werden.

**Ein hierarchisches Ebenenkonzept** wird zwar selten explizit genannt, ist aber praktisch allen Wahrnehmungsmodellen inherent. So wie man die Verarbeitung von Sinneseindrücken auf unterschiedlichen neuronalen Niveaus, angefangen von der Rezeptorebene über Hirnstamm und Mittelhirn bis zum Kortex, beschreiben kann, ist auf triviale Weise die Gegenüberstellung von Abstraktionsebenen der sensorischen Information möglich: ausgehend von primären Repräsentationen, daraus abgeleiteten Merkmalen und schließlich der Kombination von Merkmalen zur symbolischen Beschreibung von Objekten. Die überwiegende Mehrzahl aller visuellen Wahrnehmungsmodelle betreffen im Sinne einer Objekterkennung solche Mechanismen, die neurologisch den visuellen Kortexarealen zugeschrieben werden. Eine frühe Ebene der Merkmalsrepräsentation bilden dabei die topographischen Karten für Intensität, Farbe, Bewegung und orientierte Kanten in den Arealen V1 und V2. Auf höherem und abstrakterem Niveau vollzieht sich die vorrangig retinotopie Verarbeitung in einem dorsalen Projektionsweg (Areale V3A, V5(MT), PP) und die nicht topographische Kodierung von translations-, skalierungs- und orientierungsinvarianten Objektprototypen in einem ventralen Pfad (Areale V4, IT). Sichtbare Manifestationen der aufmerksamen Wahrnehmung sind motorische Reaktionen wie die Steuerung der Blickrichtung. Ihre Modellierung erfordert desweiteren, eine subkortikale Ebene einzubeziehen und retinotopie Strukturen des Mittelhirns und deren Interaktionen mit Thalamus und Kortex zu analysieren.

In der zentralen Hörbahn ist die Verbindung zwischen Rezeptor und kortikaler Repräsentation sehr viel indirekter. Viele Merkmale eines Schallereignisses – sowohl räumliche als auch spektrale oder temporale – müssen zunächst aus monauralen oder binauralen Zeitsignalen extrahiert werden. Zu den grundlegenden Ebenen der auditorischen Verarbeitung zählen die Frequenzzzerlegung und Rezeption im Innenohr, die primäre tonotopie Kodierung im Hörnerv, die cochlearen Kerne sowie der superior olivare Komplex im Hirnstamm, die auditorischen Kerne in Mittelhirn und Thalamus und schließlich der auditorische Kortex. Oft wird eine in den cochlearen Kernen begin-

nende Trennung in temporale und intensitätsbezogene Informationen, sowie allgemein eine zum Kortex hin steigende Komplexität der Merkmale beschrieben. Bei der auditorischen Modellbildung spielt die subkortikale Extraktion verschiedenartiger Merkmale offensichtlich eine weitaus wichtigere Rolle als im visuellen System.

Die Fusion verschiedener sensorischer Modalitäten ist auf höherer neuronaler Ebene allgemein anerkannt und kann als Bestandteil der ventralen Repräsentation von Objekten nachgewiesen werden [PFR<sup>+</sup>04]. In Ermangelung multisensorischer Rezeptoren existiert jedoch auch kein multisensorischer Projektionsweg mit einer dem visuellen oder auditorischen System vergleichbaren Vorverarbeitung spezifischer Merkmale. Ohne eine schlüssige Hierarchie von Merkmalskodierung und symbolischer Repräsentation können frühe multisensorische Mechanismen demnach nicht direkt zur Objekterkennung beitragen. Gleichwohl wird ihre Rolle bei der räumlichen Orientierung im Rahmen dieser Arbeit von besonderem Interesse sein.

**Bottom-Up/Top-Down Konzept:** Mit dem Paradigma der Abstraktionsebenen und der im Verlauf der neuronalen Pfade steigenden Generalisierung ließe sich die sensorische Informationsverarbeitung als streng hierarchischer und konvergenter Prozess beschreiben – eine Sichtweise, die bis zur Mitte des letzten Jahrhunderts üblich und befriedigend war [Zek93]. Dem Denkansatz einer Bottom-Up Strategie mit ausschließlich afferenten Projektionen entsprechen letztendlich eine Vielzahl von Modellen und technischen Lösungen zur Objekterkennung, in deren Zusammenhang höchstens das spätere Auslösen einer motorischen Reaktion als efferente Komponente interpretierbar ist. Eine Unterbrechung zwischen den seriellen Verarbeitungsstufen eines hierarchischen Systems führt zwangsläufig zum kompletten Versagen der Wahrnehmung und zum Ausfall der motorischen Koordination. Beobachtungen, dass bei spezifischen Läsionen im visuellen System nur bestimmte Leistungen verlorengehen, andere aber erhalten bleiben, stehen im Widerspruch mit einem solchen seriellen Bottom-Up Ansatz. Einen bekannten Befund stellen beispielsweise Verletzungen des primären visuellen Kortex dar, infolgedessen die scheinbar erblindeten Patienten Objekte zwar nicht sehen, aber dennoch orten und mit ihnen interagieren können [MG95]. Nicht alle motorischen Efferenzen sind demnach an eine erfolgreiche Objekterkennung gebunden. Andererseits ist die Repräsentation von Objekten nur unter Beteiligung von sensomotorischen rekurrenten Mechanismen denkbar: Die ständige Verlagerung des begrenzten fovealen Bereichs des Gesichtsfeldes durch sakkadische Augenbewegungen ermöglicht erst eine detaillierte Analyse objektspezifischer Merkmale. Anstelle des vereinfachten Bottom-Up Regimes der klassischen Bildverarbeitung basieren natürliche visuelle Leistungen

auf komplexen Regelkreisen, in denen neben ventralen und dorsalen Kortexarealen auch Bereiche des Thalamus und Mittelhirns integriert sind. Rekurrenzen, die notwendigerweise über die beschriebenen Hierarchieebenen hinweg reichen, stellen dem Bottom-Up Paradigma der Objekterkennung einen Top-Down Pfad gegenüber.

Efferente Projektionen spielen nicht nur bei der Weitergabe von Motorkommandos als beobachtbare Verhaltensleistungen eine Rolle. Ähnliche Regelkreise mit ventralen, dorsalen, kortikalen und thalamischen Komponenten nehmen als verdeckte Wahrnehmungsleistungen auch direkten Einfluss auf sensorische Repräsentationen. Im Kontext der Aufmerksamkeitssteuerung bilden diese inzwischen detailliert untersuchten Efferenzen die Grundlage für die Mechanismen der visuellen Suche, bei der die lokale Aktivierung in einer Merkmalskarte über eine Top-Down Projektion die Fokussierung der sensorischen Vorverarbeitung bewirkt. Im auditorischen System mit seiner aufwendigen subkortikalen Verarbeitung spielen efferente Projektionen eine kritische Rolle. Insbesondere die Analyse komplexer Schallereignisse oder der Einfluss von Störgeräuschen und ungünstigen akustischen Bedingungen erfordern rekurrente Adaptions- und Selektionsmechanismen, schon bei der Kodierung primärer auditorischer Merkmale [Kin97]. Einen kurios anmutenden Beweis dafür, dass der Top-Down Pfad im auditorischen System selbst die Rezeptorebene erreicht, stellt die Messung otoakustischer, das heißt in der Hörbahn generierter, efferenter Signale im Ohr dar.

**Parallele Verarbeitung und das „Was?“ vs. „Wo?“ Konzept:** Mit der immer detaillierteren Beschreibung der topographisch basierten dorsalen Verarbeitung und der orientierungsinvarianten Merkmals- und Objekterkennung im ventralen visuellen Kortex etablierte sich eine Doktrin des dorsalen „Wo“- und ventralen „Was“-Pfades [MUM83]. Gestützt wurde die Theorie einer solchen Aufgabenteilung bei der Wahrnehmung durch Läsionsexperimente, bei denen die entstehenden Defizite, räumliche Desorientiertheit oder die Unfähigkeit, Objekte zu erkennen, genau der Verletzung von Regionen im jeweiligen Pfad entsprachen [Poh73]. Bislange ungeklärt ist die Frage, ob die Spezialisierung auf räumliche Informationen oder Farb- und Form-Merkmale erst im Kortex beginnt, oder ob „Was“- und „Wo“-Pfad nur die kortikale Verlängerung der parvozellulären bzw. magnozellanulären Projektionen aus der Retina sind [LH87]. Wichtiger als für die Initiierung des afferenten Bottom-Up Pfades von der Retina über den Corpus geniculatum laterale (LGN) nach V1 erscheint diese Fragestellung in Bezug auf die Verbindungen zwischen Kortex, Thalamus und Mittelhirn. Strukturen wie der Superior Colliculus (SC) werden aus Arealen beider Pfade angesprochen, dienen aber überwiegend der räumlichen Orientierung bzw. der Koordination von Bewegungen. Das

Konzept der „Was“- und „Wo“-Pfade stellt auch insofern eine starke Vereinfachung der neuronalen Architekturen dar, als dass damit weder die komplexen kortiko-kortikalen Verbindungen zwischen Arealen unterschiedlicher Pfade [Zek93] noch der genaue Ursprung der somatomotorischen Efferenzen zu erklären sind.

Ungeachtet seiner Schwächen ist das „Was“/„Wo“-Konzept zumindest als abstraktes Wahrnehmungsmodell so populär, dass es in der jüngeren Vergangenheit auch auf die auditorische Verarbeitung angewandt wurde [KVV02]. Untersuchungen an Primaten belegten die Spezialisierung bestimmter auditorischer Kortexareale für die spektralen und temporalen Merkmale von Vokalisierungen der eigenen Spezies, und zwar unabhängig von der Lokalisation [RT00]. Während dabei die Projektionswege u.a. über den primären auditorischen Kortex A1 bis in den ventralen Teil des thalamischen Corpus geniculatum mediale (MGB) zurück verfolgt wurden, konnte gleichzeitig der dorsale Teil des MGB als Ursprungsort räumlicher Abbildungen nachgewiesen werden. Auch ohne diesen Befund liegt eine derartige Spezialisierung innerhalb des Kortex nahe, um beispielsweise beim Menschen räumliches Hören und Sprache zu realisieren. Dass diese Aufgaben wiederum nicht vollkommen unabhängig bearbeitet werden, zeigt die Erfahrung, dass die Lokalisiertheit eines Sprachsignals oftmals eine Hilfe für das Erkennen seines Inhalts ist.

Wie zuvor im visuellen System kann man fragen, auf welchem neuronalen Niveau eine entsprechende funktionelle Unterteilung in „Was“- und „Wo“-Pfad beginnt. Schließlich existieren bereits im Inferior Colliculus (IC) des Mittelhirns komplizierte Merkmalskarten. Die Mechanismen zur Detektion und Repräsentation der binauralen Laufzeit könnten tatsächlich als exklusive „Wo“-Strukturen angesehen werden. Die meisten auditorischen Karten bilden jedoch Informationen ab, die insbesondere, wenn sie spektralen Charakter haben und tonotop organisiert sind, sowohl zur Ortung als auch zur Identifikation einer Schallquelle beitragen. Im Gegensatz zu den Stäbchen- und Zapfen-Photorezeptoren der Retina sind die weniger spezialisierten Haarzellen der Cochlea kaum in der Lage, gezielt räumliche oder objektspezifische Informationen zu kodieren. Ihre einzige Selektivität resultiert aus der vorangegangenen mechanischen Frequenzerlegung im Innenohr. Ein zumindest indirekter Zusammenhang zwischen auditorischen Rezeptoren und Verarbeitungspfaden ließe sich höchstens aus dem Umstand ableiten, dass verschiedene Frequenzbereiche im Verlauf ihrer tonotopen Projektion unterschiedliche Gebiete der cochlearen Kerne innervieren, und später zum Teil spezifische Lokalisations- und Klassifikationsmechanismen ansprechen.

Den drei genannten Konzepten gemeinsam ist ihr Ursprung in der Untersuchung und Erklärung der visuellen Wahrnehmung und ihre spätere Übertragung auf das



auditorische System. Wiederholt wurden dabei unmittelbare Parallelen zwischen den sensorischen Verarbeitungsprinzipien beim Sehen und Hören unterstellt. In seinem als Standardwerk geltenden „*Auditory scene analysis*“ [Bre90] formuliert BREGMAN den Grundgedanken, dass den von ihm beschriebenen auditorischen „Streams“ eine ähnliche Bedeutung wie den visuellen Objekten zukommt. BREGMANs Idee, alle Komponenten einer auditorischen Repräsentation, die durch dieselbe Schallquelle verursacht werden, in einem gemeinsamen „Stream“ zusammenzufassen, erweist sich in der Praxis jedoch als außerordentlich problematisch. Deutlich wird dies an den Untersuchungen von KOBOVY und VAN VALKENBURG, die in [KVV02] das Konzept auditorischer Objekte diskutieren. Sie kritisieren das Prinzip einer Gleichbehandlung auditorischer und visueller Rauminformationen als „Wo“-Eigenschaft sowie die Gegenüberstellung spektraler akustischer Informationen und optischer Merkmale wie der Farbe im Sinne des „Was“-Aspekts. Da die räumliche Unterscheidbarkeit vorrangig für die visuelle Objekterkennung wichtig sei, die Differenzierung von Frequenzen hingegen ausschlaggebend für die Separation von Geräuschen, sollte stattdessen die visuelle räumliche Abbildung mit der Repräsentation des akustischen Spektrums verglichen werden. Als Indiz wird unter anderem der Effekt angeführt, dass ein identischer Ton, der simultan aus verschiedenen Richtungen erklingt, als ein einziges akustisches Ereignis wahrgenommen wird, währenddessen zwei verschieden hohe Töne aus derselben Richtung unterscheidbar sind. Der Vergleich von Ort und Tonhöhe ist nicht neu – in [Dan76] und [War82] werden Verdeckungsexperimente beschrieben, bei denen Töne mit konstanter oder sich gleichmäßig ändernder Frequenz von Störgeräuschen unterbrochen werden. So wie ein sich bewegendes Objekt von einem visuellen Hindernis kurzzeitig verdeckt wird, werden die genannten Töne über die Unterbrechung hinaus als kontinuierliche, einheitliche Schallereignisse empfunden.

Ungeachtet der Popularität, die derartige Vergleiche und Experimente nach wie vor genießen, erscheint eine kritische Wertung ihrer Aussagekraft geboten. Reine Töne sind kein Bestandteil unserer natürlichen akustischen Umwelt, und trotz der Separierbarkeit simultaner Frequenzen ist gerade die spektrale Komposition von Klängen und Geräuschen in ihrer Gesamtheit oft typisch für *ein* auditorisches Objekt. Die Analogie von topographischer Organisation der Retina und tonotoper Repräsentation im Hörnerv bedeutet sicherlich, dass ähnliche neuronale Strukturen zur sensorischen Filterung und Kodierung zur Anwendung kommen, rechtfertigt deshalb aber noch keine funktionellen Vergleiche der grundverschiedenen visuellen und auditorischen Merkmale.

### 1.2.2 Aufmerksamkeitsmodelle

Wie bereits die hierarchischen oder parallelen Konzepte der Wahrnehmungsmodelle gehen auch die Bemühungen um eine Definition des Aufmerksamkeitsbegriffs in aller Regel auf visuelle Mechanismen und Experimente zurück. Bei der Verallgemeinerung einer solchen Definition für die multisensorische Wahrnehmung können daher ähnliche Bedenken wie bei der unmittelbaren Gegenüberstellung von visuellen und auditorischen Merkmalen oder Objekten formuliert werden. Unter dem Vorbehalt der eingangs erwähnten Unterschiede in der räumlich–zeitlichen Natur visueller und auditorischer Reize kann aber dennoch eine Reihe universeller Aufmerksamkeitsmodelle zitiert und zur Einordnung multisensorischer Mechanismen benutzt werden.

#### Räumliche Selektion vs selektives Sehen

Als möglicher Ausgangspunkt für die Diskussion von Aufmerksamkeitsmodellen wurde in [Sch02d] auf die unterschiedliche Bedeutung von Objekt und Ort eingegangen. Man könne fragen, ob der Selektionsaspekt der Aufmerksamkeit zwangsläufig räumliche Bereiche der Umgebung betrifft oder ob ebenfalls eine direkte Auswahl diskreter Objekte möglich ist. Ein Modell mit dem Charakter der räumlichen Selektion kann intuitiv und plausibel als Spotlight–Aufmerksamkeit beschrieben werden, wobei es zunächst keine Rolle spielt, auf welche Weise oder durch welche Merkmale sich ein mehr oder weniger begrenzter Ort auszeichnet. Ein einfacher experimenteller Beleg für die räumlich selektive Aufmerksamkeit ist die Verringerung der Reaktionszeit auf einen Reiz, wenn eine Erwartung über den Ort seines Auftretens besteht. In [DP85] und [PSD80] wird eine solche Erwartung beispielsweise durch einen vorangegangenen Reiz an der gleichen Position oder die Markierung einer Region hervorgerufen.

Schwieriger zu beantworten ist die Frage, ob Selektionsmechanismen auch ohne eine gleichzeitige räumliche Fokussierung auftreten. Tatsächlich wurden solche nicht–räumlichen Aspekte in einem erweiterten Aufmerksamkeitsbegriff als selektives Sehen bezeichnet. Bei der Präsentation zweier räumlich komplett überlagerter visueller Szenen ist demnach die Konzentration auf eine der beiden Teilszenen möglich, was außerdem zur Folge hat, dass Details in der anderen Szene unbewusst bleiben [NB75]. Auch dieses Phänomen kann unabhängig von spezifischen visuellen Merkmalen beobachtet werden und ein Vergleich mit dem konzentrierten Zuhören unter schwierigen akustischen Verhältnissen ist naheliegend: Selbst in Mono–Aufnahmen ohne Lokalisationsmerkmale können überlagerte Sprachsignale separiert und als einzelne Streams (vergl. [Bre90]) zum besseren Verstehen selektiert werden.

### Parallele und serielle Suche

Während die Unterscheidung von räumlicher Selektion und selektivem Sehen eine Trennung von „Wo“- und „Was“-Aspekten darstellt, geht die Beschreibung der Wahrnehmung als parallele oder serielle Suchaufgabe auf verschiedene Abstraktionsebenen der sensorischen Verarbeitung ein. Demnach haben primäre Verarbeitungsschritte einen eher präattentiven und parallelen Charakter, höhere und komplexere Mechanismen laufen dagegen attentiv und seriell ab [Nei67]. Jüngere Untersuchungen zeigen allerdings auch, dass eine strikte Trennung zwischen präattentiver Vorverarbeitung und attentiver Wahrnehmung schwierig ist, da bereits einfache visuelle Merkmale durch die Beteiligung von Aufmerksamkeitsmechanismen schneller oder genauer bestimmt werden [LKB97, BJ98]. Auch parallele und serielle Suche lassen sich in visuellen Versuchsanordnungen und mit Hilfe der Messung von Reaktionszeiten demonstrieren: Unterscheidet sich ein visuelles Ziel nur in *einer* einfachen Eigenschaft von seiner Umgebung, so ist die Dauer der in diesem Fall parallelen Suche weitgehend unabhängig von der Komplexität der Szene. Zeichnet sich ein Ziel jedoch durch Konjunktionen auffälliger Merkmale aus, erfordert sein Auffinden eine serielle Suche und eine mit der Anzahl weiterer Objekte proportional steigende Suchzeit.

### Auffälligkeit und selektive Bahnung

Neben den phänomenologischen Beschreibungen von Wahrnehmungsleistungen gibt es natürlich auch Bemühungen, aus den eher abstrakten experimentellen Befunden konkrete und simulierbare Architekturen abzuleiten. Ein einfaches Modell der (visuellen) Aufmerksamkeit benötigt nach KOCH parallel erzeugte, elementare Merkmalskarten, die in einer gemeinsamen topographischen Auffälligkeitskarte fusioniert werden [KU85]. In der Auffälligkeitskarte soll ein Selektionsprozess bestimmen, an welcher Position eine signifikante Ausprägung eines oder mehrerer Merkmale vorliegt, woraufhin dieser Bereich in eine zentrale Repräsentation projiziert wird. Der Selektionsvorgang selbst kann als datengetriebener Winner-Take-All-Prozess (WTA) realisiert werden, bei dem sich eine oder mehrere dominante Regionen gegen konkurrierende topographische Bereiche durchsetzen. In Situationen, in denen nur ein Merkmal zur Kodierung der Auffälligkeit ausreicht, wird anhand dieses Kriteriums eine parallele Suche realisiert und eine einzelne Gewinnerregion in der Auffälligkeitskarte gebildet. Müssen hingegen Kombinationen von Merkmalen gefunden werden, führt dies zu mehreren auffälligen Regionen, die in einer anschließenden seriellen Suche auszuwerten sind. In dem sehr ähnlichen Merkmalsintegrationsmodell [TG80] wird ebenfalls von einer aus Merkmalen gewonnenen topographischen Abbildung der Auffälligkeit ausgegangen. Allerdings

sollen besondere WTA-Prozesse schon selbst eine sequentielle Ausprägung eindeutiger Gewinnerregionen garantieren. Damit wird ein konkreter Mechanismus zur seriellen Suche vorgeschlagen und die Auswertung der Auffälligkeitskarte auch bezüglich der Konjunktion von Merkmalen realisiert.

### **Geleitete Suche**

Die selektive Bahnung durch WTA-Prozesse in einer Auffälligkeitskarte stellt, global betrachtet, zunächst eine Bottom-Up Strategie dar. In [WCF89] und [Wol94] wird im Rahmen einiger Modifikationen auch die Erweiterung durch einen Top-Down Pfad beschrieben. Anstelle einer einzigen Auffälligkeitskarte werden merkmalsbezogene Selektionsprozesse in parallelen Verarbeitungspfaden vermutet. Bei einer Top-Down gerichteten Beeinflussung der Selektionsprozesse durch die Vorgabe bestimmter Erwartungen an ein Ziel, können diejenigen Merkmale genutzt werden, die einen optimalen Kontrast zwischen Ziel und Umgebung (bzw. anderen Objekten) erzeugen. Allerdings bleibt unklar, welchen Ursprung die Zielvorgaben zur Steuerung des Bottom-Up Pfades haben und auf welcher Abstraktionsebene des hierarchischen Pfadkonzeptes ein Aufmerksamkeitsmechanismus letztendlich modelliert werden soll. Ein Beispiel für Aufmerksamkeitsmodelle auf Grundlage von WTA-Dynamiken mit der besonderen Möglichkeit von wahrscheinlichkeitsbasierten Zielvorgaben schildern KOPECZ und SCHÖNER in [KS95].

### **Paralleler und gelenkter Wettbewerb**

Mit der Attentional Engagement Theorie [DH89, DH92] wurde versucht, die grundlegenden Annahmen über WTA-Mechanismen bei der selektiven Aufmerksamkeit besser zu motivieren und die bestehenden Modelle zu verfeinern. Wie bei der geleiteten Suche wird von parallelen Merkmalsbeschreibungen, jedoch auf unterschiedlichen Auflösungsebenen, ausgegangen. Die Zielvorgabe für die Selektion erfolgt durch komplexe Merkmalschablonen (attentional templates), und als zentrale Repräsentation dient ein begrenztes visuelles Kurzzeitgedächtnis. Damit Muster dort gespeichert werden können, muss eine Auswahl durch Wettbewerbsprozesse vorausgehen – die limitierten neuronalen Ressourcen motivieren die Selektivität der Aufmerksamkeit.

Weiterführende Wettbewerbstheorien gehen davon aus, dass auch bei der spezialisierten Verarbeitung im ventralen „Was“- und dorsalen „Wo“-Pfad aufgrund einer begrenzten Kapazität Wettbewerbsmechanismen notwendig sind [DD95]. Als Ursprung von merkmalsbezogenen und objektspezifischen Zielvorgaben werden die orientierungs-invarianten Objektprototypen im Areal IT vorgeschlagen. Gleichzeitig werden aber auch die im dorsalen Pfad repräsentierte Position und Orientierung als gleichberechtig-

te Eigenschaften einer Suchschablone integriert. Als Pendant zur phänomenologischen Beschreibung der parallelen oder seriellen Suche können im Konzept der Wettbewerbsmodelle eine parallele Bottom–Up sowie eine gelenkte Top–Down Verarbeitung angesehen werden. Die wesentlichen konzeptionellen Neuerungen sind die Zusammenführung von räumlicher und objektbasierter Selektion und die explizite Beschreibung der begrenzten Kapazität als Motivation. Ein implementierbares Simulationsmodell für die wettbewerbsbasierte, zielgerichtete Selektion wird in [UN96] vorgestellt. Es schlägt eine Assembly–Kodierung von Objektprototypen im Areal IT vor und beschreibt einen WTA–Prozess zwischen den Assemblies, also innerhalb des IT.

In einer Integrated Competition Hypothesis [DHW97] wird schließlich auch ein Verhaltenskontext hergestellt und der selektive Aspekt der Aufmerksamkeit durch die Handlungsrelevanz der Wahrnehmung motiviert. Im Zusammenhang mit den Wettbewerbstheorien erscheinen auch aktuelle Arbeiten von TAYLOR interessant, in denen er sich ebenfalls auf Untersuchungen zur Top–Down Kontrolle der Aufmerksamkeit im Wahrnehmungs–Handlungs–Zyklus [HBM00, HM01] bezieht. In [TF04] entwirft er ein Overall Attention/Emotion Network, welches dorsale und ventrale Aufmerksamkeitsmodule beinhaltet, die mit einem Amygdala–Modul und so mit dem limbischen System gekoppelt sind. Das Amygdala–Modul kann mittels Projektion über den orbitofrontalen Kortex (OFC) die dorsalen „Wo“-Aspekte der Wahrnehmung inhibieren, wodurch Emotion und Attention in einen Wettbewerb um kognitive Ressourcen treten. Wie abstrakt der Aufmerksamkeitsbegriff selbst in Simulationsmodellen gefasst sein kann, zeigen die Arbeiten von TAYLOR und KASDERIDIS zu einem Paradigma aufmerksamkeitsbasierter Lernvorgänge, die schließlich zur Formulierung einer Lernziel–orientierten Aufmerksamkeit führen [KT04].

### **Objektbasierte Aufmerksamkeit**

Je nachdem, wie das Ziel oder die Aufgabe bei der Wahrnehmung bestimmt wird, ist die Aufmerksamkeit an mehr oder weniger primäre oder komplexere Merkmale geknüpft. Die gleichen Merkmale dienen typischerweise auch dazu, Objekte als unterscheidbare Einheiten in der Umgebung zu definieren – zur (Wieder–) Erkennung individueller Objekte oder aber im Sinne eines generischen Objektbegriffes zur Definition einer Klasse von ähnlichen Objekten. Infolge der engen Verbindung von Aufmerksamkeit und Objekt durch die betrachteten Merkmale wird in einigen aktuellen Veröffentlichungen eine objektbasierte Aufmerksamkeitssteuerung diskutiert [Sch02d, KVV02]. Auch anhand einer objektbezogenen Definition der Aufmerksamkeit sind Modellierung und Simulation einer visuellen Suche möglich [LD04].

In Bezug auf die erwähnte Handlungsrelevanz und das Konzept komplexer Aufmerksamkeits-Templates hat die objektbasierte Aufmerksamkeit die Eigenschaft, alle Merkmale eines Templates zu selektieren, unabhängig davon, ob sie für eine momentane Aufgabe wichtig oder irrelevant sind [OD99]. Andererseits vereinfacht eine objektbezogene Sichtweise die Übertragung der Definition einer visuellen Aufmerksamkeit auf auditorische bzw. multisensorisch basierte Mechanismen. Im Rahmen der Wettbewerbstheorien wurde die Lokalisiertheit eines Ereignisses als eine von vielen Objekteigenschaften angesehen, die auf prinzipiell gleiche Art und Weise zu behandeln ist, wie Farbe oder Form. Daraus resultiert, dass ein visuelles Kurzzeitgedächtnis zur Speicherung einer Aufmerksamkeitsschablone dezentral organisiert sein muss und nicht nur ventrale „Was“-Komponenten, sondern auch dorsale „Wo“-Aspekte beinhaltet. Wenn man ohnehin dezentrale Templates und einen in verschiedenen Pfaden parallel ablaufenden Selektionsprozess voraussetzt, erscheint es nur folgerichtig, auch auditorisch-visuelle Templates zu erlauben.

Lokalisiertheit ist schließlich keine exklusive Eigenschaft im visuellen System, sondern kann auch als wichtige Komponente auditorischer Streams verstanden werden. Einen deutlichen Hinweis auf die räumlich basierte Aufmerksamkeit und Selektion beim Hören stellt die Fähigkeit zur Fokussierung von Richtungen beim Cocktailparty-Effekt dar [Bod93]. Um aber in Analogie zu den vermuteten Form/Farbe/Ort-Templates für die visuelle Aufmerksamkeit die relevanten Merkmale der auditorischen Streams zu bestimmen, muss die stark vereinfachende Sichtweise einer vorrangig spektralen Kodierung von KOBOVY und VAN VALKENBURG [KVV02] erweitert werden. Mit der in den Abschnitten 2.2.5 und 2.2.6 ausführlich diskutierten Frequenz-Zeit-Ort Abbildung im auditorischen System wird eine schlüssige und adäquate Kombination von räumlichen und spektralen Merkmalen vorgeschlagen, die auch in einfachen Simulationsmodellen umgesetzt werden kann.

## 1.3 Frühe Aufmerksamkeit als Ziel der Modellierung

### 1.3.1 Kritik an den Paradigmen der Objekterkennung

Die Gegenüberstellung einiger wesentlicher Modelle zur Erklärung der visuellen und auditorischen Wahrnehmung verdeutlicht das überwiegende Interesse, Wahrnehmungsleistungen den kortikalen Strukturen zuzuordnen und im Sinne einer Objekterkennung zu interpretieren. Dies betrifft insbesondere die multisensorische Integration, die, wenn überhaupt, vornehmlich auf Objektebene vollzogen wird. Aufgrund der begrenzten

Ressourcen der sensorischen Systeme sind bei der Bildung von Objekten Selektionsmechanismen unumgänglich, zu deren Steuerung momentan handlungsrelevante Aspekte schablonenhaft in Aufmerksamkeits-Templates zusammengefasst werden. Die einzelnen Merkmale in solchen Templates (Farbe, Form, Tonhöhe, Ort etc.) werden der Modellbildung nach in spezifischen Pfaden separat extrahiert. Entsprechend des neuronalen Niveaus der Verarbeitung sowie der Abstraktion und Komplexität der Merkmalsrepräsentation haben sensorische Informationen den Charakter primärer Elemente, die im Verlauf der visuellen und auditorischen „Was“- und „Wo“-Pfade zu (Merkmals-) Gruppen und letztendlich zu Objekten zusammengefasst werden [KVV02].

Als Schwachstelle muss in diesem Modellkonzept die erforderliche, räumliche Zielvorgabe angesehen werden, die auch über die Verhaltensrelevanz in einem Wahrnehmungs-Handlungs-Kontext nur bedingt realisierbar erscheint. Die methodisch elegante Integration von „Was“- und „Wo“-Merkmalen in einem allgemeinen Objekt- bzw. Aufmerksamkeitsbegriff löst an sich noch nicht das Dilemma der begrenzten Kapazität, zumindest wenn neue Reize an unerwarteten Positionen auftreten. Der nicht zu beherrschende Aufwand bei einer weitreichenden, parallelen Suche stellt ein Grundproblem der Bildverarbeitung und Objekterkennung dar und vereitelt eine robuste und schnelle Reaktion oder reflexhafte sensomotorische Antwort künstlicher Systeme. Der Versuch, den Rechenaufwand im parallelen Teil durch eine räumliche Zielvorgabe zu minimieren, birgt zwangsläufig die Gefahr, wesentliche Ereignisse in der Umgebung zu ignorieren. Aus dieser Überlegung lässt sich einerseits die Forderung nach einer frühen räumlichen Selektion in Bottom-Up Pfaden ableiten und dabei zugleich die Frage stellen, ob etwa frühe multisensorische Mechanismen existieren, die ohne die Bildung von Objekten eine räumliche Aufmerksamkeit auf subsymbolischem Niveau unterstützen.

Eine Alternative zum objektbasierten Verständnis von Wahrnehmung und Aufmerksamkeit beschreibt CALVERT, der in [CC01] eine ganze Reihe aktueller Untersuchungen von multisensorischen Effekten auf verschiedenen neuronalen Ebenen diskutiert. Sowohl in mehreren kortikalen Arealen (STS, IPS, Insula posterior, parieto-preokzipitales und frontale Areale) als auch im Superior Colliculus (SC) des Mittelhirns sei eine anatomische multisensorische Konvergenz nachgewiesen worden. CALVERT nimmt unter anderem Bezug auf vorausgegangene Untersuchungen von STEIN, MEREDITH und WALLACE [SW96, MWS92, WWS96], die detailliert die multisensorischen Eigenschaften schon auf Ebene des subkortikalen SC beschreiben. Darüber hinaus entwirft er ein multimodales Konzept, dass mächtiger als die Aufmerksamkeits- und Objekt-Templates erscheint. Multisensorische Effekte und Be-

funde lassen sich demnach drei allgemeinen Kategorien zuordnen: *Crossmodal Matching*, *Integration* und *Learning*. Während Crossmodal Matching die kognitive Fähigkeit beschreibt, separate unisensorische Repräsentationen von Objekten miteinander zu vergleichen (etwa den visuellen und den taktilen Eindruck von Gegenständen [SM93]), sind hier insbesondere multisensorische Integrationsvorgänge interessant. Die Integrationsaspekte untergliedert CALVERT in die Teilleistungen Identifikation und Lokalisation – vergleichbar mit der Verarbeitung in „Was“- und „Wo“-Pfad. Außerdem nutzt er Zeitpunkt und Verlauf eines Reizes als „Wann“-Eigenschaft sowie die Neuheit eines Stimulus als zusätzliche Merkmale. In Abbildung 1.1 sind die vier Merkmalsgruppen (Was, Wo, Wann und Neuheit), ihre Repräsentationsorte sowie deren Verbindung zum primären visuellen und primären auditorischen Kortex dargestellt. Nach [CC01] sind einige Verbindungen nicht statisch, sondern unterliegen Lernvorgängen. In Konditionierungsexperimenten [MCL98, GSD00] wurden multisensorische Lerneffekte für die kortiko-kortikalen Verbindungen des „Was“- und des „Neuheits“-Zweiges dokumentiert, nicht jedoch für die auf- oder absteigenden Projektionen zwischen primären Kortexarealen und SC.

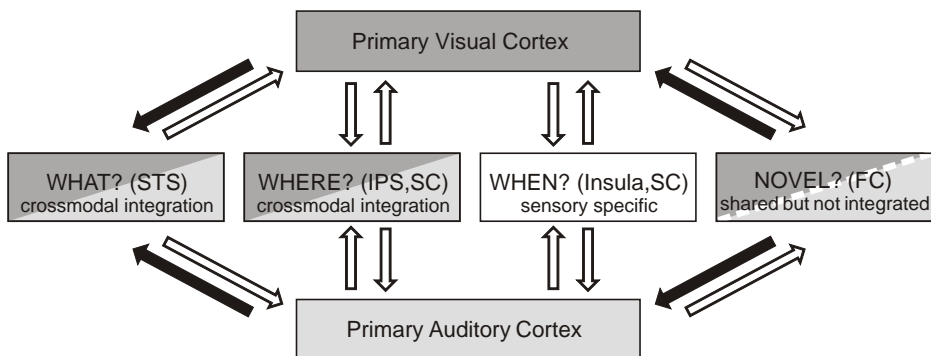


Abbildung 1.1: Multisensorische Verarbeitung auf unterschiedlichen Ebenen und in verschiedenen Pfaden. Dunkle Pfeile markieren Verbindungen, die Lernvorgängen unterliegen. (STS: Sulcus temporalis superior, IPS: Intraparietaler Sulcus, SC: Superior Colliculus, FC: frontale Kortexareale, nach [CC01]).

Desweiteren weist CALVERT wie zuvor bereits STEIN darauf hin, dass bei multisensorischen Lokalisationsleistungen unterschiedliche Zeitfenster eine Rolle spielen. Er unterscheidet deshalb zwischen einer direkten multisensorischen Integration auf dem Niveau einer einzelnen neuronalen Struktur (SC, STS) und einer allgemeineren räumlichen Aufmerksamkeit, die mehrere Hierarchieebenen einschließt. Die wesentlich größeren Zeitfenster, in denen Effekte der räumlichen Aufmerksamkeit zu beobachten sind [Ber94], seien kaum mit Strukturen innerhalb eines Areals zu bewerkstelligen



und erfordern Rückkopplungen, beispielsweise zwischen primärem Kortex und Mittelhirn. Insgesamt stützen die Arbeiten von STEIN, WALLACE und CALVERT eindeutig die These einer schnellen und fest verschalteten Projektion und multisensorischen Verarbeitung räumlicher Informationen schon auf subkortikalem Niveau. Die dokumentierten Befunde erfüllen alle Voraussetzungen für einen Bottom-Up Pfad, in dem zusätzlich zu den vermuteten Vorgängen bei der Objekterkennung neue Zielregionen für Aufmerksamkeits-Templates generiert werden können. Das Manko der objektbasierten Aufmerksamkeit beim Auftreten neuer und unerwarteter Reize wäre so auf plausible Weise zu kompensieren. Eine äußerst interessante Frage ergibt sich zudem aus der zentralen Rolle, die der SC als mutmaßlicher multisensorischer Integrationsort bei der Steuerung der Blickrichtung durch Augen- und Kopfbewegungen spielt [GW72, DR93, GS92, HN99, DPM97]. Sind zum Auslösen motorischer Reaktionen als Manifestation von Aufmerksamkeit und Verhalten komplexe Merkmalsbeschreibungen im Sinne einer Objekterkennung überhaupt erforderlich? Oder kann die Generierung bestimmter, eher reflexartiger Reaktionen bereits auf Basis der subkortikalen multisensorischen Repräsentationen modelliert werden? Der Vergleich der Zeitfenster für multisensorische Interaktionen mit den sensomotorischen Reaktionszeiten, die beispielsweise bei der Untersuchung von Gap-Effekten gemessen wurden, bietet diesbezüglich einen relativ großen Raum für Interpretationen und lässt noch keine abschließende Aussage zu [SRZ00, Spa99, FK03].

### 1.3.2 Methodik, Randbedingungen und Lösungsansätze

Aus den Indizien für eine subkortikale multisensorische Verarbeitung und der angeführten Kritik am Paradigma der objektbasierten Aufmerksamkeit leiten sich Anliegen und Methodik dieser Arbeit ab. Es soll nicht versucht werden, ein weiteres Simulationsmodell zur Objekterkennung zu entwerfen oder ein gesamtes technisches System zur audio-visuellen Ortung zu realisieren. Es ist vielmehr das Ziel, motiviert durch die Schwächen bekannter Bildverarbeitungs- und Schallortungsansätze, nach robusten Mechanismen zur Steuerung der räumlichen Aufmerksamkeit zu suchen.

Die Diskussion von Aufmerksamkeit und Objektbegriff führte zur Frage nach der frühestmöglichen multisensorischen Verknüpfung und gestattet nun die Formulierung einiger Prämissen bei der Modellierung ausgewählter Mechanismen einer primären räumlichen Wahrnehmung:

- Es werden nur sensorische Merkmale betrachtet, die unspezifisch für konkrete Objekte oder Aufgaben sind.

- Aufmerksamkeit und Selektionsmechanismen sind im Rahmen der Modellierung prinzipiell Bottom–Up gerichtet.
- Multisensorische Mechanismen werden beispielhaft an auditorisch–visuellen Modellen untersucht.
- Es wird von kompatiblen räumlich–zeitlichen Repräsentationen beim Sehen und Hören ausgegangen.

Bei dieser Herangehensweise wird der Begriff des Objekts durch den der Quelle ersetzt, die abgesehen von ihrer Lokalisiertheit keine spezifischen Merkmale besitzen muss. Anstelle der klassischen Frage in der Bildverarbeitung und Objekterkennung: „Welches Objekt ist an welcher Stelle zu sehen?“, tritt die allgemeinere Formulierung: „Wann und wo geschieht *etwas?*“.

Als Ausgangspunkt für eine konkrete technische Anwendung ist diese Fragestellung gerade in ihrer Allgemeinheit wenig geeignet. Die untersuchten Modelle können jedoch als Komponenten zur Lösung komplexer Wahrnehmungsaufgaben genutzt werden. In potentiellen Anwendungsgebieten wie autonomen Robotiksystemen, Videokonferenzen und multimedialen Mensch–Maschine Schnittstellen besteht einerseits ein dringender Bedarf, enorme Datenmengen zu reduzieren, was durch eine räumliche Fokussierung sehr gut möglich ist. Andererseits bedeutet das Ignorieren mancher Ereignisse und das Verharren an falschen Objekthypothesen meist ein fatales Fehlverhalten der künstlichen Systeme, weshalb der Bottom–Up Charakter eines Modells in Verbindung mit einem multisensorischen Ansatz wesentlich zur Flexibilität und Robustheit einer Anwendung beitragen kann. Um die Relevanz der aus der Biologie entliehenen Verarbeitungsprinzipien und Architekturkonzepte für technische Systeme zu verdeutlichen, werden die experimentellen Untersuchungen in Kapitel 5 beispielhaft in möglichst universellen und aussagekräftigen Szenarien realisiert. Ausgewählte Komponenten der Simulationsmodelle wurden darüberhinaus in Gestalt eines Demonstrators auf einer Roboterplattform implementiert.

Ein zentrales Problem bei der Modellierung und Implementierung stellte die Wahl einer geeigneten Modellarchitektur sowie der konkreten mathematischen Beschreibungsmittel zu ihrer Realisierung dar. In den vergangenen Jahren führten sowohl die zunehmend detaillierteren neurophysiologischen Untersuchungen als auch der geschilderte Entwicklungsbedarf der immer anspruchsvolleren multimedialen Anwendungen dazu, dass erste Computermodelle entstanden, die zur Simulation einer frühen multisensorischen Wahrnehmung dienen – oder, wenn sie technische Applikationen betreffen,

doch zumindest als solche interpretiert werden können. Unabhängig vom neurologischen oder technischen Hintergrund lassen sich dabei bislang zwei sehr verschiedene Paradigmen unterscheiden:

**Mit Künstlichen Neuronalen Netzen (KNN)** wird eine mehr oder weniger direkte strukturelle Umsetzung des Konzepts der topographischen Karten versucht. Seinen Ursprung hat dieser Ansatz in auditorischen Lokalisationsmodellen, bei denen ausgehend von funktionellen und algorithmischen Beschreibungen der binauralen Schallortung [MK81, CK90] Netzwerkstrukturen zur Detektion und Verarbeitung interauraler Zeitunterschiede abgeleitet wurden. Solche Netzmodelle eignen sich aufgrund ihres meist regulären Aufbaus aus einheitlichen, neuronalen Basiselementen nicht nur für die Simulation, sondern auch für spezifische Hardware-Implementierungen [Mea89, Zor95]. Die inzwischen als Neuromorphic Engineering bezeichneten Techniken ermöglichen eine effektive sensorische Verarbeitung im parallelen Teil des Bottom-Up Zweiges. Einige Beispiele sind die VLSI-Implementierung von Innenohrfiltern, Korrelator-Strukturen, topographischen Karten und WTA-Netzen [LM95, vSFVM97]. Eine Erweiterung des KNN-Ansatzes für multisensorische Systeme ist leicht möglich, wenn man wie in [SM93] vermutet, von der Möglichkeit der multisensorischen Integration auf dem Niveau einzelner Neurone ausgeht und im Modell synaptische Terminals zu den topographischen Karten verschiedener Modalitäten herstellt. Aus der Notwendigkeit einer räumlichen und zeitlichen Korrelation der zu kombinierenden Abbildungen wird in einigen Arbeiten die Motivation für Kalibrierungs- und Lernvorgänge abgeleitet [RWE00, AP03].

**Probabilistische Modelle** stellen ein zweites grundlegendes Paradigma zur multisensorischen Integration dar. Auditorische und visuelle Ereignisse werden hier als stochastische Prozesse angesehen, die über ihre bedingten Wahrscheinlichkeiten im Sinne des Bayesschen Gesetzes in Beziehung zu den sensorischen Beobachtungen gesetzt werden können [APBB00]. Ein kritisches Problem der Bayesschen Ansätze stellt die Approximation der Wahrscheinlichkeitsdichtefunktionen (PDF) der stochastischen Zustandsgrößen dar. Als eine mögliche Realisierung wurden Mixture-Models (auch generative Modelle) vorgeschlagen, die als probabilistische Graphen darstellbar sind und deren Zustandsvariablen und Modellparameter mittels Expectation Maximization Algorithmen (EM) geschätzt werden können [BJA02, BJA03]. Eine weitere Gruppe probabilistischer Methoden beschreibt die Erweiterung und Anwendung von Partikel Filter Modellen, die als diskrete Punktwolke im Zustandsraum beliebige

PDFs approximieren und deren Steuerung mit Monte Carlo Techniken erfolgen kann [BGPV01, VGBP01, ZDD02].

In Bezug auf ein biologisch und psychologisch verankertes Verständnis von Wahrnehmung und Aufmerksamkeit stellen die typischerweise anwendungsorientierten, probabilistischen Verfahren oftmals Ansätze zur Objekterkennung dar, was weniger der Art und Weise der multisensorischen Integration als vielmehr einem heterogenen Umgang mit sensorischen Merkmalen zu verdanken ist. Während praktisch in allen Arbeiten in der Audio-Komponente eine primäre räumliche Repräsentation auf Grundlage binauraler Laufzeiten Verwendung findet, werden im Video-Modell neben dem ebenfalls primären Merkmal Bewegung [BJA03] auch Farbe [ZDD02] oder sogar komplexe Konturen herangezogen [BGPV01], um potentielle Objektpositionen zu kodieren. Der erforderliche Aufwand, um beispielsweise eine beleuchtungsunabhängige Hautfarbe-Erkennung oder eine orientierungs- und skalierungsinvariante Konturdarstellung zu realisieren, zeigt, dass die multisensorische Kombination auf einer eher abstrakten, symbolischen Ebene vollzogen wird.

Abgesehen von der expliziten Erwähnung des SC in [APBB00] und [PA03] lassen sich in den probabilistisch orientierten Arbeiten keine Bezüge zu konkreten neuronalen Arealen, Ebenen oder Pfaden herstellen. Ebenso wurde bislang kaum der Versuch unternommen, KNN-basierte und probabilistische Ansätze direkt zu vergleichen. Lediglich ANASTASIO und PATTON [AP03, PBBA02] diskutieren den Bayesschen Charakter der multisensorischen Integration im SC bzw. kortiko-tektalen Strukturen und erörtern informationstheoretische Aspekte von Befunden und Modelleigenschaften. Ein Versuch, das Verhalten eines simulierten neuronalen Netzes nach diesem Vorbild als Bayessche Filterung zu interpretieren [DP04], endet mit eher abstrakten und allgemeinen Bezügen zu verschiedenen subkortikalen und kortikalen Befunden.

In den folgenden Kapiteln soll ausgehend von einem netzwerkorientierten, auditorischen Modell ein homogener, eigener Ansatz zur frühen multisensorischen Integration bei der räumlichen Wahrnehmung entwickelt werden. Darüberhinaus wird ein probabilistisches Referenzsystem so adaptiert, dass es auf den gleichen objektunspezifischen, sensorischen Repräsentationen aufsetzt, wodurch eine vergleichende Untersuchung möglich wird. Ziel dieses Vergleichs ist es, die Sichtweisen der Neurowissenschaften, der Informatik und der Stochastik in einem ingenieurwissenschaftlichen Kontext zusammenzuführen. Die Diskussion der neurophysiologischen Phänomene und Befunde sowie die gleichzeitige Einbettung der Experimente in anschauliche Szenarien soll einen alternativen Zugang zum Verständnis früher, subsymbolischer Wahrnehmungsleistungen als notwendige Ergänzung zum objektbasierten Paradigma aufzeigen.

### 1.3.3 Orientierung am biologischen Vorbild

Bereits die interdisziplinäre Einordnung der zu untersuchenden Simulationsmodelle und die Diskussion von Aufmerksamkeit und Objektbegriff zeigten, wie neurobiologische Befunde einen methodischen Rahmen für Aufgabenstellungen aus der Informatik vorgeben können. Auch potentielle Anwendungen und konkrete technische Lösungen für Objekterkennungs- und Trackingprobleme können mit einem abstrakten Verständnis von Merkmalen und Selektionsmechanismen und der Kenntnis prinzipieller visueller, auditorischer und multisensorischer Verarbeitungsstrategien kritisch hinterfragt werden.

Beim Entwurf einer Modellvariante mit künstlichen neuronalen Netzen ermöglicht die Orientierung am biologischen Vorbild über die makroskopische Beschreibung einer Modellarchitektur hinaus auch konkrete strukturelle und funktionelle Lösungen. Auf der Abstraktionsebene des einzelnen Neurons stellt sich zunächst die Frage nach einer adäquaten Approximation von neuronalen Potentialverläufen und Zuständen. Vor- und Nachteile von spike- oder ratenbasierten Kodierungen können im Kontext der räumlichen und zeitlichen Natur der sensorischen Informationen diskutiert werden. Bei der Konstruktion von Netzwerkstrukturen liefern die in rezeptiven Feldern manifestierte räumliche Sensitivität und die darauf aufbauenden topographischen Abbildungen eine Vorgabe zur Modellierung sensorischer Karten. Neben der bloßen Repräsentation von Merkmalen kann in einigen Fällen die Kenntnis von spezifischen synaptischen Verschaltungen auch Aufschluss über die Funktion einer Struktur geben. So kann bestimmten gegenläufigen axonalen Bahnen und Koinzidenzzellen im auditorischen Hirnstamm eindeutig die Realisierung einer binauralen Kreuzkorrelation zugeordnet werden. Auch die erwähnten WTA-Netze zur Manipulation von Merkmalsmustern und zur Realisierung von Selektionsmechanismen sind ein Beispiel für die Nachbildung einer Funktion durch die Simulation der neuronalen Verschaltungsstruktur.

Eine weitere Gelegenheit, Nutzen aus dem Vergleich mit dem biologischen Vorbild zu ziehen, bietet sich bei der Gestaltung von audio-visuellen Experimenten. Im hier vertretenen Paradigma früher Wahrnehmungsleistungen führt der Verzicht auf Objekt- und Handlungskontext zu einem schwerwiegenden Problem bei der Evaluierung der Modelle. Von richtigem und falschem Modellverhalten, im Sinne der Objekterkennung oder handlungsbasierten Wahrnehmung, kann in Verbindung mit Vorgängen, die sich auf einer unbewussten Ebene abspielen, nicht gesprochen werden. Stattdessen gilt es einerseits, geeignete akustische und optische Reizkombinationen zu bestimmen und zum anderen objektive Kriterien zu finden, anhand derer das Ergebnis einer Simula-

tion eingeschätzt werden kann. Dazu werden zunächst einfache Szenarien entworfen, in denen ganz bewusst Bewegungen und Gesten realer Personen und natürliche Geräusche und Lautäußerungen auftreten – ohne dass dabei eine Objekterkennung oder ein Personen-Tracking beabsichtigt wird. Die Szenarien verankern die Modelle in der Umwelt, da sich die Randbedingungen für die evolutionäre Entwicklung und Optimierung der biologischen Vorbildmechanismen in gewisser Weise auch in den potentiellen Anwendungen wiederfinden. Personen in einem Szenario der Mensch-Maschine Interaktion unterliegen anatomischen, physiologischen und physikalischen Eigenschaften und Gesetzmäßigkeiten, die in Versuchen mit realen Personen immer eingeschlossen sind. Solche räumlich-zeitlichen Dynamiken der natürlichen Umwelt können in artifiziellen Lampe-Lautsprecher Anordnungen [RWE00] oder gänzlich virtuellen Experimenten kaum berücksichtigt werden. Wenn es nun außerdem gelingt, aus den qualitativen und quantitativen Befunden zur multisensorischen Integration eine Reihe von Kriterien zur Bewertung der Experimente abzuleiten, ist eine elegante Lösung des Evaluierungsproblems möglich: Je besser das Ergebnis einer Testreihe diesen Kriterien entspricht, umso plausibler sollte das Modellverhalten bei einer tatsächlichen Anwendung sein. Für die vergleichende Untersuchung probabilistischer Verfahren bedeutet dieses Konzept, dass weder spezifische Merkmale einbezogen noch Trackingaufgaben gelöst werden. Stattdessen kommt die gleiche Kombination von einfachen audio-visuellen Szenarien und neurobiologisch motivierten Bewertungskriterien zur Anwendung.

# Kapitel 2

## Primäre Mechanismen beim räumlichen Hören

### 2.1 Spezifik auditorischer Rauminformationen

Im Kontext der Aufmerksamkeitssteuerung und Bildung von Objekt-Templates wurden in Kapitel 1 Vergleiche visueller und auditorischer Merkmale zitiert und zum Teil kritisiert. Für spezifische Probleme bei der Klassifikation von Objekten mag es hilfreich sein, Korrespondenzen zwischen Ort und Tonhöhe oder Farbe und akustischem Spektrum herzustellen – allgemeine und biologisch plausible Konzepte sind solche konstruierten Vergleiche aber kaum. Wie sieht das bei der Ortung von Objekten bzw. Signalquellen aus? Die Vermutung liegt nahe, dass die Lokalisiertheit eines Reizes eine universelle Eigenschaft ist, in deren Folge die Strategien zur räumlichen Orientierung im visuellen und auditorischen System Parallelen aufweisen. Stellt man Überlegungen zu einem multisensorischen Wahrnehmungsmodell an, so stößt man tatsächlich schnell auf die grundlegende Forderung nach einer räumlich und zeitlich kompatiblen Abbildung der Umwelt in den beteiligten sensorischen Modalitäten.

### Deklaration eines biologisch plausiblen Koordinatensystems

In der technischen, meist kartesischen Welt wird das Problem kompatibler multisensorischer Abbildungen pragmatisch gelöst: Beschreibt man den Raum möglichst unabhängig von den Eigenschaften der Sensorik ohne weitere Transformation mit dreidimensionalen orthogonalen Koordinaten, lassen sich die Positionen der Quellen (visuelle und akustische Objekte) sowie die der Sensoren (Kameras, Mikrofone) auf triviale Weise direkt und absolut darstellen. An primären Mechanismen der natürlichen Wahrnehmung fällt zunächst auf, dass solche globalen Karten oder absolute Koordinaten

nicht benutzt werden können. Stattdessen führt ein Individuum seine persönlichen Koordinatensysteme für die an der Wahrnehmung beteiligten Sinne mit sich und ändert mit jeder eigenen Bewegung den Koordinatenursprung, den Bezug zu Objektpositionen und den Ausschnitt der Umgebung, der überhaupt wahrgenommen wird. Zum einen ist es mit den meisten Sinnesorganen schlichtweg nicht möglich, die komplette Umgebung gleichzeitig abzubilden. Außerdem stellt die physikalische Limitierung auf bestimmte räumliche Bereiche eine erste, präattentive Selektion aus dem reichen Informationsangebot der Umwelt dar. Beim Sehen führt diese erste Selektion zur Begrenzung eines Gesichtsfeldes, in dem darüberhinaus eine unterschiedliche Wichtung der visuellen Information im peripheren und fovealen Bereich stattfindet. Das räumliche Hören schließt auch Bereiche außerhalb des Gesichtsfeldes ein – allerdings mit vergleichsweise geringer Genauigkeit. Eine exakte Schallortung ist bei binauralen Spezies analog zur visuellen Verarbeitung nur in einem zentralen Bereich der räumlichen Abbildung möglich.<sup>1</sup> Anders als die in der Neuroanatomie der Retina begründete Fovea ist die akustische Fokussierung aber bereits physikalisch vorgegeben und resultiert allein schon aus der unterschiedlichen, zeitlichen Disparität in Stereosignalen, wenn Schallquellen zentral oder seitlich positioniert sind.

Die weitere Verarbeitung der räumlichen Informationen in der zentralen Hörbahn zeigt schließlich eine für das auditorische System spezifische Strategie. Im Gegensatz zum visuellen System, in dem verschiedene Richtungen nahezu gleichwertig und gleichartig behandelt werden, unterscheidet sich die Verarbeitung von horizontalem und vertikalem Einfallswinkel einer Schallquelle ganz erheblich. Die Ursachen sind auch hier evolutionäre Erfordernisse und die daran angepassten anatomischen und neuroanatomischen Möglichkeiten. Für die Mehrzahl der an Land lebenden Spezies ist die Orientierung in der horizontalen Ebene von besonderer Bedeutung und wird folglich auch am genauesten und zuverlässigsten realisiert. Dazu stehen mit den verbreiteten binauralen Systemen offenbar besonders effektive Werkzeuge zur Verfügung. Nichts spricht dagegen, dieses Konzept der Konzentration der neuronalen Ressourcen auf die wichtigsten Lokalisationsaufgaben bei der Anwendung von Schallortungsmodellen im Umfeld des Menschen zu adaptieren.

Prinzipiell erscheint ein Polarkoordinatensystem, in dem ein Ort durch Azimut, Elevation und Entfernung bestimmt ist, besser zur natürlichen Repräsentation der

---

<sup>1</sup>Bisweilen wird der Bereich des höchsten, räumlichen Auflösungsvermögen beim Hören als *akustische Fovea* bezeichnet. Diese Veranschaulichung birgt leider die Gefahr von Missverständnissen, da der gleiche Begriff schon wiederholt im Zusammenhang mit Mechanismen zur Verstärkung der Frequenzsensitivität im frühen Verlauf der Hörbahn gebraucht wurde [BS80].



Umgebung geeignet als kartesische Abbildungen. Die spezifischen auditorischen Merkmale für die horizontale und vertikale Ausrichtung von Schallereignissen können dann direkt zur Kodierung des Ortes verwendet werden. Eine hinreichende Korrespondenz zur visuellen Repräsentation ist gegeben, da sich die unserem bildhaften Eindruck der Welt zugrundeliegenden retinotopen Projektionen ob ihres topologischen Charakters ebenfalls leicht in einer Azimut–Elevation Form darstellen lassen. Wird eine solche zweidimensionale Abbildung durch eine Entfernungsschätzung ergänzt, erfolgt die Beschreibung von Richtung und Distanz einer Signalquelle analog zum auditorischen Raum mit egozentrischen Polarkoordinaten.

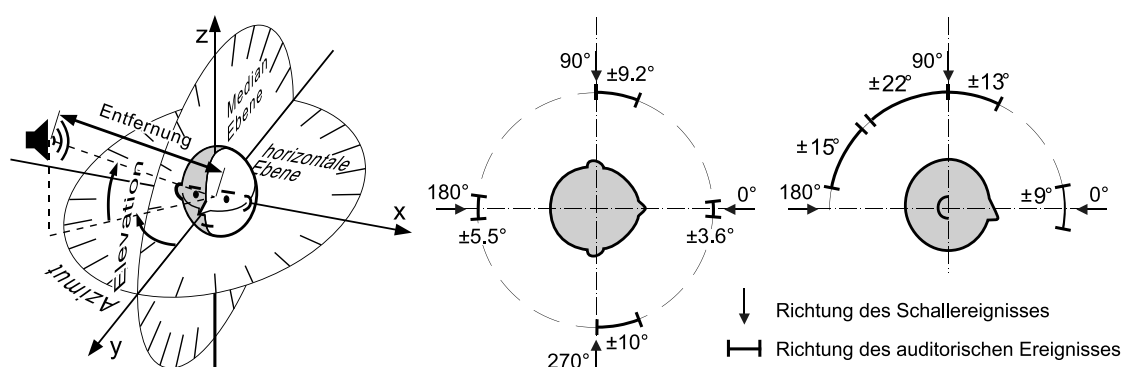


Abbildung 2.1: Links: Deklaration eines Polarkoordinatensystems zur auditorischen Beschreibung räumlicher Informationen. Mitte, Rechts: Lokalisation von Rauschen und Sprache durch den Menschen. Die Genauigkeit der Lokalisation ist richtungsabhängig und in der horizontalen Ebene wesentlich größer als in der Medianebene (nach [Bla96]).

Solange von einer topologieerhaltenden Verarbeitung ausgegangen wird, ist die formale Transformation zwischen visuellen und auditorischen Repräsentationen oder eine Projektion in die kartesische Welt unproblematisch. Für ein Simulationsmodell erscheint es daher zweitrangig, ob nun kartesische, Polarkoordinaten oder eine möglichst exakte Nachbildung bestimmter rezeptiver Felder des biologischen Vorbilds zur Anwendung kommen. Der Unterschied zwischen diesen Darstellungen ist weniger geometrischer Natur, sondern vielmehr begründet in der Wichtung verschiedener Orientierungen und den unterschiedlichen Informationen und Algorithmen zu deren Bestimmung. Dass die daraus resultierenden Abhängigkeiten der auditorischen Lokalisationsleistung von den akustischen Bedingungen, der Art des Geräusches und der Richtung selbst sowohl im Rahmen der Akustik, der Biologie und der Informatik interessant sind, belegen eine Vielzahl von Untersuchungen (vergl. [Bla96] und [vSJC99]). In Abbildung 2.1 wird ein Polarkoordinatensystem deklariert, an dem sich die folgenden Ausführungen und die spätere Modellierung orientieren.

## Konzept der Computational Maps

Anders als das Sehen mit einer topologisch organisierten Retina basiert Hören auf der Verarbeitung von eindimensionalen Zeitsignalen<sup>2</sup>. In solchen, mit Mikrofonsignalen vergleichbaren Stimuli sind Merkmale der Schallquellen, des Raumes und Informationen über die relative Position der Quelle bezüglich des Hörers enthalten. Um die primär zeitliche Kodierung der akustischen Informationen in eine mit der räumlich-zeitlichen Repräsentation der Retina kompatible Form zu überführen, sind eine Reihe von Filterungen und Transformationen notwendig. Da das Gros der topologischen Repräsentationen im auditorischen System anhand von Spektrum, Modulationen, Amplitude oder Phase der Signale erst *berechnet* werden muss, hat sich das Modellkonzept der *Computational Maps* etabliert. Die ersten dieser berechneten Karten findet man bereits im oberen olivaren Komplex (SOC) des Hirnstammes. Hier werden zunächst spezifische Informationen wie binaurale Phasen- oder Intensitätsunterschiede in teilweise parallelen Verarbeitungspfaden ausgewertet, bevor im Inferior Colliculus des Mittelhirns erstmals der komplette auditorische Raum repräsentiert wird.

Die zur Realisierung der räumlichen Karten herangezogenen auditorischen Merkmale lassen sich verschiedenen Kategorien zuordnen. Zunächst ist die Unterscheidung monauraler und binauraler Informationen möglich. Beim monauralen Hören kann das akustische Spektrum oder dessen Änderung infolge einer Kopfbewegung Aufschluss über die Einfallsrichtung des Schalls geben. Außerdem liefern spektrale Informationen und des Verhältnis zwischen direkt und indirekt eintreffendem Schall Hinweise auf die Entfernung einer Schallquelle [MK99]. Wesentlich erleichtert wird die Ortung jedoch durch die Einbeziehung binauraler Unterschiede. Diese wiederum werden auf Basis der relativen Amplitude und Phase ermittelt. Da die Verarbeitung im auditorischen System von Beginn an parallel in spezifischen Frequenzkanälen erfolgt, gibt die Amplitudenauswertung nicht nur pauschal Auskunft über binaurale Pegeldifferenzen, sondern beschreibt genau genommen Unterschiede zwischen rechtem und linkem Spektrum. Somit kann eine Vielzahl akustischer Effekte in Form auditorischer Merkmale

---

<sup>2</sup>Auf Signalebene hat ein visueller Stimulus massiv parallelen und relativ statischen Charakter. Bei der Aufnahme von Videosignalen ist daher eine hohe Anzahl von Bildpunkten bei langsamer Bildfolge (z.B. 30Hz) erforderlich. Demgegenüber entspricht ein eindimensionales Mikrofonsignal einem seriellen Konzept mit einem deutlich höheren Signaltakt (z.B. 44.1kHz). Bereits auf Rezeptorebene erfolgt auch die auditorische Verarbeitung parallel, nämlich durch eine Vielzahl frequenzsensitiver Haarzellen. Die parallele Projektion des Sehfeldes einerseits und des akustischen Spektrums andererseits ist eine Motivation für den Vergleich der visuellen Kategorie des Ortes mit dem auditorischen Merkmal Tonhöhe ([Dan76, War82, KVV02]). Da einzelne, reine Töne jedoch keinen Aufschluss über ihre Lokalisiertheit geben, erscheint ein derartiges Konzept für die Ortungsproblematik nicht sinnvoll.

quantifiziert und für die Generierung einer räumlichen Abbildung verwendet werden. Die Relevanz, Genauigkeit und Zuverlässigkeit der genannten Merkmale ist abhängig von der Art des Geräusches, der Situation und nicht zuletzt von der Einfallsrichtung des Schalls selbst. Breitbandige Geräusche lassen sich meist besser orten als schmalbandige, die weniger aussagekräftige spektrale Eigenschaften besitzen. Insbesondere sehr tiefe und sehr hohe Töne können kaum lokalisiert werden, da für solche Schallereignisse, abgesehen von ihren spärlichen spektralen Merkmalen, auch keine sicheren binauralen Phasendifferenzen bestimmt werden können. Innerhalb der Medianebene treten schließlich fast gar keine binauralen Unterschiede auf. Die Ortung ist hier folgerichtig ungenauer, aber anhand monauraler Informationen prinzipiell möglich.

### **Konsequenzen für multisensorische Aufmerksamkeitsmodelle**

Auf Basis kompatibler, visueller und auditorischer Koordinatensysteme ist es formal möglich, einen auditorischen Fokus ähnlich einer visuellen Spotlight-Aufmerksamkeit zu beschreiben oder die Schallortung als räumliche Hypothese in Aufmerksamkeits-Templates zu integrieren. Aufgrund des Aufwandes zur Erzeugung der berechneten auditorischen Karten ist aber zu beachten, dass im auditorischen System verschiedene räumliche Orientierungen erst auf unterschiedlichen neuronalen Niveaus bereitgestellt werden. Die besonders schwer zu ermittelnde Entfernungsinformation steht beispielsweise für frühe Bottom-Up gerichtete Mechanismen noch nicht zur Verfügung – sie scheint andererseits für einfache Aufgaben wie die grobe Steuerung der Blickrichtung oder für reflexhafte Bewegungen auch keine wichtige Rolle zu spielen.

## **2.2 Modellentwurf am Vorbild der Hörbahn**

Das Konzept der Computational Maps und die Vielzahl an sensorischen Merkmalen und Verarbeitungsschritten führten bei der Modellierung des räumlichen Hörens zu einer breiten Palette von Algorithmen und Implementierungsoptionen. Angefangen von der Beurteilung der akustischen Voraussetzungen zur Kodierung richtungsspezifischer Schalleigenschaften bis hin zu einer eher abstrakt-mathematischen oder detaillierten und biologisch motivierten Beschreibung der berechneten räumlichen Abbildungen gilt es, eine Reihe von Einschränkungen und Randbedingungen zu beachten. Beim folgenden Modellentwurf sollen am Vorbild der zentralen Hörbahn wichtige Voraussetzungen sowie Vor- und Nachteile von ausgewählten, publizierten Modelloptionen diskutiert werden. Notwendige Einschränkungen in der Geometrie der Schallortung und konkrete

algorithmische Lösungen in dem für die weitere Anwendung vorgeschlagenen Simulationssystem werden durchgehend anhand bekannter Wahrnehmungsexperimente und neurologischer Befunde motiviert.

### 2.2.1 Periphere Schallverarbeitung

Hören beginnt nicht erst im Innenohr. Würden wir über Kopfhörer eine Stereoaufnahme hören, der ein unserer Kopfbreite entsprechender Mikrofonabstand zugrunde liegt, hätten wir, abhängig von der Geräuschart und der Raumakustik bei der Aufnahme, gravierende Probleme beim Lokalisieren der Schallquellen. In dieser unnatürlichen Situation der „Abwesenheit“ unseres Körpers bei der Rezeption der Geräusche bliebe der räumliche Eindruck auf eine grobe Lateralisation beschränkt. Allein schon die Gestalt unseres Körpers, insbesondere die Form von Kopf und äußerem Ohr, bewirken aufgrund verschiedener akustischer Effekte wie Reflexion, Dämpfung, Streuung, Beugung und Interferenz eine richtungsspezifische Übertragung von Schallereignissen. Deshalb sind neben Intensitäts- und Phasendifferenzen in binauralen Reizen, die bei seitlicher Auslenkung von Quellen entstehen, bereits im Schallsignal eines einzelnen Ohres räumliche Schallfeldmerkmale kodiert. Zur Beschreibung des akustischen Übertragungsverhaltens am Kopf werden Head-Related Transfer-Functions (HRTF) herangezogen, die richtungsabhängige Komponenten (Directional Transfer Functions, DTF) beinhalten. HRTFs können am Individuum gemessen aber auch in Simulationsmodellen berechnet werden. In verschiedenen Untersuchungen wurde deutlich, dass nur eine genaue Anpassung der Modelle an die Spezifik einzelner Hörer zu einem exakten Raumeindruck führt [CW99a, CW99b, JvSBC03]. In Abbildung 2.2 sind exemplarische HRTF-Messungen dargestellt. Neben einem als Primary Spectral Notch bezeichneten Frequenzeinbruch, der sich abhängig vom Elevationswinkel der Quelle zwischen etwa 6 und 8 kHz auswirkt, sind insbesondere im oberen Frequenzbereich komplexe richtungsspezifische Merkmale zu erkennen. Infolge der Ausprägung der spektralen Charakteristik sowohl für Elevation- als auch für Azimutwinkel müssen bei HRTF-basierten Ortungsmechanismen Mehrdeutigkeiten aufgelöst werden. Dazu schlagen VAN OPSTAL und HOFMAN vor, beim binauralen Vergleich die Spektren in Abhängigkeit von einem eindeutig kodierten Azimutwinkel zu wichten [HvO03].

Abgesehen von der gleichzeitigen Kodierung von Elevation und Azimut gibt es bei der Auswertung von HRTF-beaufschlagten Mikrofonsignalen ein weiteres, grundlegendes Problem: das empfangene akustische Spektrum geht aus der richtungsabhängigen Filterung des originalen Spektrums der Quelle hervor. Zur Lokalisation müssten dem-

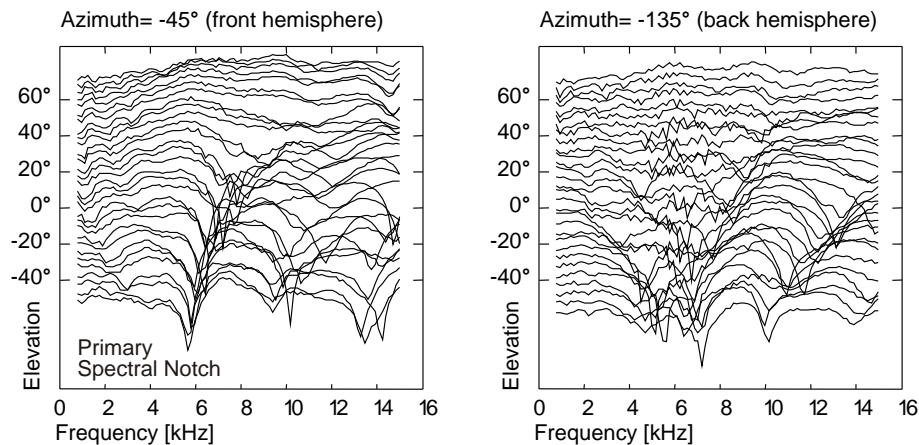


Abbildung 2.2: Head-Related Transfer Function (HRTF) eines Menschen für verschiedene Azimut- und Elevation-Winkel (nach [JvSBC03]).

nach nicht nur die Transferfunktionen der Aufnahmeeinrichtung bekannt sein, sondern ebenfalls das Quellspektrum. Dieser heikle Punkt bleibt in vielen Untersuchungen ausgespart. Zwar wurden Messungen und Simulationen von HRTFs realisiert – die Evaluation und die eigentliche Lokalisationsleistung erfolgte aber durch Versuchspersonen, denen natürliche und simulierte Stimuli mittels Kopfhörer dargeboten wurden [JvSBC03]. Die Probanden können das Dilemma des unbestimmten spektralen Ursprungs der Schallereignisse nicht anders als durch ihre Hörerfahrung lösen. Ein erwachsenes Individuum wird in seiner natürlichen Umgebung kaum mit gänzlich unbekanntem Geräuschen konfrontiert werden und kann offensichtlich anhand seiner Erwartung an vertraute Klangfarben eine Vielzahl von Geräuschen und Richtungseindrücken einordnen. Sind Geräusche hinreichend kontinuierlich, können zusätzlich auch die Änderungen des Spektrums infolge einer Kopfbewegung (sofern diese im Experiment zugelassen wird) zu Lokalisierungshypothesen beitragen. Die Komplexität solcher Wahrnehmungsleistungen sowie die angesprochene Sensitivität eines Hörers gegenüber Abweichungen der simulierten HRTFs von seinen individuellen Transferfunktionen, stellen ernsthafte Hindernisse für die Modellbildung und Simulation dar. Es erscheint wenig erfolgversprechend, ohne irgendeine akustische Referenz, etwa nur anhand des primären Frequenzeinbruches, eine HRTF-basierte Lokalisation zu bewerkstelligen.

Neben monauralen Richtungsmerkmalen können in der Natur auch binaurale Unterschiede in Amplitude, Phase und Spektrum nur anhand der HRTF-gefilterten Signale des rechten und linken Ohres bestimmt werden. Die binaurale Auswertung spektraler Eigenschaften löst zum Teil das Problem der fehlenden akustischen Referenz, da nun zwei gefilterte Instanzen des originalen Schallereignisses zur Verfügung stehen. Auch eine generelle Pegeldifferenz ist in natürlichen binauralen Signalen aussagekräftiger als

in Stereoaufnahmen ohne HRTF. Für die Bestimmung der interauralen Laufzeitdifferenz stellen richtungsabhängige Transferfunktionen hingegen ein Hindernis dar: Je mehr die spektralen Merkmale und Pegelunterschiede durch HRTFs verstärkt werden, umso geringer ist die Korreliertheit der Zeitsignale und damit die Ausprägung der interauralen Phase. Letztendlich stellt sich in der Praxis auch die Frage, ob HRTFs in konkreten technischen Lösungen, etwa durch Verwendung eines Kunstkopfmikrofons, überhaupt realisiert werden können. Im Rahmen dieser Arbeit bestand diese Möglichkeit nicht. Der erforderliche technische Aufwand sowie die noch unzureichenden Befunde über die zugrundeliegenden, komplexen neuronalen Mechanismen verhindern bislang die Einbeziehung von HRTF-basierten Merkmalen in ein biologisch motiviertes Modellkonzept. Andere, vor allem phasenbezogene Informationen, werden dadurch aber nur wenig beeinträchtigt und ermöglichen zumindest eine sinnvolle Untersuchung von Teilproblemen des räumlichen Hörens. Zur Beurteilung der Modelle und Verfahren sollte die Bedeutung der akustischen Transferfunktionen für die tatsächlichen Lokalisationsfähigkeiten von Probanden oder Versuchstieren jedoch nicht unerwähnt bleiben.

Kaum richtungsspezifische Eigenschaften besitzen der äußere Gehörgang und das Mittelohr, in dem die Schallreize via Trommelfell und Gehörknöchelchen das ovale Fenster der Cochlea erreichen. Wesentliche Funktionen dieser Bereiche des Ohres sind ein frequenzabhängiger Schalldruckgewinn und die Impedanzanpassung vom Medium Luft an die mit Flüssigkeit gefüllte Cochlea, wodurch die hohe Empfindlichkeit und der große Dynamikbereich unseres Gehörs erst ermöglicht werden [Zen94]. Unter der Annahme, dass im Mittelohr keine für die folgenden Modelle relevanten Transformationen stattfinden, bilden die ungefilterten Mikrofonssignale den Ausgangspunkt der Simulation ausgewählter Strukturen und Funktionen der zentralen Hörbahn.

### 2.2.2 Innenohr und Hörnerv

Als rezeptive Schnittstelle zwischen zentraler Hörbahn und mechanischer Schalltransformation in äußerem und Mittelohr ist das Innenohr, die Cochlea, verantwortlich für die eigentliche mechano-elektrische Transduktion der akustischen Stimuli in auditive Reizmuster. Bevor aber Schallereignisse die Haarzellen der Cochlea erreichen und durch diese Aktionspotentiale in den Ganglien des Hörnervs auslösen können, wird mit zunächst nur mechanischen Mitteln eine bemerkenswerte Transformation erreicht. Aufgrund ihrer spezifischen anatomischen Eigenschaften realisieren die cochlearen Membranen eine kontinuierliche Frequenzanalyse und legen damit die Grundlage für das fundamentale neuronale Kodierungs- und Verarbeitungsprinzip der Tonotopie.

Das tonotope Konzept, also die parallele, mit dem Spektrum des sensorischen Signals korrespondierende Organisation neuronaler Reize, ist typisch für weite Bereiche des auditorischen Systems. Sie ist Voraussetzung für die Detektion von Grundfrequenzen, das Hören von Klangfarben oder das Erkennen von Stimmen und Erfassen von Sprache. Im Zusammenhang mit dem räumlichen Hören ermöglicht die Tonotopie die Auswertung richtungsspezifischer Transferfunktionen sowie die Separation von spektral unterscheidbaren Quellen.

### Anatomische und physiologische Befunde

Wird die Cochlea über das ovale Fenster erregt, löst die Druckänderung in ihrem Inneren eine sogenannte Wanderwelle aus, die sich entlang der zur cochlearen Trennwand gehörenden Basilarmembran bewegt.<sup>3</sup> Entsprechend der Frequenzanteile des erregenden Schalls kommt es an definierten Orten der cochlearen Trennwand zur Resonanz mit einer charakteristischen Frequenz (CF). Da die Basilarmembran am ovalen Fenster steif und schmal geformt ist, während sie zum Ende, dem Helikotrema hin, breiter und plastischer wird, verändert sich in ihrem Verlauf kontinuierlich die Resonanzfrequenz. Die spektrale Zusammensetzung komplexer Schallformen kann so logarithmisch in einem Bereich von weniger als 100 Hz (nahe dem Helikotrema) bis ca. 20kHz (am ovalen Fenster) auf die Basilarmembran des Menschen abgebildet werden (Abbildung 2.3).

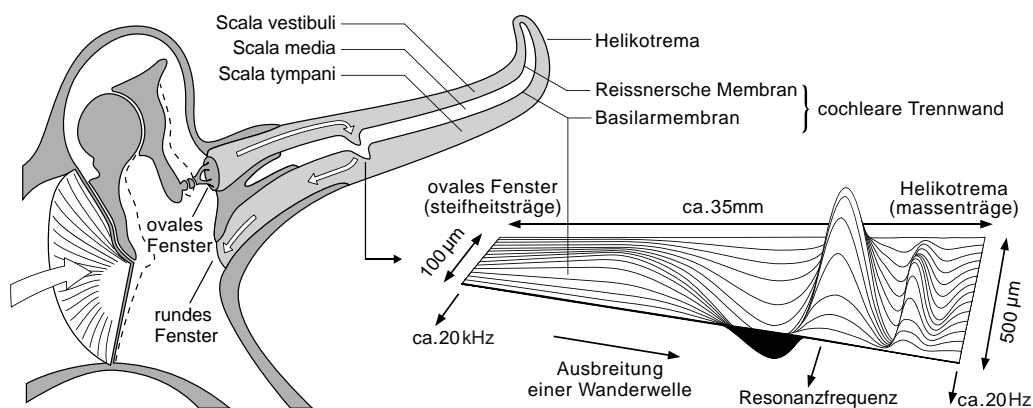


Abbildung 2.3: Schallinduzierte Auslenkung in Mittel- und Innenohr (links) und tonotopes Auftreten von Resonanzen bei der Ausbreitung einer passiven Wanderwelle in der cochlearen Trennwand (rechts) nach [Zen94].

<sup>3</sup>Die beiden Hauptkammern der Cochlea (Scala vestibuli und Scala tympani) werden durch die Reissnersche Membran und die Basilarmembran getrennt. Diese Membranen schließen wiederum die Scala media und das Cortische Organ mit den Haarzellen ein. So entsteht eine mehrschichtige, mechanische Trennwand in der Cochlea. Stellvertretend für die Schwingung dieser gesamten Trennwand wird oft nur die Auslenkung der Basilarmembran beschrieben.

Die beschriebene, schallinduzierte Wanderwelle erzwingt eine Relativbewegung der verschiedenen Bestandteile der cochlearen Trennwand, insbesondere zwischen Tektorialmembran und dem mit der Basilarmembran verbundenen Cortischen Organ. Die Haarzellen des Cortischen Organs erfahren dadurch einen dem Schallereignis adäquaten, mechanischen Reiz. Bei der exzitatorischen Reizung der Sinneshärchen der inneren Haarzellen setzen diese Neurotransmitter frei und bewirken letztendlich Aktionspotentiale (Spikes) in den korrespondierenden Ganglienzellen des Hörnervs. Die vorwiegend efferent wirkenden äußeren Haarzellen führen als aktive, cochleare Verstärker zu ausgeprägten Resonanzspitzen in der sonst passiven und flachen Wanderwelle.

Das Ergebnis der cochlearen Transformation der eindimensionalen Schallsignale in eine tonotope, d.h. ortskodierte Frequenzrepräsentation, ist mit der Frequenz–Zeit–Darstellung in einem Spektogramm vergleichbar und spiegelt sich unmittelbar in der zeitlichen und räumlichen Verteilung der afferenten Aktionspotentiale im Hörnerv wider. Der Schalldruck eines akustischen Ereignisses wird durch die mittlere Entladungsrate in den Nervenfasern kodiert, wobei die Differenz zwischen Spontanaktivität und maximaler Entladungsrate einer einzelnen Zelle nicht ausreicht, um den wahrnehmbaren Dynamikumfang von etwa 120dB abzubilden. Nachdem zuerst die Nervenfasern mit korrespondierender, charakteristischer Frequenz in einen Sättigungsbereich geraten, werden bei weiter steigendem Schalldruck auch benachbarte Fasern rekrutiert, um die Pegelinformation zu übertragen. Im Rahmen der biochemisch bedingten, refraktären Eigenschaften des Haarzellen–Ganglien–Komplexes ist eine Phasenkopplung von Entladungsfolge und kodiertem Signal bis in den kHz–Bereich hinein zu beobachten.<sup>4</sup> Diese mit dem Signal korrespondierende Periodizität der Entladungsmuster ist einerseits die Grundlage für die Repräsentation einer binauralen Phasendifferenz. Gleichzeitig dient sie der Frequenzkodierung, wenn bei mittleren und hohen Schallpegeln die Rezeptorpotentiale in einer zunehmend breitbandigen Umgebung gesättigt sind und die Tonotopie allein keine scharfe Frequenzabbildung mehr ermöglicht [Zen94].

## Modellierung

Die Fähigkeit der Cochlea zur Frequenzanalyse des eintreffenden Schalls stellt als Grundlage der tonotopen Organisation der zentralen Hörbahn die erste neurobiologi-

---

<sup>4</sup>Hier soll ausdrücklich zwischen dem Begriff der Phasenlage, wie er in der Nachrichtentechnik gebraucht wird, und Phasen bezüglich der Periodizität des Signals (positive und negative Halbwellen) unterschieden werden. Aufgrund des gleichen Phasengangs von linkem und rechtem Kanal macht die Angabe einer Phasenlage für monaurale Signale wenig Sinn. Im Folgenden bezeichnet *Phasenkopplung* die übereinstimmende Periodizität von Spikefolge und Signal, während die interaurale Laufzeit in Stereosignalen durch die *Phasendifferenz* beschrieben wird.



sche Modelloption als Alternative zur direkten Verarbeitung der Mikrofon-signale dar. Da sowohl ein breiter Frequenzbereich verarbeitet werden soll, gleichzeitig jedoch eine hohe zeitliche Auflösung nötig ist [Sie70], scheidet die zunächst nahe liegende Anwendung der schnellen Fouriertransformation (FFT) aus. In den für die Abbildung tieferer Frequenzen erforderlichen, großen Zeitfenstern wäre die Kodierung zeitlicher Merkmale wie der interauralen Laufzeitdifferenz zu träge und eine Unterscheidung zwischen Originalsignal und Echos oder Raumresonanzen schwierig. Eine elegante Alternative bietet die Verwendung spezieller Bandpass-Filterkaskaden, die online Momentanfrequenzen berechnen können. Unter der Bezeichnung Gammatone<sup>5</sup> Filter (GTF) und One-Zero Gammatone Filter (OZGF) wurden von FLANAGAN erstmals einfache Simulationsmodelle für die Schwingungsmechanik der Basilarmembran vorgeschlagen [Fla60]. Das hier benutzte Filter ist eine Implementierung des Modells von LYON, der ausführliche Studien zur Simulation der cochlearen Abstimmkurven betrieb [LM88]. Das All-Pole-Gammatone-Filter (APGF)  $n$ -ter Ordnung beschreibt er in [Lyo97] im Laplacebereich durch:

$$H(s) = \frac{K}{[(s-p)(s-p^*)]^N} \quad (2.1)$$

mit einer über die Konstante  $K$  justierbaren Verstärkung sowie dem komplexen Polpaar  $p$  und  $p^*$ . Neben der Pol-/Nullstellen Darstellung in der  $s$ -Ebene kann zur einfacheren Parametrisierung eine kartesische Notation mit der Resonanzfrequenz  $\omega_r$  und der Bandbreite  $b$  angegeben werden.

$$H(s) = \frac{K}{[(s+b)^2 + \omega_r^2]^N}; \quad K = b^2 + \omega_r^2 \quad (2.2)$$

Das Übertragungsverhalten des APGF (Abbildung 2.4a) zeigt mit seiner asymmetrischen Gestalt eine recht realistische Näherung der Tuningkurven, die bei der Auslenkung der Basilarmembran in der Cochlea zu beobachten sind. Die Möglichkeit, das Filter ohne Grundverstärkung oder Dämpfung zu betreiben ( $H(0)=1$ ), vereinfacht zudem die Kaskadierung der Bandpässe. Da Signalanteile unterhalb der charakteristischen Frequenz nicht gedämpft werden, bildet eine APGF-Kaskade in Analogie zur Cochlea zuerst Resonanzen mit hoher CF und zeitlich verzögert tiefere Frequenzen ab. Das Filter ist effizient zu berechnen, einfach zu parametrisieren und hat sich als Standardlösung für Innenohrmodelle etabliert. SLANEY stellt mit seiner Auditory Toolbox [Sla98] vielfach bewährte Implementierungen des APGF nach LYON zur Verfügung.

<sup>5</sup>Zur Berechnung von Spektrogrammen für die Spracherkennung suchte FLANAGAN ein Bandpassfilter dessen Kennlinie der Gestalt der Gamma-Verteilung (auch Erlang-Verteilung) ähnelt. Andere Autoren benutzten die Bezeichnung Gammatone zur Beschreibung der Hüllkurve der Impulsantwort des Filters.

Neben der passiven Ausbreitung der Wanderwelle modelliert das APGF mit der Ausprägung von Resonanzen implizit auch die aktive Komponente der äußeren Haarzellen [LM88, SL93]. Die Filterantwort wird als proportional zur Feuerrate der inneren Haarzellen angesehen [Sla98].

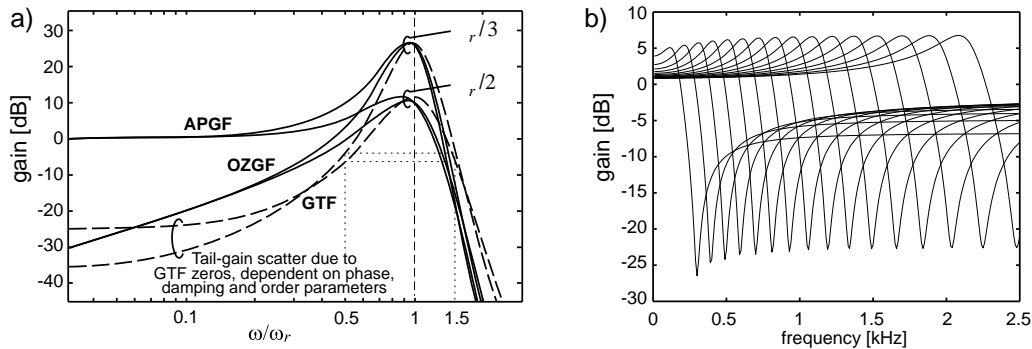


Abbildung 2.4: a) Transferfunktionen verschiedener Gammatone-Bandpässe. b) Verhalten der verwendeten APGF-Kaskade (Berechnung mit SLANEY's Auditory Toolbox [Sla98])

Im Zusammenhang mit Befunden zu frühen Lateralisationsmechanismen erfolgt die Beschreibung von Reizmustern oftmals nicht anhand der ehemals analogen periodischen Signale oder der korrespondierenden Feuerrate, sondern durch detaillierte Spikekodierungen [Col73, Col77]. Nach der Frequenzanalyse in einem adäquaten Cochleafilter ist eine binäre Kodierung der Reizmuster auf dem Niveau einzelner Spike-Zeitpunkte eine zweite wesentliche Modelloption, deren Konsequenzen auf die Lokalisationsleistung des Gesamtsystems am Ende des Kapitels diskutiert werden sollte. Um eine frühe auditorische Spikekodierung zu realisieren, ist es notwendig, das Zusammenwirken der inneren Haarzellen (mechano-elektrische Transduktion) und der primär-sensorischen, den Hörnerv innervierenden Ganglienzellen, zu simulieren. Die enge Koppelung der axonlosen Haarzellen mit den spikegenerierenden Ganglienzellen legt nahe, beide in einem vereinfachten Rezeptormodell zu integrieren. Anders als bei der Problematik der Frequenzanalyse mit Hilfe des etablierten APGF kann hier jedoch nicht auf *ein* allgemein anerkanntes Standardverfahren zurückgegriffen werden. Die Überführung der kontinuierlichen Signale in Spikefolgen kann mit sehr unterschiedlichem Aufwand bewerkstelligt werden, was an dieser Stelle die grundsätzliche Frage nach dem Abstraktionsniveau bei der Modellierung neuronaler Reizmuster aufwirft.

Als einfaches Spikemodell kann bereits die Detektion der Nulldurchgänge am Filterausgang angesehen werden. Die Schwächen dieser trivialen Lösung sind die komplett fehlende Kodierung der Signalamplitude sowie das Auftreten kaum auswertbarer Spikefolgen bei sehr geringen Pegeln und verrauschten Signalen. Ebenfalls nachteilig erweist sich der Umstand, dass tiefe Frequenzen durch extrem spärliche, hohe Fre-

quenzen dagegen durch sehr dichte Spikemuster repräsentiert werden, was im Kontext der neurobiologischen Plausibilität kaum nachvollziehbar ist. Das letztgenannte Problem bleibt auch bestehen, wenn statt der Nulldurchgänge das Überschreiten einer konstanten Reizschwelle ermittelt wird. Eine sorgfältigere Betrachtung der neurophysiologischen Gegebenheiten erscheint unumgänglich.

Mit der mehr oder weniger standardisierten mathematischen Beschreibung der künstlichen neuronalen Netze haben sich neben Ansätzen mit Feuerraten auch verschiedene spikeorientierte Modelle verbreitet. Mit diesen werden neuronale Potentialverläufe entweder eher abstrakt beschrieben (Leaky Integrate-and-Fire Modell, Spike Response Modell) oder detailliert und aufwendig über Ionenströme definiert (Hodgkin-Huxley Modell) [GK02]. Wie komplex aber ein Rezeptor- bzw. ein Neuronenmodell sein muss, um sich adäquat zu verhalten, ist letztendlich im Rahmen der Aufgabenstellung zu entscheiden. Da sich die Anforderungen an die Modelle im visuellen und auditorischen System sowie auf verschiedenen Ebenen der Merkmalsrepräsentation unterscheiden, werden im Anhang A geeignete Notationen sowohl für die Potentialverläufe von Spiekodierungen als auch für ratenbasierte Modelle angegeben.

Für die Simulation der primären Kodierung in den Hörnervenfaser müssen zunächst eine hohe temporale Auflösung sichergestellt und ferner elementare refraktäre Eigenschaften der Rezeptoren beachtet werden. Das hier vorgeschlagene Modell einer Rezeptorzelle interpretiert den Ausgang des APGFs ihres Frequenzbandes als Generatorpotential und reagiert bei Erreichen einer Reizschwelle mit einem Spikeimpuls. Das Refraktärverhalten des Rezeptors wird durch den zeitlichen Verlauf eines After-Hyperpolarization Potentials (AHP):  $g(t) = e^{-\frac{t}{\tau}}$  definiert. Über die Zeitkonstante  $\tau$  in der Abklingfunktion der AHP kann schließlich eine maximale Spikerate in der Hörnervenfaser und damit der Grad der Phasenkopplung der resultierenden Spikefolge eingestellt werden. Im Modell wird eine Kodierung vereinfachend als phasengekoppelt bezeichnet, wenn jede Periode eines hinreichend starken Reizes mit charakteristischer Frequenz (CF) zur Auslösung mindestens eines Spikes führt. In Fasern mit höheren CF kodieren Spikes die Periodizität des Reizes nicht mehr eindeutig, da mehrere positive Signalphasen auftreten können, bevor das abklingende AHP der Rezeptorzelle ein erneutes Feuern zulässt. Das Maß der Phasenkopplung im Hörnerv scheint mit dem durch interaurale Zeitunterschiede ortbaren Frequenzbereich zu korrespondieren. Das Refraktärverhalten der modellierten Rezeptoren wurde deshalb so gewählt, dass im höchsten Frequenzband (2.5 kHz) gerade noch phasengekoppelte Spikefolgen zu beobachten sind. Neben der Abbildung von Frequenzkomponenten und binauralen Phasendifferenzen gewährleistet das so parametrisierte Modell außerdem die Kodie-

zung von Amplitude und Hüllkurve über die mittelfristige Spikerate. Im Gegensatz zum trivialen Nullstellen- oder Schwellen-Verfahren bleibt dabei die Proportionalität von Schallenergie und neuronaler Aktivierung in allen Frequenzbereichen erhalten.

Die Simulation der Rezeptorzellen erfolgte mit Hilfe der in Anhang A definierten zeitdiskreten Filter. Die Präzision der berechneten Spike-Zeitpunkte korrespondiert folglich mit dem Simulationstakt, der einheitlich für Lyonfilter und Rezeptormodell an die Samplingrate der mit 44,1 kHz digitalisierten Audiosignale gekoppelt wurde.

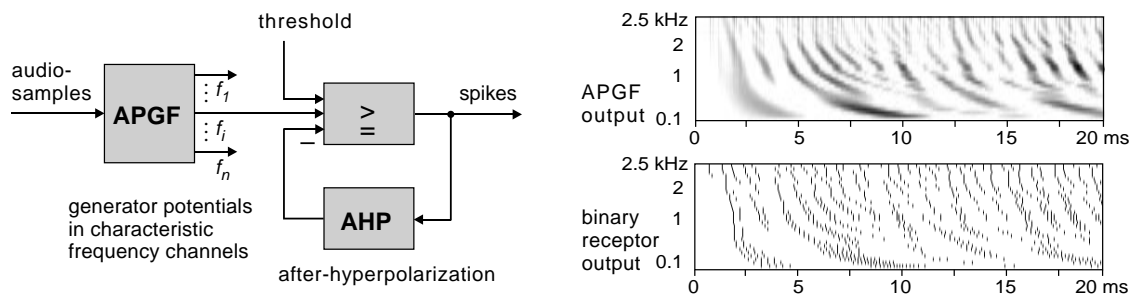


Abbildung 2.5: Schematische Darstellung des implementierten Innenohrmodells mit APGF-Filter für die Frequenzanalyse und Haarzellensimulation zur nachfolgenden Spiekodierung. Die Diagramme zeigen ein exemplarisches Spektrogramm des Innenohrfilters (oben) und dessen Spiekodierung im Rezeptormodell (unten).

### 2.2.3 Auditorische Pfade in Hirnstamm und Mittelhirn

Die im Hörnerv auftretenden Spikemuster bilden die gemeinsame Grundlage für alle auditorischen Verarbeitungsleistungen. Neben der in der Cochlea initiierten, tonotopen Organisation sowie der Kodierung von Phase und Amplitude durch Spikezeitpunkt und Spikerate sind auf diesem frühen neuronalen Niveau noch keine Repräsentationen komplexerer Signaleigenschaften zu finden. Im Unterschied zu anderen sensorischen Systemen werden in der zentralen Hörbahn solche primären sensorischen Reizmuster nicht direkt in kortikale Areale projiziert, sondern über mindestens fünf bis sechs nacheinander geschaltete Neurone vorverarbeitet. Dabei ist allgemein eine zunehmende Spezialisierung der Neurone auf immer komplexere Signaleigenschaften zu beobachten. Tatsächlich werden nur abstrakte Informationen wie die spektrale Zusammensetzung, Amplituden- und Frequenzmodulationen oder auch Schallrichtungen einer kortikalen Beurteilung zugeführt [Zen94].

Die anatomische und funktionale Spezialisierung der sensorischen Neurone beginnt im Projektionsgebiet des Hörnervs, dem Nucleus cochlearis, der eine im Bereich des Hirnstammes massiv divergente Projektion initiiert. Noch im Hirnstamm findet eine parallele Extraktion und Repräsentation auditorischer Merkmale in morphologisch klar

unterscheidbaren, neuronalen Kernen statt. Das dabei eingesetzte Repertoire neuronaler Verarbeitungsmechanismen beinhaltet komplexe rezeptive Felder, exzitatorische und inhibitorische Wechselwirkungen innerhalb und zwischen Arealen sowie kontralaterale und selbst efferente Verschaltungen. Im Gebiet des Inferior Colliculus konvergieren schließlich Projektionen von mehr als 20 identifizierten Neuronentypen aus zehn auditorischen Kernen des Hirnstammes [Irv92], bevor über den Thalamus in die kortikalen Projektionsfelder innerviert wird. Aus Sicht der Informationstechnik und Informatik fordern die morphologischen und funktionellen Befunde über Zelltypen und die relativ detaillierte Kenntnis der Projektionswege zwischen den neuronalen Arealen den Entwurf von modularen Simulationsmodellen geradezu heraus. Neben der Betrachtung der neuronalen Kerne als abgegrenzte Module werden die Beschreibungen der Projektionswege als Pfade oft im Sinne serieller und damit leicht zu berechnender Verarbeitungsketten adaptiert. Beispielhaft geschieht dies zur Untersuchung der Computational Maps der räumlichen Schallinformationen.

Im Verlauf intensiver Forschungen am auditorischen System der Eule etablierte KONISHI das strukturell und funktionell außergewöhnlich klar gegliederte Modellkonzept von Timing und Intensity Pathway. Im Nucleus angularis (NA) der Eule ist eine Amplitudenkodierung zu beobachten – der Ausgangspunkt des Pfades, in dem intensitätskodierte Richtungsmerkmale verarbeitet werden. Er verläuft auf relativ direktem Weg zum kontralateralen Inferior Colliculus (IC) und mündet dort in einer Abbildung der vertikalen Schallrichtung. Im Nucleus magnocellularis (NM) beginnt die phasengenaue Kodierung des Timing Pathway, der zunächst eine bilaterale Verknüpfung über den Nucleus laminaris (NL) herstellt, um dann jeweils kontralateral in einer Region des IC eine Karte horizontaler Richtungen zu projizieren. Im externen Inferior Colliculus (ICx) treffen beide Pfade zusammen und ermöglichen eine durch Elevation und Azimut definierte, komplette räumliche Abbildung [Kon93]. Das Konzept der auditorischen Pfade der Eule wurde als Vorbild zahlreicher Simulationsmodelle aufgegriffen, die LAZZARO bis zur Hardwareimplementierung der Komponenten des Timing-Pfades verfolgte [LM95]. Auch RUCCI verweist in der Beschreibung seines Demonstrators zur multisensorischen Ortung explizit auf KONISHIs Modell, wenngleich er ebenfalls nur zeitliche Merkmale berechnet [RWE00]. Während sich die Implementierungen von LAZZARO und RUCCI ausdrücklich auf die Neuroanatomie der Eule beziehen, drängt sich gleichwohl die Frage auf, ob die Befunde spezifisch für diese als nachtaktiver Jäger hochspezialisierte Tierart sind. Eine Verallgemeinerung des Konzepts für das Lokalisationsvermögen anderer Tierklassen ist hier auch deshalb interessant, weil die in Kapitel 4 erörterten Mechanismen der multisensorischen Integration häufig an Säugetieren un-

tersucht werden [SM93, WWS96]. KONISHI, als einer der Begründer des Timing und Intensity Pathway Modells, versuchte schließlich selbst eine Gegenüberstellung der Befunde der Eule mit dem auditorischen System des Menschen [Kon00]. Ein Vergleich der Kerne NA und NM mit den ventralen Bereichen des Nucleus cochlearis sowie die Gegenüberstellung der Timecode–Areale NL mit dem medial–superioren Nucleus olivaris (MSO) der Säuger erscheint anhand korrespondierender Antwortmuster legitim. Der

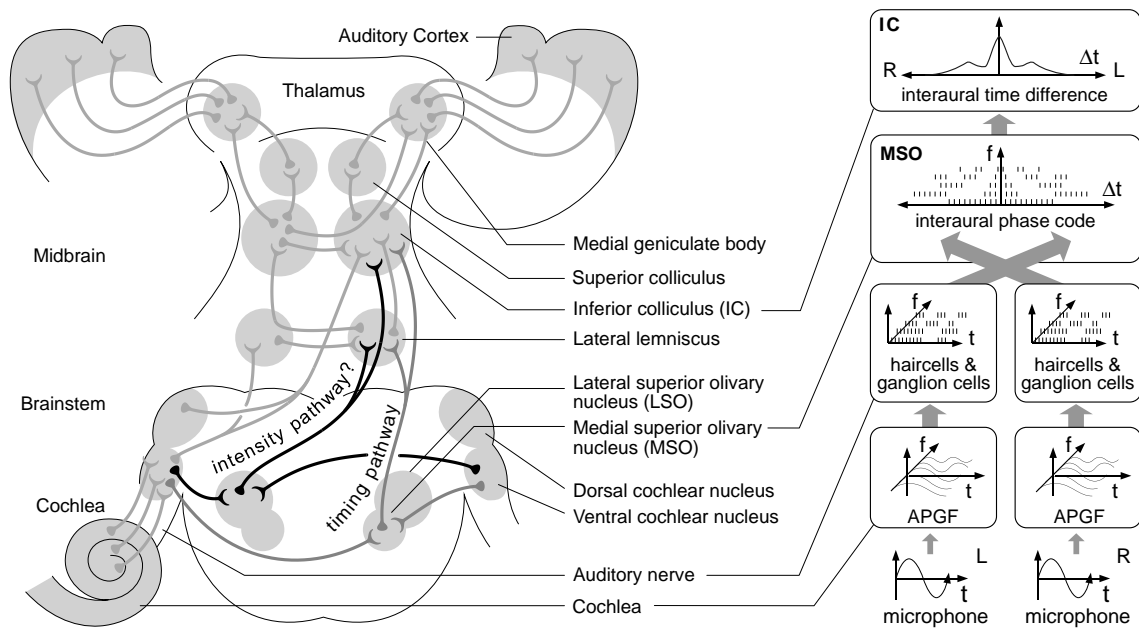


Abbildung 2.6: Links: Wesentliche afferente Komponenten der Hörbahn bei Säugern (nach [Zen94] und [HA97]). Die nur einseitig dargestellten Verknüpfungen existieren generell auch in lateral gespiegelter Form. Unter Beachtung der abweichenden Projektionswege im ursprünglichen Timing–Intensity–Konzept KONISHIS lassen sich die kontralateralen Bahnen zwischen den ventralen cochlearen Kernen und MSO und die folgende ipsilaterale Projektion in den zentralen IC als Timing Pathway interpretieren. Der über den ipsilateralen LSO zum kontralateralen IC verlaufende Weg hat den Charakter eines Intensitäts–Pfad. Vereinfachend kann dem Timing–Pfad die Auswertung binauraler Phaseninformationen und die Abbildung horizontaler Schallrichtungen zugeschrieben werden. Im mutmaßlichen Intensity–Pfad tragen binaurale Amplitudendifferenzen zur Generierung von horizontalen *und* vertikalen Abbildungen bei. Weitere afferente Verknüpfungen der cochlearen Kerne bestehen zu den ipsilateralen Nuclei lemnisci lateralis (NLL) und zum kontralateralen Inferior Colliculus (IC). Auch auf Ebene von NLL und IC sind kontralaterale Verbindungen zu finden. Zu erkennen ist außerdem eine deutliche Konvergenz im Inferior Colliculus, der in den multisensorischen Superior Colliculus und über den Corpus geniculatum mediale (MGB) letztendlich in den auditorischen Kortex projiziert. Für weitere abstrakte Pfadmodelle ist eine Untergliederung in binaurale und monaurale bzw. in tonotope und nicht–tonotope Merkmale möglich.

Rechts: Nach der monauralen Frequenzanalyse und Spikekodierung sind ausgewählte Funktionen des Timing–Pfad in den folgenden Abschnitten 2.2.4–2.2.6 Vorbild für die Komponenten des Simulationsmodells zur binauralen, laufzeitbasierten Schallortung.

weitere Verlauf der Pfade weist allerdings einige Unterschiede auf: die vorrangig zeitbasierten Merkmale des MSO werden in den ipsilateralen IC projiziert, der Timing-Pfad der Eule kreuzt dagegen vom NL zum kontralateralen IC. Die binaurale Auswertung der Amplitudeninformation erfolgt bei Säugern bereits im superior-olivaren Komplex, also auf der gleichen Ebene, auf der auch die primären Phaseninformationen binaural verknüpft werden. Aufgrund der direkten Projektion von NA zum IC bewertet die Eule binaurale Pegelunterschiede erst mit Hilfe einer kontralateralen Verknüpfung ihrer IC-Areale. Die Ursache dieser unterschiedlichen Projektionswege scheint tatsächlich eine ausgeprägte Spezialisierung im Orientierungsvermögen der Eule zu sein, die dazu führte, dass Azimut und Elevation jeweils exklusiv durch binaurale Zeit- bzw. Pegeldifferenzen kodiert werden. Demgegenüber liefert bei Säugern auch die Amplitudenauswertung einen Beitrag zur Lateralisation und ist durch die Einbettung im SOC weniger scharf von der Verarbeitung zeitbasierter Informationen getrennt. Dieser Befund erschwert insbesondere eine am Hirnstamm der Säuger orientierte Modellierung der intensitätsbasierten Lokalisationsmechanismen. Die strikte Trennung zwischen Timing und Intensity Pathway in KONISHIS Darstellungen kann demnach nicht als universelle auditorische Strategie angesehen werden. Dennoch ist die in Abbildung 2.6 veranschaulichte konzeptionelle Unterscheidung von vorrangig phasen- oder amplitudenbasierten Mechanismen im Sinne abstrakter Verarbeitungspfade plausibel.

## 2.2.4 Spezifische Kodierungen im Nucleus cochlearis

### Anatomische und physiologische Befunde

Der Nucleus cochlearis (CN) ist das erste Kerngebiet der zentralen Hörbahn. Hier enden sämtliche primär-auditorischen Axone des Hörnervs. Als Ursprungsort der auditorischen Pfade nimmt der CN eine Umkodierung der primären Reizmuster des Hörnervs vor und initiiert die Verarbeitung spezifischer Signalmerkmale. Unter Beibehaltung der tonotopen Organisation lassen sich eine Reihe von Subarealen mit differenzierten neuronalen Erregungsmustern unterscheiden – ein Befund, der aus dem Vorhandensein vielfältiger Zelltypen und einer komplexen synaptischen Organisation resultiert [Zen94, RA97]. Nach anatomischen und funktionalen Gesichtspunkten werden drei Hauptgebiete unterteilt. Während Hörnervfasern mit tiefen charakteristischen Frequenzen (CF) vorrangig im anteroventralen Gebiet enden (AVCN), innervieren die basalen Fasern (mit hohen CF's) auch einen posteroventralen Teil (PVCN) sowie einen dorsalen Kernbereich (DCN) [RA97]. Da für die auditorischen Lokalisationsmechanismen die Auswertung der unteren Frequenzbereiche unverzichtbar ist, kann bereits

aufgrund dieses Befundes eine Beteiligung des AVCN an der Kodierung ortungsrelevanter Merkmale vermutet werden. Um die Bedeutung verschiedener CN-Bereiche im Sinne des Timing vs. Intensity Konzeptes zu klären, ist jedoch eine genauere Charakterisierung der Neurone und Subareale erforderlich, wozu dank umfangreicher, elektrophysiologischer Ableitungen detaillierte Befunde zur Verfügung stehen.

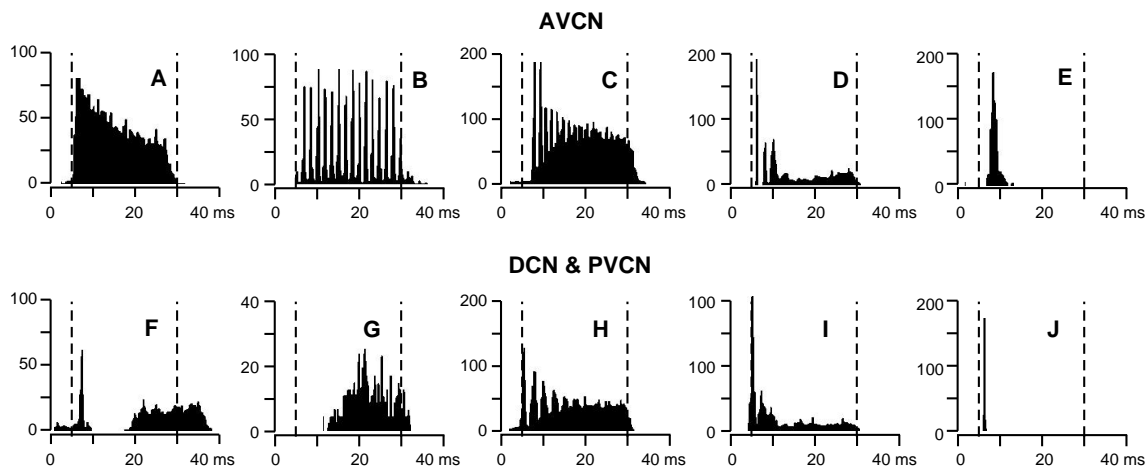


Abbildung 2.7: Post-Stimulus-Zeit-Histogramme typischer Erregungsmuster im cochlearen Kern als Antwort auf einen 25ms Ton-Burst [RA97]. Im anteroventralen cochlearen Kern (AVCN) findet man A: *Primary like*, B: *Phase locked*, C: *Sustained chopper*, D: *Onset chopper*, E: *Onset response*. Im posteroventralen (PVCN) und dorsalen (DCN) Teil sind F: *Pauser*, G: *Buildup*, H: *Sustained chopper*, I: *Onset sustained* und J: *Onset transient response* anzutreffen.

Auskunft über Amplituden- und Phasenkodierung geben die in Abbildung 2.7 dargestellten *Post-Stimulus-Zeit-Histogramme* (PSTH). Von einigen Zellen werden die Reize aus dem Hörnerv offensichtlich ohne erkennbare Manipulation weitergegeben (*Primary like Response*). Andere kodieren relativ unabhängig von der Amplitude sehr genau die Phase eines Stimulus (*Phase locked Response*), was allerdings nur für hinreichend niedrige Frequenzen möglich ist. Weitere typische Antwortmuster sind die *Onset-Form* und die von der Stimulusfrequenz unabhängige *Chopper Response*. Die komplexeren *Pauser*- und *Buildup*-Muster sind ausschließlich im dorsalen CN zu beobachten. PST-Histogramme werden in Bezug auf konkrete auditorische Merkmale nur sehr vorsichtig interpretiert. Prinzipiell scheinen *Phase locked* und *Onset Response* unter anderem für die Kodierung interauraler Zeitdifferenzen verantwortlich zu sein, während phasenunabhängige *Primary*- und *Chopper*-Antworten die Amplitudeninformation übertragen. *Buildup* und *Pauser Response* sind möglicherweise sensitiv gegenüber Frequenz- und Amplitudenmodulationen [RA97].

Die PST-Histogramme liefern alleine nur ein unvollständiges Bild der Spezifik der Responseformen und geben insbesondere keinen Aufschluss über die spektrale Gestalt



einer Aktivierung. Neben der allgemeinen, tonotopen Organisation ist aber auch die Frage von Interesse, ob und wo die eher breitbandige Aktivierung der Hörnervenfaser in eine detailliertere, spektrale Repräsentation überführt wird. Ein Indiz für eine schärfere Abbildung spektraler Merkmale sind laterale, inhibitorische Verschaltungen innerhalb des tonotopen Verbundes, deren Wirkung sich in Frequenz-tuning-Kurven veranschaulichen lässt. Setzt man schließlich die zeitlichen Response-Muster im PSTH und die Tuning-Kurven in Bezug zu den Projektionsgebieten der CN-Zellen (siehe Abbildung 2.8), wird die Rolle der CN-Subareale bei der Initiierung der auditorischen Pfade transparenter.

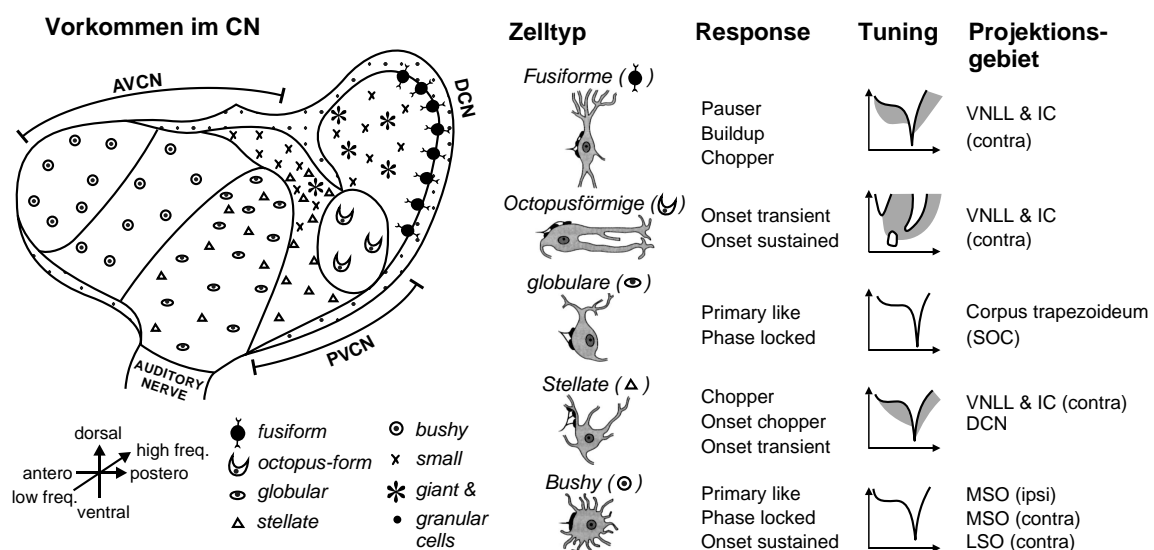


Abbildung 2.8: Verbreitung, Eigenschaften und wesentliche Projektionen der Zelltypen im CN nach [Rou97, DMS96, RG92]. Schraffierte Bereiche in den Tuning-Diagrammen indizieren einen inhibitorischen Einfluss benachbarter Frequenzen.

Zunächst fällt auf, dass die ausschließlich im AVCN vorkommenden globularen und Bushy-Zellen als einzige in den superior olivaren Komplex (SOC) projizieren und die ausgeprägteste Primary like und Phase locked Response zeigen. In Anbetracht seiner tonotop bedingten Präferenz niedriger Frequenzen, der Antwortmuster und der Verbindung zum SOC kodiert der AVCN zweifellos ortungsrelevante Informationen [RG92]. Als wesentlicher Ausgangspunkt für den Timing- und den Intensity-Pfad des auditorischen Lokalisationssystems überträgt er die Reizmuster des Hörnervs schnell und ohne aufwendige Manipulation. Da gerade bei globularen und Bushy-Zellen kaum inhibitorische Wechselwirkungen mit benachbarten Frequenzbändern bestehen, scheint die binaurale Amplituden- und Phasenauswertung im SOC noch keine scharfe Abbildung des Spektrums vorauszusetzen. Weniger eindeutig sind die Befunde im posteroventralen und dorsalen CN, in denen zunehmend höhere Frequenzen und komplexere neu-

ronale Muster übertragen werden. Sowohl die Ausprägung inhibitorischer Frequenzbänder als auch die Fähigkeit, Amplituden und Frequenzmodulationen zu kodieren, setzen eine komplexere synaptische Organisation als im AVCN voraus. Es liegt nahe, im PVCN eine Kodierung spektraler Richtungsmerkmale zu vermuten, die infolge der kopfbezogenen Transferfunktionen entstehen.

### **Modellierung**

Abgesehen von den komplizierten Buildup- und Pauser-Mustern sind für die Simulation der Responseformen im CN keine aufwendigen Algorithmen erforderlich. Bereits das in 2.2.2 vorgestellte Rezeptormodell zeigt mit einer typischen Parametrisierung Phase locked Response bis etwa 2.5kHz und Chopper Response für höhere Frequenzen. Eine bewusst übermäßig groß gewählte, refraktäre Zeitkonstante kann auch eine Onset-Kodierung erzwingen. Zur Nachbildung des tonischen und phasischen Anteils in Primary- und Chopper-Antworten wäre die Berechnung eines zweiten Potentials erforderlich, das parallel zur Steuerung einer absoluten Refraktärphase durch das schnell abklingende AHP noch eine mittelfristige Anhebung der Reizschwelle modelliert. Die in einigen Tuning-Kurven sichtbaren Off-on-off Gebiete innerhalb der tonotopen Organisation können in beliebiger Gestalt durch lateral inhibitorische Projektionen in die Zielregionen realisiert werden [Sha85, WBS96]. Die Funktion des PVCN und DCN als tonotope Hochpassfilter scheint hauptsächlich für die spektral basierte Objekterkennung, wie die Verarbeitung von Lauten und Sprache, eine Rolle zu spielen. Entgegen den zitierten Befunden wurde eine schärfere Abbildung des Spektrums aber auch in binauralen Lokalisationsmodellen angestrebt [SSG89].

Für die nachfolgende Simulation binauraler Ortungsmechanismen wird auf ein separates AVCN-Modell verzichtet (s. Abbildung 2.6). Die hier relevanten elementaren Reizmuster unterscheiden sich nur wenig von der Kodierung im Hörnerv und können allein mit Hilfe der spezifizierten Rezeptoren erzeugt werden.

## **2.2.5 Extraktion räumlicher Informationen im Hirnstamm**

### **Neuroanatomische und physiologische Befunde**

Der obere olivare Komplex (SOC) bildet ein Cluster auditorischer Kerne im Hirnstamm, der afferent vom anteroventralen cochlearen Kern (AVCN) sowie efferent vom Inferior Colliculus (IC) angesprochen wird. Sein afferenter Input erfolgt bilateral, wodurch erstmals im aufsteigenden Verlauf der Hörbahn Informationen von rechtem und linkem Ohr miteinander verknüpft werden können [HA97, GZ96]. Diese Position prä-

destiniert den SOC, grundlegende auditorische Merkmale für das räumliche Hören zu extrahieren, wobei auf dem Niveau des Hirnstammes Einfallsrichtungen des Schalls detektiert werden, noch nicht jedoch sein exakter Ursprungsort [Zen94]. Die konkrete Ausprägung der binauralen Strukturen in verschiedenen Tierklassen erfordert eine spezifische Auslegung des Modellkonzeptes von Intensity und Timing Pathway. Während KONISHI eine der spezialisierten Lebensweise der Eule geschuldete Dominanz interauraler Zeitdifferenzen (ITD) beschreibt, wertet der SOC der Säuger gleichzeitig auch Intensitätsunterschiede (IID) aus.<sup>6</sup>

In der afferenten Hörbahn lassen sich drei Hauptabteilungen des SOC unterscheiden: der mediale Nucleus des Trapezkörpers (MNTB), der laterale Nucleus olivaris superior (LSO) sowie der mediale Nucleus olivaris superior (MSO). Jeder dieser Kerne weist intern eine klare, tonotope Organisation auf. In der efferenten Bahn sind die periolivaren Kerne mit Projektionen in den dorsalen CN und bis ins cortische Organ der Cochlea zu finden.

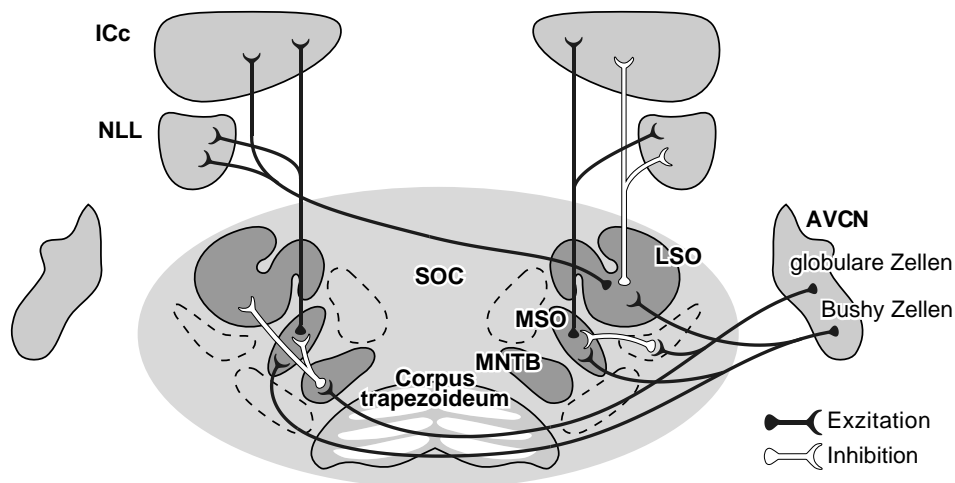


Abbildung 2.9: Afferente Komponenten des SOC nach [HA97].

### *Medialer Nucleus des Corpus trapezoideum (MNTB)*

Der MNTB wird afferent von globularen und Bushy-Zellen des kontralateralen AVCN exzitatorisch erregt. Er zeigt eine präzise Tonotopie mit niedrigen CFs im lateralen und höheren Frequenzen im medialen Bereich [HA97]. Die häufigsten im MNTB anzutreffenden Neurone, seine Principal cells, zeichnen sich durch einfache Primary like Response aus, wie sie bereits für AVCN oder Hörnerv beschrieben wurde. Die Principal

<sup>6</sup>Auch ohne explizite Bezüge in einigen der zitierten Quellen kann in den hier geschilderten Befunden vom allgemeinen Fall der simultanen Verarbeitung von ITD- und IID-Merkmalen im Säugerhirnstamm ausgegangen werden.

Neurone des MNTB projizieren ipsilateral in den LSO und den MSO, wo sie inhibitorische Wirkung zeigen. Aufgrund seines kontralateralen, exzitatorischen Inputs realisiert der MNTB in seinem Projektionsgebiet eine vom CN ausgehende kontralaterale Inhibition, die vermutlich auch die Lokalisationsleistung des SOC beeinflusst.

#### *Nucleus olivaris lateralis superior (LSO)*

Der LSO ist einer der beiden Kerne, die im superior-olivaren Komplex für die Detektion binauraler, auditorischer Merkmale verantwortlich gemacht werden. Er zeigt bei Säugetieren eine mehrfach gefaltete Form, dabei aber eine durchgehend tonotope Organisation. Seine vorwiegend fusiformen, bipolaren Neurone erhalten exzitatorische Reize direkt von den Bushy-Zellen des ipsilateralen AVCN, wogegen sie vom kontralateralen CN über den erwähnten Umweg des MNTB gehemmt werden. Sie antworten phasisch oder mit Chopper Response bevorzugt auf ipsilaterale Stimulation und besitzen im Vergleich zu anderen superior-olivaren Zellen schmalbandige Abstimmkurven. Da den Bushy-Zellen im AVCN ein eher breitbandiges Verhalten bescheinigt wurde, könnte die genaue Frequenzabstimmung erst durch eine lateral inhibitorische Projektion oder Wechselwirkung innerhalb der tonotopen Organisation des LSO entstehen. In Abhängigkeit von Intensität und Onset-Dauer des kontralateralen Stimulus ist eine monoton zunehmende Hemmung der Neurone zu beobachten. Dieser Effekt wird durch die Kombination von kontralateral-inhibitorischen und ipsilateral-exzitatorischen rezeptiven Feldern (RF) mit gleicher Bestfrequenz verursacht. Es entstehen sogenannte *IE-Units*, in deren Abstimmkurven eine vollständige Überlappung der ipsi- und kontralateralen RF sichtbar ist [HA97]. Als Detektoren binauraler Informationen sind die bipolaren LSO-Neurone sensitiv gegenüber transienten Zeitdifferenzen (ITD) und einsetzenden Intensitätsunterschieden (IID). Die im LSO konzentrierten IE-Units sind insbesondere für die IID-basierte Lokalisation hochfrequenter Schallanteile von Bedeutung [Irv92]. Tatsächlich weist der gesamte LSO-Komplex eine Betonung höherer Frequenzen auf und ist bei Spezies besonders gut entwickelt, deren Gehör auf hohe Töne oder sogar Ultraschall spezialisiert ist [HA97, GZ96].

LSO-Neurone projizieren in tonotoper Organisation bilateral: inhibitorisch in den ipsilateralen Lemniscus lateralis und exzitatorisch in den kontralateralen Inferior Colliculus centralis (ICc) [HA97].

#### *Nucleus olivaris medialis superior (MSO)*

Zwischen MNTB und LSO befindet sich der vor allem bei Säugetieren gut entwickelte MSO-Kern. Auch der MSO ist an der Lokalisation von Schallereignissen beteiligt,

zeigt dabei aber eine Spezialisierung auf den, auch vom Menschen gut ortbaren, unteren Teil des akustischen Spektrums [HA97]. In seiner tonotopen Organisation wurde eine überproportionale Repräsentation niedriger Frequenzen im dorsalen Bereich nachgewiesen [GZ96]. Ähnlich wie im LSO sind auch im MSO vorrangig bipolare Zellen zu finden, die bilateral vom AVCN innerviert werden und außerdem unter inhibitorischem Einfluss des MNTB stehen. Allerdings bilden die meisten unter ihnen bilateral exzitatorische Einheiten (EE-Units) und werden von der größtenteils phasengekoppelten Primary-like Response des AVCN beider Seiten erregt. EE-Units reagieren stärker auf interaurale Zeit- als auf Intensitätsunterschiede, was nicht nur ihrer Exzitation durch rechten *und* linken CN, sondern vor allem ihrer Anatomie zuzuschreiben ist. Aussagekräftige Befunde für ein neuronales Substrat zur Detektion von ITDs ergaben erstmals die Untersuchungen am auditorischen System der Eule [Kon93], deren anatomische Gegebenheiten im Nucleus laminaris an dieser Stelle auf den MSO der Säuger übertragbar sind [Irv92].

Die Neurone des besagten Lamellenkerns werden bilateral von entgegengesetzten Seiten erregt. Ipsilaterale, afferente Fasern innervieren von dorsal nach ventral, kontralaterale Axone verlaufen umgekehrt. Die ipsilaterale Pfadlänge zu einem bipolaren Neuron und die damit verbundene Verzögerung seiner Erregung erhöht sich mit zunehmend ventraler Position. Gleichzeitig verringert sich die kontralaterale Ansprechzeit. Die EE-Units werden schließlich am stärksten erregt, wenn Reize von beiden Seiten gleichzeitig eintreffen. Durch die beschriebene Verzögerung und die phasengekoppelte Kodierung ist dies nur bei bestimmten interauralen Phasendifferenzen der Fall. Je nach Abstand der beiden Ohren kodieren Phasendifferenzen in hinreichend niedrigen Frequenzbändern aber genau die durch eine laterale Auslenkung der Quelle entstandene interaurale Zeitdifferenz.

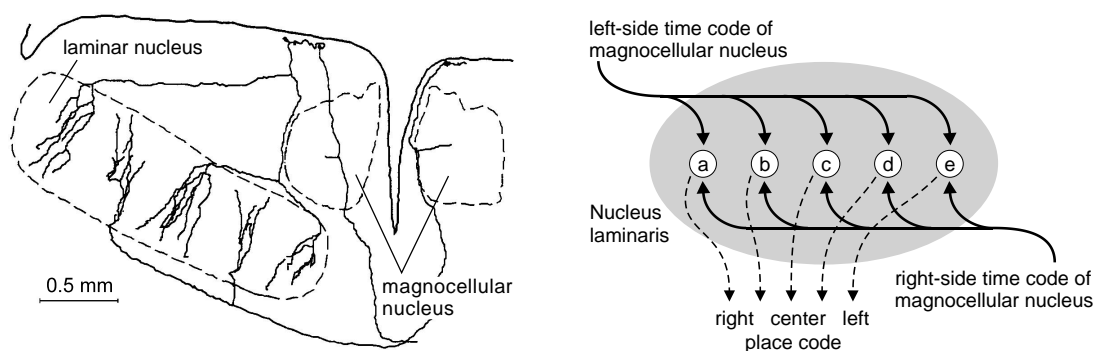


Abbildung 2.10: Neuroanatomischer Befund und modellhafte Darstellung der Detektion von interauralen Zeitdifferenzen (ITD) im Hirnstamm der Eule nach [Kon93].

Die geschilderte, anatomische Struktur existiert parallel für verschiedene charakteristische Frequenzen und korrespondiert mit dem Koinzidenzmodell von JEFFRESS [Jef48, YC90, Irv92, HA97]. Neben der Sensitivität für eine charakteristische Frequenz im Rahmen der tonotopen Organisation besitzen sowohl die Neurone des Lamellenkerns der Eule als auch die Zellen im MSO der Säuger ein charakteristisches Delay (CD). Diese bevorzugte Verzögerung ist gleichbedeutend mit der interauralen Zeitdifferenz und unabhängig vom Frequenzband [HA97]. Der horizontale Raumwinkel (Azimut), der sich im zeitlichen Versatz der phasengekoppelten AVCN-Muster manifestiert, wird im MSO in einen Ortscode umgewandelt und orthogonal zur tonotopen Organisation topologisch abgebildet [DMS96].

Die in Abbildung 2.11a) sichtbare Mehrdeutigkeit in der Reaktion einer MSO-Zelle wird durch die Periodizität ihres phasengekoppelten Inputs aus dem AVCN hervorgerufen. Die Periodendauer des Stimulus ist für hohe Frequenzen klein gegenüber den zu detektierenden ITDs und führt zur Ausbildung mehrerer Erregungsmaxima entlang der dorsal-ventralen Verzögerungsachse. Aus eben diesem Grund ist die Kodierung von ITDs durch EE-Units auf einen unteren Frequenzbereich beschränkt. Neurone mit höherer CF bilden verstärkt Kombinationen von exzitatorischen und inhibitorischen Feldern aus und werden auf ähnliche Weise sensitiv für interaurale Intensitätsunterschiede, wie es im LSO zu beobachten war. Dieser Befund ist ein Indiz dafür, dass die Schallortung niedriger Töne stärker an die Detektion von temporalen Merkmalen gebunden ist, die Lokalisation hochfrequenter Schallanteile dagegen eher auf einer Intensitätsauswertung basiert. Die Beobachtung, dass bei Tieren mit der Fähigkeit zur Ultraschallortung die oberen Frequenzbereiche im LSO besser entwickelt sind, bei den

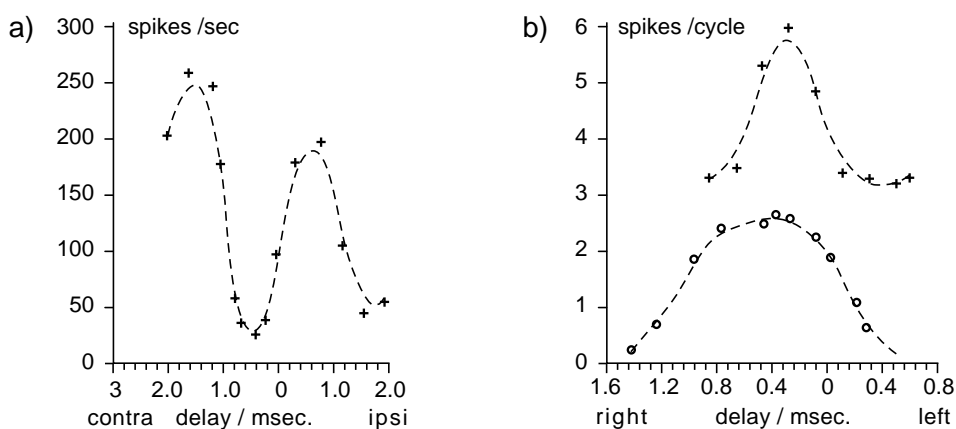


Abbildung 2.11: Aktivierung in ITD-sensitiven Neuronen des MSO. a) Periodische Änderung der Spikerate als Funktion der interauralen Verzögerung, b) gleiches Delay bei unterschiedlicher Lateralisationsschärfe in verschiedenen Frequenzbändern (nach[HA97]).

meisten Säugern mit konventionellem Gehör dagegen die tiefen Frequenzen im MSO, stützt diese Theorie. Zu erwähnen bleiben die afferenten Projektionen des MSO, die, wiederum tonotop organisiert, hauptsächlich ipsilateral direkt oder über den Lemniscus lateralis in den zentralen Inferior Colliculus reichen (vergl. Abbildung 2.6).

## Modellierung

Als Resümee der Diskussion der komplexen Strukturen im SOC kann die Detektion elementarer Richtungsmerkmale in Form binauraler Intensitäts- und Zeitdifferenzen unterstrichen werden. Befunde über die gleichzeitige Kodierung dieser beiden Merkmale im LSO zeigen, dass deren Verarbeitung abweichend von der Intensity vs. Timing Doktrin im Detail nicht vollkommen unabhängig voneinander geschieht. Die ITD-basierten Mechanismen profitieren von einem direkten physikalischen Zusammenhang zwischen interauraler Laufzeit und horizontaler Schallrichtung. Im Gegensatz zur plausiblen Erklärung der ITD-Detektion durch neuronale Delay Lines und Koinzidenz-Zellen im MSO sind die Aussagen zur Bewertung von Intensitätsmerkmalen weniger konkret. Möglicherweise vereitelt die besondere Beeinflussung der Amplitudeninformation durch kopfbezogene Transferfunktionen (HRTF) eine einfache Zuordnung zwischen interauraler Pegeldifferenz und Einfallsrichtung der Schallereignisse. Die zitierte, detailliert-tonotope Abbildung im vergleichsweise schmalbandig abgestimmten LSO legt nahe, dass neben der monauralen auch die binaurale Bewertung von Intensitätsmerkmalen nicht anhand allgemeiner Pegel, sondern durch einen spektralen Vergleich bewerkstelligt wird. Für eine derartige Verarbeitungsleistung müssten folglich korrespondierend zur spektralen Richtcharakteristik der HRTF räumliche und tonotope Register kodiert werden. Da im Rahmen der in Kapitel 5 diskutierten Experimente keine Kunstkopfaufnahmen realisiert werden konnten, war eine solche Personalisierung der IID-Kodierung physikalisch ausgeschlossen.

Auch wenn konkrete, biologisch inspirierte IID-Modelle auf technische Anordnungen mit akustisch frei positionierten Mikrofonen nicht anwendbar sind, ist doch die Frage von Interesse, welchen Anteil ITD und IID an der Abbildung bestimmter Richtungen haben. Welche Einbußen sind beispielsweise bei der Lateralisation von Schallereignissen zu erwarten, wenn ausschließlich interaurale Zeitdifferenzen und keine Pegelunterschiede ausgewertet werden? Bei der Deklaration eines biologisch plausiblen Koordinatensystems wurde zu Beginn dieses Kapitels Bezug zu BLAUERTS Untersuchungen der Lokalisationsgenauigkeit genommen. Abbildung 2.12 verdeutlicht, wie sich das beobachtete Ortungsvermögen des Menschen bei der Schätzung des Azimut allein durch eine diskrete Approximation der ITD und den geometrischen Zusammenhang

zwischen Laufzeit und Winkel beschreiben lässt. In einem allgemeinen Fall, in dem gewöhnliche, breitbandige Geräusche unter hinreichend günstigen akustischen Bedingungen geortet werden, scheinen die nachgewiesenen binauralen Intensitätsmerkmale höchstens die Sicherheit der Azimut-Schätzung zu erhöhen, nicht jedoch deren Genauigkeit. Solange keine Schätzung von Elevation und Entfernung der Schallquellen angestrebt wird, sind auch nach der hier unumgänglichen Beschränkung auf ITD-Modelle kaum Abstriche bei der Lokalisationsleistung zu befürchten. Die in 2.2.3 angeführten Implementierungsbeispiele ([LM95, RWE00]) belegen einen gewissen Konsens über diese konzeptionelle Limitierung auf Elemente des Timing-Pfades.

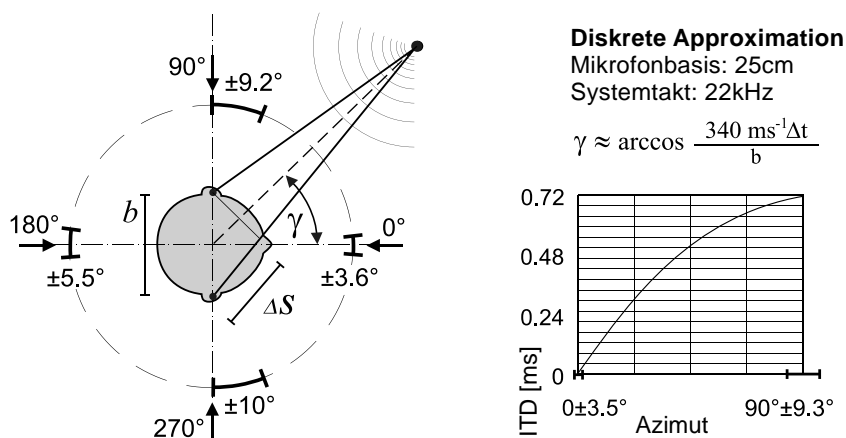


Abbildung 2.12: Zusammenhang zwischen horizontaler Lokalisationsschärfe beim Menschen und Genauigkeit einer diskreten Winkelberechnung anhand der interauralen Laufzeit (Darstellung der Lokalisationsfehler nach [Bla96]).

Mit der interauralen Laufzeit als zentralem Richtungsmerkmal steht bei der Modellbildung auf Ebene der superior-olivaren Kerne der binaurale MSO im Mittelpunkt (s. Abbildung 2.6). Die Untersuchungen am MSO verschiedener Tierklassen [YC90, Kon93] bestätigten eine schon sehr viel früher aufgestellte Hypothese: bereits 1948 prognostizierte JEFFRESS eine auf Delay Lines und Koinzidenzzellen beruhende Überführung der Zeitverzögerung zwischen ipsi- und kontralateralen Reizmustern in den Ortscode einer neuronalen Karte [Jef48]. Das außergewöhnliche an JEFFRESS' funktioneller Beschreibung einer neuronalen Struktur ist ihre direkte, mathematische Interpretation. Anhand KONISHIs Konkretisierung des ursprünglichen JEFFRESS-Modells in Abbildung 2.10 ist ersichtlich, dass die Detektion von Koinzidenzen in binauralen Spikemustern elementare Züge der Kreuzkorrelationsfunktion (KKF) in Form des Verschiebeproduktes der Stereosignale<sup>7</sup> trägt:  $r_{xy} = \int_{-\infty}^{\infty} x(t - \tau)y(t)dt$ . Entspre-

<sup>7</sup>Offensichtlich unabhängig von der Diskussion neuronaler Mechanismen wurde für das Problem der Ortung und Störschallunterdrückung auch ein im Frequenzbereich arbeitender Algorithmus ent-



chend der maximalen lateralen Verzögerung werden die Korrelationswerte kontinuierlich in einem anatomisch plausiblen Zeitfenster ermittelt. In einer Gegenüberstellung der ITD–Auswertung bei verschiedenen Tierarten schildert CARR, wie sich im Laufe der Evolution spezifische Realisierungen des Koinzidenzkonzeptes etablierten:

Klasse, bzw. Spezies	physiologische Spezialisierung zur Phasenkopplung	neuronales Delay	ortskodierte Abbildung
Säuger	Phase locked Spikes bis 3kHz, kurze Refraktärzeiten	einseitige Delay Lines	Eine ITD–Karte je Fre- quenzband
Schleiereule	Phase locked Spikes bis 8kHz, spezielle Glutamat–Rezeptoren	doppelseitige Delay Lines	Mehrfache ITD–Karten je Frequenzband

Tabelle 2.1: Kodierung und Detektion von ITDs in verschiedenen Arten (Auszug aus [CF99]).

Heute existieren zahlreiche Simulationsmodelle, in denen anstelle der abstrakten mathematischen Notation der KKF konkrete Implementierungen von Delay Lines und Koinzidenzzellen eingesetzt werden. Einen gemeinsamen Ausgangspunkt solcher biologisch motivierten Lösungen bildet die Vorverarbeitung der Mikrofonssignale durch ein als Innenohrmodell interpretierbares Frequenzfilter. Oft werden die Signale der Frequenzbänder anschließend binär als Spikemuster kodiert, die Delay Lines durch diskretes Verschieben im Simulationstakt realisiert und die gesamte Funktionalität der Koinzidenzzellen auf eine binäre UND–Verknüpfung reduziert. Die „Zellen“ des MSO–Modells zeigen dabei keine weiteren neuronalen Eigenschaften – sie übernehmen die Responseformen der vorgeschalteten Rezeptoren, was für die Abbildung des ITD–Ortscodes aber zweitrangig ist. In LAZZAROs analoger Hardware–Implementierung [LM95] kommt prinzipiell die gleiche Architektur zur Anwendung – allerdings bei einer quasi–kontinuierlichen Kodierung der Verzögerungsglieder und neuronalen Potentiale. Die ausschlaggebende Funktion als neuronaler Kreuzkorrelator wird dadurch nicht verändert. Die spezielle Signalkodierung und parallele Verarbeitung sind vielmehr inherente Effekte des innovativen VLSI–Designs.

Immer wieder wurden Erweiterungen des ursprünglichen Koinzidenzmodells von JEFFRESS vorgeschlagen, oft mit der Absicht, nicht nur die interaurale Laufzeit, sondern gleichzeitig auch Pegeldifferenzen auszuwerten [ST95]. In Einklang mit den Hypothesen von BLAUERT [BC78, Bla80] entwarf LINDEMANN ein Delay Line Modell mit kontralateraler Inhibition [Lin86a, Lin86b]. In seiner inhibierten Kreuzkorrelation der Frequenzbandsignale wird die laterale Hemmung im ITD–Ortscode durchwickelt. Obwohl es umstritten ist, den ITD–sensitiven Strukturen eine spektrale Wichtung störender Schallanteile zu unterstellen, wurde die von KNAPP beschriebene Generalized Cross–Correlation [KC76] mit der Koinzidenzdetektion in parallelen Frequenzbändern verglichen [RWE00].

die kontralaterale Amplitude gesteuert und soll zu einer deutlicheren ITD–Abbildung führen. Ganz ähnliche Ansätze wurden unter anderem von WOLF und GAIK verfolgt [Wol91, Gai93]. WOLF wendet die Idee der kontralateralen Inhibition in trivialer Weise auf Spike–kodierte Signale an: Er erzeugt einen inhibitorischen Vorgang schlichtweg durch das Löschen von Spikes, wenn diese eine Koinzidenz ausgelöst haben. WOLF schildert weiterhin, wie Spikes infolge des diskreten und synchronen Verschiebevorgangs in ipsi– und kontralateraler Delay Line aneinander vorbeilaufen, wenn ihr Delay ein geradzahliges Vielfaches des Simulationstaktes beträgt. Der Effekt der „Missing Spikes“ bewirkt, dass die zeitliche Auflösung bei der Koinzidenzdetektion nur halb so groß wie bei der ITD–Kodierung ist. Eine elegante Lösung findet WOLF mit der Einführung halbseitiger Delay–lines (Abbildung 2.13), die vermutlich unabsichtlich mit der Mehrzahl der von CARR verglichenen Befunde korrespondieren.

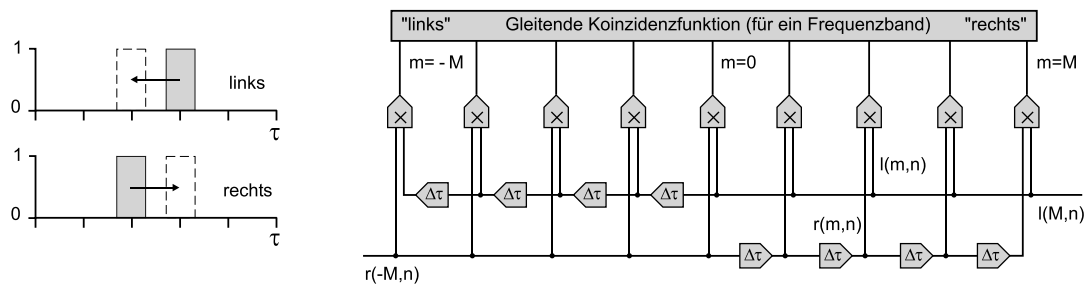


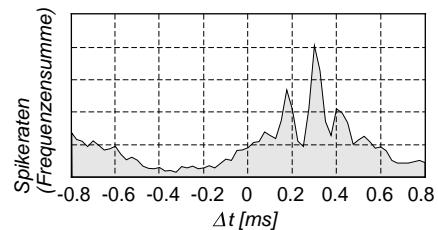
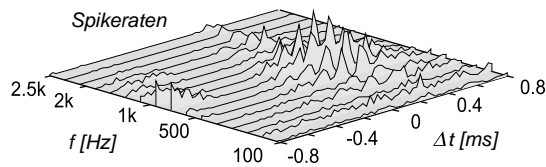
Abbildung 2.13: Das dargestellte Problem der „Missing Spikes“ wird im Koinzidenzmodell mit halbseitigen Delay Lines vermieden (vergl. [Wol91]).

Im Rahmen dieser Arbeit wurde ein Koinzidenzdetektor mit halbseitigen Delay Lines, jedoch ohne kontralaterale Hemmung entworfen. LINDEMANN und WOLF versuchen mit den vorgeschlagenen Inhibitionsmechanismen die in der Periodizität der Signale und Spikefolgen bedingte Mehrdeutigkeit im Korrelationsergebnis zu unterdrücken. Bereits WOLF erkannte, dass eine der Inhibition vergleichbare Wirkung auch durch spezielle Kodierungen der Rezeptormuster erzeugt werden kann. Naheliegender wäre es, anstelle seiner binären Inhibition spärlich kodierte Muster zu generieren und dabei die Refraktärphase der Rezeptoren auf die Dauer der maximalen Verzögerung der Delay Lines auszudehnen.

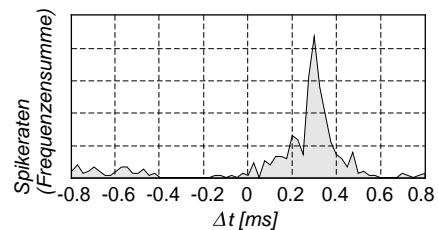
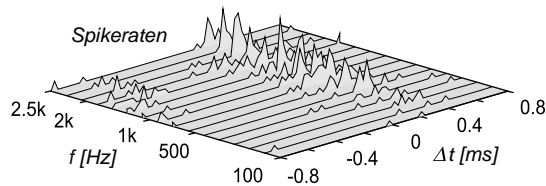
Abbildung 2.14 verdeutlicht beispielhaft die Wirkung verschiedener Spike–Kodierungen auf das Korrelationsergebnis im MSO–Modell. Ein biologisch motiviertes Koinzidenzmuster entsteht, wenn die Phase locked Response der Bushy–Zellen des AV–CN nachgeahmt wird (2.14a). Die tonotope Repräsentation des Korrelationsergebnisses beinhaltet in diesem Fall zwei periodische Komponenten: eine mit längerer Perioden–

dauer in den Frequenzbändern zwischen 0.5 und 2kHz, deren Ursache die stimmhaften Anteile des akustischen Stimulus sind, und eine kürzere Periode, die synchron in mehreren Frequenzkanälen Seitenbänder des Korrelationsmaximums generiert. Diese zweite Komponente korrespondiert zum Chopper-Verhalten der Rezeptorzellen, wenn ihre Refraktärzeit kleiner als die Periodendauer des kodierten Signals ist. Passt man die refraktäre Zeitkonstante der Rezeptoren an die charakteristische Periodendauer des jeweiligen Frequenzbandes an, verschwinden die Chopper-Muster und die von ihnen verursachten dichten Nebenmaxima im Korrelationsergebnis (2.14b). Das Haarzellenmodell verhält sich dann aber ähnlich einem Nulldurchgangsdetektor und kodiert tiefe Frequenzen spärlich, während hohe Frequenzen überrepräsentiert werden. Die tonotopie Aktivierung weicht nun stark von der spektralen Zusammensetzung des Stimulus ab und tiefere Frequenzen tragen kaum zur Lokalisationsleistung bei. Mit der von WOLF empfohlenen Onset-Kodierung kann in einer Delay Line gleichzeitig nur noch ein einziger Spikeimpuls auftreten. Mehrdeutige Korrelationsmuster in höheren Frequenzbändern, wie sie als Antwort auf phasengekoppelte Spikefolgen entstehen, sind damit unmöglich – im Beispiel entsteht ein einzelnes Maximum im ITD-Ortscode (2.14c).

a) MSO-Response für Phase locked Coding



b) MSO-Response für Frequency tuned Coding



c) MSO-Response für Onset Coding

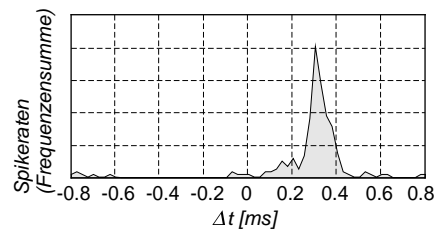
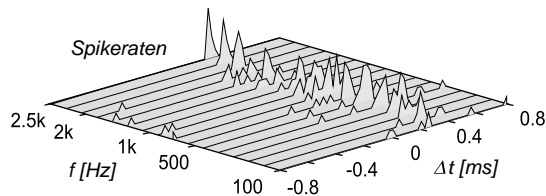


Abbildung 2.14: Koinzidenzstatistiken des binären MSO-Modells für ausgewählte refraktäre Parameter der in Abschnitt 2.2.2 entworfenen Rezeptorzellen. Als breitbandiger akustischer Stimulus diente ein kurzes Sprachsignal [Sch98].

Die Beeinflussung der koinzidenzsensitiven Zellen durch eine kontralateral gesteuerte Inhibition konnte im biologischen Vorbild tatsächlich nachgewiesen werden [BCPR05]. Als Modellkonzept sind solche hemmenden Mechanismen oder die Simulation der Onset Response aber problematisch, da im Gegensatz zum realen auditorischen System nur eine geringe Anzahl an Rezeptoren und Delay Lines berechnet werden kann. Die schärfere Abbildung des ITD–Ortscodes auf der Ebene der Koinzidenzdetektion wird im Modell aufgrund der spärlicheren Kodierung mit einer steigenden Unsicherheit des Ergebnisses erkauft. Unter realen akustischen Bedingungen besteht die Gefahr, dass sich rechte und linke Rezeptoren nicht mehr sicher auf eine erste Wellenfront synchronisieren und die Korreliertheit von ipsi– und kontralateralem Muster sinkt. Das hat fatale Abweichungen in der binauralen Onset–Kodierung und eine drastische Verschlechterung der Lateralisation zur Folge (vergl. Abschnitt 2.3.3).

### 2.2.6 Auditorische Karten im Inferior Colliculus

Nach der divergenten Initialisierung der auditorischen Pfade im cochlearen Kern und der massiv parallelen Extraktion sensorischer Merkmale in den spezifischen Hirnstammarealen stellt sich die Frage, wie insbesondere die berechneten, räumlichen Karten bei der weiteren afferenten Projektion behandelt werden. Mit der Erörterung der Intensity– und Timing–Pfadmodelle wurde vorweggenommen, dass keine der beschriebenen IID– oder ITD–Repräsentationen des superior–olivaren Komplexes einer direkten kortikalen Bewertung oder multisensorischen Verknüpfung zugeführt wird. Die beiden an der räumlichen Kodierung beteiligten Kerne LSO und MSO projizieren zunächst in den Inferior Colliculus (IC) oder den Lemniscus lateralis (LL), der seinerseits mit dem IC verschaltet ist (s. Abbildung 2.6). Auch die Befunde zu den übrigen auditorischen Bereichen des Hirnstammes belegen, dass *alle* auf diesem Niveau gewonnenen, sensorischen Informationen, ungeachtet ihres bisherigen Projektionsweges, den IC als Integrationsgebiet durchlaufen [HA97, GZ96]. Im folgenden Abschnitt wird die Vermutung bestätigt werden, dass der Inferior Colliculus, den EHRET als “Rangierbahnhof der auditorischen Informationsverarbeitung” bezeichnete [Ehr97], auch für die Schallortung von Bedeutung ist

#### Neuroanatomische und physiologische Befunde

Anhand seiner Zelltypen, der Schichtung und dem Verzweigungsmuster der Axone und Dendriten lassen sich im IC ein zentrales (ICc), ein kleineres dorsales (ICd) sowie ein laterales, auch extern genanntes Subareal (ICx) unterscheiden. Das Gros des afferenten

Inputs projiziert in den zentralen Teil und von diesem aus weiter in den ICx. In efferenter Richtung werden ICd und ICx sowohl ipsi- als auch kontralateral von auditorischem Kortex, Corpus geniculatum mediale (MGB) und Superior Colliculus (SC) angesprochen [HA97]. Eine plausible interne Beschreibung anhand von Response-Formen und Tuning-Kurven ist im IC nicht möglich. Die erstmals im DCN beobachtete, deutlich von der Primary-Kodierung des Hörnervs abweichende Differenzierung in den PSTH-Mustern erreicht im IC eine neue Qualität. Gemessen wurden Kombinationen von Chopper Response mit Buildup- und Pauser-Eigenschaften, lange Latenzzeiten, nichtlineare Intensitätskennlinien sowie komplizierte Abstimmkurven für Frequenzen. Dabei sind Variationen einzelner Eigenschaften scheinbar ohne topologische Ordnung über den gesamten IC verteilt [Ehr97]. Erst die makroskopische Sicht auf topologische Ordnungsprinzipien der Merkmalsrepräsentation kann Befunde liefern, die als Modellvorgabe interpretierbar sind. Angesichts der Konvergenz der gesamten auditorischen Pfade im ICc überrascht es kaum, dass dessen räumliche Struktur tatsächlich in verschiedenen Richtungen von neuronalen Merkmalskarten durchzogen ist. Ausgehend von einer tonotop organisierten, laminaren Grundstruktur lassen sich innerhalb von Isofrequenzflächen beispielsweise Topologien für Modulationsfrequenzen und Azimut-Winkel nachweisen [Ehr97]. Offensichtlich ist der im ICc repräsentierte Merkmalsraum aber von höherer Ordnung als das naturgemäß nur dreidimensionale, neuronale Substrat, in dem er kodiert wird. Das hat zur Folge, dass sich die Merkmalskarten entsprechend ihrer jeweiligen Ausdehnung zwangsläufig schneiden und überlappen, dass also gleiche Zellverbände an der Kodierung unterschiedlicher Merkmale beteiligt sind. Klassifikationsversuche auf der Ebene einzelner Zellantworten müssen allein deshalb schwer fallen, da diese Zellen sensitiv für viele Schallereignisse sein sollten, um ihrer Repräsentationsaufgabe gerecht zu werden.

Der IC stellt durch seine exklusive Position einerseits die Schnittstelle dar, über die höhere Hirnregionen Zugang zu aufbereiteten, auditorischen Merkmalen erhalten. Außerdem lassen komplexe Response-Formen und die simultane Kodierung aller verfügbaren Informationen im ICc eine Verknüpfung der Repräsentationen zu qualitativ neuen Merkmalen vermuten. Ein anschauliches und für die Schallortung besonders relevantes Beispiel ist die Kombination von Azimut und Elevation zu einer kompletten räumlichen Abbildung im externen IC der Eule. Infolge der im vorangegangenen Abschnitt entschiedenen konzeptionellen Beschränkung auf zeitbasierte binaurale Mechanismen kann das hier vorgestellte Modell von einer solchen Zusammenführung des Intensity- und Timing-Pfades nicht profitieren. Allerdings liefert allein die Manipulation der ITD-Repräsentationen im zentralen und externen IC wertvolle Anhalts-

punkte, wie mit dem Koinzidenzmuster des MSO weiter zu verfahren ist. Wie alle auditorischen Merkmale wird der im MSO noch tonotop verteilte ITD-Ortscode nicht direkt an höhere Areale übermittelt. Angesichts der Vielzahl von Frequenzbändern und der hohen Winkelauflösung wäre dazu eine erhebliche Datenmenge zu verarbeiten. Für welche Wahrnehmungsleistungen wird die auditorische Richtungsinformation aber letztendlich genutzt? Ob als Ortungshypothese in abstrakten Aufmerksamkeits-Templates oder zur frühen multisensorischen Fusion auf subkortikalem Niveau – in beiden Konzepten ist eine spektrale Darstellung der Geräusche irrelevant! Tatsächlich realisiert der externe IC auf Basis des ITD-Ortscodes eine frequenzunabhängige Azimut-Karte [VK89, Irv92, Kon93] und schont damit die neuronalen Ressourcen seiner Projektionsgebiete. Eine Gegenüberstellung der Frequenz- und ITD-Sensitivität von Zellen im zentralen und externen IC (Abbildung 2.15a–f) belegt nicht nur diese Funktion, sondern bescheinigt darüberhinaus eine qualitative Verbesserung der Richtungsabbildung.

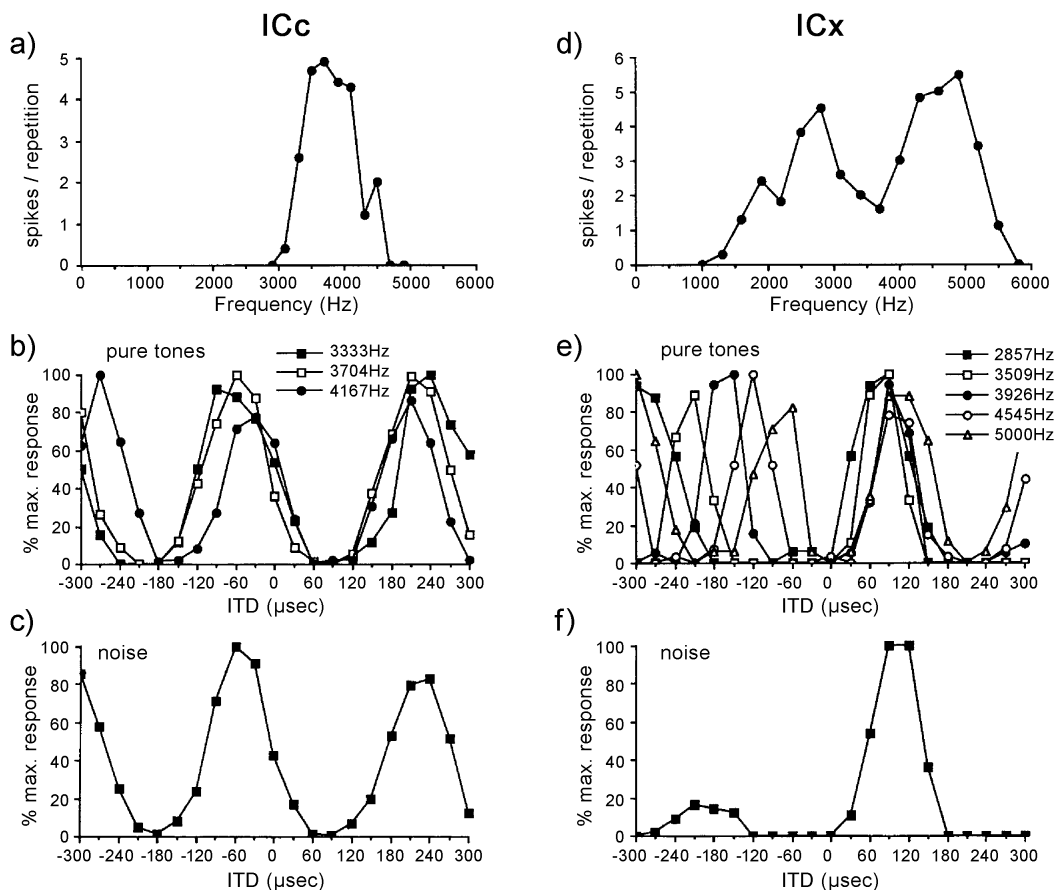


Abbildung 2.15: Einzelzell-Messungen zum Vergleich der Frequenz- und ITD-spezifischen Antworten in ICc und ICx der Eule [VK89]. Vermutlich wurde im ICc ein Neuron mit einem relativen charakteristischen Delay von  $-60\mu\text{s}$  abgeleitet, während die gewählte Position im ICx eine Verzögerung von ca.  $100\mu\text{s}$  repräsentiert.

ITD-sensitive Zellen im ICc antworten aufgrund ihres Inputs aus dem MSO zwar nicht besonders schmalbandig, aber entsprechend ihrer tonotopen Position zumindest innerhalb eines gewissen Frequenzbandes (Abbildung 2.15a). Die periodischen Komponenten in der ITD-Abbildung hängen primär von der charakteristischen Periodendauer in diesem Frequenzband ab. Die Zellen reagieren auf reine Töne nahe ihrer best-matching Frequenz (b) kaum anders als auf breitbandige Stimulation (c). Die Neurone der ITD-Karte im externen IC sind hingegen im gesamten für die laufzeitbasierte Ortung relevanten Frequenzbereich aktivierbar (d). Zwar zeigt auch ihre ITD-Sensitivität bei Stimulation mit reinen Tönen mehrdeutige Phasen – der Abstand der Seitenbänder zum eigentlichen ITD-Maximum variiert jedoch mit der Periodendauer hoher und tiefer Töne (e). Im Ortscode breitbandiger Geräusche kann sich deshalb die eindeutige interaurale Zeitdifferenz gegen mehrdeutige Phasendifferenzen (IPD) durchsetzen (f).

Ohne Zweifel wird die Unterdrückung der periodischen IPDs durch eine Rekombination der tonotop verteilten Kodierung erreicht. Nicht endgültig geklärt ist aber, ob dabei einfach eine Addition der Aktivierungen einzelner Frequenzbänder auftritt oder ob nichtlineare Mechanismen eine zusätzliche Verminderung von Phasenmehrdeutigkeiten bewirken. Während für den ICx der Katze eine lineare Summation der tonotopen ICc-Muster beschrieben wurde [YC90], gibt es einmal mehr Untersuchungen an der Eule, die nichtlineare Prozesse vermuten lassen [WTK87]. Weiterhin wurde spekuliert, die abweichenden Befunde könnten allein dadurch zustande kommen, dass die periodischen IPD-Muster zwar im Rahmen der physiologischen Möglichkeiten der Eule auftreten, bei der Katze aber außerhalb des durch Phase Locking und ICc-Tonotopie begrenzten Bereiches liegen [Irv92, Maz98]. Vergleicht man die bei einer Spezies möglichen, maximalen Zeitdifferenzen mit den zitierten Phase Locking Befunden und Frequenzbereichen der Koinzidenzdetektion im MSO, können mehrdeutige ITD-Kodierungen bei weiteren Tierarten aber kaum ausgeschlossen werden.

## Modellierung

Als Vorleistung zur multisensorischen Integration in Kapitel 4 markiert der von mehrdeutigen Phasenartefakten befreite ITD-Ortscode in seiner kompakten und nicht mehr tonotop verteilten Darstellung im externen IC die Zielstellung der auditorischen Modellierung. Die topologische Aktivierung bildet auf diesem neuronalen Niveau klare Maxima für ein breites Spektrum natürlicher Geräusche aus und ist direkt und eindeutig als Richtungsinformation interpretierbar. Abbildung 2.16 verdeutlicht, dass bereits eine lineare Integration über den abgebildeten Frequenzbereich die interaurale Lauf-

zeit gegenüber anderen Phasendifferenzen (IPD) betont. Die mehrdeutigen Maxima der IPDs treten entsprechend der unterschiedlichen, charakteristischen Periodendauer der Frequenzbänder topologisch verstreut auf und ergeben in der Summe eine diffuse Aktivierung. Dagegen wird die interaurale Laufzeit frequenzunabhängig immer in dieselbe spektrale Kolumne projiziert und stellt bei hinreichend breitbandigen Stimuli das globale Maximum im eindimensionalen Ortscode. Hier zeigt sich ein konzeptioneller Vorteil des ITD-Modells im Vergleich zur Intensitätsauswertung: Die einfache additive Rekombination der ITD-Abbildungen mehrerer Frequenzbänder funktioniert identisch für alle horizontalen Richtungen und unabhängig von der Amplitude des Stimulus. Die Bewertung von tonotop kodierten Intensitätsunterschieden setzt aufgrund der kopfbezogenen Transferfunktionen aber eine Wichtung mit richtungsspezifischen spektralen Templates voraus. Hinzu kommt, dass die IID-Kodierung im LSO stark von der allgemeinen, monauralen Amplitude abhängt, weshalb eine zusätzliche Invarianzleistung bei der Projektion in den IC erbracht werden muss [PKHG04]. Die spektralen und pegelabhängigen Mehrdeutigkeiten der IID-Abbildung aufzulösen, erscheint deshalb ungleich aufwendiger als die Unterdrückung der Phasenartefakte im ITD-Ortscode.

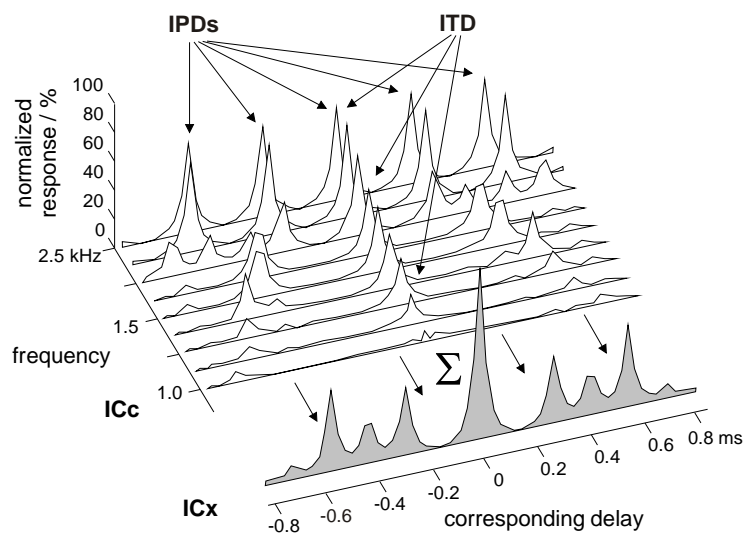


Abbildung 2.16: Interaurale Phasendifferenzen (IPD) in den Korrelogrammen einzelner Frequenzbänder und Betonung der interauralen Laufzeit nach einer linearen Summation der Aktivierung über den Frequenzbereich.

Nach der Summation über alle Frequenzbänder könnte die ITD-Abbildung angesichts der scheinbar linearen Mechanismen im ICx einiger Tierarten [YC90] als Endergebnis der Lateralisation gelten. Die experimentelle Evaluation des hier entworfenen Modells offenbart jedoch eine Reihe unerwünschter Effekte. Unter realen Bedingungen verschlechtern Raumakustik und Störgeräusche oft die Korreliertheit zwischen ipsi-



und kontralateralen Reizmustern. Tatsächlich sind in den meisten geschlossenen Räumen außerhalb des Hallradius nur in den Anhallphasen der Geräusche aussagekräftige Korrelogramme zu erwarten<sup>8</sup>. Auch bei schmalbandiger Stimulation sind zur effektiven Unterdrückung periodischer Phasenbilder weitere, nichtlineare Inhibitionsprozesse unverzichtbar. Gleichzeitig könnten solche Mechanismen die Ausprägung eines Maximums in der unscharfen ITD–Abbildung tiefer Töne verbessern [FK01].

#### *Winner–Take–All Prozesse im ITD–Ortscode*

LAZZARO motivierte eine weitere Manipulation der ITD–Karte mit der für die Eule nachgewiesenen, nichtlinearen Transformation zwischen ICc und ICx. Ohne auf detaillierte Befunde über die interne synaptische Struktur zurückgreifen zu können, leitete er aus der Forderung einer gezielten lateralen Inhibition der ITD–Seitenbänder ein Winner–Take–All (WTA) Verhalten des ICx ab [LM95]. WTA–Prozesse selektieren in einer universellen Weise dominante Merkmale in einer topologischen Abbildung und erscheinen formal zur Hervorhebung des ITD–Peaks in mehrdeutigen Korrelationsmustern geeignet. Mit den von AMARI beschriebenen dynamischen Feldern [Ama77] standen LAZZARO etablierte Standardmodelle für die Simulation von WTA–Prozessen zur Verfügung. Allerdings existiert mittlerweile ein Arsenal an unterschiedlichen WTA–Architekturen, und es wäre wünschenswert, durch weitere Befunde Anhaltspunkte für eine spezifische Implementierung zu erhalten. Tatsächlich bestätigen und konkretisieren einige Untersuchungen die WTA–Hypothese, ohne dabei den in der Neuroinformatik populären Begriff der Winner–Take–All Netze selbst zu benutzen. Generell scheint eine zur Intensität proportionale Inhibition die ITD–Kodierung weitgehend unabhängig von der Schallamplitude zu machen [PVAK96]. Der Umstand, dass die Intensität in der ITD–Abbildung verteilt kodiert wird, sowie die sinnvolle Forderung, ein einziges Maximum zu selektieren, korrespondiert mit AMARIs Konzept der globalen Hemmzelle. Auch wurde beschrieben, dass Inhibitionsmechanismen die seitlichen IPD–Maxima stärker hemmen als den zentralen ITD–Peak [FK91, PK00], was leicht mit Hilfe einer lokalen Off–on–off Verschaltung realisierbar ist. Da es in der nicht tonotopen ICx–Kodierung keine charakteristische Periodendauer gibt, müssten aber die Breite dieser Verschaltung und der laterale Abstand der inhibitorischen Bereiche pauschalisiert wer-

---

<sup>8</sup>Bei der Interpretation der auditorischen Ortungsmechanismen wird oft verschwiegen, dass unsere heutige akustische Umgebung als potentieller Anwendungsort der Modelle stark von den Rahmenbedingungen bei der Evolution der auditorischen Systeme abweicht. Dies gilt insbesondere für Artefakte durch Nachhall und Raumresonanzen, die in Verbindung mit den idealisierten Stimuli der neurophysiologischen Untersuchungen nicht auftreten.

den. In Kombination mit einer globalen Hemmzelle ist es jedoch vorstellbar, auf lokale Inhibitionen gänzlich zu verzichten.

Noch ein weiterer Befund steht zumindest in indirektem Zusammenhang mit einem WTA-Konzept. Bei Experimenten, die eine Beteiligung des IC der Säugetiere beim Auftreten von Präzedenzeffekten und der Unterdrückung von Echos klären sollten, wurde eine lange andauernde Inhibition von ITD-sensitiven Regionen beobachtet. Schon die Realisierung eines wenige Millisekunden langen Inhibitionsintervalls kann allerdings nicht mehr einzelnen Neuronen oder einfachen Feed-Forward-Strukturen unterstellt werden. Es liegt nahe, dass an der Entstehung solcher zeitlichen Phänomene auch IC-interne Mechanismen beteiligt sind, wenngleich diese bislang nicht näher spezifiziert wurden [Yin94]. Betrachtet man den ITD-Ortscode des IC, die Echo-Unterdrückung und den Präzedenzeffekt in einem gemeinsamen Kontext [Yin94, KT96, LY98a, LY98b], erlaubt die WTA-Hypothese eine Interpretation der verschiedenen Befunde in einem übergreifenden abstrakten Modell.

Dazu ist zunächst ein WTA-Prozess erforderlich, der nicht nur die räumliche Manipulation der ITD-Abbildung sondern auch eine zeitliche Integration der Aktivierungen ermöglicht. Sowohl AMARI als auch KOHONEN beschreiben WTA-Netze, die neben der Rückkopplung der globalen Inhibition noch über ein lokales Feedback verfügen. Im Gegensatz zu der von AMARI vorgeschlagenen „Difference of Gaussian“-Wichtung (DoG) von lokal exzitatorischen und lateraler inhibitorischen Bereichen [Ama77], verzichtet KOHONEN auf laterale Projektionen und beschränkt die lokale Verschaltung auf eine direkte exzitatorische Rückführung der Ausgänge auf dieselben Zellen [Koh93, KK94]. Beide Methoden realisieren einen hochgradig nichtlinearen Selektionsprozess, in dem sich eine topologisch begrenzte Region mit dominanter Aktivierung gegen konkurrierende Bereiche durchsetzt. Außerdem initiieren die Feedback-Komponenten einen Hysterese-Effekt: Erst nach einer Abklingphase oder beim Einsetzen eines stärkeren Stimulus an einer anderen Position kann ein neues Maximum ausgeprägt werden. Generell birgt ein zu starkes exzitatorisches Feedback natürlich die Gefahr, dass sich die Netzaktivität unkontrolliert ausbreitet oder nach der Verlagerung des Schwerpunktes im Eingabemuster an einer alten Position verharret. Um dieses als schwaches WTA-Verhalten bezeichnete Problem sicher zu unterbinden, sehen alle zitierten Autoren eine Erweiterung ihrer Architekturen durch geeignete Resetmechanismen vor.

Die Anwendung des AMARI- oder KOHONEN-WTA auf den ITD-Ortscode kann einerseits die nichtlineare Inhibition der Phasen-Mehrdeutigkeiten im ICx modellieren. Gleichzeitig bewirkt die Hysterese-Eigenschaft eine Unterdrückung von Echos, die normalerweise eine vom Originalsignal abweichende ITD-Signatur aufweisen, sich infolge

ihrer geringeren Energie aber nicht im WTA-Prozess durchsetzen können. Erst wenn eine eindeutige Gewinnerregion in der ITD-Karte abgeklungen ist, endet das Inhibitionsintervall und erlaubt neue Richtungshypothesen. Bei geeigneter Dimensionierung von Feedback-Stärke und Abklingzeit wird so aber gerade auch ein Präzedenzeffekt verursacht: ist die Pause zwischen zwei Präzedenz-Klicks kürzer als das Inhibitionsintervall, dominiert die Richtungsinformation des zuerst eintreffenden Geräusches.

Für die folgenden auditorischen und multisensorischen Experimente wurde im Rahmen dieser Arbeit ein modifiziertes WTA-Filter implementiert, das im wesentlichen auf dem Modell von AMARI basiert (Abbildung 2.17). Eine globale Hemmzelle integriert die Aktivität in der gesamten ITD-Karte und projiziert mit einheitlicher Wichtung auf alle Positionen inhibitorisch zurück. Als Antagonist im Selektionsprozess stellt eine Exzitation lokaler Bereiche die Bevorteilung des ITD-Maximums gegenüber den lateral positionierten Phasenbildern sicher. Nach der Rekombination der ITD-Muster aus den Frequenzbändern treten jedoch keine spezifischen Phasendifferenzen mehr auf, und auch die Position, an der sich Echos oder Resonanzen im Ortscode abbilden, kann kaum vorhergesagt werden. Ohne Informationen über die räumliche Gestalt der ITD-Signaturen erscheinen konkrete laterale Wichtungsvektoren nicht sinnvoll. Anstelle der DoG-Wichtung AMARIS mit ihren typischen inhibitorischen Randbereichen ist der hier angewandte, schmale und rein exzitatorische Feedback-Vektor eher mit den Rückkopplungen im KOHONEN-Modell vergleichbar. Zumindest unmittelbar benachbarte Zellen sollten allein schon aufgrund drohender Jitter-Effekte im diskreten Ortscode miteinander verknüpft werden. Falls eine interaurale Laufzeit genau zwischen die repräsentierten ITDs zweier benachbarter Zellen fällt, können diese einen lateral-exzitatorischen Verbund bilden, anstatt im WTA-Prozess zu konkurrieren.

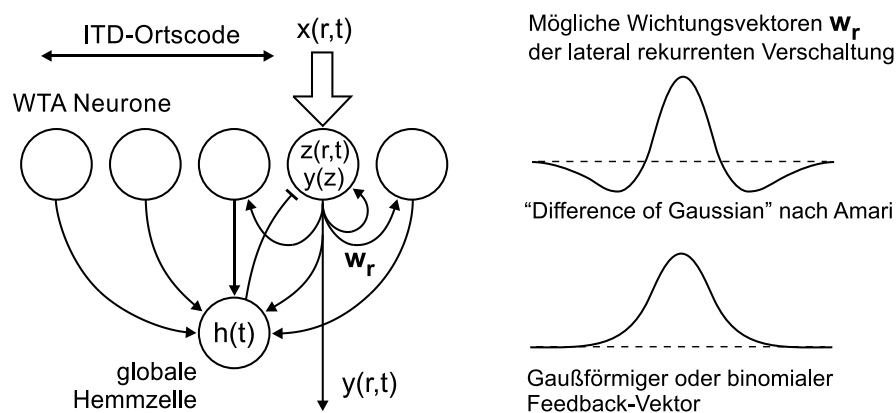


Abbildung 2.17: Neuronales Feld nach AMARI. Die Zellen berechnen intern Zustände und Ausgänge, bevor eine Verknüpfung von Netzeingang, lokal exzitatorischem und global inhibitorischem Feedback erfolgt.

*WTA mit Spike Response Neuronen*

Schließlich muss ein adäquates Abstraktionsniveau für die Beschreibung des Zustandsraumes der modellierten WTA-Zellen bestimmt werden. Vergleichbar mit dem Hardware-orientierten Ansatz von LAZZARO fanden frühere Untersuchungen vor dem Hintergrund einer Implementierung in analoger VLSI-Technik statt [ISP99]. Dabei wurde nach mathematisch einheitlich beschriebenen Modellen gesucht, die im Sinne einer neuronalen Basisbibliothek möglichst universell in den technischen Designprozess integriert werden konnten. Zum Konsens der Modellbildung zählte unter anderem eine uniforme spikebasierte Kodierung und die Simulation mit einem globalen Systemtakt. Anhand dieser Vorgaben wurde ein Spike Response Modell deklariert [SP99]. In Analogie zur Rezeptorzelle des Innenohrmodells erzwingt es durch die Kombination von Reizschwelle  $\theta$  und Refraktionspotential  $ahp(t)$  eine binäre Antwort (Gleichung 2.4). Um die Spikefolgen anderer Zellen plausibel verarbeiten zu können, müssen eintreffende Impulse zunächst durch postsynaptische Potentiale (PSP) zeitlich integriert werden. Wenn ein Neuron  $i$  mit  $n$  Synapsen spikeförmige Reize  $x_j$  empfängt, lassen sich einzelne postsynaptische Potentiale  $z_j(t)$  als Ergebnis der Faltung:

$$\begin{aligned} z_j(t) &= w_{ij} \int_0^t g(\tau) x_j(t - \tau) d\tau \\ &= w_{ij} \cdot x_j(t) * g_j(t) \end{aligned} \quad (2.3)$$

notieren. Als Transferfunktion der Synapse  $g_j(t)$  bietet sich die von Integrate and Fire Modellen bekannte Form  $\frac{t}{\tau} e^{1-\frac{t}{\tau}}$  an [GRvH93, GK02]. Der statische Gewichtungsfaktor  $w_{ij}$ , mit dem die synaptische Verbindung an der Entstehung eines Somapentials beteiligt ist, entscheidet mit seinem Vorzeichen über die exzitatorische oder inhibitorische Natur der Synapse. Entsprechend der WTA-Topologie beschreibt die räumliche Integration von Eingangs-, Inhibitions- und Feedback-PSP ein Summenpotential der Zelle  $z_i(t) = \sum_{j=1}^n z_j(t)$ . Die binäre Ausgabefunktion  $y(t)$ , die Reizschwelle  $\theta$  und das Refraktionspotential  $ahp(t)$  werden in der vom Rezeptormodell bekannten Weise berechnet:

$$y_i(z_i(t), ahp(t), \Theta) = \begin{cases} 1 & : z_i(t) + ahp(t) > \Theta, \\ 0 & : \text{sonst} \end{cases} \quad (2.4)$$

Unter der Maßgabe eines globalen Systemtaktes erfolgte die Simulation synchron zum Innenohrmodell mit den in Anhang A beschriebenen zeitdiskreten Filtern. In einer Reihe von experimentellen Untersuchungen konnte ein robustes und weitgehend Amplituden-unabhängiges WTA-Verhalten im ITD-Merkmalsraum demonstriert werden. Darüberhinaus war es möglich, durch eine adäquate Parametrisierung

der Feedback-Dynamik physiologisch plausible Inhibitionsfenster für Präzedenzeffekte zu schaffen. Als sichtbare Folge konnten auch einige stark gestörte ITD-Signaturen, die unter schwierigen akustischen Bedingungen entstanden waren, anhand der Anhallphasen korrekt bewertet werden [SZPG00a, SZPG00b].

#### *WTA mit Dynamischen Neuronen nach AMARI*

Ohne die Randbedingungen der Hardware-Implementierung bezüglich der einheitlichen Modellierung neuronaler Potentiale und der synchronen Simulation aller Komponenten liegt es nahe, die aufwendige Spikekodierung im WTA-Netz zu hinterfragen. Außer Diskussion steht, dass zur Koinzidenzdetektion im Mikrosekundenbereich die hohe zeitliche Präzision bei der Aufnahme und der primären Kodierung der binauralen Signale unumgänglich ist. Nach erfolgreicher Transformation des ITD-Timecodes in den Ortscode einer neuronalen Karte erscheint eine plausible Merkmalsrepräsentation aber auch mit einem deutlich niedrigeren Simulationstakt möglich. Schließlich wird sich die Richtung von Schallquellen nicht binnen weniger Millisekunden ändern, und auch Echos und Resonanzen in geschlossenen Räumen treffen mit Verzögerungen ein, die wesentlich größer als die ermittelten interauralen Laufzeiten sind. Die zur weiteren Manipulation der ITD-Abbildung notwendige Auflösung im Zeitbereich ist demnach weniger von der Stereo-Geometrie als vielmehr von den akustischen Bedingungen abhängig. Tatsächlich stehen sich im Falle der Einbeziehung von Bildsequenzen im multisensorischen Modell die Audiosignale mit einer Abtastrate von 44.1kHz und die visuelle Domäne mit der Videofrequenz von nur 30Hz gegenüber. Vorteilhaft wäre hier ein Konzept, in dem sich die Schrittweite der zeitdiskreten numerischen Verfahren an der Dynamik der Daten orientiert und für die verschiedenen auditorischen und multisensorischen Module angepasst werden kann.

Die dynamischen Feldgleichungen des ursprünglichen AMARI-WTA bieten dazu eine elegante Lösung: Netzein- und -ausgänge modellieren mittlere Feuerraten, die im vorliegenden Ansatz auf beliebige Zeitmaßstäbe umgerechnet werden können. Für die in Abbildung 2.17 dargestellte WTA-Architektur lautet die nichtlineare Notation eines dynamischen AMARI-Feldes:

$$\begin{aligned} \tau \frac{d}{dt} z(r, t) = & -z(r, t) + x(r, t) - c_i \int y(z(r, t)) dr \\ & + c_n \int w(r - r') y(z(r', t)) dr' \end{aligned} \quad (2.5)$$

Der Zustand  $z(r, t)$  eines Neurons an der Position  $r$  ist abhängig von drei Komponenten: Den Eingang  $x(r, t)$  liefert der ITD-Ortscode der Position  $r$  als Summe der

Korrelationsergebnisse aus den Frequenzbändern. Die globale Inhibition weist kein separates Zeitverhalten auf und entspricht der durch  $c_i$  gewichteten Summe des Netzausganges. Das lokale Feedback kann durch einen schmalen binomialen Wichtungsvektor (z.B.  $w=[0.25 \ 0.5 \ 0.25]$  oder  $w=[0.0625 \ 0.25 \ 0.375 \ 0.25 \ 0.0625]$ ) gesteuert werden. Alle WTA-Zellen besitzen sigmoide Ausgänge, deren Limitierung durch die Fermi-Funktion bestimmt wird:

$$y(z(r, t)) = \frac{1}{1 + \exp(-\sigma \cdot z(r, t))} \quad (2.6)$$

Im Vergleich zum spikebasierten Algorithmus konnte mit dem ratenkodierten WTA-Netz für Simulationstakte von 500Hz bis 4kHz ein makroskopisch gleichwertiger Selektionsprozess in den ITD-Mustern demonstriert werden. Unter einer sehr groben zeitlichen Auflösung leidet allerdings die Detektion der Anhallphase von echo- und resonanzbehafteten Signalen, weshalb der Simulationstakt in Abhängigkeit von den akustischen Bedingungen nicht unter 1-2kHz gewählt werden sollte. Exemplarisch zeigt Abbildung 2.18 die Verarbeitung eines realen Rauschsignals, das mit seinem extrem gestörten Korrelationsmuster ein besonders kritisches Testgeräusch darstellt. Die Mikrofonanordnung wurde während der Geräuschkdauer von einer Sekunde annähernd

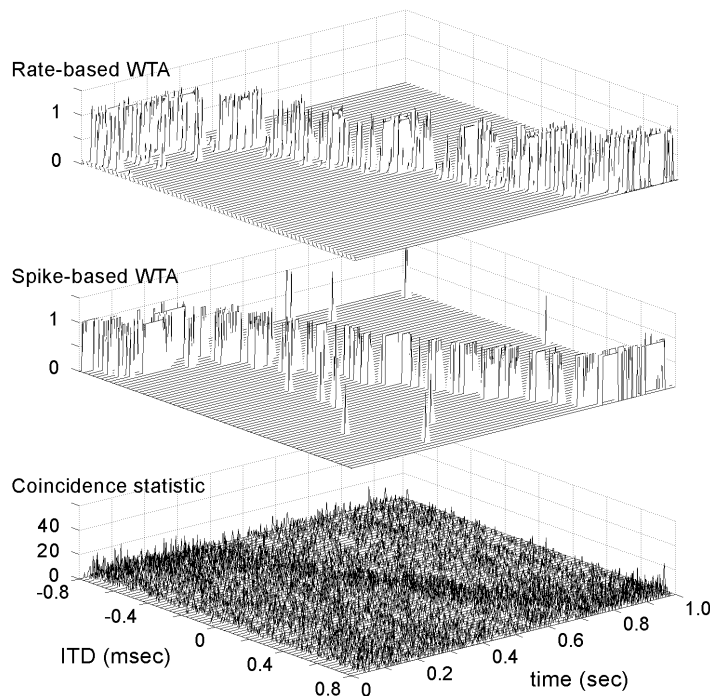


Abbildung 2.18: Ein Rauschsignal bewegt sich innerhalb einer Sekunde von  $90^\circ$  links nach  $90^\circ$  rechts. In der Koinzidenz-Statistik (unten) ist eine vage Richtungsrepräsentation zu erahnen. Beide WTA-Realisierungen mit Spike Response Neuronen (Mitte) bzw. dynamischen Amari-Zellen (oben) können das extrem gestörte Korrelationsmuster sicher auswerten.

gleichförmig um 180 Grad gedreht, was einer schnellen Bewegung der Schallquelle von 90 Grad links nach 90 Grad rechts gleichkommt. Die Ergebnisse der spike- und ratenbasierten Simulation können gleichermaßen als sehr stabil beurteilt werden. Trotz der gewünschten Hysterese in der gefilterten ITD-Abbildung folgt der WTA-Fokus beinahe kontinuierlich der korrekten Richtung des Schallereignisses. Im Gegensatz zu sogenannten schwachen WTA-Prozessen, die zur Aktivitätsverlagerung explizite Resetmechanismen benötigen, kann hier im Rahmen der WTA-Terminologie ein starkes Selektionsverhalten attestiert werden, was die Robustheit der Algorithmen unterstreicht.

## 2.3 Vergleich und Wertung der Modelle

### 2.3.1 Optionen der Modellierung

Wie in zahlreichen anderen Simulationsmodellen und Implementierungen [LM95, RWE00, BRS02] wird auch in dieser Arbeit die Aufgabe der Schallortung auf die Auswertung der Laufzeit im Stereosignal reduziert. Die Motivation für eine solche Vereinfachung ist gleichermaßen in der physikalisch bedingten Limitierung der Audioaufnahmen wie in der unterschiedlichen Komplexität von laufzeit- und intensitätsbasierten Mechanismen begründet. Selbst wenn der technische Aufwand zur Erzeugung plausibler interauraler Intensitätsunterschiede mittels kopfbezogener Transferfunktionen betrieben würde, wäre eine adäquate Bewertung der tonotop repräsentierten Amplitudenmerkmale schwierig. Im Gegensatz zum eindeutigen Zusammenhang zwischen Lateralisation und ITD-Ortscode müssen bei der Kodierung und Auswertung der IID eine Reihe von Invarianzen gegenüber Spektrum und Pegel der Geräusche realisiert werden. So können beispielsweise den IID-Karten des LSO ohne Kenntnis der monauralen Amplitude keine konkreten Winkel zugeordnet werden: Eine topologische Position kann mit einem eher zentralen und lauten akustischen Ereignis korrespondieren – ebenso gut aber auch mit einem leisen, lateralen Stimulus [PKHG04]. An der Simulation von IID-Modellen [Zah03] wird deutlich, dass, abgesehen von der Problematik konkreter kopfbezogener Transferfunktionen, allein schon diese pauschale, amplitudenabhängige Verlagerung im Ortscode ein prinzipielles Manko der IID-Ansätze darstellt.

Die Auswertung der interauralen Laufzeit hat im Vergleich zu IID-Verfahren aber nicht nur den Vorteil, weitgehend unabhängig von Lautstärke, spektraler Zusammensetzung und Entfernung der Schallereignisse zu sein. Im Sinne eines Timing-Pfades lie-

fern die Befunde zum auditorischen Hirnstamm auch ein vollständiges Modellkonzept: von der phasengenauen zeitlichen Kodierung über die Koinzidenzdetektion bis hin zur nichtlinearen Filterung des ITD–Ortscodes im IC (Abbildung 2.6). Als klare Motivation für einen neurobiologisch inspirierten, aber auf zeitliche Merkmale beschränkten Ansatz kann die horizontale Lokalisationsleistung sehr genau anhand der variablen ITD–Präzision im Rahmen der Stereo–Geometrie abgeleitet werden (Abschnitt 2.2.5, Abbildung 2.12). In Einklang mit dieser Hypothese belegen psychoakustische Experimente eine zentrale Bedeutung der ITD–basierten Lateralisation bei der multisensorischen Steuerung der Aufmerksamkeit [SHB00]. Schließlich kann angenommen werden, dass sich primäre Wahrnehmungs– und Aufmerksamkeitsmechanismen für typische Situationen im menschlichen Umfeld auch mit einer eindimensionalen Richtungskodierung innerhalb der horizontalen Ebene demonstrieren lassen<sup>9</sup>.

Selbst mit der konzeptionellen Einschränkung eines ITD–dominierten Timing–Pfades werden im vorgeschlagenen Modell eine Fülle von Befunden zur Kodierung, Extraktion und Manipulation auditorischer Merkmale aufgegriffen. Während die prinzipielle Bedeutung der interauralen Laufzeit für die Schallortung phänomenologisch nachvollziehbar und durch psychoakustische Experimente belegt ist, bietet der Beitrag einzelner Mechanismen an der Generierung des ITD–Ortscodes durchaus Raum für unterschiedliche Interpretationen. Sowohl zum Verständnis der natürlichen Wahrnehmungsleistungen als auch zur Bewertung des Simulationsmodells kann hinterfragt werden, ob die neuronalen Strukturen des Timing–Pfades optimale evolutionäre Lösungen einer speziellen Kodierungs– oder Verarbeitungsaufgabe darstellen. Sollten die beschriebenen Mechanismen hingegen Kompromissen unterliegen, die der simultanen Verarbeitung vieler auditorischer Merkmale im Hirnstamm geschuldet sind, wären im technischen System andere und einfachere Algorithmen anzuraten. Spielt beispielsweise die Tonotopie für laufzeitbasierte Lokalisation eine signifikante Rolle oder ist die tonotope Koinzidenzdetektion im MSO und die entsprechende Kodierung im ICc nur

---

<sup>9</sup>Im multisensorischen Konzept markiert die Möglichkeit, Positionen außerhalb des Blickfeldes zu erfassen, eine wesentliche Stärke der Schallortung. Aus anwendungsorientierter Sicht könnte man hier einwenden, dass die bloße horizontale Lateralisation nicht zwischen vorderer und hinterer Halbebene unterscheiden kann. Eine solche Ortungsleistung ist in der Natur tatsächlich erst unter Einbeziehung des richtungsspezifischen peripheren Schalltransfers realisierbar. Um die Anwendbarkeit des ITD–basierten Modells nicht einzuschränken, wurde im Rahmen der experimentellen Untersuchungen ein patentiertes Verfahren zur 360–Grad Ortung ohne Kunstkopfaufnahmen entwickelt [SPG01a, Sch02b]. Die dabei gefundene technische Lösung ist für die Erörterung natürlicher Wahrnehmungsleistungen jedoch irrelevant. Der ITD–Ortscode allein korrespondiert bereits mit einem Winkelbereich von ca. 200 Grad (vergl. Abbildung 2.12) – was über typische, visuell erfasste Ausschnitte der Umgebung deutlich hinausgeht und zur Demonstration multisensorischer Effekte ausreicht.



Folge der allgemeinen spektralen Repräsentation im Hörnerv und in den cochlearen Kernen? Wie wirkt sich die Spikekodierung des Haarzellenmodells gegenüber der kontinuierlichen Darstellung analoger Signale im interauralen Korrelationsergebnis aus? Die Analogie der binauralen Koinzidenzdetektion mit der Kreuzkorrelationsfunktion eröffnet zudem die außergewöhnliche Möglichkeit, die neuronalen Mechanismen unmittelbar mit formalen mathematischen Methoden zu vergleichen. Offensichtlich gibt es eine Reihe von Optionen für eine eher biologisch inspirierte oder technische Strategie. Um eine Wertung des aufwendigen auditorischen Modells vorzunehmen, sollen unter der Vielzahl denkbarer Modellvarianten folgende Realisierungen einem exemplarischen Vergleich unterzogen werden:

- XC: Eine triviale Lösung zur Laufzeitbestimmung stellt die Kreuzkorrelationsfunktion (KKF) dar, die in einem diskreten Zeitfenster für die Stereo-Mikrofonsignale ermittelt wird.
- XC-f: Um gezielt den für die ITD-Ortung physiologisch relevanten Frequenzbereich zu betrachten, kann eine Korrelation der Frequenzbandsignale des Innenohrfilters berechnet und anschließend linear über die spektralen Komponenten summiert werden.
- CD: Unter Verwendung von Innenohrfilter und Rezeptormodell können die simulierten Hörnervmuster einer binären Koinzidenzdetektion zugeführt werden. Auch hier wird das Ergebnis durch die Rekombination der ITD-Karten der Frequenzbänder gebildet.
- CD-i: Als eine Variante der binären Koinzidenzdetektion ist die kontralaterale Inhibition nach WOLF denkbar, bei der Spikes nach dem Auslösen einer Koinzidenz aus den kontralateralen Delay Lines gelöscht werden.
- XC+WTA: Die Manipulation des ITD-Ortscodes durch eine WTA-Verarbeitung ist prinzipiell mit allen primären ITD-Karten realisierbar und kann beispielsweise anhand des Korrelationsergebnisses der Mikrofonsignale demonstriert werden.
- CD+WTA: Die Koinzidenzdetektion der Hörnervmuster mit anschließender WTA-Filterung entspricht schließlich dem vollständigen, neurobiologisch motivierten Gesamtmodell mit allen in Abbildung 2.6 vorgestellten Optionen des Timing-Pfades.

MLE: Als derzeit mächtigstes Referenzverfahren kann eine Maximum–Likelihood Schätzung für ein probabilistisch definiertes Delay herangezogen werden. Anstelle der häufig zitierten, spektralen bzw. generalisierten KKF nach KNAPP wurde hier eine äquivalente Implementierung im Zeitbereich von MODDEMEIJER verwendet [Mod88].

### 2.3.2 Experimenteller Ansatz

#### *Akustische Szenarien*

Die häufig gestellte Frage nach einem Qualitätsmaß oder einer Erfolgsrate bei der Schallortung ist nicht leicht zu beantworten. In einem einfachen Szenario, in dem eine einzige Schallquelle unter akustisch vorteilhaften Bedingungen wahrgenommen wird, sind bei keinem der aufgezählten Algorithmen fehlerhafte Ortungsergebnisse zu erwarten. Sobald simultane Quellen oder eine ungünstige Raumakustik mit Echos und Resonanzen in die Überlegungen einbezogen werden, erschweren akustische Interferenzen eine systematische, experimentelle Untersuchung. Abgesehen von kurzen Onset–Intervallen führt die lokale Ausprägung eines Interferenzmusters zu erheblichen Störungen der binauralen Phaseninformation. Die Resultate charakterisieren dann vor allem die Spezifik des Szenarios und lassen kaum Rückschlüsse über das Verhalten der Algorithmen in anderen Situationen zu. Gleichwohl widmen sich einige Arbeiten gezielt komplexen Szenarien. Wiederholt wurde die Idee diskutiert, die tonotopen ITD–Abbildungen simultaner Quellen anhand ihrer spektralen Signatur zu trennen oder die generalisierte Kreuzkorrelation im Bildbereich mit dem inversen Spektrum von Störgeräuschen zu wichten (vergl. [KC76, BRS02]). Im Kontext der Aufmerksamkeitssteuerung überwindet ein derartiger Ansatz die konzeptionelle Trennung von primärer, räumlicher Abbildung und Objekterkennung.

Bei der hier angestrebten Wertung der spezifizierten Modellvarianten sind Szenarien mit simultanen Quellen unter mehreren Aspekten kritisch zu bewerten. Einerseits müssten sich die Spektren der zu separierenden Quellen deutlich unterscheiden, was für breitbandige Stimuli wie z.B. Sprachsignale nicht zutrifft. Zum anderen scheint im biologischen Vorbild eine spektrale Gewichtung oder eine separate ITD–Kodierung simultaner Quellen gar nicht angestrebt zu werden. Sowohl die begrenzte Frequenzauflösung im Koinzidenzdetektor des MSO als auch die nichtlineare Ausprägung einzelner ITD–Maxima bei der Rekombination der Frequenzbänder im ICx deuten eher auf eine sequentielle Richtungsabbildung hin. Tatsächlich wechseln in den Korrelogrammen komplexer und echo–behafteter Szenen einzelne aussagekräftige Anhallphasen und län-

gere Intervalle mit diffusen ITD–Mustern. In den folgenden experimentellen Untersuchungen wurde auf konkurrierende Signalquellen oder willkürliche Kombinationen von Nutz– und Störgeräuschen verzichtet. Sinnvoller erschien es, einzelne Schallereignisse in akustisch mehr oder weniger vorteilhafte Umgebungen einzubetten und zu fragen, ob sich die vermuteten Vor– und Nachteile der Modellvarianten in der beobachteten Ortungsleistung widerspiegeln.

### *Testsignale*

Mit dem Ziel, verallgemeinerbare Aussagen treffen zu können, wurde versucht, eine hinreichende Menge repräsentativer Schallereignisse unter einigen typischen akustischen Bedingungen zu reproduzieren. Die Auswahl der Testsignale sollte sicherstellen, dass der prinzipielle Einfluss der Breitbandigkeit sowie der Gestalt der Hüllkurve der Geräusche beurteilt werden konnte. Neben verschiedenen artifiziellen Rauschsignalen kam eine Reihe natürlicher Geräusche zum Einsatz. Besonderes Augenmerk wurde auf Sprachsignale gelegt, die aufgrund eines meist unscharfen Anfalls und ihrer inherenten Variabilität in spektraler Zusammensetzung und Hüllkurve hohe Ansprüche an die Lokalisationsalgorithmen stellen. Aufnahmen reiner Töne wurden nicht benutzt, da diese infolge der Interferenzbildung für schmalbandige Signale in geschlossenen Räumen weder aussagekräftig noch für potentielle Anwendungen relevant sind. Die Testsignale zur vergleichenden Untersuchung der auditorischen Modellvarianten beinhalten im Einzelnen:

- Weißes Rauschen mit verschiedenen Hüllkurven (Burst, Rechteck, Gauß).
- Farbigen Rauschen (pink noise) mit entsprechenden Hüllkurven.
- Ein kurzes (schmalbandiges) Klick–Geräusch, wie es für Präzedenzexperimente verwendet wird.
- Ein in freier Umgebung aufgenommenes Händeklatschen (breitbandig).
- Einzelne ein– oder zweisilbige Wörter im Abstand von einer Sekunde (Englische Zahlwörter 'one', 'two', ..., 'eight').
- Ein kontinuierliches, ca. 5 Sekunden langes Sprachsignal.

Um Unterschiede im Lokalisationsverhalten für längere Signalabschnitte und die mehr oder weniger ausgeprägten Onsets der Geräusche zu beurteilen, wurden in den Aufnahmen manuell Anhallphasen indiziert. Auf diese Weise entstand ein Set von 32

Geräuschen, das vom kurzen Noise Burst bis zu einem kontinuierlichen, fünf Sekunden langen Sprachsignal eine erhebliche Bandbreite an Lokalisationsaufgaben bereithält.

### *Beispielhafte akustische Bedingungen*

Die beschriebenen Algorithmen zur Bestimmung des ITD–Ortscodes setzen eine ausreichende Korreliertheit der Stereokanäle voraus. In einer akustisch freien Umgebung oder in speziell präparierten, hallarmen Räumen werden sich rechtes und linkes Signal einer freien Mikrofonanordnung jedoch so ähneln, dass jedwede Korrelationsverfahren die interaurale Laufzeit korrekt ermitteln. Um aber relevante Eigenschaften und Unterschiede zwischen den Modellvarianten herauszustellen, sollte eben die Korreliertheit der binauralen Signale durch den Einfluss einer realen Raumakustik beeinträchtigt werden. Dazu wurden die zuvor in einer hallarmen Umgebung aufgenommenen Testgeräusche als Monosignale über eine hochwertige Wiedergabeeinrichtung in verschiedenen Räumen reproduziert. Die Aufnahme der exemplarischen Schallereignisse erfolgte durch zwei Kondensatormikrofone mit Kugelcharakteristik bei einem Basisabstand von 15cm. Da selbst bei frei positionierten Mikrofonen die Korreliertheit der Stereokanäle mit der seitlichen Auslenkung der Quelle abnimmt, wurde die Anordnung aus mehreren Winkeln (0, 30 und 70 Grad) beschallt. Die folgenden Räume und Aufnahmesituationen sollen typische akustische Szenarien darstellen:

- Ein großer Hörsaal mit ca. 1 Sekunde Nachhallzeit. 4m Abstand zwischen Mikrofon und Schallquelle.
- Ein spartanisch eingerichteter Seminarraum mit außerordentlich schlechter Akustik. 2m Abstand zwischen Mikrofon und Schallquelle.
- Derselbe Seminarraum. 6m Abstand zwischen Mikrofon und Schallquelle.
- Ein kleines Arbeitszimmer mit dem für mehrere, eingeschaltete PCs typischen Hintergrundgeräusch. 1m Abstand zwischen Mikrofon und Schallquelle.

Nach der Reproduktion der Testsignale aus drei Azimut–Winkeln in vier akustischen Aufnahmesituationen standen 384 Schallereignisse zur Auswertung bereit. In Verbindung mit einer Script–basierten Simulation der sieben Modellvarianten konnten demnach 2688 elementare Lokalisationsexperimente bewertet werden.

### 2.3.3 Ergebnisse und Wertung

#### *Analyse im Zeitbereich*

Zur Beurteilung der ITD-Kodierung kommen zwei prinzipielle Sichtweisen auf die Simulationsergebnisse in Frage: die Auswertung eines gemittelten Ortscodes oder die Analyse des zeitlichen Signalverlaufes an einer konkreten ITD-sensitiven Position. Für das biologische System und für spikekodierte Modelle versuchen KEMPTER und GERSTNER eben diese Bewertung des Koinzidenzverhaltens im Zeitbereich. Wie schon COLBURN in seinem Hörnervmodell [Col73] gehen sie von Poisson-verteilten Spikezeiten aus. Den theoretischen Rahmen ihrer Überlegungen bilden Kohärenzmaße für die Spikefolgen vor und nach der Koinzidenzdetektion. Aus dem Verhältnis der Kohärenzmaße leiten KEMPTER und GERSTNER einen Qualitätsfaktor ab und benutzen diesen als Optimierungskriterium für neuronale Modellparameter [KGWvH99]. Am Beispiel des Integrate-and-fire Modells von GERSTNER zeigen sie, dass die Reizschwelle der Koinzidenzzellen in einem großen Interval kaum Einfluss auf die Qualität der ITD-Kodierung hat. Dieses Ergebnis stellt aber gleichzeitig die Notwendigkeit komplexer neuronaler Modelle für die Koinzidenzdetektion in Frage und kann eher als Rechtfertigung von idealen binären Koinzidenzzellen angeführt werden.

Die spezifischen Kohärenzmaße für Spikemuster sind nicht direkt auf die hier benutzten Modellvarianten mit kontinuierlicher Kodierung übertragbar. Sowohl für die binäre UND-Verknüpfung im idealen Koinzidenzdetektor als auch bei der Kreuzkorrelationsfunktion kann der Ansatz von KEMPTER und GERSTNER ohnehin nicht greifen, da sich die Kohärenzeigenschaften der Eingangs- und Ausgangssignale nicht unterscheiden. Darüberhinaus wurde bereits in Abschnitt 2.2.6 deutlich, dass nach erfolgter Überführung der zeitlichen Phasenkodierung in den ITD-Ortscode eine unvermindert hohe zeitliche Präzision der Kodierung nicht mehr zu motivieren ist. Von einer weitreichenden Analyse des Modellausganges im Zeitbereich wird daher abgesehen.

In zumindest indirektem Zusammenhang mit der Beurteilung des zeitlichen Verlaufes der ITD-Abbildung steht die überraschende Beobachtung, dass mit der Dynamik der rekurrenten WTA-Prozesse die typischen Zeitfenster von Präzedenzeffekten (abhängig vom Geräusch ca. 5–20ms, vergl. [Moo97]) realisiert werden können.

#### *Bewertung im ITD-Ortscode*

Greift man die heikle Frage nach prozentualen Erfolgsraten der Lokalisationsexperimente auf, bedeutet dies, dass eine Bewertung der Ergebnisse im Merkmalsraum des ITD-Ortscodes vorzunehmen ist. Als pragmatische Lösung kann das Maximum in der

über die Dauer des Experiments gemittelten ITD–Abbildung bestimmt werden. Von einer exakten Lateralisation kann gesprochen werden, wenn die Position des Maximums mit der zur tatsächlichen Richtung des Geräusches korrespondierenden ITD übereinstimmt. Um dem diskreten Charakter des Ortscodes und eventuellen Ungenauigkeiten im experimentellen Aufbau Rechnung zu tragen, müssen auch unmittelbar benachbarte Winkelintervalle als korrekte Ergebnisse gewertet werden.

Für viele Aufgaben ist eine hohe Präzision bei der Winkelbestimmung gar nicht erforderlich. Insbesondere im Rahmen primärer multisensorischer Mechanismen wird sich zeigen, dass aufgrund räumlicher Disparitäten zwischen auditorischer und visueller Repräsentation eine grobe Richtungsschätzung ausreicht. Zur Steuerung der Aufmerksamkeit kann es unter Umständen schon als erfolgreiches Verhalten gewertet werden, wenn eine Signalquelle nach einer vagen auditorischen Ortung wieder ins Blickfeld gebracht wird. Gerade im multisensorischen Kontext erscheint es daher hilfreich, ein Histogramm der Ortungsfehler heranzuziehen und neben der Anzahl exakter Lokalisationen einen lokalen Bereich zulässiger Abweichungen zu definieren. Im konkreten Fall könnte sich der maximale erlaubte Fehler an der Geometrie des multisensorischen Szenarios orientieren. Außerhalb des festgelegten Bereiches ist der Ortungsversuch als missglückt zu beurteilen. Ein Beispiel für eine abgestufte Erfolgsbewertung mit Hilfe von Fehlerhistogrammen ist in Abbildung 2.19 dargestellt.

Wie nicht anders zu erwarten war, traten unter günstigen akustischen Bedingungen (Abbildung 2.19 links) bei keinem Algorithmus nennenswerte Lokalisationsfehler auf. Im Zusammenhang mit den Analyse phasengekoppelter Informationen wie der ITD kann ein akustisches Szenario pauschal als vorteilhaft eingestuft werden, wenn sich die Mikrofone im Hallradius der Schallquelle befinden. Nur dann nämlich dominiert das direkte Schallfeld der Quelle die ITD–Abbildung. In den Experimenten traf dies auf die Anordnungen im Büroraum und im Hörsaal zu.

Die akustischen Rahmenbedingungen änderten sich grundlegend bei den Aufnahmen im Seminarraum, der für seine problematische Halligkeit und schlechte Hörsamkeit bekannt war und als „Worst Case Szenario“ prädestiniert erschien. Schon in den Signalen der nur zwei Meter von der Quelle entfernten Mikrofone überwog der diffuse Schallanteil, wodurch verlässliche Phaseninformationen nur im Anhall der Geräusche auftraten. Folgerichtig stellten Sprachsignale mit ihrer eher unscharfen Anhallphase den Problemfall bei der Ortung dar. Selbst unter Einbeziehung der manuell markierten Onset–Intervalle bescheinigen die ITD–Fehlerhistogramme ein Versagen der einfachen Kreuzkorrelation der Mikrofonensignale (XC, Abbildung 2.19, rechts). Die Korrelation der Frequenzbandsignale des Innenohrfilters schneidet sogar noch etwas schlechter ab:

ihr potentieller Vorteil der Beschränkung auf einen ITD-relevanten Frequenzbereich kommt nicht zum Tragen, weil die störenden Raumresonanzen gerade im unteren Teil des akustischen Spektrums auftreten.

Interessant ist, dass bereits die binäre Kodierung der binauralen Signale durch Spikefolgen den Anteil der größeren Ortungsfehler gegenüber den Korrelationsverfahren mit kontinuierlichen Signalen etwa halbiert. Voraussetzung dafür scheint eine geeignete Phase locked Response des Rezeptormodells zu sein, denn die Resultate einer spezifischen Onset-Kodierung oder der inhibierten Koinzidenzdetektion nach LINDEMANN und WOLF (CD-i) sind unter realen Bedingungen indiskutabel. Wie in Abschnitt 2.2.5 vermutet wurde, hat der Versuch, eine schärfere Abbildung im ITD-Ortscode bereits auf Korrelator-Ebene herbeizuführen, gerade bei einer ungünstigen Raumakustik eine spärliche und unsichere Kodierung der Phasenlage zur Folge.

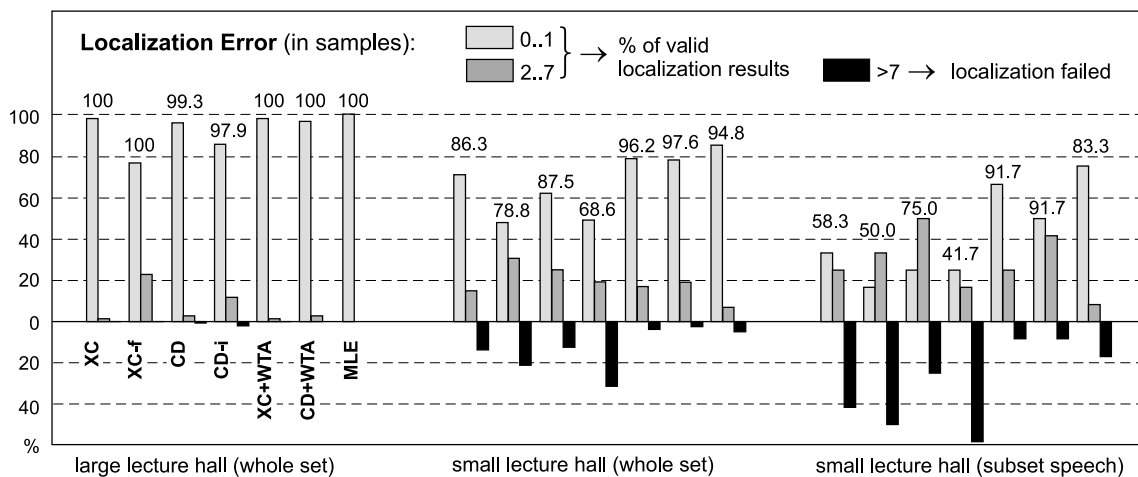


Abbildung 2.19: Histogramme der Lokalisationsergebnisse für ausgewählte akustische Bedingungen und Testsignale. Die Histogramm-Container entsprechen den Fehlerintervallen bei der ITD-Schätzung in Zeitschritten der Simulation. Die Auflösung des Ortscodes betrug bei 0.15m Mikrofonbasis und 44.1kHz Systemtakt 39 Samples (1 Sample = 22.7 $\mu$ sec). Exemplarisch wurden Abweichungen bis zu 7 Samples als gültige, größere Fehler (8–38 Samples) als erfolglose Ortung gewertet.

- XC: Kreuzkorrelation der Mikrofonssignale
- XC-f: Kreuzkorrelation der Frequenzbandsignale des Innenohrfilters
- CD: binäre Koinzidenzdetektion der simulierten Hörnervmuster
- CD-i: Koinzidenzdetektion mit kontralateraler Inhibition
- XC+WTA: Kreuzkorrelation der Mikrofonssignale und WTA-Verarbeitung
- CD+WTA: Koinzidenzdetektion und WTA-Verarbeitung
- MLE: Maximum-Likelihood Schätzung nach MODDEMEIJER

Erst die Filterung des ITD–Ortscodes im rekurrenten WTA–Netz verbesserte die Erfolgsrate bei der Lateralisation der Sprachsignale drastisch. Offensichtlich unterdrückt die Hysterese in der ITD–Repräsentation selbst unter schwierigen Bedingungen effektiv den Einfluss von Echos und Resonanzen. Ob bei der vorangegangenen Korrelation die kontinuierlichen Audiosignale oder die spikekodierte Frequenzbänder des Innenohrmodells verwendet werden, scheint im konkreten Fall zweitrangig zu sein. Eine deutliche Streuung der Ergebnisse um die exakte Position im ITD Ortscode, insbesondere bei der Korrelation der spikekodierte Signale, ist möglicherweise der geringen Anzahl der Delay Lines geschuldet. Entsprechend der 16 Frequenzbänder der APG–Filterkaskade wurden auch nur ebenso viele Delay Line Paare implementiert. Bisweilen wird für Koinzidenzmodelle vorgeschlagen, in jedem Frequenzband eine *große Zahl* von Rezeptoren und Delay Lines unter Einbeziehung von Rauschen zu simulieren. Die Generierung der Spikemuster soll auf diese Weise den Charakter eines stochastischen Prozesses erhalten, der für die Abweichung einzelner Spike–Zeitpunkte bei der Phasenkodierung weniger anfällig ist [Wol91]. Zumindest für die hallbehafteten Aufnahmen konnte dieser von WOLF beabsichtigte Vorteil, anhand einer parallelen Berechnung von Haarzellen, deren Reizschwelle mit einem 10–20%–igen Rauschen beaufschlagt wurde, nicht nachvollzogen werden. Auch sind der Idee, stochastische Eigenschaften durch massive Parallelität und Systemrauschen im Modell nachzubilden, allein schon durch den resultierenden Berechnungsaufwand Grenzen gesetzt.

Sinnvoller lassen sich wahrscheinlichkeitsbasierte Ansätze in der prinzipiellen Form einer Maximum–Likelihood Schätzung formulieren, bei der die interaurale Laufzeit die Rolle des zu optimierenden Parameters übernimmt. Zu diesen Methoden zählen die generalisierte Kreuzkorrelation [KC76] oder die im Zeitbereich deklarierte informationstheoretische Delay–Schätzung nach MODDEMEIJER [Mod88]. Letztere benutzt die mutuelle Information<sup>10</sup> in zwei Signalvektoren als Kontrastfunktion, deren Minimum in Abhängigkeit von einem Verschiebeparameter ermittelt wird. Sie ist unmittelbar auf Stereosignale anwendbar und diene hier als Referenzmethode. Aus Sicht der Informationstheorie werten Korrelationsverfahren nur den unmittelbaren Zusammenhang zwischen Wertepaaren  $[x(t), y(t+\tau)]$  aus, weshalb sich die Korreliertheit zweier Signale bereits anhand der Kovarianz als statistisches Moment zweiter Ordnung beschreiben lässt. Probabilistische Verfahren bieten den prinzipiellen Vorteil, mit einer adäquaten

---

<sup>10</sup>Der mutuelle Anteil der Information in Stereosignalen wird durch beide Kanäle gemeinsam kodiert. Entspricht der Delay–Parameter in einem probabilistischen Signalmodell der tatsächlichen Stereolaufzeit, weisen rechtes und linkes Signal die höchste Abhängigkeit auf und der Informationsgehalt, der nur durch die Kombination der Kanäle kodiert wird, ist am geringsten (s. Anhang B.1).



Likelihood-Funktion auch komplexere Abhängigkeiten als die Korrelation der Signale zu bewerten. Beispielsweise ist mit dem Algorithmus von MODDEMEIJER die Berechnung der mutuellen Information mit beliebig hoher Ordnung möglich.

Die theoretische Überlegenheit der probabilistischen Referenzmethode im Vergleich zu den weniger mächtigen Verfahren der Kreuzkorrelation oder Koinzidenzdetektion spiegelt sich erwartungsgemäß in den Histogrammen der Lokalisationsergebnisse wider. Umso bemerkenswerter ist aber das gute Abschneiden der WTA-gefilterten Korrelationsmuster. Zwar ermittelte der Maximum-Likelihood Schätzer etwas häufiger das exakte Delay, seine falschen Hypothesen streuen jedoch gleichmäßig über den kompletten ITD-Ortscode. Die abweichenden Ergebnisse bei der WTA-Filterung liegen meist dicht neben der korrekten Position. Auf diese Weise erhöhte sich im dargestellten Benchmark die Rate der gültigen Lateralisationsergebnisse auf 91.7%, während MODDEMEIJERS Algorithmus immerhin 83.3% erreichte. Angesichts der äußerst ungünstigen akustischen Bedingungen sind diese Resultate beeindruckend, insbesondere da in keiner Modellvariante eine Optimierung oder gezielte Anpassung der Parameter an die einzelnen Aufnahmesituationen durchgeführt wurde.

Zur Definition von Erfolgsraten bei der Lokalisation wurde in den Fehlerhistogrammen ausschließlich die Position der maximalen Aktivierung im ITD-Ortscode ausgewertet. Darüberhinaus hatten die Modellerweiterungen zur schärferen ITD-Abbildung im Koinzidenzdetektor oder zur nichtlinearen Inhibition im ICx eine Manipulation der Gestalt der Abbildung zum Ziel. Um sowohl die Schärfe der ITD-Abbildung als auch eventuelle Störungen durch mehrdeutige Phasenbilder oder Echos und Raumresonanzen zu beurteilen, kann ein einfaches Qualitätsmaß eingeführt werden. Es setzt den Wert des Maximums im Ortscode ins Verhältnis zur Summe der gesamten Aktivierung der ITD-Karte. In Abbildung 2.20 ist dargestellt, wie beispielsweise infolge der binären Kodierung schon eine Unterdrückung mehrdeutiger Phasendifferenzen und eine schärfere Ausprägung der maximalen Aktivierung bewirkt wird. Demzufolge wird der Ortscode des binären Koinzidenzdetektors (CD) durch das  $max/\sum$ -Qualitätsmaß höher bewertet als das Ergebnis der Korrelation der kontinuierlichen Signale (XC-f).

Da es in der Natur der WTA-Netze liegt, die Energie der Aktivierung auf einen lokalen Bereich zu konzentrieren, kann mit dem Kriterium gleichzeitig überprüft werden, ob in den Simulationen reguläre Selektionsprozesse stattgefunden haben. Anhand der Gegenüberstellung der Qualitätsmaße aller Modellvarianten (Abbildung 2.20, rechts) kann dies für die diskutierten Experimente bestätigt werden.

Abgesehen vom Manko der etwas ungenaueren Ergebnisse unter extrem schlechten akustischen Bedingungen lieferten die Kreuzkorrelationsverfahren in Verbindung mit

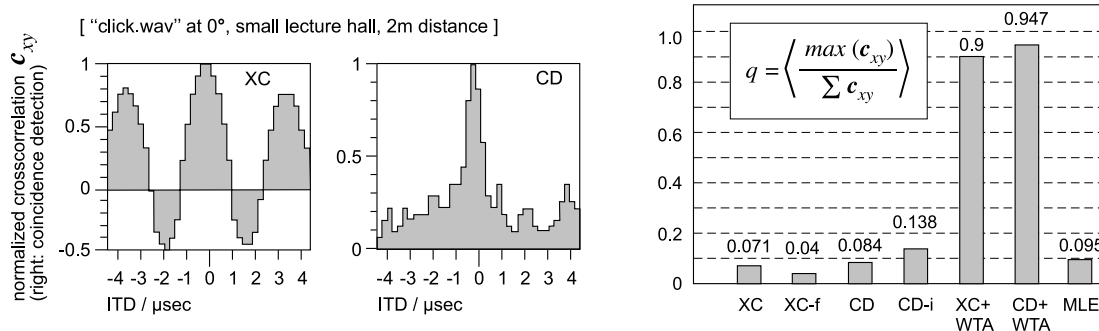


Abbildung 2.20: Links: Zwei exemplarische, mittels Kreuzkorrelation und binärer Koinzidenzdetektion berechnete ITD-Muster für dasselbe schmalbandige Klick-Geräusch weisen unterschiedlich ausgeprägte Phasenartefakte auf.

Rechts: Während die Lokalisations-Histogramme die Zuverlässigkeit der Verfahren veranschaulichen, stellen die  $\max/\sum$ -Qualitätsmaße der getesteten Algorithmen eine Art Signal-to-Noise Wertung der Richtungsabbildung dar. Höhere Werte entsprechen eindeutigeren Lokalisationsergebnissen und bestätigen außerdem das Selektionsverhalten der WTA-Filterung.

einem WTA-Filter bemerkenswert gute Resultate. Mit der Zielstellung, eine notfalls grobe, aber dafür robuste, räumliche Orientierung für primäre Aufmerksamkeitsmechanismen zu realisieren, erscheinen die WTA-Modelle dem probabilistischen Referenzverfahren in nichts nachzustehen. Im Vergleich zu dem ähnlich leistungsstarken Maximum-Likelihood Schätzer MODDEMEIJERS bietet die Kombination der Kreuzkorrelation der Zeitsignale mit einem WTA-Filter (XC+WTA) die Vorteile eines effizienten Online-Algorithmus. Die aufwendigere Simulation von Innenohrfilter und tonotop verteilter Korrelation konnte die Anzahl der deutlichen Lokalisationsfehler nicht weiter verringern. Zwar würde der tonotope Ansatz bei einer ebenbürtigen ITD-Verarbeitung die Integration zusätzlicher, spektral kodierter Merkmale in einem erweiterten auditorischen Modell erleichtern. Solange aber die spektrale Merkmalsanalyse im Timing-Pfad noch nicht beherrscht wird, ist der Nutzen der beschriebenen, tonotopen Simulationsmodelle, gemessen an ihrem Berechnungsaufwand, vergleichsweise gering.

# Kapitel 3

## Primäre visuelle Verarbeitung

### 3.1 Auge, Retina und Sehnerv

Die elementaren Unterschiede zwischen Sehen und Hören als komplementäre Sinneswahrnehmungen sind bereits auf Rezeptorebene offensichtlich. Dem seriellen Konzept des akustischen Zeitsignals steht die parallele bildhafte Repräsentation der Umgebung gegenüber. Den Beginn der auditorischen Wahrnehmung bildet die mechanische Transformation des seriellen Schallsignals in die parallele Kodierung des Spektrums entlang der Basilarmembran. Räumliche Informationen werden schließlich erst im Verlauf der zentralen Hörbahn als abgeleitete Merkmale in berechneten Karten abgebildet. Die parallele Perzeption der Retina ermöglicht hingegen die primäre Kodierung sowohl räumlicher als auch spektraler Informationen.

Im Vergleich zur unspezifischen mechano-elektrischen Transduktion im Haarzellen-Ganglien-Komplex des Innenohrs weisen die Zellen der Retina (bei den allermeisten Tierarten) eine Reihe neuroanatomischer und physiologischer Spezialisierungen auf. Bereits auf dem Niveau der Photorezeptoren wird funktionell differenziert die Intensität in Stäbchen-Zellen und die Farbinformation in verschiedenen pigmentierten Zapfen-Zellen verarbeitet.

Als weiterer Unterschied zum auditorischen System existiert in der Retina anstelle einer unmittelbaren Verknüpfung zwischen Rezeptoren und Ganglienzellen ein aufwendiges neuronales Netzwerk. Es stellt über Bipolarzellen die Verbindung zum Sehnerv her und realisiert mittels Horizontal- und Amakrin-Zellen zusätzliche laterale Verschaltungen. Eine deutliche Konvergenz und die Ausprägung von On- und Off-Bipolaren in der vertikalen Projektion bilden zusammen mit der horizontalen Verschaltung lateraler Bereiche die Grundlage der ersten rezeptiven Felder (RF). Die Ganglienzellen können nun einerseits die Eigenschaften der rezeptiven Felder aus der

bipolaren Verschaltung übernehmen. Darüberhinaus besitzen sie selbst eine differenzierte Physiologie und spezifische Koppelung an die verschiedenen Typen von Photorezeptoren.

**Parvozellulare Ganglien (P-Zellen)** kommen vor allem im Retinazentrum vor. Sie haben insbesondere im fovealen Bereich kleine rezeptive Felder und liefern damit eine Grundlage zur scharfen Abbildung visueller Details und zur Erkennung von Formmerkmalen. P-Zellen antworten auf kontinuierliche Reize mit einer tonischen Spikefolge. Da sie vorrangig von Zapfen-Bipolaren des gleichen Pigmenttyps angesprochen werden, kodieren sie neben der Helligkeit vor allem lokale Farbinformationen.

**Magnozellulare Ganglien (M-Zellen)** integrieren Reize über größere rezeptive Felder und kommen zur Retinaperipherie hin immer häufiger vor. Sie reagieren auf Signale verschiedener Zapfentypen und zeigen deshalb keine Farbempfindlichkeit. Dafür ermöglichen die große Anzahl an Photorezeptoren sowie der höhere Anteil an Stäbchen, mit denen M-Ganglien verknüpft sind, eine bessere Helligkeits- und Kontrastauflösung. M-Zellen antworten schnell und dabei phasisch und kodieren auf diese Weise zeitliche Veränderungen der Stimulusintensität, wie sie beispielsweise als Folge von Objektbewegungen auftreten.

**Koniozellulare Ganglien (K-Zellen)** sind gleichmäßig über die ganze Retina verteilt. Anstelle abrupt einsetzender Spikebursts ist das Antwortverhalten der koniozellularen Ganglien durch eine langsame Veränderung ihrer Spontanaktivität geprägt. Sie haben weit verzweigte, aber lockere Dendritenbäume und bilden komplexe RF aus, die nicht pauschal durch eine Differenz of Gaussian Form beschrieben werden können. Die großen und komplizierten RF-Formen verursachen eine geringe Ortsauflösung, erlauben zugleich aber die Kodierung von Bewegungsrichtungen für hinreichend starke und ausgedehnte Stimuli. K-Ganglien sollen unter anderem zur Steuerung reflexhafter Blickrichtungsänderungen von Bedeutung sein [RC93, RP95].

Die P-, M- und K-Ganglien der Primaten entsprechen der allgemeinen Klassifikation von X-, Y- und W-Zelltypen der Wirbeltiere. Wie es ihre differenzierte Merkmalskodierung bereits vermuten lässt, besitzen sie spezifische Projektionsgebiete und können folglich als Ursprung funktionell getrennter visueller Pfade angesehen werden. P-Ganglien innervieren den parvozellularen Corpus geniculatum laterale (PLGN), der wiederum die Verbindung zum primären visuellen Kortex (V1) herstellt. Auch M-Zellen aktivieren V1 über eine separate Genuculatumschicht, den

magnozellularen LGN. Außerdem projizieren sie in einen Kern des Thalamus (Pulvinar) und ins Mittelhirn (Superior Colliculus, SC). K-Ganglien sind schließlich sowohl über LGN mit V1 als auch direkt mit den äußeren Schichten des SC verschaltet [Ber88b, Ber88a, LV93, KMMS94].

## 3.2 Retinotope Organisation

Der einfache optische Aufbau unseres Auges mit fester Brennweite und die topographische Anordnung von Rezeptoren und Ganglien in der Retina führen zu einer direkten und eindeutigen Richtungskodierung visueller Reize. Für eine einzelne Ganglienzelle kann ungeachtet ihres P-, M- oder K-Typs eine räumliche Spezifik anhand ihres rezeptiven Feldes beschrieben werden. Sie wird stets von denselben Photorezeptoren angesprochen und antwortet damit immer auf einen bestimmten Bereich des Gesichtsfeldes. Nebeneinander liegende Ganglien reagieren auf benachbarte Richtungen im Gesichtsfeld und projizieren in dieser retinotopen Organisation in die nächsten neuronalen Strukturen. Die strenge räumliche Ordnung der retinotopen Abbildung findet sich folglich in den unmittelbaren Projektionsgebieten der Ganglien, dem LGN, dem SC und Bereichen der Pulvinar wieder.

Primäre rezeptive Felder haben im Rahmen der Retinotopie des visuellen Systems und der Tonotopie der Hörbahn offenbar unterschiedlichen Charakter. Die ersten RF im auditorischen System dienen einer groben Kodierung des akustischen Spektrums im Hörnerv, die erst, wenn es bestimmte Wahrnehmungsleistungen erfordern, mittels lateraler Inhibition in eine detailliertere Abbildung überführt wird. Demgegenüber nimmt die Größe der visuellen RF auf höherem neuronalen Niveau zu und der Abstraktionsgrad der Repräsentation steigt. Dieser Befund belegt, dass die räumliche Abbildung an sich beim Sehen und Hören verschiedene Bedeutung hat. In der zentralen Hörbahn wird die Repräsentation des Raumes als sekundäres Merkmal aus der primären sensorischen Kodierung abgeleitet. Die aufwendige Generierung der meist binauralen berechneten Karten (ITD, IID) geschieht dabei mit dem exklusiven Ziel der räumlichen Orientierung. Zur Klassifikation von Geräuschen leisten Richtung oder Ort als auditorische Merkmale keinen unmittelbaren Beitrag. So lassen sich beliebige akustische Ereignisse in sinnvoller Weise auch ohne räumlichen Kontext etwa als Monosignal reproduzieren. Anders wird im visuellen System eine Reihe von objektspezifischen Eigenschaften wie Größe, Form und Textur räumlich kodiert, was sich auch in der zwangsläufig zunehmenden Größe der rezeptiven Felder manifestiert. Die retinotope Abbildung erfüllt somit zwei verschiedene Funktionen: Sie kodiert zunächst unabhängig von der Anwe-

senheit von Objekten den Ort, und zwar sowohl für Mechanismen in Thalamus und Mittelhirn wie auch für die dorsale topographische Kartierung im Kortex. Gleichzeitig bildet sie die Grundlage für eine Vielzahl kortikaler Verarbeitungsleistungen zur formbasierten Objekterkennung im ventralen Pfad.

Für die zweidimensionale, retinotopische Richtungskodierung sind abgeleitete Merkmale in berechneten Karten nicht erforderlich. Ebenso ist die Rolle der binokularen Auswertung im primären visuellen Kortex nicht mit den massiven binauralen Mechanismen bei der subkortikalen Extraktion auditorischer Merkmale vergleichbar. Während die Verknüpfung der rechten und linken Hörnervmuster für die horizontale und vertikale Ortung unverzichtbar ist, dient das stereoskopische Sehen vor allem der Tiefenwahrnehmung. Bereits bei der Diskussion des auditorischen Modells wurde vermutet, dass die Entfernung von Objekten nur eine untergeordnete Rolle bei der Steuerung der Aufmerksamkeit spielt. Zwar wurden auch im SC als mutmaßlichen multisensorischen Integrationsort binokulare Effekte beschrieben – diese reichen jedoch aufgrund der limitierten Auflösung in der vorrangigen Ortskodierung über M- und K-Ganglien nicht aus, um Entfernungen exakt abzubilden. Vielmehr wird vermutet, dass die vergleichsweise ungenauen binokularen Mechanismen im SC, als Teil der motorischen Kontrolle des Augenpaares, zur Fixation hinreichend naher Objekte dienen, währenddessen die eigentliche Tiefenwahrnehmung nur mit den detaillierteren Repräsentationen im V1 möglich ist [BVB<sup>+</sup>98]. Zumindest für reflexhafte Augen- und Kopfbewegungen, die als elementare motorische Reaktionen womöglich schon auf subkortikalem Niveau ausgelöst werden, sollte nur die Richtungsinformation und nicht die Entfernung ausschlaggebend sein.

Wenn bei der Modellierung der primären multisensorischen Integration die binokulare Auswertung vernachlässigt werden kann, reduziert sich die Aufgabe der adäquaten räumlichen Repräsentation auf eine bloße topologische Transformation der retinalen Abbildung. Retinotop organisierte Karten können vielerorts im visuellen System nachgewiesen werden. Abbildung 3.1 stellt exemplarisch das Gesichtsfeld eines Auges den Repräsentationen im primären visuellen Kortex (V1) sowie, auf subkortikalem Niveau, im SC des Mittelhirns gegenüber. Als Voraussetzung für die multisensorische Modellierung ist es erforderlich, eine kompatible räumliche Abbildung der auditorisch und visuell ermittelten Richtungsinformationen zu garantieren. Diese Aufgabe ist insofern nicht trivial, da sich das Auflösungsvermögen beim Sehen und räumlichen Hören vom zentralen zum peripheren Bereich in spezifischer Weise ändert. Außerdem muss aufgrund von Augenbewegungen von einem variablen Bezug zwischen visuellen und auditorischen Koordinaten ausgegangen werden. Nachdem in Kapitel 2 die Geometrie der

auditorischen Richtungsabbildung beispielsweise mit Hilfe der Azimutwinkel–Karte des ITD–Ortscodes exakt zu beschreiben war, sollen nun die Eigenschaften der retinotopen Projektionen erörtert werden.

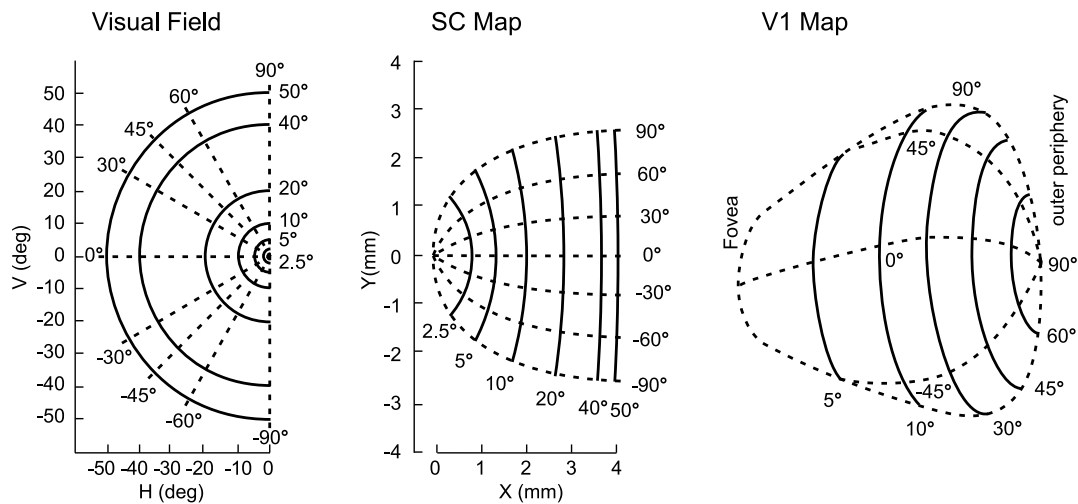


Abbildung 3.1: Gesichtsfeld und retinotopie Geometrie im SC des Affen [AR98, QAOR98] und im primären visuellen Kortex (V1) des Menschen [Hub90]. Beim Menschen umfasst der foveale Bereich der höchsten Ortsauflösung ca. 1 Grad; der zentrale Teil des Gesichtsfeldes, in dem die Zapfen–Photorezeptoren das Farbsehen ermöglichen, ist auf etwa 5 Grad begrenzt. Die restlichen lateralen Positionen werden zum peripheren Gesichtsfeld gezählt [Zek93].

Jede Spezies hat im Laufe der Evolution ein an Lebensraum und Lebensweise angepasstes Gesichtsfeld entwickelt. Dessen absolute Größe ist für die Modellierung nur insofern von Belang, als dass im multisensorischen Konzept typischerweise von einer begrenzten visuellen Abbildung und einer im peripheren Bereich weiterreichenden auditorischen Repräsentation ausgegangen wird. Demnach kann ein konkreter Winkelbereich im visuellen Modell für bestimmte Anwendungen und Szenarien arbiträr festgelegt werden. Interessanter ist der allgemeine Befund, dass sich das Auflösungsvermögen vom zentralen zum peripheren Bereich meist deutlich verschlechtert, und dass in den retinotopen Karten eine überproportionale Repräsentation der „Bildmitte“ zu beobachten ist. Hier drängt sich der Vergleich mit der Lokalisationsschärfe beim räumlichen Hören auf, die im medialen Bereich der horizontalen Ebene am höchsten ist. Bei einer solchen Gegenüberstellung muss jedoch darauf hingewiesen werden, dass die bessere Auflösung zentralen Richtungen bei der Schallortung schlichtweg der lateralen Positionierung der Ohren geschuldet ist. Am Beispiel der geometrischen Zusammenhänge bei der Entstehung der Stereolaufzeit wurde deutlich, wie eine Auslenkung der Quelle um einen bestimmten Azimutwinkel nahe der Medianebene die größtmögliche Differenzierung binauraler Merkmale bewirkt (vergl. Abbildungen 2.1 und 2.12).

Anders beschreibt die Nichtlinearität in der visuellen Abbildung, unabhängig von der optischen Geometrie des Auges, die anatomischen Eigenschaften der Retina und der retinotopen Projektionen. Während sich die auditorische Winkelauflösung in einem größeren zentralen Abschnitt nur allmählich ändert und erst peripher abfällt, ist im visuellen System eine klare Unterscheidung zwischen dem begrenzten fovealen Bereich mit hoher Schärfe und einer deutlich geringeren peripheren Auflösung möglich. Die Ausdehnung der Fovea, in der die größte Dichte von P-Ganglien und die kleinsten rezeptiven Felder vorkommen, ist in verschiedenen Spezies sehr unterschiedlich. Die Fovea des Menschen umfasst nur etwa ein Grad des Gesichtsfeldes. In einem etwas weiteren Segment von fünf Grad, der macula lutea, ist der Großteil der Zapfen-Photorezeptoren konzentriert. Nur in diesem Teil ist die Abbildung visueller Details und die Verarbeitung von Farbkontrasten möglich [Zek93].

Zwar ist die Ausprägung des zentralen Gesichtsfeldes im Tierreich durchaus vielfältig, dass der Bereich des scharfen Sehens eher klein ist und durch die binaurale Schallortung kaum noch aufgelöst werden kann, scheint aber keine Spezifik der Wahrnehmung beim Menschen zu sein. Die Betonung eines zentralen Bereiches ist in den kortikalen Repräsentationen ebenso zu beobachten wie in den visuellen Karten des SC (Abbildung 3.1). Im Gegensatz zur unumstrittenen Bedeutung der detaillierten Abbildung in V1 für die spätere Formanalyse ist die funktionelle Beurteilung der fovealen Repräsentation im SC schwieriger. In Verbindung mit der binokularen Sensitivität einiger SC-Neurone könnte die überproportionale Kodierung des zentralen Gesichtsfeldes eine exklusive Komponente von Fixationsmechanismen sein. Zur multisensorischen Fusion topographischer Karten erscheint eine hohe Ortsauflösung gerade in der Bildmitte überflüssig. Einerseits ist die Differenzierung zwischen einem wenige Grad breiten fovealen Segment und der peripheren Abbildung zumindest mit der sehr groben Auflösung im multisensorischen SC nicht mehr möglich (s. Abschnitt 4.2). Zum anderen sind insbesondere zur Initiierung reflexhafter Augen- und Kopfbewegungen gerade die peripheren Positionen wichtig. Für die multisensorische Integration und zur Steuerung einer primären Aufmerksamkeit muss die detaillierte foveale Topologie demnach irrelevant sein.

Als Resümee der Diskussion retinotoper Projektionen kann prognostiziert werden, dass die subkortikale auditorische Karte im ICx und die visuellen Repräsentationen im SC infolge ihrer topographischen Organisation prinzipiell kompatibel sind. Die Konstruktion einer multisensorischen Azimutwinkel-Karte sollte sich mittels einfacher geometrischer Transformationen und ohne weitere Differenzierung fovealer Eigenschaften modellieren lassen.



### 3.3 Subkortikale Aspekte der Aufmerksamkeit

Im visuellen System ist eine umfangreiche subkortikale Extraktion sensorischer Merkmale, wie sie für die Hörbahn diskutiert wurde, nicht erforderlich. Als primäre bildhafte Informationen können Intensität und Farbe in topographisch geordneter Form direkt im primären visuellen Kortex kodiert werden. Im kurzen Projektionsweg zwischen Retina und Kortex entscheidet der thalamische LGN als „attentional filter“ in Abhängigkeit vom kognitiven Zustand über die Weitergabe der Reizmuster [SK90, SG96]. Anders als in den neuronalen Kernen der Hörbahn findet im LGN aber keine weitreichende Manipulation und Extraktion sensorischer Merkmale statt. Gleichwohl lassen sich mannigfaltige Verknüpfungen von visuellen und motorischen Gebieten im Kortex, Thalamus, Mittelhirn und Hirnstamm in afferenter und efferenter Projektionsrichtung verfolgen (Abbildung 3.2). Vermutlich sind solche Befunde Ausdruck dafür, dass viele detaillierte sensomotorische Aufgaben, wie die Fixation von Objekten oder Greifbewegungen, auf Basis der primär räumlichen, visuellen Repräsentation besonders einfach und sicher realisiert werden können.

Die außerordentlich hohe Konnektivität und Parallelität der subkortikalen visuellen Pfade deutet darauf hin, dass auf diesem niedrigen neuronalen Niveau nicht nur motorische Befehle ausgeführt werden. Neben der Ausgestaltung von Greifbewegungen oder Augensakkaden als beobachtbare Verhaltenskomponenten könnten Thalamus und Mittelhirn mit ihrer räumlichen und motorischen Kompetenz den Kortex auch bei verdeckten kognitiven Leistungen, wie der Verlagerung der Aufmerksamkeit oder der Planung von Bewegungen, unterstützen. An dieser Stelle sind die Differenzierung verschiedener visueller Mechanismen und die Interpretation ihres Beitrags zur Aufmerksamkeitssteuerung auf unterschiedlichem neuronalen Niveau wünschenswert. Bereits in Abschnitt 1.2 wurden visuelle Aufmerksamkeitsmodelle vorgestellt und ihre enge Bindung an die Objekterkennung kritisch bewertet. Nach einer genaueren Betrachtung der primären retinalen Reizmuster können die in Kapitel 1 vorangestellten Überlegungen nun anhand einer Diskussion der wesentlichen kortikalen und subkortikalen Projektionswege des visuellen Systems konkretisiert werden.

*P- und M-Pfade & die Doktrin der dorsalen „Wo?“ und ventralen „Was?“-Verarbeitung*

Angesichts der kaum überschaubaren Masse an Befunden zum visuellen System verheißt die strikte Unterteilung zwischen räumlicher Orientierung und ortsinvarianter, objektspezifischer Repräsentation ein Mindestmaß an Ordnung und Übersicht. Tatsächlich korrespondieren die charakteristischen Kodierungen und Projektionsgebiete

der Ganglien-Typen mit der von MISHKIN ET AL. vertretenen Doktrin einer dorsalen „Wo?“ und ventralen „Was?“-Verarbeitung [Zek93]. Unterscheidet man pauschal zwischen farbspezifischen (color opponent, co) und breitbandigen (bb) Projektionen, lässt sich beispielsweise die Farbverarbeitung im ventralen Pfad nachzeichnen: ausgehend von den Schichten 2 und 3 des V1 direkt oder über V2 nach V4 und weiter in den inferior-temporalen Kortex (IT), der als „*höchstes neuronales Niveau der Objekterkennung*“ bezeichnet wurde [MUM83]. Während die Kodierung der Farb-Opponenten ausschließlich über P-Ganglien erfolgt, ist die breitbandige Aktivierung ein Merkmal von M- und K-Zellen. Ausgehend vom Projektionsziel der M-Ganglien (Schicht 4B in V1) sind breitbandige Kodierungen ihrerseits charakteristisch für den dorsalen „Wo?“-Pfad. Dieser soll direkt und indirekt über V2 nach V5 und von dort aus weiter in den parietalen Kortex verlaufen.

ZEKI kritisierte das Dogma der strikten Trennung zwischen dorsalen und ventralen Wahrnehmungsleistungen. Weder könne im Verlauf der kortikalen Verarbeitung von einer exklusiven Projektion der M- und P-Kodierung ausgegangen werden, noch sei eine funktionelle und anatomische Unabhängigkeit der dorsalen und ventralen Areale gegeben [Zek93]. Schon bei der Diskussion der retinotopen Abbildungen, die auch in ventralen Arealen verbreitet sind, wurde deutlich, dass die Unterscheidung von „Wo?“- und „Was?“-Aspekten nicht etwa die Unabhängigkeit der Objekterkennung von der räumlichen Wahrnehmung bedeutet. Gerade zur Detektion von Form und Textur sowie zur Erzeugung von Skalierungs- und Rotationsinvarianzen müssen zunächst räumliche Zusammenhänge ausgewertet werden. Auch die an räumliche Repräsentationen gekoppelte Bewegungskodierung ist trotz ihres Ursprungs in den Reizmustern der M- und K-Ganglien keine exklusive Domäne der dorsalen Verarbeitung. Eine visuell erfasste Bewegung kann neben der Lokalisation eines Objektes ebenso zu dessen Formerkennung in IT beitragen [SWD<sup>+</sup>03]. Allerdings weisen dorsale und ventrale Verarbeitung einen signifikanten Unterschied auf. Die Topographie der dorsalen Bahn ist egozentrisch organisiert, um Relationen zwischen dem eigenen Körper und Objekten zu kodieren und deren zielgerichtete Manipulation zu ermöglichen. Demgegenüber sind die ventralen Abbildungen allozentrisch: sie geben Aufschluss über die Gestalt und die Relation zwischen Objekten einer visuellen Szene [SCF04].

Die konzeptionellen Stärken und Schwächen der „Was?“ vs. „Wo?“ Doktrin stehen in engem Zusammenhang mit den in Abschnitt 1.2 skizzierten Aufmerksamkeitsmodellen. Beispielsweise korrespondiert die Annahme einer unabhängigen dorsalen „Wo?“- und ventralen „Was?“-Verarbeitung unmittelbar mit der Trennung zwischen räumlicher Selektion und selektivem Sehen. Die Modelle zur parallelen und seriellen Suche

lassen sich wiederum mit der ventralen Merkmalsanalyse vergleichen. Dabei entsprechen parallele Komponenten der Kodierung von Farbe, Orientierung und Form in V1, V2 und V3, während die serielle Analyse von Merkmalskonjunktionen anhand der Objektklassen in V4 und IT realisiert wird. Dieses abstrakte Schema der seriellen und parallelen Suche lässt außer Acht, dass bereits auf subsymbolischer Ebene, etwa bei der Extraktion von Formeigenschaften, eine im Detail serielle Informationsverarbeitung zu beobachten ist [SWD<sup>+</sup>03]. Ebenso bleibt die Frage des Zusammenspiels der topographischen Kodierung einzelner Merkmale mit der ortsinvarianten Repräsentation der Objektprototypen letztendlich ungeklärt.

In Wettbewerbsmodellen und bei der Formulierung einer objektbasierten Aufmerksamkeit werden dorsale und ventrale Elemente durch die Integration der Objektposition in abstrakten Merkmalschablonen (Attentional Templates) kombiniert. Auch hier bedarf aber die Forderung nach einer Verknüpfung von dorsaler, egozentrischer und ventraler, allozentrischer Repräsentation einer konkreten Lösung. Einen Modellansatz könnten dazu Befunde über das FEF (Frontal Eye Field) liefern, das sowohl mit dorsalen Komponenten (posterior parietaler Kortex, PP) als auch mit dem Areal V4 des ventralen Pfades verschaltet ist. Die Repräsentationen im FEF haben den Charakter von Auffälligkeitskarten [TB05], in denen Inhibition of Return Mechanismen eine serielle Auswahl von Zielpositionen ermöglichen [RFC03, SMKB95, SH98]. Anhand dieser Eigenschaften sind die wesentlichen Ideen der Theorie einer seriellen Suche und das Modellkonzept der selektiven Bahnung nachvollziehbar. Darüber hinaus spielt das FEF aufgrund seiner Lage in efferenten Projektionswegen [MTS01] eine wichtige Rolle bei verschiedenen Top-Down gerichteten Projektionen der selektierten Zielpunkte. Einerseits kann die visuelle Suche als verdeckte Wahrnehmungsleistung (covert Behavior) durch einen direkten Wiedereintritt der Zielvorgaben in den ventralen Verarbeitungspfad demonstriert werden [TMS<sup>+</sup>01, TSE92, Ham03, Ham05]. Davon unabhängig lösen subkortikale Efferenzen motorische Reaktionen in Form von Augensakkaden aus (overt Behavior) [OMCW04, SW00, MJCW03, Sch02a]. Analog zur verdeckten visuellen Suche können auch Sakkaden zur Fovealisierung neuer Regionen im Gesichtsfeld als Formen der aufmerksamen Wahrnehmung angesehen werden [GG89].

Um eine kohärente topographische Organisation der Wahrnehmung zu garantieren, muss die Auswirkung einer geänderten Blickrichtung umgehend in den dorsalen und ventralen Abbildungen berücksichtigt werden. Tatsächlich ist der vom FEF angesprochene Superior Colliculus des Mittelhirns nicht nur für die Ausgestaltung von Augenbewegungen verantwortlich, sondern stellt über den Thalamus eine topographische Rückkopplung zu kortikalen Arealen her [LHP94, GPFM02, SW04b, SW04a].

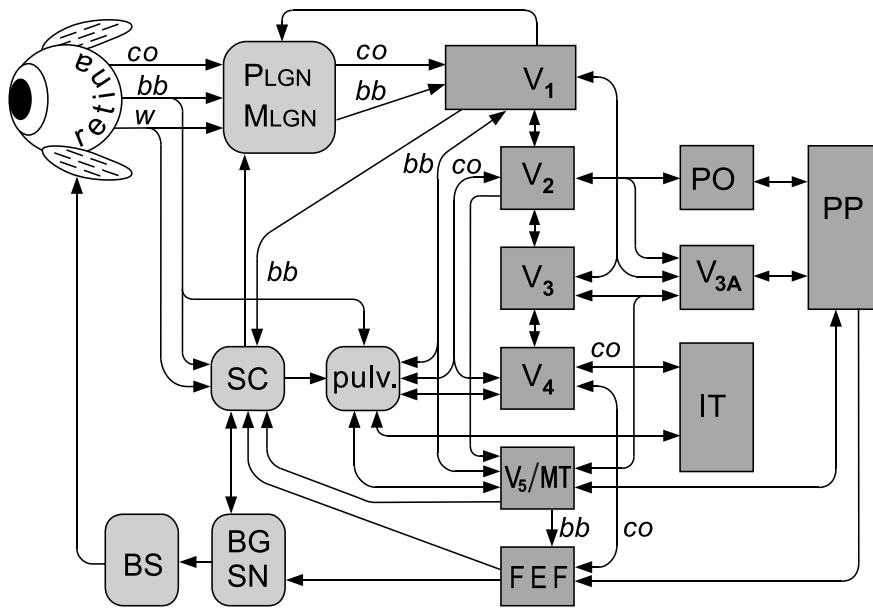


Abbildung 3.2: Übersicht der visuellen Verarbeitungspfade. Kortikale Areale sind als dunkel schraffierte Rechtecke dargestellt, helle, abgerundete Rahmen symbolisieren subkortikale Zentren. An der Ausführung der Pfeile sind reziproke oder vorrangig unidirektionale Projektionen sowie die afferente oder efferente Natur einer Verknüpfung zu erkennen. Mit der Retina fungiert das Auge einerseits als Rezeptor, andererseits verweist die Efferenz zum Augenmuskel auf sensomotorische Reaktionen. Entsprechend der unterschiedlichen Ganglientypen sind Projektionen mit P-Kodierung (color opponent, co), M-Kodierungen (broadband, bb) oder K-Zellen Response (w-like) gekennzeichnet. Sofern bekannt, wurde auch der Verlauf von M- und P-Pfad über den Corpus geniculatun laterale (LGN) hinaus auf diese Weise markiert. Ein **dorsaler Pfad** führt vom primären visuellen Cortex (V1) sowohl direkt als auch indirekt über Areal V2 zum medial temporalen Areal MT (oft als V5 bezeichnet). Außerdem werden von V2 aus das parieto-occipitale Areal (PO) sowie von V1 und V2 aus das Areal V3A angesprochen. PO und V3A projizieren weiter in den posterior parietalen Komplex (PP). Zum **ventralen Pfad** zählen das über V2 angesprochene Areal V4 und der inferior-temporale Kortex (IT). Direkte kortikale Projektionswege über V1 werden als **striäre Sehbahn** bezeichnet. Eine **extra-striäre Bahn** verläuft über den im Mittelhirn gelegenen Superior Colliculus (SC) und die thalamische Pulvinar zum parietalen Kortex. An der motorischen Steuerung von Augensakkaden beteiligt sind Efferenzen über das Areal FEF (Frontal Eye Field), SC, Basal-Ganglien (BG), die Substantia nigra (SN) und Hirnstamm (BS). Einzelne Befunde werden in [MG95, Lab95, Joh95, CS94, Sch85, VEO94] erörtert.

### *Striärer und extra-striärer Pfad*

Die Afferenzen vom Mittelhirn über den Thalamus in verschiedene Kortexareale erklären unter anderem den Effekt des Gaze Locking in V3, der von ZEKI als Beweis topographischer Mechanismen im ventralen „Was?“-Pfad angeführt wurde. SC und Bereiche der Pulvinar dienen aber nicht einfach zur Rückmeldung, welche Zielpositionen der FEF-Auffälligkeitskarte letztendlich eine motorische Reaktion ausgelöst haben. Mit

ihrem direkten retinalen Input über M- und K-Ganglien steht parallel zum LGN ein eigenständiger afferenter Projektionsweg zur Verfügung. Die bislang erörterten Affenzen von LGN nach V1 und die folgenden dorsalen und ventralen Pfade werden unter Bezugnahme auf die Neuroanatomie und die streifenartig organisierten Merkmalskarten im primären visuellen Kortex einer striären Sehbahn zugeordnet. Dementsprechend kann ein zweites, anfangs subkortikal verlaufendes visuelles System als extrastriäre Sehbahn bezeichnet werden [ZvC79]. Im Gegensatz zu den umfassenden und detaillierten Kenntnissen über den primären visuellen Kortex sind die strukturellen und funktionellen Zusammenhänge der extrastriären Affenzen bislang weniger gut verstanden. Der unbewusste Charakter subkortikaler Wahrnehmungsleistungen erschwert verhaltensbasierte Experimente, und die genauen Funktionen einiger thalamischer Bereiche wie der Pulvinar sind noch nicht geklärt. Ungeachtet vieler Parallelen zur kortikalen Architektur sei es unwahrscheinlich, dass in dieser Region nur eine mehr oder weniger redundante Replikation kortikaler Mechanismen stattfindet. Vielmehr scheint die Pulvinar aktiven Einfluss auf die räumliche, merkmalspezifische oder objektbezogene Wichtung der Aufmerksamkeit im Kortex auszuüben [Shi03]. Insbesondere bei der Repräsentation von Bewegungen im Areal MT (V5) wurde der extrastriären Bahn über den SC des Mittelhirns schließlich auch eine selbständige Merkmalskodierung unterstellt [RGA89, RGA90], die eine Ursache dafür ist, dass MT schneller auf neue Reize reagieren kann als beispielsweise Areal V4 im ventralen Pfad [PMA88, RXMO99].

Einen wichtigen Zugang zum Verständnis der extrastriären Wahrnehmung liefern pathologische Befunde über Läsionen einzelner visueller Hirnbereiche und die damit einhergehenden Defizite. Der Verlust höherer dorsaler oder ventraler Funktionen lässt sich zunächst im Einklang mit der Doktrin der separaten „Wo?“- und „Was?“-Verarbeitung erklären. Bei Patienten mit parietalen Läsionen sind massive Probleme bei visuell geführten Sakkaden, Zeige- und Greifbewegungen zu beobachten (visuelle Ataxie, [BMFM<sup>+</sup>00, BMC02]). Sind hingegen ventrale Areale wie die inferior temporale Region geschädigt, tritt eine Formagnosie ein: die Patienten können sich nahezu normal orientieren und bewegen, haben aber Schwierigkeiten bei der Klassifikation von Objekten [ZK98, FS84, DBDU93].

Problematischer ist die Interpretation sogenannter Blindsight-Befunde [CS91, AC01]. Patienten mit Läsionen in V1 können trotz intakter Retina und LGN die Informationsverarbeitung im ventralen Pfad nicht über die striäre Bahn initiieren. Sie sind nicht fähig, elementare visuelle Merkmale oder gar Objekte zu erkennen, noch ihre Umgebung in irgendeiner bildhaften Form bewusst wahrzunehmen und gelten als klinisch blind. Dennoch reagieren sie reflexhaft oder sogar mit gezielten motorischen

Reaktionen auf bestimmte visuelle Stimuli. Es überrascht wenig, dass die Wahrnehmungsleistungen von Blindsight-Patienten mit den Eigenschaften der subkortikalen Projektionsgebiete von M- und K-Ganglien korrespondieren. Ohne die detaillierte P-Typ-Kodierung des fovealen Bereiches in V1 kann lediglich eine grobe, topographische Bewegungsanalyse realisiert werden. Die Fähigkeit zur gezielten motorischen Reaktion deutet darauf hin, dass die extrastriäre Sehbahn nicht nur eine alternative Afferenz zu höheren Arealen bereitstellt, sondern efferente Strukturen anspricht, die im gesunden Sehsystem auch der bewussten Interaktion mit der Umwelt dienen.

### *Subkortikale Funktionen*

Die Generierung von Zielpunkten für die visuelle Suche oder zur Initiierung von Bewegungen wurde bislang stets in Verbindung mit höheren dorsalen und ventralen Mechanismen diskutiert, die letztendlich auf der exakten topographischen Merkmalskodierung im striären Kortex basieren. Ohne striären Input müssen die bei Blindsight-Patienten erhalten gebliebenen Leistungen aber allein auf der räumlichen Kodierung in SC und Pulvinar beruhen. Als zentrale Frage bei der funktionellen Beurteilung der subkortikalen Zentren gilt es nun zu klären, an welchem Ort im extrastriären System Ziele selektiert und Motorkommandos ausgelöst werden können.

Verfolgt man zum einen die extrastriären sensorischen Afferenzen und zum anderen die motorischen Efferenzen vom FEF, fällt auf, dass beide Bahnen im Mittelhirn den Superior Colliculus (SC) als gemeinsames neuronales Zentrum durchlaufen. Die mutmaßlichen Funktionen des SC sind die Steuerung von Augensakkaden, die Kodierung der Blickrichtung für ein Korrektursignal in kortikalen Repräsentationen und die indirekte Inhibition of Return zur visuellen Suche im FEF [DKEM02]. Diese Aufgaben alleine bedürfen nicht zwingend einer Verknüpfung von sensorischen Informationen und Motorkommandos und könnten ebenso in getrennten Projektionswegen gelöst werden. Die wohl herausragendste Eigenschaft des SC ist jedoch gerade die massive Überlagerung sensorischer und motorischer Karten [EG01]. Es gibt kaum plausible Gründe, warum eine deutliche Aktivierung des SC durch einen visuellen Stimulus nicht auch eine direkte Befehlskodierung in den korrespondierenden prämotorischen Neuronen bewirken kann. Die Voraussetzungen für die Bestimmung von motorischen Zielpunkten sind mit retinotopen Karten und weitreichenden lateralen Verknüpfungen auch im SC erfüllt [MI98, MF02]. Schon im SC können Signale zur Selektion von Zielpunkten und zur Kodierung von Motorkommandos unterschieden werden [HN99, HN01], und es ist unwahrscheinlich, dass dieser Befund nur eine redundante Ausprägung der Mechanismen im FEF beschreibt. Indes konnte der experimentelle Nachweis einer ei-

genständigen motorischen Leistung, etwa anhand von Latenzzeiten zwischen Stimulus und Augensakkaden, infolge der komplexen Wechselwirkung mit anderen Arealen noch nicht erbracht werden [Spa99, SFCG01].

Umstritten ist auch die Interpretation der relativ geringen räumlichen Auflösung in den sensorischen Karten des SC. Im Ergebnis der Kodierung über M- und K-Ganglien erscheint die Größe der extrastriären rezeptiven Felder zur exakten Sakkadensteuerung ungeeignet. Auch ein Modell zur verteilten Kodierung [McI91] kann dieses Problem nicht vollständig lösen. Die Vermutung, dass die thalamische Pulvinar-Region die Genauigkeit bei der Generierung von Zielpositionen erhöht [RM89, RP92], wurde durch jüngere Untersuchungen eher widerlegt als bestätigt. Offensichtlich reagieren auch die Neurone der Pulvinar nur auf räumlich ausgedehnte Stimuli und tragen somit kaum zur Steuerung präziser Motorkommandos bei [SS00]. Je exakter eine motorische Kontrolle der Augen geführt werden soll, umso wichtiger erscheint die Rolle der akkuraten, striären Kodierung und deren noch nicht vollständig verstandene Projektion von V1 über die dorsale Bahn zum FEF [ST03].

Angesichts des heterogenen disziplinären Hintergrundes der neurobiologischen, klinisch-medizinischen und psychologischen Untersuchungen ergibt die schier überwältigende Zahl an Befunden zur striären und extrastriären Verarbeitung ein anspruchsvolles Puzzle. Ein allgemeines und einheitliches Modell, das allen Formen von Aufmerksamkeitsmechanismen und allen Aspekten der Sakkadensteuerung gerecht wird, kann schwerlich entworfen werden. Insbesondere die funktionelle Interpretation des SC vereinfacht sich aber wesentlich, wenn bei der motorischen Kontrolle der Augen Teilaufgaben formuliert werden.

- Der Fixation einer Position durch Mikrosakkaden sollte zwingend eine präzise topographische Kodierung in V1 [MCMH00] und eine ventrale, merkmalsbezogene Zielauswahl vorausgehen. In diesem spezifischen Zustand der Aufmerksamkeit fungiert der SC lediglich als Werkzeug zur Umsetzung der motorischen Zielvorgaben aus dem FEF.
- Auch die Ausgestaltung größerer Sakkaden kann, als beobachtbares Pendant zur verdeckten visuellen Suche, der kortikalen Kontrolle unterliegen: um verschiedene Regionen in merkmals- oder objektbasierten Auffälligkeitskarten sequentiell zu fokussieren, aber auch um den Aufbau solcher Karten überhaupt erst zu ermöglichen.
- Im Falle des unerwarteten Auftretens neuer visueller Stimuli ist es schließlich vorstellbar, dass eine unwillkürliche Zuwendung der Blickrichtung unmittelbar

in den sensomotorischen Karten des SC ausgelöst wird [Ken86]. Da sich eine wie auch immer geartete Neuheit meist in Form von Bewegung, zumindest aber als Intensitätsänderung manifestiert, wäre die phasische Kodierung der M- und K-Ganglien im extrastriären Pfad zu diesem Zweck ohnehin prädestiniert [WWB<sup>+</sup>04]. Selbst mit einer groben Ortsauflösung der retinotopen Abbildung würde die ungefähre Fokussierung eines neuen Stimulus seine folgende Analyse im ventralen Pfad beschleunigen. Zeichnet sich der so gefundene Bereich durch einzelne Merkmale oder seine Übereinstimmung mit einem Attentional Template aus, kann seine Position in Auffälligkeitskarten integriert werden und in die Aufmerksamkeitsmodi der Fixation oder der visuellen Suche zurückgekehrt werden.

Die hier formulierte These einer mehrfachen Rolle des SC bei der entweder bewussten oder unbewussten Wahrnehmung wird im Rahmen verschiedener Untersuchungen zwar nicht explizit vertreten, aber zumindest durch die angeführten Befunde gestützt. Beschrieben wurde unter anderem ein Wettbewerb sensorischer Afferenzen mit efferenten motorischen Mechanismen [TDMK01] sowie eine mit der Fixation einsetzende Kontrolle des FEF über den SC [SW00]. Außerdem gelang es, verschiedene räumliche und dynamische Eigenschaften von Augensakkaden mit einem einfachen Hirnstamm-SC-Regelkreis zu modellieren [WSG05]. Ungeachtet der im Detail noch nicht verstandenen Mechanismen gilt es als wahrscheinlich, dass der SC einen eigenständigen und alternativen Beitrag zur Sakkadensteuerung leistet [RMKP91, MK02].

## 3.4 Modellierung

### 3.4.1 Bewegung als universelles Merkmal

Die grundsätzliche Frage, welche visuellen Merkmale sich in der räumlichen Repräsentation niederschlagen sollen, kann anhand der neurobiologischen Befunde diskutiert und aus dem Modellansatz der frühen Aufmerksamkeitssteuerung abgeleitet werden. Untersuchungen zur Klärung der multisensorischen Effekte in Teilen des SC bescheinigen, dass die Neurone in dessen visuellen Subarealen ausschließlich auf Intensitätsänderungen und somit auf bewegte Stimuli reagieren [WWS96]. Diese Beobachtung steht in völligem Einklang mit der funktionellen Interpretation des SC als extrastriärer Neuheitsdetektor, der motorische Reflexe bei plötzlichen Veränderungen im visuellen Stimulus auslöst. Seine phasisch kodierten Afferenzen über M- und K-Ganglien können auch gar keine anderen Merkmale übertragen: Farbinformationen fehlen im SC, da er nicht mit P-Ganglien und damit nicht mit pigmentierten Photorezeptoren verbunden



ist. Orientierungsvarianzen oder detaillierte Konturmerkmale gehen infolge der großen rezeptiven Felder, zumindest im peripheren Bereich, in einer unscharfen Abbildung verloren. Auch über seinen efferenten Input scheint der SC entweder nur Bewegungsmerkmale (aus dem Areal MT) oder abstrakte räumliche Informationen (vom FEF) zu erhalten, die nicht unmittelbar an spezifische Farb- oder Formeigenschaften der ventralen Abbildung gekoppelt sind.

Die Unabhängigkeit von objektspezifischen Eigenschaften ist letztendlich aber auch eine logische Konsequenz aus der Aufgabe, die subkortikale Repräsentationen bei der Sicherstellung einer primären Aufmerksamkeit erfüllen. In gefährlichen Situationen kann beispielsweise das schnelle Auslösen von Reflexen noch vor der Objektklassifikation kostbare Reaktionszeit sparen. Und auch, um die Wichtung von ventralen Merkmalen in Attentional Templates an neue Objekte anzupassen, müssen die höheren Aufmerksamkeitsmechanismen des Kortex zunächst mit unabhängigen Zielpositionen überschrieben werden.

Aus pragmatischer Sicht zeichnet sich Bewegung als robustes Merkmal im extrastriären Bottom-Up Pfad durch äußerst vorteilhafte Invarianzen aus. Klassische Probleme der Bildverarbeitung, die in Verbindung mit der Orientierung von Objekten, deren Textur, der Struktur eines Bildhintergrundes, oder bei variablen Beleuchtungsverhältnissen auftreten, brauchen nicht explizit behandelt zu werden. Ob in einer visuellen Szene helle oder dunkle, einfarbige oder strukturierte Objekte in Aktion treten, ist zunächst unwichtig – allen gemein ist die Ausprägung von Intensitätsunterschieden, sobald sie sich bewegen.

### 3.4.2 Abstraktion im Simulationsmodell

Im Vergleich zum auditorischen Modell, in dem binaurale, zeitlich kodierte Richtungsmerkmale zunächst in eine topographische Darstellung überführt werden müssen, ist die Aufbereitung der visuellen Daten unkompliziert. Die hier ausreichenden monokularen Informationen sind von Natur aus räumlich organisiert und zuverlässig. Während akustische Probleme, wie die Halligkeit von Räumen, laute Störgeräusche oder simultane Quellen, die Qualität der auditorischen Richtungsmerkmale massiv beeinträchtigen, leidet unter visuellen Störeinflüssen nicht die räumliche Abbildung sondern die Objekterkennung. Darum betreffen klassische Aufgaben der visuellen Vorverarbeitung die Realisierung vieler Invarianzleistungen gegenüber Farbschwankungen, Ausrichtung und Gestaltmerkmalen. Die Konzentration auf die Detektion von Bewegung als objektunspezifisches Merkmal macht derartige Berechnungen überflüssig. Zudem entkräftet

der grundsätzliche Verzicht auf die Auswertung von Farbinformationen ein wesentliches Argument für die detaillierte Modellierung der retinalen Physiologie. Die Eigenschaften spezifisch pigmentierter Photorezeptoren oder eine folgende Konstruktion adäquater Farbräume ist für die Abbildung im Superior Colliculus irrelevant.

### *Topographie*

Zur Festlegung der topographischen Modelleigenschaften müssen die allgemeinen Prinzipien der retinotopen Organisation letztendlich arbiträr auf die konkrete Geometrie einer technischen Versuchsanordnung übertragen werden. Für die folgenden Experimente stand auf Basis einer omnidirektionalen Optik ein Gesichtsfeld mit einem verzerrungsfreien und beliebig großen horizontalen Winkelbereich zur Verfügung. Exemplarisch wurde ein ca.  $\pm 100$  Grad umfassendes Bild über gaußförmig gewichtete rezeptive Felder in die modellierte retinotop Karte projiziert. Die in Abbildung 3.3 dargestellte Geometrie orientiert sich an der ungefähren Angabe einer RF-Größe von 1–15 Grad im SC der Primaten [WWS96]. Mit zunehmend peripherer Position nimmt nicht nur die Breite, sondern auch der Abstand der rezeptiven Felder zu, wodurch die überproportionale Repräsentation des zentralen Gesichtsfeldes berücksichtigt werden soll. Für eine Näherung der topographischen Verhältnisse im SC (vergl. Abb. 3.1) werden Breite und Position der RF anhand von Exponentialfunktionen ermittelt:

$$RF_i \sim \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2} \frac{(x-\mu_i)^2}{\sigma_i^2}}; \quad \text{mit} \quad \sigma_i \sim 1 - e^{\alpha i} \quad \text{und} \quad \mu_i \sim e^{\beta i - 1} \quad (3.1)$$

Über den Parameter  $\alpha$  wird die Zunahme der RF-Größe bestimmt, während  $\beta$  die Positionierung der RF-Zentren steuert. Die Parameterpaare  $(\sigma_i, \mu_i)$  der einzelnen gaußförmigen RFs sind wiederum proportionale Werte, die entsprechend der Auflösung des diskreten Bildes zu skalieren sind. Die topographische Betonung des zentralen Gesichtsfeldes ist im Kontext einer frühen Aufmerksamkeitssteuerung schwer zu interpretieren. Möglicherweise ist eine besondere visuelle Bewertung in diesem Bereich vornehmlich für Fixationsmechanismen von Bedeutung und stellt eine Modelloption dar, die zur Simulation SC-interner sensomotorischer Leistungen eine untergeordnete Rolle spielt.

Im Verlauf der retinotopen Projektionen tritt die stärkste topographische Nichtlinearität beim Übergang vom fovealen in den peripheren Bereich auf. Anders wird im auditorischen ITD-Ortscode ein sehr viel größerer zentraler Bereich nahezu gleichmäßig hoch aufgelöst, bevor die Lokalisationsschärfe in der äußeren Peripherie deutlich sinkt. Um eine exakte Übereinstimmung der topographischen Register zu garantieren, sind vor der multisensorischen Fusion weitere geometrische Transformationen unumgänglich.

In Anbetracht der Limitierung des auditorischen Modells auf die Abbildung horizontaler Richtungen liegt es schließlich nahe, auch bei der visuellen Verarbeitung vertikale Richtungen nicht separat zu bewerten und das Ergebnis einer Bewegungskodierung vor der Anwendung der RF-Geometrie für jede Pixel-Spalte des Bildes zu summieren. Die vorgeschlagene Geometrie der rezeptiven Felder hat folglich eindimensionalen Charakter, wodurch weder die prinzipielle Diskussion multisensorischer und prämotorischer Mechanismen noch die experimentelle Evaluierung der Simulationsmodelle beeinträchtigt wird.

### *Primäre Bewegungskodierung*

Da eine Beteiligung von pigmentierten Photorezeptoren und P-Ganglien an den extrastriären Projektionen ausgeschlossen wurde, markiert die Verwendung von Grauwertbildern den Ausgangspunkt der Modellierung. Die primäre phasische Kodierung von Intensitätsunterschieden durch On- und Off-Bipolare kann auf triviale Weise mit der Berechnung von Differenzbildern abstrahiert werden. Die letztendlich eindimensionale Richtungskodierung legitimiert außerdem eine spaltenweise Summation der detektierten Pixeldifferenzen über die gegebene Bildhöhe. Im Ergebnis repräsentieren Differenzvektoren mit einer gleichmäßig hohen räumlichen Auflösung das mehr oder weniger ausgeprägte Auftreten von Bewegungen in verschiedenen horizontalen Richtungen (Abbildung 3.4b). Die anschließende topographische Transformation bewirkt entsprechend der Geometrie der rezeptiven Felder neben der Nichtlinearität im zentralen Bereich eine erste (räumliche) Tiefpass Filterung (Abbildung 3.4c).

### *Zeitliche Kodierung*

Die dynamischen Merkmale der visuellen Repräsentation können nur in einem vergleichsweise groben Zeitraster gestaltet werden. Ein typischer Videotakt von 30Hz erscheint einerseits ausreichend, um z.B. menschliche Bewegungsmuster abzubilden – auf eine Simulation detaillierter neuronaler Rezeptormodelle, etwa auf dem Niveau der Spike-Kodierung, kann jedoch zweifellos verzichtet werden. Gleichwohl können aus einer makroskopischen Betrachtung der zeitlichen Kodierung in der extrastriären Sehbahn weitere relevante Modelleigenschaften abgeleitet werden. Zum einen ist die Latenz der neuronalen Antworten visueller Neuronen im SC mit ca. 100ms vergleichsweise hoch [WWS96] und selbst bei einem groben Simulationstakt nicht zu vernachlässigen. Es liegt nahe, diesen Befund vor dem Hintergrund der langsameren Schallausbreitung und der daraus resultierenden Verzögerung der akustischen Reize

zu sehen. Mit großen visuellen Latenzzeiten wäre es prinzipiell möglich, eine zeitliche auditorisch-visuelle Disparität zu kompensieren. Gegen einen solchen direkten Zusammenhang spricht wiederum die Variabilität der Verzögerung infolge einer beliebigen und dem Modell nicht bekannten Entfernung der Quellen. Allein die vielgestaltige und nicht zwingend synchrone Ausprägung von Geräuschen und Bewegungen erfordert aber ohnehin schon ein robustes Konzept der zeitlichen Kodierung. Neben der Latenz stellt dabei die Dauer einer neuronalen Antwort einen interessanten Parameter dar. So könnte die Länge der Spikebursts von Ganglien oder SC-Zellen gerade für die Realisierung größerer multisensorischer Zeitfenster von besonderer Bedeutung sein. Obwohl auf diese Fragestellung in aktuellen Veröffentlichungen nicht explizit eingegangen wird, belegen die Post-Stimulus-Zeit-Histogramme visueller Ganglien, dass die Burst-Dauer bei der phasischen Kodierung nicht unerheblich ist. Schon HARTLINE zeigte mit seinen beispielhaften elektrophysiologischen Ableitungen, wie die Aktivität von Off-sensitiven Zellen bis zu 400ms über das Ende der visuellen Stimulation hinaus anhält [Har38, Wie59]. Die bewegungssensitiven Neurone des SC zeichnen sich offenbar durch ein vergleichbares Verhalten aus [HN01].

Im Transferverhalten des modellierten visuellen Projektionsweges lassen sich Off-set- und Latenz-Eigenschaften abstrakt in einer gemeinsamen Notation darstellen: Die hier vorgeschlagene zeitliche Filterfunktion  $\frac{t}{\tau}e^{1-\frac{t}{\tau}}$  wurde bereits zur Simulation der synaptischen Potentiale im Spike-Response Modell benutzt und kann über ihre Zeitkonstante so parametrisiert werden, dass die visuelle Aktivierung ihr Maximum nach 100ms erreicht und in weiteren 400-500ms wieder abklingt (Abbildung 3.3, rechts).

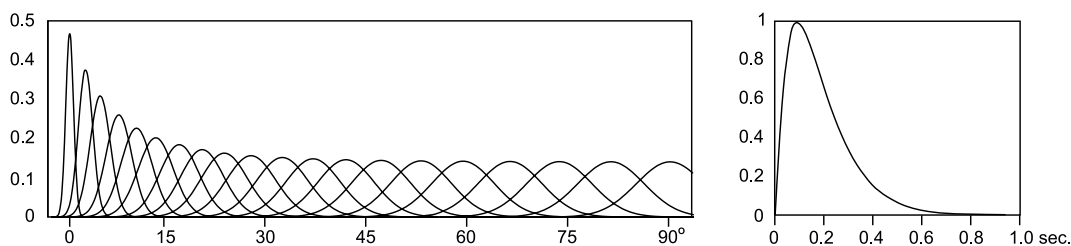


Abbildung 3.3: Rezeptive Felder (links) und Impulsantwort der visuellen Kodierung (rechts).

### *Visueller Dynamikbereich*

Visuelle und auditorische Wahrnehmung zeichnen sich durch eine beachtliche Empfindlichkeit und einen extremen Dynamikumfang aus, gegenüber dem die sensomotorischen Karten im SC in gewissem Maße unempfindlich sein sollten. Im auditorischen Modell können die limitierte Ausgabefunktion der Neuronenmodelle und die binäre Kreuzkor-

relation zur Erzeugung des ITD–Ortscodes eine weitgehend pegelunabhängige Richtungsabbildung garantieren. Bei der visuellen Verarbeitung kommt es, abgesehen von der unvermeidlichen Beschränkung durch die verwendete Kamera, auch infolge der Differenzbildung zwischen aufeinander folgenden Bildern zu einer Limitierung der Dynamik. Schließlich können Amplitude und Dynamik der visuellen Bewegungskarte mit der aus dem auditorischen Modell bekannten Fermi–Funktion in eine sigmoide Form transformiert werden. Als Pendant zur auditorischen topographischen Karte im Modell des Inferior Colliculus soll das Simulationsergebnis in Abbildung 3.4 d) der Bewegungsrepräsentation in den visuellen Subarealen des Superior Colliculus entsprechen.

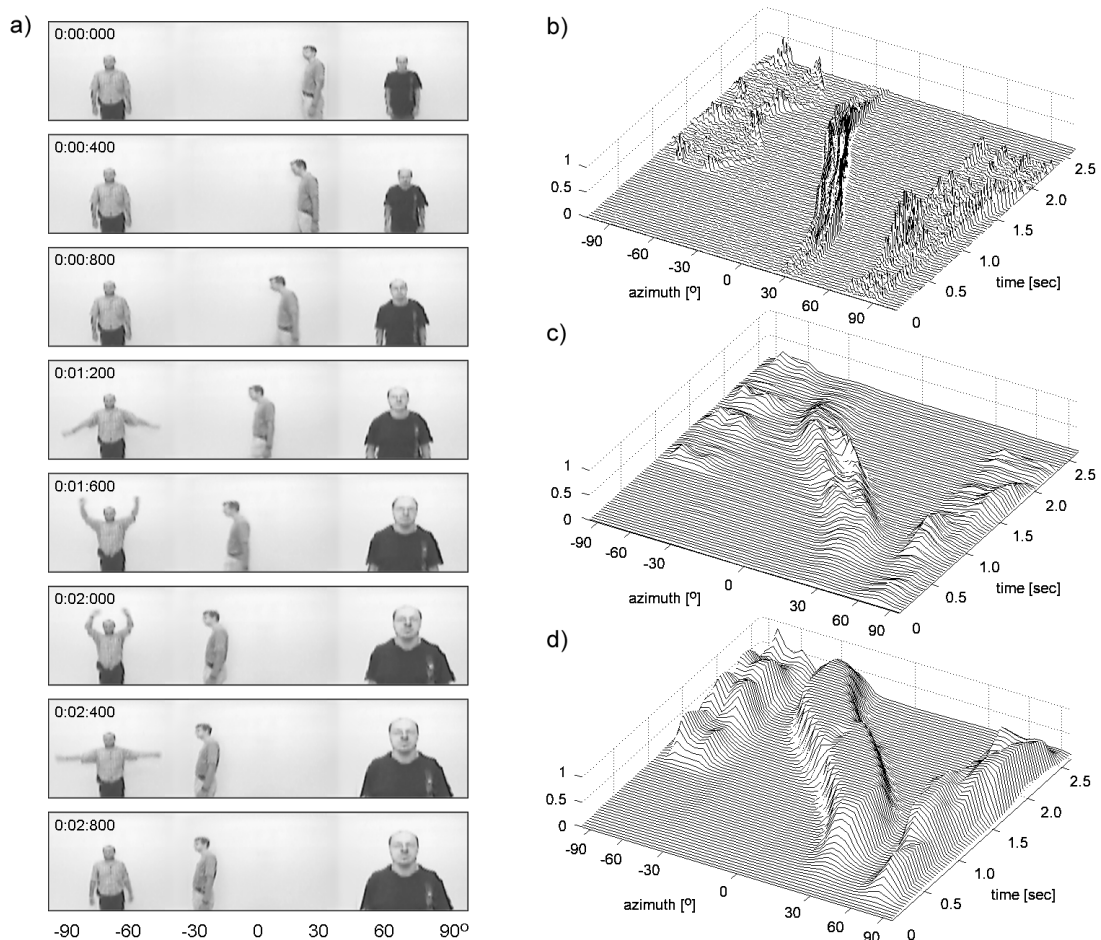


Abbildung 3.4: a) In einem 200–Grad Segment der Bildfolge einer omnidirektionalen Kamera erzeugen 3 Personen exemplarische Bewegungsmuster. b) Spaltensumme der Differenzbilder. c) Anwendung der rezeptiven Felder. d) Zeitliche Filterung und Ausgabe–Normierung.

An der Verarbeitung der exemplarischen visuellen Szene aus Abbildung 3.4 wird deutlich, dass die Bewegungskomponenten unterschiedlicher Objekte gleichberechtigt und unabhängig voneinander kodiert werden. Im Gegensatz zur WTA–Filterung im auditorischen Modell wurde im Rahmen der visuellen sensorischen Repräsentation auf

einen separaten Selektionsmechanismus verzichtet. In der auditorischen Richtungsabbildung war die Bestimmung einer Gewinnerregion unumgänglich, da bereits einzelne Schallquellen mehrdeutige Phasenbilder erzeugen können und der störende Einfluss von Echos und Raumresonanzen unterdrückt werden musste. Die Gefahr, dass in diesem Auswahlprozess weitere Quellen ausgeblendet werden und bei der Aufmerksamkeitssteuerung unberücksichtigt bleiben, ist eher gering. Schallereignisse treten einerseits selten simultan auf und können selbst dann aufgrund ihres meist nicht kontinuierlichen Charakters sequentiell lokalisiert werden. In Anbetracht der Eindeutigkeit und Robustheit der Bewegungsabbildung ist ein WTA-Netz zur Verbesserung der visuellen Repräsentation überflüssig. Die Bewegungsmuster einzelner Quellen sind bereits aussagekräftig, und im Falle simultaner Aktivierungen an mehreren Positionen der Karte könnte sich eine solche Filterung sogar als kontraproduktiv erweisen. Würde in einem visuellen Selektionsprozess unter verschiedenen, auffälligen Richtungen gerade diejenige unterdrückt werden, aus der ein Geräusch zu hören ist, wäre keine sinnvolle multisensorische Fusion mehr möglich. Wenn mehrere sensorische Modalitäten zur verdeckten Steuerung der Aufmerksamkeit oder zu motorischen Reaktionen beitragen, sollte ein weiteres räumliches Selektionsverfahren deshalb erst in der multisensorischen Karte realisiert werden [SP99, SPG01b].

Aus dem Blickwinkel der klassischen Bildverarbeitung oder Objekterkennung muss das Simulationsergebnis ob seiner unscharfen und zeitlich verschliffenen Darstellung erschreckend ungenau anmuten. Tatsächlich sind im Diagramm der Abbildung 3.4d) keinerlei Details mehr zu erkennen – diese sind schließlich für die angestrebten multisensorischen Leistungen auf subkortikalem Niveau auch nicht relevant. Im Gegenteil: die räumliche und zeitliche Glättung, mit der die anatomischen und physiologischen Befunde zum SC in abstrakter Weise modelliert werden sollen, erfüllt diverse konkrete Aufgaben. Beispielsweise werden Bewegungssäume geschlossen und so unstrukturierte oder nahe Objekte eindeutiger dargestellt. Die zeitlich leicht verzögerte und verlängerte Aktivierung soll helfen, große Zeitfenster für multisensorische Effekte zu realisieren. Auch die vollständige Bedeutung der geringen Ortsauflösung wird sich schließlich erst im multisensorischen Kontext und im Verlauf der audio-visuellen Experimente in Kapitel 5 klären.

# Kapitel 4

## Frühe auditorisch–visuelle Integration

### 4.1 Klassifizierung multisensorischer Effekte

In den vorangegangenen Kapiteln wurden räumliche auditorische und visuelle Wahrnehmungsleistungen im Kontext einer primären Aufmerksamkeit und im bewussten Gegensatz zu Klassifikationsaufgaben diskutiert. Die dabei aufgeworfenen Fragestellungen zum Aufmerksamkeits- und Objektbegriff sind prinzipiell auch bei der Fusion mehrerer sensorischer Modalitäten interessant. Zwar betreffen multisensorische Mechanismen in Ermangelung „multisensorischer Rezeptoren“ zwangsläufig mehr oder weniger abstrakte Merkmalsrepräsentationen – sie bleiben aber keinesfalls einer symbolischen, kognitiven Ebene vorbehalten [Ste98, KC01, CC01]. Im Rahmen der Wettbewerbsmodelle zur Aufmerksamkeitssteuerung können multisensorische Aspekte auf einem frühen Niveau in Merkmalschablonen (Attentional Templates) oder auf einer höheren, ventralen Ebene in Objektprototypen integriert werden. Als erster Versuch einer Einordnung dieser Effekte wurde in Kapitel 1 die Klassifikation in Crossmodal Matching, Integration und Learning nach STEIN und CALVERT zitiert. STEINs renommierte Darlegung multisensorischer Wahrnehmungsleistungen [SM93] und CALVERTs Bestandsaufnahme der jüngeren Forschungsergebnisse [CC01] sollen nun auch den Ausgangspunkt für eine detaillierte Erörterung bilden, an deren Ende neurologische Vorbildstrukturen und Modellanforderungen benannt werden.

Eine Annäherung an multimodale Wahrnehmungskonzepte über die Definitionen von Crossmodal Matching und Integration führt schnell zu der Erkenntnis, dass die beiden scheinbar plausiblen Begriffe durchaus unterschiedlich ausgelegt werden. STEIN beschreibt Crossmodal Matching als kognitive Leistung, bei der beispielsweise visuelle und taktile Objekteigenschaften verglichen werden. Dabei könne entweder eine der beteiligten unisensorischen Repräsentationen als Referenz fungieren, die durch die jeweils

andere gewichtet wird. Denkbar sei aber auch, dass multisensorische Informationen in einer separaten, amodalen Form gespeichert werden. Als Grundlage einer amodalen Kodierung formuliert STEIN in Anlehnung an die Arbeiten von STEVENS und PRICE eine allgemeine These: *„Alle Stimuli können ungeachtet ihres sensorischen Ursprungs auf kontinuierlichen Skalen bezüglich ihrer Intensität, Größe, Anzahl, Dauer oder ihres räumlichen Ursprungs eingestuft werden: leise und laute Geräusche lassen sich in gleicher Weise ordnen wie schwache und helle Lichtreize oder unterschiedlich kräftige Druckempfindungen an den Fingerspitzen. Solche Skalen können demnach als allgemeine Metriken dienen...“* [SM93, Ste75, Pri88].

CALVERT schließlich stößt beim Versuch, konkrete Hirnareale für Matching-Effekte verantwortlich zu machen, auf Schwierigkeiten. Er beruft sich auf Untersuchungen, in denen die Rolle bekannter multisensorischer Kortexregionen bei Matching-Experimenten nicht eindeutig geklärt werden konnte und stattdessen die Beteiligung einer Struktur des Frontallappens, der Insula-Clastrum, vermutet wurde [EW90]. Generell betreffe Crossmodal Matching eher höhere und abstraktere kognitive Leistungen und ließe sich folgendermaßen von Integrationsmechanismen unterscheiden: Beim Matching können auch Merkmale verschiedener und womöglich nacheinander wahrgenommener Objekte verglichen werden – die multisensorische Integration verknüpft hingegen unbedingt die aktuellen Repräsentationen desselben Objekts [SM93, Rad94]. CALVERTs vorrangiges Interesse gilt den unmittelbaren Integrationsmechanismen, zu denen er beispielsweise eine multimodale Lokalisation zählt. In Analogie zum Crossmodal Matching grenzt er auch bei Lokalisationsleistungen komplexe Vorgänge mit großen multisensorischen Zeitfenstern von der konkreten Integration auf einem lokalen, neuronalen Niveau ab. Seine Zusammenfassung der long-term-Mechanismen als Crossmodal Spatial Attention korrespondiert nur bedingt mit der Terminologie der in Kapitel 1 genannten Aufmerksamkeitsmodelle. Wenn man etwa einen Beitrag der Crossmodal Spatial Attention an den Aufmerksamkeitsschablonen der Wettbewerbstheorien unterstellt und andererseits annimmt, dass mittels Crossmodal Localization räumliche Objekthypothesen unterstützt werden, bleiben frühe Integrationsorte wie der Superior Colliculus weitgehend unberücksichtigt.

Nach der noch vagen Matching-Definition STEINS scheinen auch die von CALVERT zusammengetragenen jüngeren Befunde bislang kein übergreifendes, multimodales Konzept und insbesondere keine schlüssige Einbindung der frühen, subkortikalen Effekte in abstraktere Wahrnehmungsmodelle zu erlauben. Auf welche Weise könnten subkortikale Regionen wie der SC besser in ein multisensorisches Gesamtbild integriert und die mitunter verwirrend heterogene Verwendung der Begriffe Matching und



Integration vermieden werden? Mit der in dieser Arbeit proklamierten Unterscheidung zwischen Objektklassifikation bzw. bewusstem Tracking einerseits und objektunabhängigen, primären Aufmerksamkeitsmechanismen andererseits, sollen folgende pragmatischen Thesen formuliert werden:

- Der Begriff Matching beschreibt zunächst nicht eine kognitive Leistung, sondern die Tatsache, dass amodale Informationen von unterschiedlichen Sinnesorganen kohärent abgebildet werden können – und zwar ungeachtet des Abstraktionsgrades oder des neuronalen Niveaus ihres Repräsentationsortes.
- Sobald ein multisensorisches Matching vom Gehirn ausgewertet werden soll, müssen zwangsläufig auch mehr oder weniger direkte Integrationsmechanismen auf Basis der von STEIN benannten Metriken realisiert werden.
- Im extrastriären und frühen dorsalen Pfad im Kortex beinhalten die amodalen Metriken ausschließlich unspezifische Informationen (Ort, Intensität, Dauer), auf höherer (ventraler) Ebene dagegen auch Merkmale, die typisch für Objekte oder Objektprototypen sind<sup>1</sup>.

Mit der Anwendung der Idee der Common Metric auf konkrete Integrationsmechanismen und der Unterscheidung von Metriken für Aufmerksamkeit und Objekterkennung kann ohne die widersprüchliche Diskussion des Matching-Begriffes eine Einordnung von Befunden auf verschiedenen neuronalen Niveaus vorgenommen werden. Beispielhafte Leistungen zur Objektdetektion und Klassifikation sind demnach die visuell-taktile Kodierung von orientierten Kanten in V4 [MSND91], die multisensorische objektbasierte Repräsentation im ventralen Pfad beim Menschen [PFR<sup>+</sup>04] oder die Kodierung von Gesichtsausdrücken, Hand- und Körpergesten in Verbindung mit Geräuschen im Areal STS [BXB<sup>+</sup>05]. Typische Aufmerksamkeitsmechanismen betreffen die Fusion räumlicher Informationen wie die auditorische Aktivierung von Positionen in V1, die außerhalb des fovealen Bereiches liegen [FCBK02], oder eben die massiv multisensorischen Eigenschaften des Superior Colliculus (SC) im Mittelhirn [SM93, PBW93, WWS96].

Angesichts der Beteiligung des SC an den komplexen Mechanismen der visuellen Suche und der Sakkadensteuerung wäre es wünschenswert, ein multisensorisches Pendant zu den in Kapitel 3 beschriebenen, visuellen Leistungen zu finden. Befunde wie

---

<sup>1</sup>Objektspezifische, amodale Informationen können wiederum unterschiedliche kognitive Abstraktionsgrade beschreiben – angefangen von elementaren visuell oder taktil erfassten Form- und Struktureigenschaften bis hin zu visuellen oder auditorischen Einschätzungen über Alter, Geschlecht oder emotionalem Zustand von Personen.

die auditorische V1-Aktivierung [FCBK02], die Projektion aus dem FEF in multisensorische SC-Regionen [Mer99] oder die Beteiligung des SC an der Fixation auditorischer Zielpunkte [PB97] deuten solche Leistungen an, ohne dass mit dem derzeitigen Kenntnisstand ein tragfähiges Modell entworfen werden kann. Bei der vermuteten eigenständigen, sensomotorischen Funktion des SC steht die Bedeutung der multisensorischen Integration jedoch außer Frage. Die Diskussion entsprechender Befunde im folgenden Abschnitt 4.2 und deren anschließende Modellierung und Simulation wird zeigen, wie durch die auditorisch-visuelle Fusion schneller und sicherer auf neue Reize reagiert werden kann, oder wie Ereignisse in der Umgebung mit Hilfe der auditorischen Wahrnehmung überhaupt erst ins Gesichtsfeld gebracht werden, damit visuelle und multisensorische Mechanismen greifen können.

Die Unterschiede in den räumlich-zeitlichen Konzepten der verschiedenen sensorischen Systeme führen aber nicht nur zum Informationsgewinn bei der multisensorischen Integration – sie stellen auch eine Herausforderung an die Kodierung in amodalen Metriken dar. Am häufig zitierten Bauchrednereffekt [SM93, CC01, WRH<sup>+</sup>04] wird zum einen deutlich, dass tatsächlich eine Verstrickung von früher sensorischer Integration und symbolischer Verarbeitungsebene (Ortung und Klassifikation, Verarbeiten von Sprache und Mimik) existiert, wodurch die Interpretation der Befunde zu den multisensorischen SC-Regionen erschwert wird. Andererseits ist es gerade interessant, wie die Mechanismen der Aufmerksamkeitssteuerung und Objekterkennung koordiniert werden und welche neuronalen Ebenen und räumlich-zeitlichen Register<sup>2</sup> die Korrespondenz oder Disparität audio-visueller Ereignisse bewerten. Die Realisierung von auditorisch-visuellen Metriken ist vermutlich nicht trivial durch ein lineares Mapping der topographischen Repräsentationen möglich und wird anspruchsvollere Modelle und Algorithmen erfordern.

## 4.2 Multimodale Integration im Superior Colliculus

Dem im Mittelhirn gelegenen Superior Colliculus (SC) wurden bereits bei der Erörterung der visuell gesteuerten Aufmerksamkeit in Kapitel 3 wesentliche Funktionen in afferenten und efferenten Projektionswegen sowie eine eigenständige sensomotorische

---

<sup>2</sup>Der Begriff des Registers bezeichnet in der zitierten Literatur meist pauschal die für Integrationsleistungen notwendige räumliche Korrespondenz zwischen auditorischen, visuellen und somatosensorischen rezeptiven Feldern. Oft werden aber auch die im Detail noch unbekannt, komplexen neuronalen Verschaltungen und selektiven Projektionsmechanismen, die eine solche Korrespondenz erst realisieren, abstrakt als topographische Register beschrieben.

Kompetenz attestiert. Unter den neuronalen Arealen mit multisensorischen Eigenschaften nimmt er eine Sonderstellung ein, die ihn vor allem im Kontext früher Aspekte der Aufmerksamkeit interessant erscheinen lässt: er ist in der afferenten Projektionsrichtung schlichtweg der *erste* multisensorische Integrationsort. Dass die frühestmögliche Fusion der sensorischen Modalitäten gerade im Mittelhirn erfolgt, ist keine zufällige anatomische Gegebenheit und kann am Beispiel der visuellen und auditorischen Pfade plausibel begründet werden. Erst im Inferior Colliculus (IC), dem auditorischen Kern des Mittelhirns, werden alle verfügbaren räumlichen Schallinformationen in einer gemeinsamen topographischen Karte abgebildet. Diese universelle Karte sollte sich wesentlich einfacher mit den räumlichen Repräsentationen anderer Modalitäten verknüpfen lassen als die verteilte Kodierung der zeit- und intensitätsbezogenen Merkmale im auditorischen Hirnstamm. Ein weiteres Argument für das Mittelhirn als idealen Integrationsort liefert die in Kapitel 1 formulierte These, dass die multisensorischen Mechanismen einer frühen Aufmerksamkeit unabhängig von spezifischen Objekteigenschaften sein sollen. Aufgrund der tonotopen Verarbeitung in parallelen Frequenzbändern enthalten aber alle auditorischen Kodierungen im Hirnstamm gezwungenermaßen auch die spektrale Spezifik der akustischen Quellen. Die topographische Karte im externen IC enthält diese Information nicht mehr und stellt eine kompakte und universelle Kodierung bereit – zur kortikalen Bewertung, aber ebenso zur Fusion mit der retinalen Abbildung des benachbarten SC.

Die Übertragung der sensorischen Information zwischen IC und SC ist im Wesentlichen unidirektional. Zwar wird der IC auf einem noch nicht vollständig bekannten Weg von der motorischen Steuerung des Auges beeinflusst, um seine topographischen Repräsentationen an die Blickrichtung anzupassen [GvO99, DBSK00, GTU<sup>+</sup>01, ZVvO04], die tatsächliche auditorisch-visuelle Integration bleibt aber dem SC vorbehalten. Wenn, wie es scheint, die retinotopie Abbildung als Referenz für weitere sensorische Karten eine räumliche Metrik definiert, unterstreicht dies zunächst das etablierte Verständnis vom SC als visuellem Kern mit entsprechenden Funktionen in der extrastriären Sehbahn. Haben deshalb aber auch alle im SC kodierten Informationen vorrangig visuellen Charakter und dient der zusätzliche auditorische und somatosensorische Input nur deren Wichtung? Oder realisieren die multisensorischen Bereiche des SC eine, wie STEIN es nannte, separate amodale Kodierung, aus der man schließlich neben einer visuell basierten auch eine amodale Steuerung der Motorik ableiten könnte. Diese für die spätere Modellierung wichtige Frage soll in den folgenden Abschnitten anhand der Befunde zur internen Struktur, Verschaltung und motorischen Funktion des SC beantwortet werden.

### 4.2.1 Neuroanatomie des Superior Colliculus

Zusammen mit dem auditorischen Inferior Colliculus (IC) bildet der Superior Colliculus (SC) den dorsalen Teil des Mittelhirns, das Tectum. Bereits seine visuellen Funktionen im Rahmen einer subkortikal initiierten, extrastriären Sehbahn lassen eine komplexe Verschaltung und interne Struktur vermuten, die den SC von einer einfachen Umschaltstation der sensorischen Informationen unterscheidet. Im Vergleich zum benachbarten IC geht die neuronale Verarbeitung im SC über die Manipulation einzelner sensorischer Merkmale oder die nichtlineare Filterung topographischer Abbildungen hinaus. Dank der kontinuierlichen Arbeit von STEIN, MEREDITH und WALLACE herrscht heute ein Konsens über multisensorische und motorische Leistungen im SC und deren Bedeutung für die Steuerung von Aufmerksamkeit und räumlicher Orientierung [MS86, WS94, Ste98]. Bedenkt man, dass gerade die motorischen Efferenzen des SC ihren Ursprung in Bereichen haben, deren multisensorischer Charakter seit geraumer Zeit bekannt ist [MW47], wurden die auditorischen und somatosensorischen Aktivierungen dieser Hirnregion erstaunlich lange ignoriert [SM93].

Der SC weist eine laminare Grundstruktur mit insgesamt etwa sieben Schichten auf [WWB<sup>+</sup>04]. Die äußeren drei Schichten (Stratum zonale, Stratum griseum superficiale und Stratum opticum) bilden ein exklusiv visuelles Subareal (SCs, s=superficial), dass in Abschnitt 3.3 für die retinotop Bewegungsrepräsentation als Vorleistung der multisensorischen Integration verantwortlich gemacht wurde. Die Eigenschaften seiner visuellen Kodierung werden maßgeblich von den Projektionen der W-Ganglien in Schicht I und II sowie der Y-Ganglien in die Schichten II und III bestimmt. Außerdem wird der SCs vom primären visuellen Kortex [Ber88b, Ber88a, OMS84] und vom Areal MT (V5) [LBB03] innerviert, wodurch ihm indirekt auch Reizmuster der X-Zellen zugänglich sind [WWB<sup>+</sup>04]. Vermutlich sind die Eingangsprojektionen der äußeren Schichten prinzipiell retinotop organisiert [RMF85]. Sie übertragen ausschließlich die Bewegungskodierung durch Intensitätsschwankungen und keine farb- oder formspezifischen Merkmale [WWS96].

Die mittleren Schichten Stratum griseum intermediale und Stratum album intermediale sowie die innere Stratum griseum profundum und Stratum album profundum werden als tiefe Schichten (SCd, d=deep) zusammengefasst. Da in den meisten Veröffentlichungen nicht zwischen den Strata intermediale und Stratum profundum unterschieden wird, soll im Folgenden pauschal die Bezeichnung SCd verwendet werden. Der SCd ist neben seiner visuellen Aktivierung die Domäne auditorischer, somatosensorischer, multisensorischer und prämotorischer Kodierungen, und steht in diesem

Kapitel folglich im Mittelpunkt der multisensorischen Modellierung. Visuelle Informationen erhalten die tiefen Schichten des SC teilweise von Y-Ganglien [WWB<sup>+</sup>04], vor allem aber aus den äußeren Schichten, dem SCs [RMR<sup>+</sup>89, MK04, TDPC05]. Als afferente Quellen der somatosensorischen Aktivierung des SCd kommen Regionen des Hirnstammes wie die Kerne der dorsalen Kolumne in Frage [GUGMM04]. Sowohl die visuellen als auch die somatosensorischen Repräsentationen haben den Charakter von projizierten topographischen Abbildungen. Die auditorischen Reizmuster, die den SCd über den Inferior Colliculus erreichen, mussten demgegenüber eine aufwendige Transformation von spektralen und zeitlichen, monauralen und binauralen Kodierungen in eine berechnete Karte durchlaufen [KJM98].

Projektionen aus den primären auditorischen oder somatosensorischen Kortexarealen gibt es offenbar weder in den äußeren noch in den tiefen Schichten des SC. Vermutungen, dass beispielsweise die auditorischen Regionen AI oder AII mit dem SCd verschaltet sind [GJP68, KK79], wurden unter anderem durch die Untersuchungen von STEIN und MEREDITH widerlegt [Ste78, TRSL80, MC89]. Einigkeit besteht hingegen über die bereits in Abschnitt 3.3 erwähnte topographisch geordnete Verbindung vom frontalen Augenfeld (FEF) zum SCd [Mer99, SW00]. Tatsächlich existieren aber noch weitere kortikale Efferenzen, die gerade die multisensorischen Eigenschaften des SCd beeinflussen. Sie wurden jedoch nicht in den primären, sondern in assoziativen Regionen der sensorischen Kortexes gefunden, und zwar an einer Stelle, an der visuel-

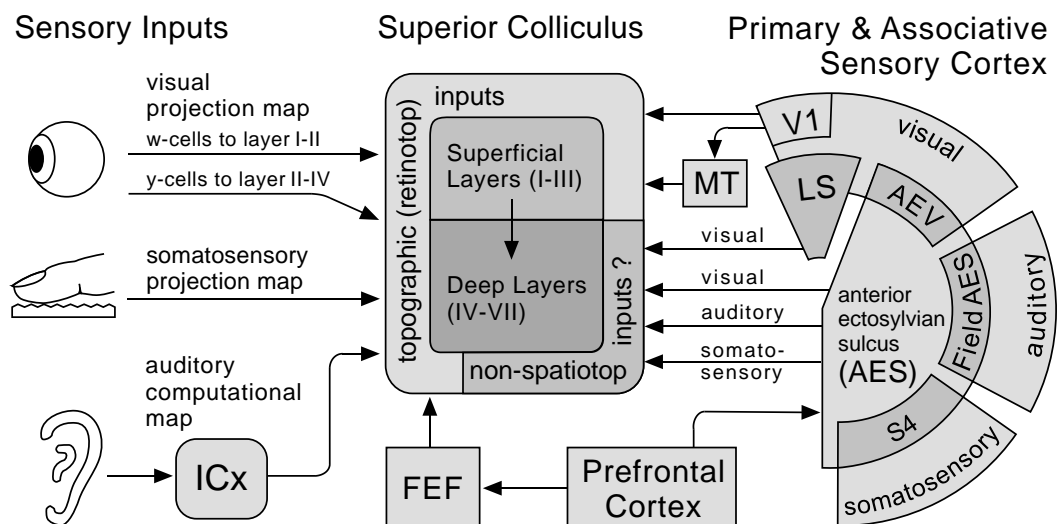


Abbildung 4.1: Übersicht der afferenten und efferenten Projektionen zum SC (ICx=externer Inferior Colliculus, V1=primärer visueller Kortex, MT=medial temporales Areal (V5), LS=lateraler suprasylvischer Sulcus, AEV/Field AES/S4=visuelles/auditorisches/ multisensorisches Areal des anterior ectosylvianischen Sulcus, FEF=frontales Augenfeld).

le, auditorische und somatosensorische Areale benachbarte Zellverbände okkupieren. Die als anterio-ectosylvischer Sulcus bezeichnete Region (AES) untergliedert sich folglich in einen visuellen Teil (AEV), einen auditorischen Bereich (Field AES) und eine somatosensorische Region (S4). Angesichts der Nähe zu verschiedenen sensorischen Kortizes überrascht es nicht, dass im AES multisensorische Aktivierungen verbreitet sind [MTA87]. Die nachgewiesenen Projektionen in den SCd erfolgen jedoch unisensorisch in separaten visuellen, auditorischen und somatosensorischen Kanälen [WMS93]. Eine auffällige Korrespondenz zu den Kodierungen im SCd stellt die Sensitivität des AES für sich bewegende Stimuli dar. Infolge des assoziativen Charakters der angrenzenden Repräsentationen werden Bewegungen aber möglicherweise nicht in einer dem SC vergleichbaren retinotopen Organisation abgebildet [MNBC82, SST<sup>+</sup>96, BECN04]. Während die räumliche Organisation in den Projektionen aus den visuellen und somatosensorischen Regionen (AEV und S4) nicht endgültig geklärt ist (vergl. [WMS93]), wurden retinotope Topographien zumindest im auditorischen Field AES ausgeschlossen [MC89].

Im Zusammenhang mit den kortikalen Verknüpfungen des SCd wird in mehreren Arbeiten auch noch das parietale visuelle Areal LS (lateraler suprasylvischer Sulcus) genannt, in dem ebenfalls multisensorische Effekte (mit noch unbekanntem Ursprung) beobachtet wurden [OMS84, JWJ<sup>+</sup>01, NSM97]. Der Einfluss, den die assoziativen sensorischen Kortexareale AES und LS auf die tiefen Schichten des SC ausüben, ist gravierend: die Deaktivierung der genannten Regionen unterbindet die multisensorischen Effekte im SCd nahezu vollständig, ohne dabei wesentliche Auswirkungen auf die unisensorischen Aktivierungen zu zeigen [WS94, WMS96, JWJ<sup>+</sup>01]. Da aber sämtliche sensorischen und kortikalen Eingänge des SC unisensorischen Charakter haben und außerdem die Latenzen im SC prinzipiell kürzer sind als im AES [SW96], scheinen die multisensorischen Befunde im SCd letztendlich doch auf internen Integrationsmechanismen zu basieren und nicht etwa auf projizierten, kortikalen Eigenschaften.

Welche genaue Funktion die multisensorischen, assoziativen Areale des Kortex im Zusammenwirken mit dem SC bei der Steuerung von Aufmerksamkeit und räumlicher Orientierung haben, konnten die zitierten Untersuchungen noch nicht klären. Das Areal AES steht, wie auch das frontale Augenfeld (FEF), in enger Verbindung zum präfrontalen Kortex [PDMNM05, CRS85], und man kann annehmen, dass beide Bereiche dazu dienen, eine eher abstrakte Verhaltensplanung in konkrete sensorische und motorische Aufgaben zu untersetzen. Während die vermutete Schleife FEF→SC→Thalamus→FEF an der visuellen Suche und der Generierung von Sakkadenzielen beteiligt ist, könnte über AES der passive oder aktive funktionelle Modus

des SC gesteuert werden. Mit diesem hypothetischen Mechanismus ließe sich erklären, wann der SC der Ausgestaltung von Sakkaden und der Rückkopplung räumlicher Informationen dient und wann ihm von höheren kognitiven Ebenen eine eigenständige sensomotorische Funktion zugestanden wird.

Bezüglich der Ausgangsprojektionen des SC sind im multisensorischen Kontext nur wenige Ergänzungen zu den in Kapitel 3 beschriebenen Befunden erforderlich. Der SC projiziert von seinen multisensorischen, tiefen Schichten aus Prämotorkommandos in den Hirnstamm und überträgt via Thalamus sensomotorische Korrektursignale für kortikale Topographien oder für die Inhibition-of-Return Mechanismen der visuellen Suche. Als Geräusch- oder Bewegungsdetektor und schnelle sensorische Afferenz kann er über einen ähnlichen thalamischen Pfad unabhängig vom kognitiven Zustand kortikale Instanzen alarmieren [WMS98]. Hinzu kommt die in den folgenden Abschnitten 4.2.2 und 4.2.3 diskutierte Verbindung zum auditorischen ICx, dessen Repräsentation von Schallrichtungen auf Augenbewegungen reagieren muss.

## 4.2.2 Sensorische und motorische Karten

Die sensorischen Projektionen in den SCd sind allesamt spatiotop organisiert. Exemplarisch wurde in den Kapiteln 2 und 3 anhand der retinalen Projektion im SCs und der berechneten Azimutwinkel im auditorischen IC gezeigt, wie die Abbildung räumlicher sensorischer Informationen in topographischen neuronalen Karten erfolgen kann. Als Grundvoraussetzung zur multisensorischen Integration in den tiefen Schichten des SC müssen die Topographien der einzelnen Modalitäten zunächst in eine kompatible Form, eine räumliche Common Metric, transformiert werden. Zwar sind alle sensorischen Afferenzen des SCd egozentrisch organisiert, doch führen Kopf- und Augenbewegungen dazu, dass visuelle, auditorische und somatosensorische Koordinaten nicht mit einer konstanten Transformationsvorschrift verrechnet werden können. Angesichts der motorischen Funktionen des SC bei der Steuerung von Augensakkaden kann angenommen werden, dass die auf die Blickrichtung bezogene Retinotopie im SCs auch die Topographie in den tiefen Schichten vorgibt. Als Ausdruck einer solchen Korrespondenz der SCd-Karten zur retinotopen Topographie belegt eine Reihe von Untersuchungen eine deutliche Überlappung der auditorischen, visuellen und somatosensorischen rezeptiven Felder von multisensorischen Neuronen [MS96, KVWS01]. STEIN und MEREDITH kennzeichnen die Fähigkeit, retinotope und intersensorisch korrespondierende rezeptive Felder auszubilden, als topographische Register der SCd-Neurone [SM93]. Abbildung 4.2 zeigt exemplarisch die rezeptiven Felder von 15 SC-Neuronen der Schichten III

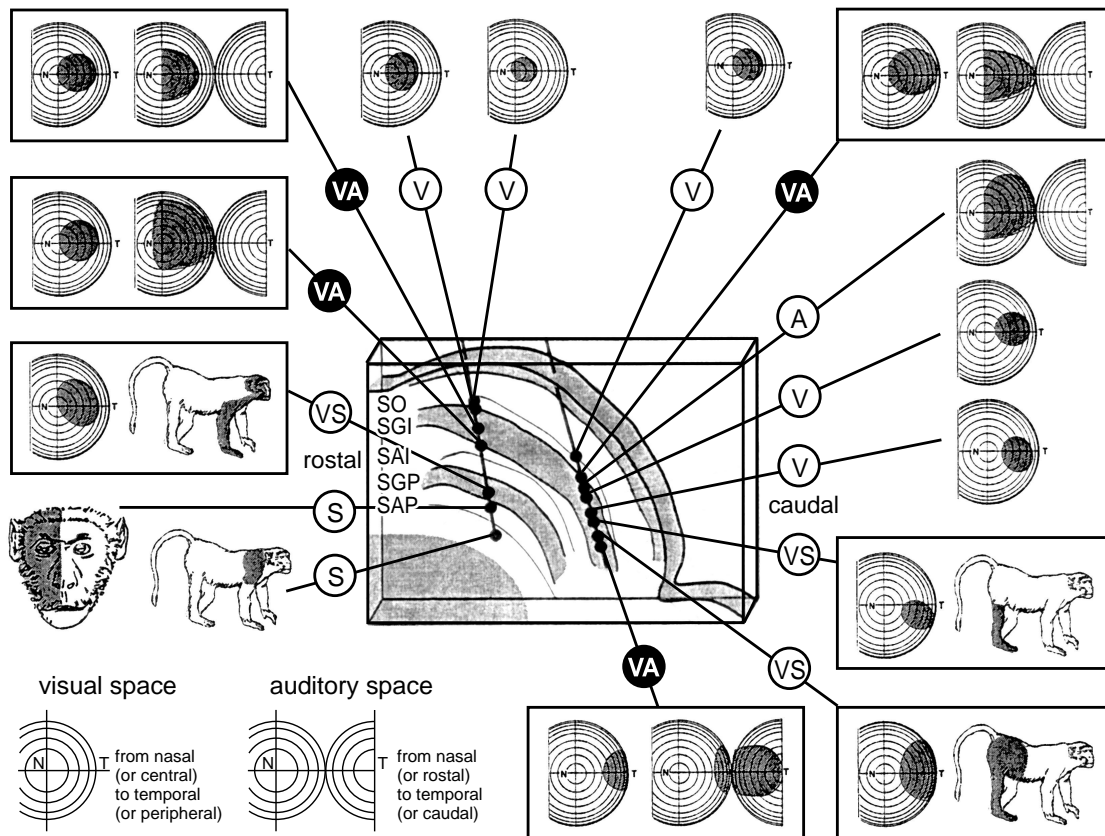


Abbildung 4.2: Veranschaulichung der topographischen Register im Superior Colliculus (SC) von Primaten. Im Verlauf zweier Messreihen konnten über die verschiedenen Schichten des SC hinweg korrespondierende rezeptive Felder und insbesondere deren Überlappung bei multisensorischen Neuronen (gerahmte Diagramme) nachgewiesen werden [WWS96]. (V,A,S,VA,VS=Messungen an visuellen, auditorischen, somatosensorischen, visuell-auditorischen oder visuell-somatosensorischen Neuronen. SO=Stratum opticum, SGI=Stratum griseum intermediale, SAI=Stratum album intermediale, SGP=Stratum griseum profundum, SAP=Stratum album profundum)

bis VII bei Primaten [WWS96]. Die Diagramme der vier visuell-auditorischen (VA) und drei visuell-somatosensorischen Zellen (VS) lassen erkennen, dass die räumlichen Register im SCd auf großen rezeptiven Feldern und einer nur groben topographischen Abbildung basieren. Nicht nur die Projektion der ohnehin unscharfen Bewegungskarte des SCs in den SCd ist mit einer weiteren Vergrößerung der rezeptiven Felder verbunden [RMR<sup>+</sup>89, MK04, TDPC05]. Auch die auditorische Topographie des SCd erscheint im Vergleich zur hohen räumlichen Auflösung im ICx wesentlich ungenauer ausgeprägt zu sein [SM93, MS96]. Diese Beobachtung ist von fundamentaler Bedeutung für die Idee der räumlichen Metrik und die Interpretation der Korreliertheit multisensorischer Ereignisse. Wie es im Zusammenhang mit der unscharfen Abbildung in den äußeren, visuellen Schichten des SC schon angedeutet wurde, haben auch die frühen multisen-



sensorischen Repräsentationen nichts mit der detaillierten retinotopen Organisation der striären Sehbahn gemein. In einigen Arbeiten geht das Verständnis dieser Unschärfe soweit, dass Ereignisse im Kontext der SCd-Repräsentation schon als räumlich korreliert betrachtet werden, wenn ihre multisensorischen Reizmuster in derselben Hemisphäre auftreten [BCMM01].

Die in Abbildung 4.2 dargestellten Messungen von WALLACE zeigen desweiteren, dass auch in den tiefen SC-Schichten neben multisensorischen Neuronen noch unimodale Zellen existieren, die ausschließlich visuell, auditorisch oder somatosensorisch aktiviert werden. Die Bedeutung dieser unisensorischen Zellen und ihre Verbindung zu multisensorischen Karten ist Gegenstand von Spekulationen. So wurde vorgeschlagen, dass ihre zusätzliche separate Kodierungsleistung im multimodalen, topographischen Verbund als unimodale Referenz die quantitativen Eigenschaften der multisensorischen Integration steuert [APBB00, PA03].

Die prämotorischen Leistungen des SC wurden in Abschnitt 3.3 im Kontext der visuellen Verarbeitung noch als These formuliert. Theoretisch schien es immerhin möglich, dass die motorische Karte zur Initiierung von Augenbewegungen ausschließlich im frontalen Augenfeld (FEF) realisiert wird, und der SC lediglich im Projektionsweg der motorischen Befehle liegt (vergl. [PB03]). Die multisensorischen Befunde zeichnen jedoch ein anderes, eindeutiges Bild. Nicht nur visuelle Ziele, wie sie unter anderem im FEF kodiert werden, sondern auch auditorisch repräsentierte Positionen können Augensakkaden auslösen [YP97, PY98, CVWMVO02]. Da aber der afferente auditorische Eingang des SC aus der sensorischen räumlichen Karte des ICx stammt und die kortikalen Efferenzen aus dem Field AES gar nicht räumlich organisiert sind, stellen die multisensorischen Schichten im SCd den einzigen Ort dar, an dem die beobachtete motorische Leistung realisiert werden kann. Auch die kurzen Latenzen der auditorisch initiierten Sakkaden schließen einen kortikalen Umweg der auditorischen Richtungsabbildung aus dem ICx aus [HRLNF94, FVOVdW95, FVO98]. Während an der Existenz einer eigenen Motorkarte des SCd [McI90, Pec96] nicht mehr gezweifelt wird, gehen die Überlegungen von STEIN und MEREDITH noch einen Schritt weiter. Sie betrachten die sensorische und motorische Kodierung angesichts der identischen topographischen Register als untrennbare Mechanismen und sprechen konsequenterweise von einer integrierten multisensorischen Motorkarte [SM93].

In scheinbarem Widerspruch zur prämotorischen Funktion des SCd steht die sehr unscharfe Abbildung in seiner sensorischen Topographie. Möglicherweise kann aber die Aktivierung in den tiefen SC-Schichten als verteilte Populationskodierung eine höhere räumliche Auflösung zur Bestimmung von motorischen Zielpunkten bereitstellen, als

es die großen rezeptiven Felder einzelner Neurone vermuten lassen [McI91, KVWS01].

Ungeachtet der Größe der auditorischen und somatosensorischen rezeptiven Felder muss deren topographische Ausrichtung nach einer deutlichen Augenbewegung im Rahmen der räumlichen Register korrigiert werden. Bei diesem Vorgang erlaubt die Annahme einer retinotopen Referenz im gesamten Superior Colliculus ein schlüssiges topographisches Konzept [McI90, PBW95, MS96]. Zu klären bleibt, welche konkreten neuronalen Mechanismen die Transformation der nicht-visuellen Karten realisieren. Im Falle des visuell-auditorischen Kartenabgleichs konnten visuelle topographische Korrektursignale im Mittelhirn der Vögel und Säugetiere nachgewiesen werden [LGW00, DBSK00]. An ihrem Projektionsziel, dem ICx, entfalten diese Signale die Wirkung einer „*topographischen Schablone, mit der die Projektion aus der auditorischen Karte gesteuert wird*“ [HE01, HE02]. Um diesen Mechanismus im Laufe der postnatalen Ontogenese zu etablieren und an das Wachstum anzupassen, sind insbesondere frühe visuelle Erfahrungen zwingend notwendig [WPHS04]. Generell lassen sich die Befunde zur Plastizität der auditorischen Topographie unter zwei Gesichtspunkten interpretieren: dem der langfristigen Kalibrierung [KSC<sup>+</sup>96, KST98, Kin99, Knu99] und dem der Reaktion auf einzelne Augenbewegungen [PBW95, PTY04]. Die kurzfristige Verschiebung der auditorischen rezeptiven Felder kann im Detail mit den derzeit bekannten Befunden und Modellen noch nicht vollständig nachvollzogen werden (vergl. [KH00]) – zu ihrer Erklärung muss die abstrakte Idee der topographischen Schablone vorerst genügen.

### 4.2.3 Multisensorische Integration auf Ebene des Neurons

Die neuroanatomischen Befunde zu Subarealen, Schichten und Projektionen des Superior Colliculus sowie die Beschreibung seiner sensorischen und prämotorischen Kodierung anhand topographischer Karten ergaben ein weitgehend schlüssiges Bild der multisensorischen Integration aus makroskopischer Sicht. Schon STEIN und MEREDITH gingen aber auch der Frage nach, mit welchen elementaren Mechanismen die Integrationsleistung im SC letztendlich bewerkstelligt wird und diskutierten die funktionelle Bedeutung des einzelnen Neurons [SM93]. Ohne Kenntnis der genauen synaptischen Struktur im SCd vermuten sie, dass bereits einzelne Zellen bi- und trimodale Verknüpfungen herstellen können und stützen ihre These mit elektrophysiologischen Ableitungen [MS86]. Als simple anatomische Voraussetzung müsste eine multipolare Nervenzelle an ihren Dendriten lediglich synaptische Terminals zu den Axonen aus verschiedenen sensorischen Repräsentationen unterhalten. Unter der

Annahme einer integrierten sensomotorischen Kodierung ist es weiterhin vorstellbar, dass dasselbe Signal am Axon einer solchen Zelle im Hirnstamm als Prämotorkommando und im Inferior Colliculus als Korrektursignal für die auditorische Topographie interpretiert wird. Abbildung 4.3 soll in einer abstrakten Weise verdeutlichen, wie auf dem neuronalen Niveau einzelner Zellen alle bisherigen sensorischen und motorischen Befunde zum SCd erklärt werden können.

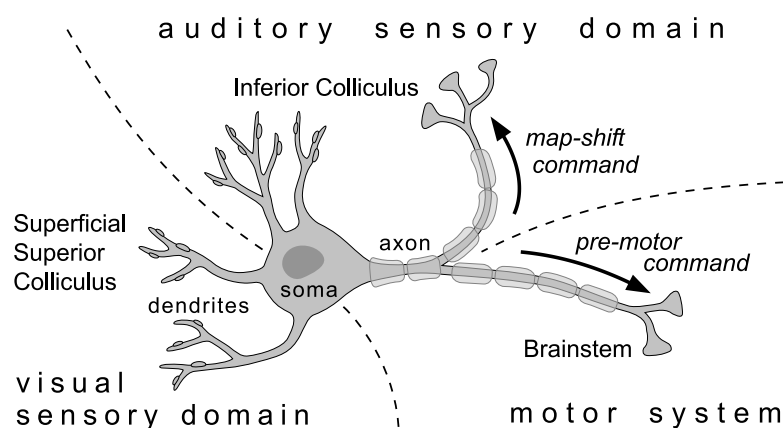


Abbildung 4.3: Auditorisch-visuelle Integration, prämotorische Kodierung und die topographische Korrektur der auditorischen rezeptiven Felder lassen sich schematisch für ein Neuron in den tiefen Schichten des Superior Colliculus darstellen.

Für das Verständnis der Rolle des Superior Colliculus bei der Steuerung von Aufmerksamkeit und räumlicher Orientierung war eine Beobachtung besonders interessant: die Aktivierung der bi- und trimodalen Zellen im SCd kann bei uni- und multisensorischer Stimulation stark voneinander abweichen und lässt sich nicht auf triviale quantitative Zusammenhänge zurückführen. Vielmehr scheinen die neuronalen Antworten in einer hochgradig nichtlinearen Weise, gleichermaßen von der Geometrie der unisensorischen, rezeptiven Felder sowie von der Intensität, der räumlichen Ausprägung und dem zeitlichen Verlauf der Stimuli beeinflusst zu werden. Um aus ihren Einzelzell-Messungen gewisse Gesetzmäßigkeiten der multisensorischen Integration abzuleiten, verfolgten STEIN und MEREDITH ein einfaches, experimentelles Regime: Nachdem eine multisensorische Zelle eindeutig lokalisiert werden konnte, wurde eine Reihe von Versuchen (Trials) mit immer denselben Stimuli wiederholt. Zu einem Trial werden jeweils die gemessenen Impulsfolgen bei separater sowie bei gleichzeitiger Präsentation der akustischen, optischen oder taktilen Stimuli zusammengefasst. Für eine Versuchsreihe mit mehreren Trials können schließlich Stimulus-Zeit-Histogramme (PSTH), mittlere Impulsraten und deren Varianzen bestimmt und mit anderen multi-

modalen Stimulus-Kombinationen verglichen werden (Abbildungen 4.4, 4.5). Als Maß zur Quantifizierung des Unterschieds zwischen uni- und multisensorischen Aktivierung führt STEIN den Begriff der Response Enhancement (RE) ein und definiert diesen über den mathematischen Zusammenhang:

$$RE = \frac{(CM - SM_{max}) \cdot 100\%}{SM_{max}} \quad (4.1)$$

Dabei sei  $CM$  die multisensorische Aktivierung und  $SM_{max}$  die Antwort auf den effektivsten unisensorischen Stimulus [SM93]. Ferner kann die Quantifizierung der neuronalen Antworten durch ein Integral über das jeweilige PSTH erfolgen. Als Resümee umfangreicher Messungen [MS86] können drei fundamentale Gesetzmäßigkeiten der multisensorischen Integration formuliert werden:

- **Response Enhancement** ist bei räumlich und zeitlich korrespondierenden Stimulus-Kombinationen zu erwarten. Auf das erforderliche Maß an räumlicher Korreliertheit kann aus den Eigenschaften der unimodalen rezeptiven Felder geschlossen werden.
- **Response Depression** wird durch eine deutliche, entweder räumliche oder zeitliche Separation der verschiedenen sensorischen Stimuli bewirkt. Die multisensorische Aktivierung ist in diesen Fällen geringer als die Antwort auf den effektiveren unimodalen Stimulus.
- **Unimodale Aktivierung und Response Enhancement verhalten sich umgekehrt proportional.** Auditorische, visuelle oder somatosensorische Reizmuster, die sich bei ihrer separaten Präsentation als ineffektiv herausstellen, verursachen durch ihre Kombination die höchsten Enhancement-Werte.

Die Abbildungen 4.4 und 4.5 veranschaulichen die Effekte Response Enhancement und Depression sowie die umgekehrte Proportionalität von unimodaler Effektivität und multimodaler Verstärkung an exemplarischen Ergebnissen der Untersuchungen von MEREDITH und STEIN. In den Post-Stimulus-Zeit Histogrammen multimodaler Neurone sind bei einigen Stimuluskombinationen auffällig lange Spikebursts zu erkennen. Eine genauere Untersuchung der zeitlichen Dynamik der Enhancement- und Depression-Eigenschaften bescheinigte den tiefen SC-Schichten multisensorisch sensitive Zeitfenster von mehreren hundert Millisekunden [MNS87]. Wie schon bei der Diskussion der nichtlinearen Filterfunktionen im auditorischen ICx (vergl. Abschnitt 2.2.6) liegt es nahe, die Ursachen einer solchen Dynamik nicht nur im einzelnen Neuron, sondern in komplexen synaptischen Strukturen und rekurrenten Verschaltungen

zu suchen. Als Kandidaten für ein zeitlich anhaltendes exzitatorisches oder inhibitorisches Feedback dürfen hier natürlich die kortikalen Efferenzen aus den assoziativen sensorischen Arealen nicht unberücksichtigt bleiben [KVV<sup>+</sup>97]. Allerdings müsste eine Rückkopplung zur Erzeugung räumlich-zeitlicher Korrelationsfenster topographisch organisiert sein, was beispielsweise im Falle der Projektion aus dem auditorischen Field AES bezweifelt wird. Aber auch im Superior Colliculus selbst wurden, ganz ähnlich wie im benachbarten IC, komplexe interne Wechselwirkungen nachgewiesen [Nor80, BK96]. Das Repertoire der neuronalen Strukturelemente reicht dabei von lokalen, exzitatorischen Verknüpfungen bis zu inhibitorischen Interneuronen [VOJ89, MI98, MR98, SI04].

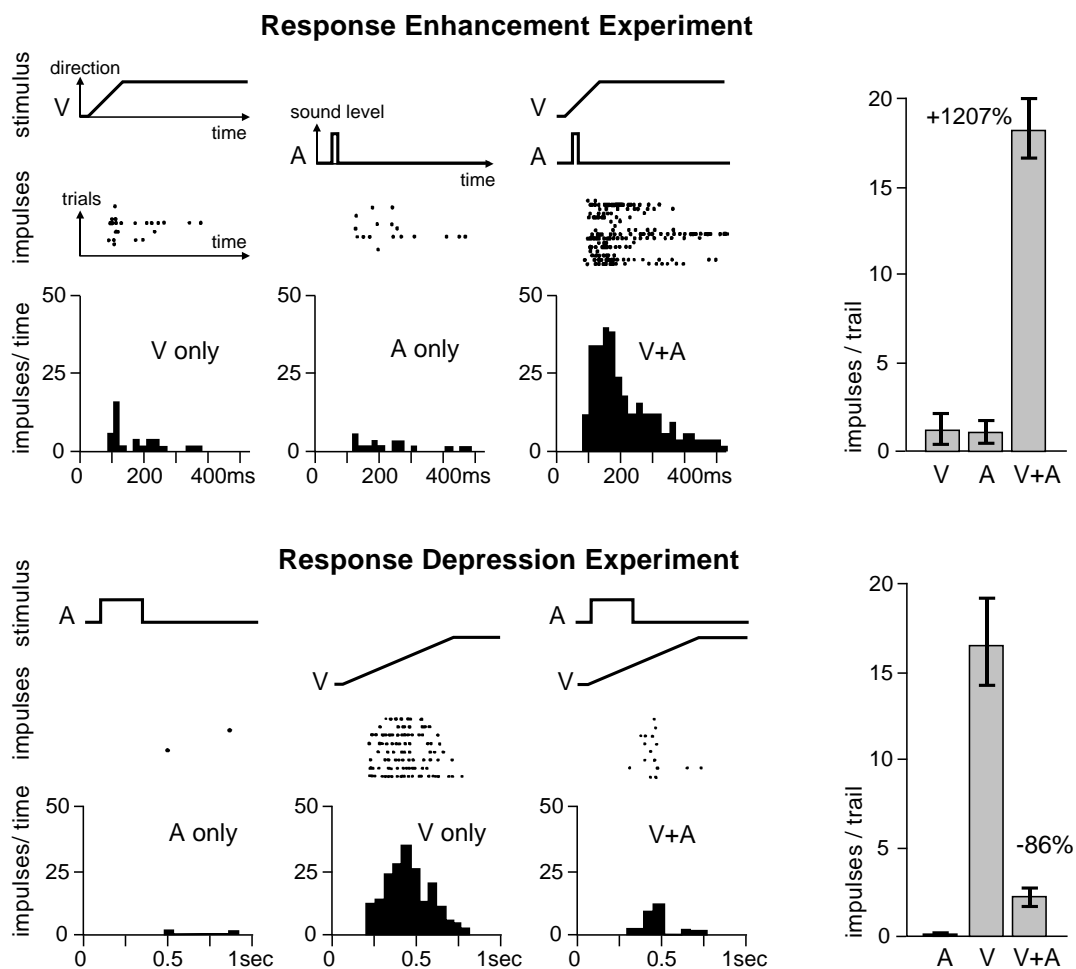


Abbildung 4.4: Demonstration von Response Enhancement und Depression in multisensorischen Zellen des SCd [MS86]. Die skizzierten Stimulusverläufe beziehen sich auf die Amplitude des akustischen Signals (A) und auf die Bewegung des visuellen Stimulus (V). Im Response Enhancement Experiment war die Stimulusdauer relativ kurz, um die räumlich-zeitliche Korrespondenz zwischen den akustischen und optischen Ereignissen zu garantieren. Eine ausgedehntere Bewegung (in den unteren Diagrammen) führt dagegen zu einer auditorisch-visuellen Disparität und zur Response Depression.

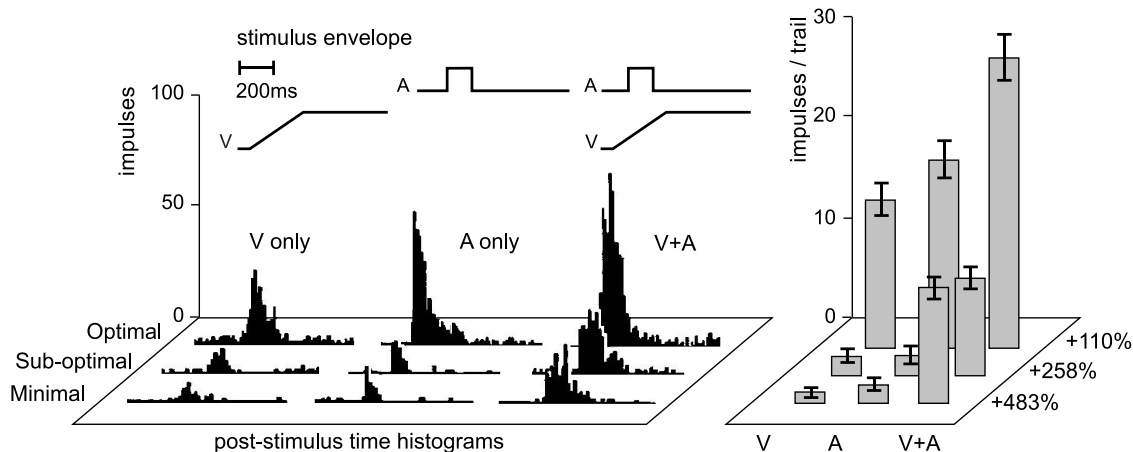


Abbildung 4.5: Demonstration der umgekehrten Proportionalität von unisensorischer Aktivierung und Response Enhancement bei multisensorischen Stimuli [MS86]. Um die Intensität der Stimuli zu variieren, wurden physikalische Parameter wie die Größe des optischen Reizes oder der Pegel des Geräusches verändert.

Neben den räumlichen und zeitlichen Randbedingungen der multisensorischen Integration und der Aufrechterhaltung der topographischen Register nach erfolgten Augenbewegungen muss im SC schließlich auch noch ein ebenso offensichtliches wie schwerwiegendes visuelles Kodierungsproblem gelöst werden. Durch die enge Kopplung von sensorischen und prämotorischen Funktionen müsste man annehmen, dass es bei der Ausführung einer Augensakkade zu einer massiven Ausprägung von visuellen Bewegungsmerkmalen kommt, die alle beschriebenen Enhancement-Effekte bei unter-schwelligem Stimuli ad absurdum führt. Schon länger ist jedoch bekannt, dass der SC zwischen realen Bewegungen und visuellen Aktivierungen als Folge selbst generierter Sakkaden unterscheiden kann [RW76, MS80]. Die triviale Lösung des Problems der Bewegungskodierung ermöglichen Inhibitionsmechanismen, die unter anderem in den mittleren SC-Schichten visuelle Antworten für die Dauer der Sakkaden unterbinden [RMKP91, WFG95]. Die schematische Darstellung eines SCd-Neurons in Abbildung 4.3 könnte somit um einen dritten axonalen Zweig ergänzt werden, der andeutet, dass der Ausgang der Zelle nicht nur als Motorkommando und topographisches Korrektursignal interpretiert werden kann, sondern auch als inhibitorische Rekurrenz in die SC-interne, visuelle Domäne.

## 4.3 Modellierung

### 4.3.1 Allgemeines Modellkonzept

Vergleichbar mit der Erörterung der visuellen und auditorischen Verarbeitungspfade in den Kapiteln 2 und 3 kamen auch im Kontext der multisensorischen Hirnfunktionen Aspekte zur Sprache, die dem Verständnis und der Einordnung dienen, die jedoch nicht unmittelbarer Gegenstand eines implementierbaren Simulationsmodells sein können. Dies betrifft insbesondere die kortikalen Mechanismen und Efferenzen, die offenbar den funktionellen Modus des Superior Colliculus bestimmen und somit eher bei einer späteren Anwendung des Modells von Bedeutung sind. Im Sinne der definierten Teilaufgabe einer frühen und nicht-assoziativen Steuerung der Aufmerksamkeit steht hier die eigentliche, multisensorische Integrationsleistung der tiefen Schichten des Superior Colliculus im Vordergrund. Auf Basis der entworfenen, visuellen und auditorischen Teilmodelle soll der Aufbau einer multisensorischen Karte realisiert sowie die prämotorische Kodierung und die topographische Korrektur der Schallortung nach einer Augenbewegung vorbereitet werden. Die Anforderungen an ein adäquates, auditorisch-visuelles Modellkonzept lassen sich zunächst unabhängig von konkreten Implementierungsoptionen aus den zitierten Befunden ableiten und in Form abstrakter, funktioneller Merkmale formulieren (vergl. [SG04] und Abbildung 4.6):

- *Die tiefen Schichten des Superior Colliculus (SCd) antworten in retinotoper Organisation auf visuelle und auditorische Stimuli.* Im multisensorischen Modell konvergieren die visuelle Bewegungskodierung der äußeren SC-Schichten und das Ergebnis der Schallortung aus dem externen Inferior Colliculus. Infolge der technischen Limitierung des auditorischen Modells auf die Detektion horizontaler Richtungen soll auch im multisensorischen System eine eindimensionale Topographie realisiert werden. Die Geometrie der Abbildung sowie die Größe der rezeptiven Felder kann bei der Projektion von visueller und auditorischer Karte beliebig gestaltet werden.
- *Die meisten multisensorischen Neurone des SCd realisieren Response Enhancement, wenn sie räumlich und zeitlich korrespondierende Stimuli erhalten sowie Response Depression, wenn visuelle und auditorische Repräsentation unkorreliert sind.* Das Modell sollte ein nichtlineares Übertragungsverhalten besitzen sowie laterale und zeitliche Abhängigkeiten herstellen. Letzteres kann beispielsweise mit Hilfe von Rekurrenzen und geeigneten iterativen Beschreibungen erfolgen.

- *Response Enhancement und unimodale Effektivität sind umgekehrt proportional.* Im realen neurobiologischen System begrenzen die physiologischen Ressourcen eine multiplikative Verstärkung der multisensorischen Reizmuster. Im Modell kann eine geeignete Ausgabefunktion oder Normierung verhindern, dass die Aktivierung der multimodalen Karte über alle Maßen steigt.
- *Überlappende multisensorische und motorische Karten initiieren ein Orientierungsverhalten.* Die Repräsentation in der multisensorischen Karte wird direkt als prämotorische Kodierung interpretiert. Die Zielauswahl für ein konkretes Motorkommando ist anhand der maximalen Aktivierung in der topographischen Karte möglich und soll eine Zuwendung zum Stimulus bewirken. Demzufolge kodiert die eindimensionale Topographie relative Kopfdrehungen bzw. die Auslenkung der Augen nach rechts oder links. Kleinere Bewegungen korrespondieren folglich mit medialen, größere dagegen mit lateralen Bereichen der Motorkarte.
- *Die spezifischen, auditorischen und visuellen rezeptiven Felder bilden topographische Register.* Als Folge einer Blickrichtungsänderung, relativ zur Position der Ohren, wird die auditorische Repräsentation neu zur retinotopen Abbildung ausgerichtet. Neurobiologisch inspirierte Mechanismen zur Kalibrierung der visuellen und auditorischen Topographie oder deren Korrektur nach Augensakkaden sind vor dem Hintergrund der späteren, technischen Anwendung schwer zu motivieren. An dieser Stelle wird stattdessen auf feste Transformationsvorschriften entsprechend einer bekannten sensorischen Geometrie zurückgegriffen.

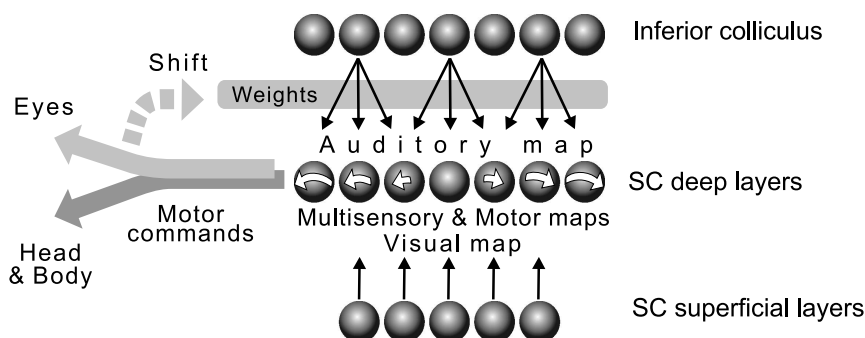


Abbildung 4.6: Schematische Darstellung des allgemeinen auditorisch-visuellen Konzepts zur Erzeugung eines einfachen Orientierungsverhaltens. Ein typischerweise begrenztes Gesichtsfeld hat zur Folge, dass die visuellen Afferenzen nur einen Teilbereich der multimodalen Karte abdecken. Wenn im Modell separate Augenbewegungen zugelassen werden, muss die resultierende Verschiebung zwischen auditorischer und visueller Abbildung bei der Projektion der detektierten Schallrichtungen kompensiert werden.



## 4.3.2 Dynamisches Neuronales Feld nach Amari

### Motivation

Der multisensorische Superior Colliculus (SC) und der in Kapitel 2 untersuchte auditorische Inferior Colliculus (IC) sind benachbarte Regionen im Mittelhirn. Die Aufgaben bei der Modellierung des IC waren die Rekombination der auf Frequenzbänder verteilten Richtungsabbildungen und die nichtlineare Filterung zur Beseitigung von Phasenartefakten aus der Detektion der Stereolaufzeit. Auch das SC-Modell soll mehrere topographische Karten kombinieren – nur eben solche, die von verschiedenen sensorischen Modalitäten erzeugt wurden. Und ebenso wie im IC sind auch im SC eine nichtlineare Aktivierung und laterale Abhängigkeiten zu beobachten – allerdings nicht, um Artefakte zu unterdrücken, sondern um zwischen korrespondierenden oder unkorrelierten Positionen von optischen und akustischen Ereignissen zu unterscheiden. Sowohl der Einfluss des IC auf Präzedenzeffekte als auch die zugegebenermaßen viel längeren multisensorischen Zeitfenster im SC belegen das Vorhandensein neuronaler Architekturen, die sensitiv für zeitliche Zusammenhänge sind und über einfache feed-forward Verschaltungen hinausgehen. Bei allen Unterschieden in den zu verarbeitenden Informationen zeigen sich demnach überraschende Parallelen in der funktionellen Beschreibung der internen, neuronalen Strukturen beider Areale. Es liegt deshalb nahe, die bei der IC-Simulation eingesetzten neuronalen Felder nach AMARI auch für die Implementierung eines SC-Modells in Erwägung zu ziehen und zu untersuchen, ob das von ihnen realisierte Winner-Take-All Verhalten (WTA) die Eigenschaften der multisensorischen Integration erklären kann.

Die neurobiologischen Indizien für eine WTA-basierte Verarbeitung sind in den tiefen Schichten des SC mindestens so überzeugend wie im auditorischen Teil des Mittelhirns. Abgesehen vom Konsens über die Existenz lateraler Verschaltungen, inhibitorischer Interneurone und SC-internen Rekurrenzen [BK96, MI98, MR98] lesen sich insbesondere die Befunde über lokal exzitatorische und global inhibitorische Verbindungen [VOJ89] wie die explizite Beschreibung eines WTA-Netzes. Die außergewöhnlich eindeutigen Befunde und die Anforderung, räumlich-zeitliche Zusammenhänge in Reizmustern in eine nichtlineare Ausgabe zu überführen, bilden eine belastbare Motivation für die Implementierung des SC-Modells als dynamisches neuronales Feld nach AMARI und KOHONEN (vergl. Abschnitt 2.2.6).

Parallel zu den hier vorgestellten Arbeiten entwarf RUCCI ein spezifisches Modell des Tectum opticum der Eule [REW99]. Unter Bezugnahme auf die Untersuchungen von CARR und KONISHI [CK90] simulierte er Komponenten des auditorischen Timing

Pfades zur Detektion horizontaler Richtungen und kombinierte diese mit visuellen Daten, um letztlich Motorkommandos auszulösen. Auch RUCCIs Implementierungen basieren auf künstlichen neuronalen Netzen, die als WTA-Architekturen interpretiert werden können. Obwohl er eine Robotik-Plattform als Demonstrator und technische Experimentieranordnung benutzt, gilt sein vorrangiges Interesse weniger den multisensorischen Response-Eigenschaften, sondern der Adaption und Kalibrierung der auditorisch-visuellen Topographie.

Auch die hier angestrebte Implementierung sollte sich als sensomotorische Komponente in artifizielle Systeme integrieren lassen, dabei aber nicht die Adaption an technische Parameter übernehmen. Im Gegensatz zu RUCCIs Arbeiten steht vielmehr die Frage im Mittelpunkt, ob und wie mit einem dynamischen neuronalen Feld die qualitativen und quantitativen multisensorischen Merkmale im Mittelhirn nachvollziehbar sind. Zwar rechtfertigen die motorischen Aspekte des abstrakten SC-Modells die Realisierung des Orientierungsverhaltens eines Hardware demonstrators (s. Abschnitt 5.3), zunächst soll aber die multisensorische Integration der simulierten SCs- und ICx-Repräsentationen in einer gemeinsamen Aufmerksamkeitskarte realisiert werden.

### Implementierung

Bereits mit einer früheren, spikebasierten WTA-Implementierung konnte demonstriert werden, wie eine um visuelle Afferenzen erweiterte Variante des im ICx-Modell eingesetzten Netzes multisensorische Integrationsleistungen erbringen kann [SP99]. Nachdem aber schon in Abschnitt 2.2.6 begründet wurde, dass nach der Berechnung der auditorischen Azimutwinkelkarte die hohe zeitliche Auflösung der spike-orientierten Simulation überflüssig ist, findet im Folgenden eine Ratenkodierung Verwendung. Als einzige Änderung der auditorischen WTA-Notation nach AMARI (Gleichung 2.5), muss ein zusätzlicher visueller Input eingefügt werden. Nimmt man an, dass auditorische und visuelle Afferenzen die SCd-Zellen weitgehend unabhängig voneinander über getrennte Dendritenzweige erreichen, sollte auch die Verknüpfung der Eingänge des WTA-Netzes voneinander unabhängig, beispielsweise in additiver Form erfolgen. Die bimodale Notation eines dynamischen neuronalen Feldes lautet dann:

$$\begin{aligned} \tau \frac{d}{dt} z(r, t) = & -z(r, t) + x_A(r, t) + x_V(r, t) \\ & -c_i \int y(z(r, t)) dr \\ & +c_n \int w(r - r') y(z(r', t)) dr' \end{aligned} \quad (4.2)$$

Wie schon in der unimodalen Form sind bei der Beschreibung des Zustandes  $z(r, t)$  die drei Grundelemente: sensorischer Eingang (hier  $x_A$  und  $x_V$ ), globale Inhibition und laterales exzitatorisches Feedback von den benachbarten Positionen  $r'$  zu erkennen. Alle Neurone besitzen weiterhin den bekannten sigmoiden Ausgang, der durch die Fermi-Funktion:  $y(z(r, t)) = (1 + \exp(-\sigma \cdot z(r, t)))^{-1}$  beschrieben wird. Die großen visuellen und auditorischen rezeptiven Felder der multisensorischen SCd-Neurone können mit einem arbiträr gestalteten, räumlichen Tiefpass am Netzeingang emuliert werden. Ebenso liegt es nahe, die Breite der lateralen exzitatorischen Rekurrenz weiter zu fassen als im auditorischen Modell. Korrespondierend zur Geometrie der experimentellen Szenarien erwiesen sich entsprechende Wichtungsvektoren mit Radien von 5–10 Winkelstufen (Neuronen) der diskreten Modell-Topographie als unkritisch.

Unter Verwendung der horizontalen Richtungsabbildungen der unisensorischen Teilsysteme aus den Kapiteln 2 und 3 zeigt das mit exemplarischen audio-visuellen

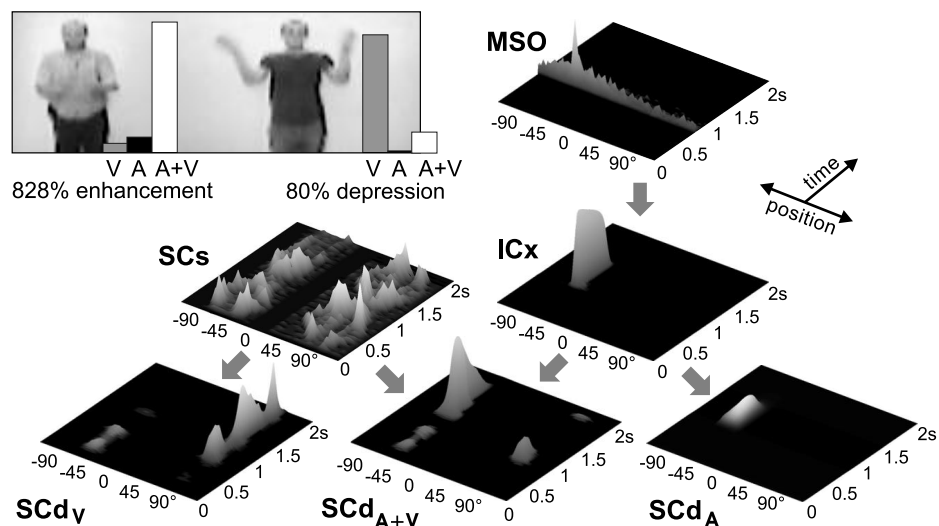


Abbildung 4.7: Visualisierung der Simulationsergebnisse in den uni- und multisensorischen Teilmodellen. In einer exemplarischen, audio-visuellen Sequenz winkt eine Person an Position  $+60^\circ$ , eine zweite klatscht einmalig in die Hände ( $-50^\circ$ ). Das Modellverhalten wurde separat für die visuelle, auditorische und multisensorische Präsentation der Szene simuliert, um an zwei ausgewählten Positionen in der von STEIN vorgeschlagenen Weise Response Enhancement und Depression zu berechnen. Die Diagramme zeigen im Einzelnen:

- SCs: retinotopie Bewegungskarte
- MSO: Ortscode der binauralen Schall-Laufzeit nach der Simulation der Koinzidenzdetektion im MSO-Kern (summiert über alle Frequenzbänder)
- ICx: WTA-Filterung der auditorischen Karte im Modell des externen Inferior Colliculus
- SCd: Ausgabe des multisensorischen WTA-Netzes zur Simulation der tiefen Schichten des Superior Colliculus im visuellen (V), auditorischen (A) und multisensorischen Experiment (A+V).

Daten beaufschlagte Simulationsmodell ein plausibles Verhalten. Wird die in Abbildung 4.7 dargestellte Szene nur visuell wahrgenommen, dominiert die rechte Person durch ihre ausladenderen Bewegungen die Aktivierung der Aufmerksamkeitskarte. Ist die Szene nur zu hören, bleibt sie dagegen verborgen und nur die Position der klatschenden Hände weiter links wird abgebildet. Bei einer multisensorischen Präsentation der Stimuli führt die Kombination der räumlich korrelierten, auditorischen und visuellen Komponenten zur Response Enhancement, weshalb sich die Position des multimodalen Ereignisses gegen die stärkste unimodale Aktivierung durchsetzt. Da die erhöhte Aktivierung im multisensorischen Experiment erst mehrere hundert Millisekunden nach dem Ende des auditorischen Stimulus abklingt, scheint die rekurrente WTA-Architektur tatsächlich auch lange multisensorische Zeitfenster zu realisieren.

Response Enhancement und Depression lassen sich intuitiv als Manifestationen eines typischen Winner-Take-All Verhaltens im simulierten dynamischen neuronalen Feld nachvollziehen. Ob auch die vom Modell erwartete umgekehrte Proportionalität von unimodaler Aktivierung und Response Enhancement erzielt wird, kann an einem einzelnen Experiment allerdings nicht diskutiert werden. Um einen Eindruck davon zu gewinnen, ob es einen systematischen Zusammenhang zwischen der Intensität der unimodalen Stimuli und dem Maß an Response Enhancement gibt, wurde in mehreren Szenen mit jeweils einem korrelierten audio-visuellen Ereignis die Intensität am Eingang des WTA-Netzes variiert. Über den Dynamikbereich des sigmoiden WTA-Ausgangs ergab sich auf diese Weise eine Enhancement-Kurve, die sich in Abhängigkeit von den unimodalen Simulationsergebnissen  $y_A$  und  $y_V$  darstellen lässt. Abbildung 4.8 zeigt diese Kurven für eine WTA-Zelle und für vier unterschiedliche Stimuli, die jeweils im Zentrum der rezeptiven Felder dieser Zelle auftraten. Alle Kurven besitzen denselben imaginären Anfangspunkt bei  $y_A + y_V = 0$  für Eingangswichtungen gleich Null. Auch der Endpunkt nahe  $y_A + y_V = 2$ , der erreicht wird, wenn bei extrem hohen Eingangswichtungen der sigmoide Netzausgang gegen 1 konvergiert, ist für alle Szenen ähnlich. In beiden Fällen ist kaum ein Enhancement-Effekt möglich. Über dieses theoretisch mögliche Intervall hinweg zeigt der Verlauf aller Kurven nicht nur unterschiedliche maximale Enhancement-Werte, die das Maß an räumlich-zeitlicher Korreliertheit der Stimuli widerspiegeln, sondern auch auffällige Gemeinsamkeiten. Offensichtlich erzeugen die WTA-Dynamik des Amarifeldes und die sigmoide Ausgabefunktion zwei typische Arbeitsbereiche: Steigt die Intensität der Aktivierung, erreicht der Arbeitspunkt der WTA-Zelle nach einem unterschwelligem Intervall mit extrem kleinen Ausgabewerten schließlich den steilen Teilbereich der sigmoiden Ausgabefunktion. Hier erzeugt die Addition der auditorischen und visuellen Eingänge die größte Verstärkung am Ausgang

des Netzes und folglich auch die höchsten Enhancement–Werte. Eine weitere Erhöhung der Eingangsintensität führt jedoch bald zur Sättigung der Aktivierung, da die Ausgangs–Sigmoide erst bei multisensorischer und schließlich auch schon bei unimodaler Stimulation die Netzausgabe begrenzt. Durch die Kombination von auditorischem und visuellem Eingang kann der Ausgang der WTA–Zellen immer weniger verstärkt werden – die Kurve der Response Enhancement sinkt. Das schraffierte Intervall in Abbildung 4.8, in dem sich unimodale Aktivierung und multimodale Response Enhancement prinzipiell umgekehrt proportional verhalten, umfasst einen Dynamikbereich von 1–100% der theoretisch möglichen Aktivierung einer WTA–Zelle.

Aufgrund dieser exemplarischen Ergebnisse kann bereits vermutet werden, dass die vorgestellte WTA–Variante die drei grundlegenden multisensorischen Aufgaben – Response Enhancement, Depression und umgekehrte Proportionalität von unimodaler Aktivierung und multimodaler Verstärkung – sehr gut zu lösen vermag. Desweiteren lassen sich Richtlinien für die Parametrisierung und die Anwendung des Simulationsmodells ableiten. Die Wichtung der auditorischen und visuellen Eingänge sollte bei gegebenen anderen Parametern so gewählt werden, dass typische Stimulusintensitäten zu Ausgangsaktivierungen führen, die in eben dem markierten Bereich liegen. Außerdem ist die Schwelle, ab der eine Aktivierung als eindeutiges Motorkommando interpretiert wird, sinnvoll in der Mitte dieses Intervalls zu positionieren. Dann nämlich kann davon ausgegangen werden, dass die Gesetzmäßigkeit der umgekehrten Proportionalität von unimodaler Aktivierung und multisensorischer Response Enhancement bei der Initiierung eines Orientierungsverhaltens auch zum Tragen kommt.

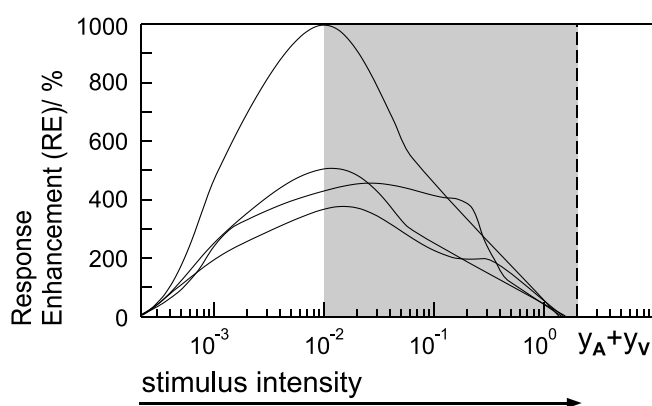


Abbildung 4.8: Vier verschiedene Versuche mit korrespondierenden audio–visuellen Ereignissen wurden mit fein abgestufter Stimulusintensität wiederholt und der resultierende Zusammenhang von unimodaler Aktivierung (hier als Summe  $y_A + y_V$ ) und Response Enhancement jeweils als Kurve für ein WTA–Neuron dargestellt. Im schraffierten Intervall zwischen etwa 1–100% des gesamten Dynamikbereichs am Netzausgang wird die gewünschte, umgekehrte Proportionalität von unimodaler Aktivierung und Response Enhancement realisiert.

### 4.3.3 Probabilistische Modelle

#### Motivation

Einem regelrechten Trend der probabilistischen Problembeschreibung in der Bild- und Audio-Verarbeitung folgend, wurde in den vergangenen Jahren versucht, auch neurophysiologische Vorgänge wie die multisensorische Integration im Superior Colliculus mit den mathematischen Werkzeugen der Statistik und Wahrscheinlichkeitsrechnung zu modellieren. Die methodische Kluft zwischen der direkten Beschreibung des Zustands eines Neurons durch dessen Aktivierung und der wahrscheinlichkeitsbasierten Notation stochastischer Prozesse teilte Biologen und Mathematiker lange in zwei wissenschaftliche Lager. Einer mittlerweile häufig zitierten Arbeit von ANASTASIO kommt deshalb die außerordentliche Bedeutung eines Brückenschlages zu. Er entwickelte, ausgerechnet in Zusammenarbeit mit dem Biologen BARRY E. STEIN ein probabilistisches Konzept, das die abstrakte mathematische Modellierung mehrerer multisensorischer Befunde erlaubt. Anhand seiner hypothetischen Überlegungen [APBB00] lassen sich im folgenden Abschnitt wesentliche probabilistische Begriffe, Definitionen und Zusammenhänge gut veranschaulichen.

Bemerkenswert ist zunächst, dass ANASTASIO am Ausgangspunkt seiner Betrachtungen bewusst den Begriff des Zieles (Target) benutzt, um Signalquellen im Allgemeinen zu bezeichnen und sie von Objekten mit spezifischen Eigenschaften bei der symbolischen Verarbeitung zu distanzieren. Der sensorische Input, den ein Target verursacht, ist das Resultat stochastischer Prozesse, weshalb die durch ihn kodierte Information unsicher ist. Man kann daher vermuten, „*dass die Neurone in den tiefen Schichten des SC die bedingte Wahrscheinlichkeit für die Existenz eines Zieles bei einer gegebenen (unsicheren) sensorischen Repräsentation berechnen*“ [APBB00]. Nach der zentralen These bei der probabilistischen Modellierung genügt diese „Berechnung“ quantitativ dem Bayesschen Gesetz: werden beispielsweise das Auftreten von Zielen sowie der visuelle sensorische Input mit den stochastischen Variablen  $T$  und  $V$  bezeichnet, dann gelte für die genannte bedingte Wahrscheinlichkeit  $P(T|V)$ :

$$P(T|V) = \frac{P(V|T)P(T)}{P(V)}; \quad (\text{Bayessche Formel}) \quad (4.3)$$

ANASTASIO diskutiert im Bayesschen Kontext den posterioren Charakter der Wahrscheinlichkeit  $P(T|V)$ , die durch eine Verknüpfung der Prioren  $P(T)$  mit der sensorischen Beobachtung  $V$  aktualisiert werde. Durch diesen Vorgang erhöhe sich die Sicherheit der Repräsentation der Umwelt – was schlechthin ein Ziel der Wahrnehmung sei. Um ein Wahrscheinlichkeits-Update nach Bayesschem Vorbild zu realisieren, müs-

sen dem Superior Colliculus die Größen  $P(V|T)$ ,  $P(T)$  und  $P(V)$  bekannt sein. Auch hierfür findet ANASTASIO in [APBB00] eine abstrakte und schwer zu widerlegende Erklärung: Die a priori Wahrscheinlichkeit für das unabhängige Auftreten eines Zieles  $P(T)$  sei eine Umwelteigenschaft, die als Erfahrungswert vom Gehirn adaptiert werde. Auch die Wahrscheinlichkeiten dafür, dass der aktuelle sensorische Input  $V$  ein Ziel kodiert ( $P(V|T)$ ) und dass unabhängig von der Situation überhaupt ein visueller Input empfangen wird ( $P(V)$ ), sollten dem Superior Colliculus als „interne Merkmale des visuellen Systems“ inherent sein.

Zur Veranschaulichung seiner Idee dient ANASTASIO ein gedankliches Experiment: Sei die Variable  $T$  binär und bedeute  $T=1$  die Anwesenheit,  $T=0$  die Abwesenheit eines Zieles. Weiterhin ergebe sich die a priori Wahrscheinlichkeit für das Vorhandensein eines Zieles  $P(T=1)$  zu 0.1 und für dessen Abwesenheit  $P(T=0)$  zu 0.9. Wie kann nun eine Notation für  $P(V|T)$  gefunden werden? In Analogie zu elektrophysiologischen Ableitungen wird der neuronal kodierte, sensorische Input  $V$  als Anzahl gemessener Spike-Impulse in einem bestimmten Zeitintervall verstanden. Geht man von der nicht seltenen Annahme aus, dass die Messung von Spikeraten poissonverteilt ist ( $P(V=v) = \lambda^v e^{-\lambda}/v!$ ), lassen sich Spontanaktivierung  $P(V|T=0)$  und getriebene Response  $P(V|T=1)$  als exemplarische Poissonverteilungen mit unterschiedlichen Parametern  $\lambda$  darstellen (Abbildung 4.9a). Nach dem Satz von der totalen Wahrscheinlichkeit kann nun die noch fehlende Größe  $P(V)$  berechnet werden:

$$P(V) = P(V|T = 1) \cdot P(T = 1) + P(V|T = 0) \cdot P(T = 0) \quad (4.4)$$

Damit ist die Bayessche Formel (4.3) vollständig bestimmt und kann in einem nächsten Schritt formal auf den bimodalen Fall angewandt werden. Entsprechend dem visuellen Input  $V$  werden die auditorische stochastische Variable  $A$  und die mit ihr verknüpften Wahrscheinlichkeiten  $P(A)$  und  $P(A|T)$  eingeführt. Dann ergibt sich die bimodale Auslegung des Bayesschen Gesetzes zu:

$$P(T = 1|A, V) = \frac{P(A, V|T = 1)P(T = 1)}{P(A, V)} \quad (4.5)$$

Unter der Annahme unabhängiger Inputs  $A$  und  $V$  kann unter erneuter Anwendung des Satzes von der totalen Wahrscheinlichkeit eine weitere Konkretisierung erfolgen:

$$P(A, V|T = 1) = P(V|T = 1) \cdot P(A|T = 1) \quad (4.6)$$

$$P(A, V) = P(V|T = 1) \cdot P(A|T = 1) \cdot P(T = 1) \\ + P(V|T = 0) \cdot P(A|T = 0) \cdot P(T = 0) \quad (4.7)$$

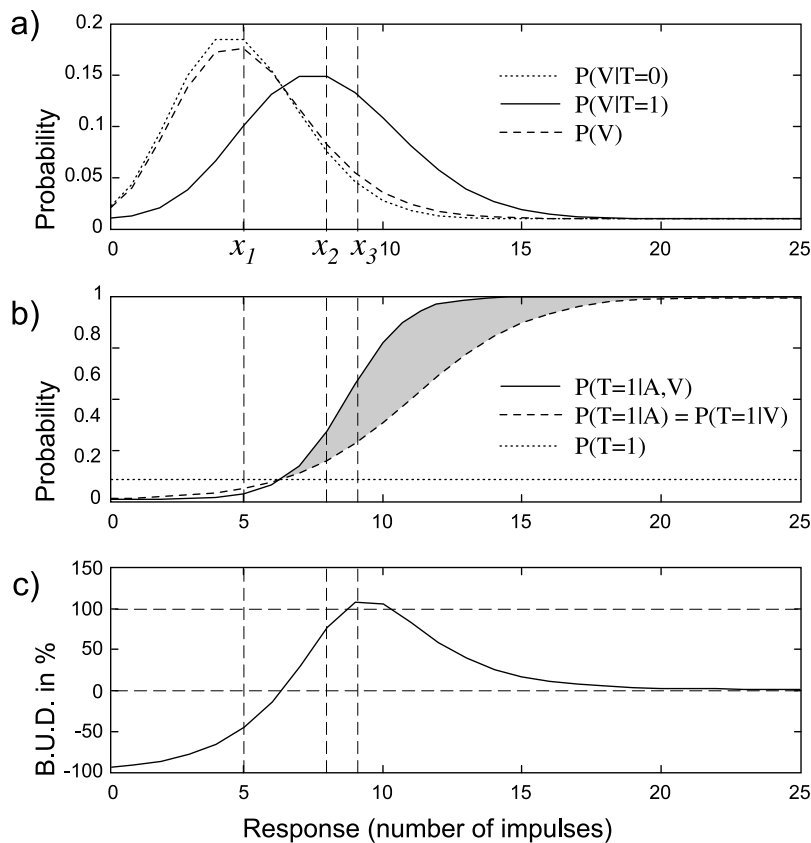


Abbildung 4.9: Visualisierung eines hypothetischen bimodalen Bayesfilters nach [APBB00]. a) Poisson-Verteilungen für unimodale Spontanaktivierungen ( $P(V|T=0), \lambda=5$ ) und getriebene Response ( $P(V|T=1), \lambda=8$ ). Die resultierende Kurve für  $P(V)$  wurde mit Gleichung 4.4 bestimmt. b) Bayessche Berechnung der unimodalen und bimodalen, posterioren Target-Wahrscheinlichkeiten bei gegebener Prioren  $P(T=1)=0.1$ . c) Bimodal-unimodale Differenz (B.U.D.) der Target-Wahrscheinlichkeiten. ( $x_1$ =Poisson-Parameter der Spontanaktivität,  $x_2$ =Poisson-Parameter der Input-getriebenen Antwort,  $x_3$ =Stelle der maximalen Differenz zwischen uni- und bimodaler Target-Wahrscheinlichkeit.)

Aus dem bimodalen Bayesschen Formelapparat (4.5–4.7) ist ersichtlich, dass im multisensorischen Superior Colliculus diverse unimodale Wahrscheinlichkeiten als Erfahrungswerte, Parameter oder Zwischenergebnisse kodiert werden müssen, damit er die posteriore, bimodale Target-Wahrscheinlichkeit berechnen kann. Dieser Umstand wird von mehreren Autoren in Verbindung mit den Befunden über unisensorische Neurone im SCd gebracht, deren Aktivierung eben diese unimodalen Größen repräsentieren soll [CD04, PBBA02, APBB00].

Abbildung 4.9 zeigt die Ergebnisse der vorgeschlagenen Bayesschen Filterung in einem Bereich möglicher visueller oder auditorischer Inputs von 1–25 Spikes im angenommenen Zeitintervall. Beide Eingänge wurden exemplarisch mit identischen Poisson-Verteilungen modelliert. Das Bayessche Filter verknüpft eine hypothetische



a priori Wahrscheinlichkeit  $P(T=1)$  mit den Wahrscheinlichkeiten der sensorischen Inputs  $P(V|T)$ ,  $P(A|T)$  und liefert die in 4.9b) dargestellten, sigmoid verlaufenden Target-Wahrscheinlichkeiten. Diese sind bei einer geringen Spike-Anzahl sehr niedrig (fast gleich Null) und für hohe Spikeraten etwa 1. Die multimodale Target-Wahrscheinlichkeit steigt dabei in einem kürzeren Intervall an als die unimodale. Deshalb kann im schraffierten Bereich analog zur Response Enhancement der neuronalen Aktivierung auch eine bimodale Verstärkung bezüglich der Wahrscheinlichkeitswerte beobachtet werden. Sie bedeutet nichts anderes, als dass mit dem bimodalen Filter eine schärfere Unterscheidung zwischen Anwesenheit und Abwesenheit eines Zieles getroffen wird. ANASTASIO interpretiert die prinzipielle Gestalt der hypothetischen Kurven vor dem neurobiologischen Hintergrund: die multisensorischen Zellen des Superior Colliculus könnten prämotorische Kommandos schneller und zuverlässiger kodieren als ihre unisensorischen Nachbarn.

Die bimodal-unimodal difference (BUD, in [PA03] auch cross-modal enhancement genannt) wird anhand der posterioren Wahrscheinlichkeiten berechnet und sollte in der Diskussion deutlicher von STEINs Definition der Response Enhancement getrennt werden als es in ANASTASIOS Ausführungen geschieht. Eine multimodale Aktivierung als Grundlage der Response Enhancement wird vom Bayesschen Filter schließlich gar nicht berechnet. Außerdem beziehen sich uni- und multisensorische Wahrscheinlichkeiten auf verschiedene, nämlich uni- oder multisensorische Neurone, die simultan ein audio-visuelles Ereignis kodieren. Response Enhancement charakterisiert hingegen die Aktivierungen ein und derselben Zelle bei aufeinander folgender uni- und multisensorischer Stimulation. Wollte man das Bayessche Filter so parametrisieren, dass höhere BUD-Werte auftreten, hätte dies steilere Anstiege in den sigmoiden Kurven der Target-Wahrscheinlichkeiten zur Folge. Dadurch würde sich aber automatisch das Intervall in der Eingangsdynamik verkleinern, in dem überhaupt positive BUD-Werte entstehen. Schon im zitierten Beispiel korrespondiert die größte bimodal-unimodale Differenz mit einer Beobachtung von neun Spikes je Zeiteinheit, also mehr als dem Poisson-Maximum der getriebenen Response (vergl. Abbildung 4.9, Stellen  $x_2$  und  $x_3$ ). Da die Spikezahl typischer Aktivierungen beim Erscheinen von Zielen in einem ungefähren Bereich zwischen spontaner und getriebener Response liegen wird, sind große BUD-Werte generell selten und gemeinhin an hohe Spikeraten gebunden. Dies widerspricht aber gerade der umgekehrten Proportionalität von unimodaler Aktivierung und Response Enhancement. Ein weiteres Manko in ANASTASIOS Konzept ergibt sich aus der isolierten Betrachtung einzelner topographischer Positionen und Zeitintervalle. Die Target-Wahrscheinlichkeiten werden unabhängig von benachbarten

Zellen berechnet, und die Gestaltung der räumlichen Merkmale von Enhancement- oder Depression-Effekten ist somit nur anhand der Geometrie der rezeptiven Felder des sensorischen Eingangs möglich. Ebenfalls offen bleibt die Frage, welche iterativen Algorithmen zur praktischen Implementierung des Filters geeignet sind, und ob die zeitliche Korrespondenz oder Dekorrelation multimodaler Stimuli außerhalb einzelner Zeitfenster überhaupt ausgewertet werden kann.

### Implementierungen

Der Bayessche Zusammenhang zur Berechnung der posterioren Target-Wahrscheinlichkeit bezieht sich in den zitierten Ansätzen [APBB00, PBBA02, PA03] auf einzelne Beobachtungen des multimodalen Systems. Die in Abbildung 4.9 dargestellten, möglichen Werte von 1–25 Spikeimpulsen sollten beispielsweise für ein hypothetisches Zeitintervall von 250ms gelten [APBB00]. Eine solche Zeitspanne ist nicht nur sehr viel länger als die sensorischen Latenzen im SC [SW96], sondern zweifellos auch ungeeignet, um anhand der Wahrscheinlichkeitswerte rechtzeitig motorische Konsequenzen einzuleiten. Um dem Simulationsmodell auf Basis eines WTA-Netzes eine vergleichbare probabilistische Implementierung gegenüber zu stellen, muss nun ein iterativer Algorithmus gefunden werden, der zumindest die zeitliche Dynamik der Bild- und Audiodaten abbilden kann.

#### *Bayessche multisensorische Tracking-Filter*

Zunächst liegt es nahe, die Anwendung einer Reihe von Algorithmen in Betracht zu ziehen, die in der Audio- und Videoverarbeitung als Bayessche Tracking-Filter bekannt sind. Immerhin scheint deren Lokisationsaufgabe und ihr Informationsgewinn durch die multimodale Fusion mit der Zielstellung der multisensorischen Modellierung verwandt zu sein. Auch das zugrundeliegende Datenmaterial – Videosequenzen mit realen Situationen – stellt einen gemeinsamen Ausgangspunkt dar. Letztendlich wird die visuelle und auditorische Vorverarbeitung aus den Kapiteln 2 und 3 auf vergleichbaren Audio- und Videosequenzen operieren.

Eine erste Gruppe von multimodalen Bayesschen Algorithmen basiert auf sogenannten probabilistischen graphischen Modellen [Cow99]. Ein exemplarischer, stark vereinfachter Graph könnte für die audio-visuelle Verknüpfung der stochastischen Größen  $A$  und  $V$  zu einer multimodalen Aktivität  $M$  eine Form mit drei Knoten und zwei gerichteten Kanten annehmen:  $(V) \rightarrow (M) \leftarrow (A)$ . Die Gesamtwahrscheinlichkeit des durch den Graphen beschriebenen Systems würde dann

$P(V, A, M) = P(V) \cdot P(A) \cdot P(M|V, A)$  lauten. In der Praxis werden probabilistische Graphen soweit verfeinert, dass durch ihren generativen Charakter beispielsweise ersichtlich wird, wie aus einer unbekanntem akustischen Quelle und der ebenfalls nicht direkt messbaren binauralen Laufzeit ein beobachtbares Stereosignal entsteht. In einem komplexen Graphen für die audio-visuelle Fusion wären außerdem Größen wie das Sensorrauschen, ein Verschiebeparameter für visuelle Objektbewegungen oder ein Maß für die räumliche Korreliertheit von optischen und akustischen Quellen enthalten – und zwar mit Instanzen für alle möglichen diskreten Richtungen [BJA03]. Als Algorithmus zur iterativen Berechnung der Systemwahrscheinlichkeit des Graphen wird die Methode der Expectation Maximization (EM) [JGJS99, RH99] vorgeschlagen. In einem am Videotakt orientierten Frame-Konzept erlaubt die EM-Methode eine iterative Bayessche Berechnung der posterioren Wahrscheinlichkeit aller beobachteten und latenten Variablen sowie aller frameunabhängigen Systemparameter. Die gleichzeitige Maximierung der Systemwahrscheinlichkeit dient nicht nur der langfristigen Optimierung der Parameter, sondern kann in Bezug auf bestimmte Variablen auch die Schätzung von audio-visuellen Zielpositionen realisieren.

Das umfangreiche Set an stochastischen Variablen und Parametern, das zur Beschreibung des zitierten audio-visuellen Modells in [BJA03] herangezogen wird, macht deutlich, dass hier über eine räumliche Aufmerksamkeitssteuerung hinaus Objekterkennung betrieben werden soll. Insbesondere der pixelorientierte Ansatz des visuellen Teilmodells ist dafür ausgelegt, unter Zuhilfenahme beliebiger statistischer Momente farb- und formbasierte Zusammenhänge in hochaufgelösten Bildfolgen zu detektieren. Die Anwendung dieses mächtigen statistischen Apparates auf die eindimensionale und unscharfe Abbildung von Bewegungsrichtungen erscheint nicht mehr im ursprünglichen Sinne der probabilistischen Graphen und deren EM-Implementierung.

Die Simulation eines umfangreichen probabilistischen Graphen ist vor allem deshalb sehr aufwendig, weil die enorme Anzahl topographischer Positionen einen riesigen Zustandsraum definiert, in dem sich die Wahrscheinlichkeitsdichten aller Variablen und Systemparameter entfalten. Eine alternative Variante Bayesscher Tracking-Algorithmien, die audio-visuelle Fusion mit Partikelfiltern [ZDD02], vermeidet dieses Problem. Im einfachsten Fall kann in Partikelfiltern die dynamische Beschreibung des Systems durch nur eine Zustandsgröße erfolgen. Sie kodiert mutmaßliche Zielpositionen über eine beliebige Wahrscheinlichkeitsdichtefunktion (PDF), die mit Hilfe einer Partikelwolke  $s_i$  in einem topographisch zu deutenden Zustandsraum  $S$  approximiert wird. Unter Einbeziehung der Beobachtung  $Z$  lautet die Beschreibung des Partikelfilters in Analogie zur Bayesschen Formel:  $P(s_i|Z) = P(Z|s_i)P(s_i)/P(Z)$ . Der Term

$P(Z|s_i)$  wird als Sensormodell bezeichnet und ist mit der Poisson-verteilter Kodierung des sensorischen Inputs in ANASTASIOS Ansatz vergleichbar.  $P(s_i)$  geht aus der Anwendung eines Bewegungsmodells hervor, mit dem das Filter an die räumlich-zeitliche Dynamik der Umwelt angepasst werden kann. Im Verlauf der iterativen Berechnung des Filters erfolgt nacheinander eine Propagation der Partikelwolke nach dem Bewegungsmodell, eine Bayessche Aktualisierung der Posterioren  $P(s_i|Z)$  und ein Resampling der Partikel zur Normierung der repräsentierten PDF. Die eigentliche multimodale Fusion soll nach [ZDD02] innerhalb des Sensormodells durch eine multiplikative Verknüpfung realisiert werden:  $P(Z|s_i) = P_v(Z_v|s_i) \cdot P_a(Z_a|s_i)$ .

Auch Partikelfilter werden gerne mit einer merkmalsbasierten Vorverarbeitung der sensorischen Daten versehen und dann zur Verfolgung von Objekten eingesetzt [BGPV01]. Während diese Besonderheit bei der Anwendung des Filters im Kontext der frühen Aufmerksamkeitssteuerung übergangen werden kann, gibt es dennoch Bedenken gegen den unmittelbaren Einsatz zur Implementierung des multisensorischen Modells. Ungeachtet der objektspezifischen oder allgemeinen Kodierung des sensorischen Inputs werden sowohl mit probabilistischen Graphen als auch mit den beschriebenen Partikelfiltern Tracking-Aufgaben gelöst. Beide Ansätze gehen von der Existenz eines Zieles aus und fragen, welche Position als Aufenthaltsort am wahrscheinlichsten ist. Die Bayessche Filterung im Modell des Superior Colliculus soll hingegen an einzelnen Positionen möglichst sichere Aussagen über die Anwesenheit von Zielen treffen. Diese beiden Aufgabenstellungen sind keineswegs identisch und führen zu einer interessanten Überlegung: Scheinbar würde die pauschale Annahme der Existenz eines Zieles ein Manko der Modellierung unabhängiger topographischer Positionen im Ansatz von ANASTASIO überwinden. Das Resampling einer Partikelwolke zur Normierung der PDF über die Zielpositionen wirkt sich beispielsweise ähnlich wie die Limitierung einer neuronalen Ressource aus. Als eine Voraussetzung für den Effekt der Response Depression müssten räumlich dekorrelierte Reize um diese Ressource konkurrieren. Allerdings führt allein die Tatsache, dass eine Repräsentation im Superior Colliculus erlischt, wenn ein Ziel verstummt und sich nicht bewegt, die Fragwürdigkeit dieser Interpretation vor Augen.

#### *Spezifisches Bayesfilter mit lateraler Wichtung*

In Anbetracht des enormen Simulationsaufwandes der diskutierten, pixelbasierten Algorithmen und ihrer besonderen Zielstellung eines Trackingverhaltens ist es überlegenswert, ob anstelle der Modifikation bekannter audio-visueller Verfahren ein eigener Ansatz zur multimodalen Bayesschen Filterung formuliert werden kann. Im Gegensatz

zu den zitierten Anwendungen von probabilistischen Graphen [BJA03] oder Partikelfiltern [ZDD02] sollen im Modell des Superior Colliculus nicht für hunderte oder tausende Bildpunkte Bayessche Berechnungen angestellt werden. Die wenigen diskreten Winkelintervalle für horizontale Richtungen lassen deren explizite Berechnung sinnvoller erscheinen als etwa die Approximation durch eine Partikelwolke. Andererseits wäre es wünschenswert, einzelne Richtungen nicht völlig unabhängig voneinander zu behandeln, sondern die aus der WTA-Implementierung bekannte nachbarschaftliche Abhängigkeit im topographischen Verbund zu berücksichtigen.

Zunächst soll analog zu ANASTASIOS Betrachtungen eine stochastische Zustandsvariable  $T$  die Anwesenheit von Zielen beschreiben – dies jedoch für den kompletten wahrgenommenen Winkelbereich  $\phi = [\varphi_1, \dots, \varphi_n]$ . Die Verteilung der Target-Wahrscheinlichkeiten über alle Winkel wird im Folgenden *Belief* des Bayesschen Filters genannt. Als Grundlage einer iterativen Berechnung soll dieser Belief vollständig durch die zurückliegenden sensorischen Beobachtungen bestimmt sein. In einem vorerst unisensorischen Fall mit stochastischen visuellen Beobachtungen  $V_\phi$  nimmt der Belief demnach die Form  $P(T_{t,\phi}|V_{t,\phi}, \dots, V_{0,\phi})$  an. Desweiteren möge der aktuelle Zustand im Sinne einer Markovschen Prozessbeschreibung nur aus seiner letzten Schätzung und der aktuellen Beobachtung hervorgehen. Wie in Anhang B.2 gezeigt wird, lässt sich unter diesen Voraussetzungen eine Rekursionsformel herleiten:

$$\begin{aligned}
P(T_{t,\phi}|V_{t,\phi}, \dots, V_{0,\phi}) &= \frac{P(T_{t,\phi}, V_{t,\phi}|V_{t-1,\phi}, \dots, V_{0,\phi})}{P(V_{t,\phi}|V_{t-1,\phi}, \dots, V_{0,\phi})}, \text{ mit } \phi = [\varphi_1, \dots, \varphi_n] & (4.8) \\
&= \alpha \underbrace{P(V_{t,\phi}|T_{t,\phi}, V_{t-1,\phi}, \dots, V_{0,\phi})}_{\stackrel{!}{=}P(V_{t,\phi}|T_{t,\phi}) \text{ (lt. Markovbed.)}} \cdot P(T_{t,\phi}|V_{t-1,\phi}, \dots, V_{0,\phi}) \\
&= \alpha \underbrace{P(V_{t,\phi}|T_{t,\phi})}_{\text{Sensormodell}} \int \underbrace{P(T_{t-1,\phi}|V_{t-1,\phi}, \dots, V_{0,\phi})}_{\text{alter Belief}} \cdot \underbrace{P(T_{t,\phi}|T_{t-1,\phi})}_{\text{Bewegungsmodell}} dT_{t-1,\phi}
\end{aligned}$$

Die grundlegenden Komponenten einer Bayesschen Ermittlung der posterioren Target-Wahrscheinlichkeit können nun in Bezug auf ihre iterative Anwendung interpretiert werden. Die Beobachtung  $V_{t,\phi}$  wird mit Hilfe eines Sensormodells  $P(V_{t,\phi}|T_{t,\phi})$  anhand der Eigenschaften der sensorischen Kodierung bewertet. Gleichzeitig beschreibt ein Bewegungsmodell  $P(T_{t,\phi}|T_{t-1,\phi})$  die Dynamik der Ziele. Der Belief aus dem vorangegangenen Zeitschritt  $P(T_{t-1,\phi}|V_{t-1,\phi}, \dots, V_{0,\phi})$  stellt die bislang geschätzten (prioren) Target-Wahrscheinlichkeiten dar und wird in einem Faltungsintegral mit dem Bewegungsmodell verknüpft. Offensichtlich ist aber die Integrationsvariable  $T_{t-1,\phi}$  mit  $2^{|\phi|}$  möglichen Belegungen der Winkel mit  $T \in \{1, 0\}$  noch zu mächtig für eine konkrete Simulation. Um die allgemeine Iterationsvorschrift 4.8 praktisch berechenbar zu machen, können zum Zeitpunkt  $t$  die bedingten Target-Wahrscheinlichkeiten (aktueller und

letzter Belief) für unabhängige Winkel  $\varphi_i$  betrachtet werden. Die nachbarschaftliche Abhängigkeit wird in diesem Fall ausschließlich über das Bewegungsmodell gesteuert. Setzt man außerdem eine hinreichende physikalische Trägheit der Ziele voraus, muss das Bewegungsmodell lediglich einen lokalen Bereich berücksichtigen (beispielsweise  $\phi = [\varphi_i - 2, \dots, \varphi_i + 2]$ ). Somit reduziert sich die Rekursionsformel auf eine konkrete Form mit expliziten Winkelangaben (vergl. Anhang B.2):

$$P(T_{t,\varphi_i} | V_{t,\phi}, \dots, V_{0,\phi}) = \beta \overbrace{P(V_{t,\varphi_i} | T_{t,\varphi_i})}^{\text{Sensormodell}} \cdot \int \underbrace{P(T_{t,\varphi_i} | T_{t-1,\phi})}_{\text{Bewegungsmodell}} \underbrace{\prod_{m=i-2}^{i+2} P(T_{t-1,\varphi_m} | V_{t-1,\phi}, \dots, V_{0,\phi})}_{\text{alter Belief}} dT_{t-1,\phi}$$

Nun kann nach einer adäquaten Gestalt von Sensor- und Bewegungsmodell gefragt werden. In seinem hypothetischen auditorisch-visuellen Bayesfilter benutzte ANASTASIO anstelle eines Bewegungsmodells eine arbiträre Konstante  $P(T=1)=0.1$  und definierte im Sinne eines Sensormodells poissonverteilte Spikeraten für  $P(V|T=1)$  und  $P(V|T=0)$ . Zur Implementierung der Iterationsvorschrift 4.9 ist keine dieser Annahmen geeignet. Beispielsweise kommen bei der in den Kapiteln 2 und 3 beschriebenen sensorischen Vorverarbeitung keine stochastischen neuronalen Modelle zum Einsatz, weshalb eine Poisson-PDF insbesondere zur Modellierung der Spontanaktivierung ausscheidet. Um zunächst die Eigenschaften des Sensormodells festzulegen, erscheint ein experimentelles Vorgehen sinnvoll. In der sensorischen Kodierung spiegeln sich nicht nur algorithmische Aspekte wie die Bewegungsdetektion und die Topographie der rezeptiven Felder wider, sondern auch technische Einflüsse (Sensorrauschen) und physikalische Merkmale der Umgebung (z.B. Beleuchtung und Kontrastverhältnisse). Erstellt man ein Histogramm der Aktivierungen in der simulierten Bewegungskarte, und zwar für eine großen Anzahl exemplarischer Ziele, werden automatisch alle der genannten sensorischen und umweltspezifischen Eigenschaften erfasst. Es überrascht wenig, dass ein solches Histogramm (Abbildung 4.10a) ein Sensormodell beschreibt, dessen Gestalt deutlich von den Poissonverteilungen in ANASTASIOS Ansatz abweicht.

Auch zur Realisierung des Bewegungsmodells könnte eine datenbasierte Verfahrensweise erwogen werden, um die Charakteristik der Ziele in Verbindung mit dem Abstand und der Geometrie der sensorischen Anordnung zu beurteilen. Allerdings erweist es sich als problematisch, einen aussagekräftigen statistischen Zusammenhang zwischen der Dynamik in der Bewegungskarte und den binären Belegungen der stochastischen Targetvariable  $T$  herzustellen. Stattdessen soll an dieser Stelle ein vorerst arbiträres Bewegungsmodell eingeführt werden, das im Rahmen einer späteren, experi-

mentellen Optimierung (Abschnitt 5.2.4) an die Dynamik eines Anwendungsszenarios angepasst wird. In Abbildung 4.10b) ist dargestellt, wie durch die Verknüpfung einer exemplarischen Normalverteilung mit den möglichen lokalen Belegungen der binären Targetvariable  $T$  das Bewegungsmodell tabelliert wird.

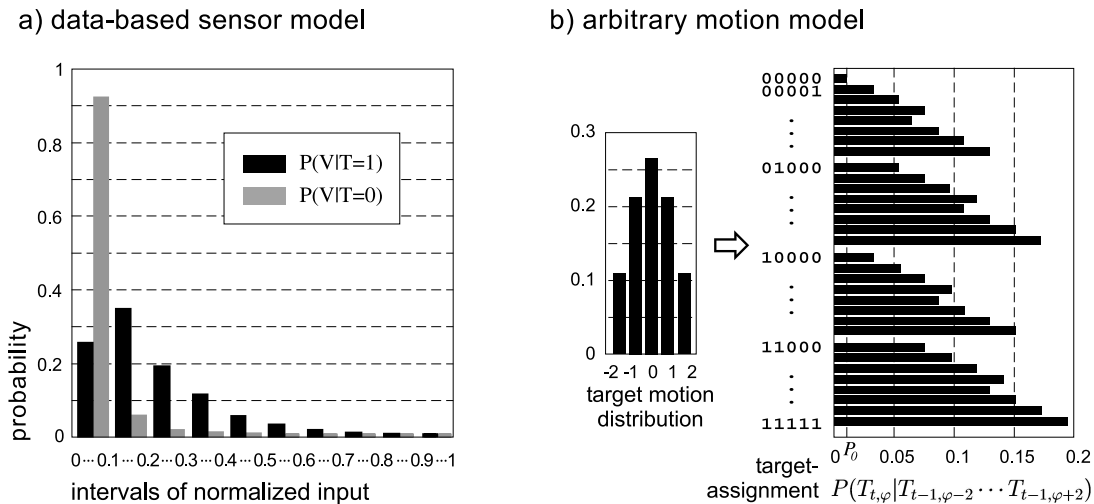


Abbildung 4.10: a) Zur Charakterisierung des Sensormodells der visuellen Kodierung wurde ein Histogramm der Bewegungsabbildung herangezogen. Bei der Schätzung der Kodierungswahrscheinlichkeit bei vorhandenen Zielen ( $P(V|T=1)$ ) gingen an einer Zielposition  $\varphi=0$  Grad die sensorischen Repräsentationen von 35 Gesten aus einer Szenendatenbank (Abschnitt 5.2.1) ein. Bei Abwesenheit der Ziele ( $P(V|T=0)$ ) wirkte sich lediglich das Pixelrauschen der verwendeten Kamera auf das Ergebnis der Bewegungskodierung aus.

b) In einem exemplarischen, stochastischen Bewegungsmodell werden die Geschwindigkeiten der Ziele als normalverteilt ( $\sigma=1.5$ ) in einem Bereich von  $\pm 2$  Winkelschritten betrachtet. Die Faltung dieser Verteilung mit allen denkbaren Belegungen der letzten Zielpositionen  $\varphi-2, \dots, \varphi+2$  ergibt eine tabellierte, bedingte Wahrscheinlichkeit für das Vorhandensein von Zielen. Ein Grundbetrag in  $P(T_t=1|T_{t-1})$  erlaubt auch Reaktionen auf Ziele, die neu erscheinen.  $P(T_t=0|T_{t-1})$  nimmt die Werte  $1 - P(T_t=1|T_{t-1})$  an.

Die Überlegungen und statistischen Auswertungen beim Entwurf von Sensor- und Bewegungsmodell können prinzipiell sowohl für die visuelle als auch für die auditorische Repräsentation der Umwelt angestellt werden. Eine derartige Unterscheidung der sensorischen Modalitäten könnte von Vorteil sein, wenn beispielsweise die auditorische und die visuelle Repräsentation eines gestikulierenden Sprechers verschiedene Bewegungsmodelle vermuten lassen. Auch die stochastischen Eigenschaften der sensorischen Kodierung werden in den jeweiligen Modalitäten spezifisch ausgeprägt sein, denn neben dem Rauschen der Video- und Audioaufnahmen wirken sich desweiteren algorithmische Unterschiede zwischen der visuellen Projektion und der auditorischen Berechnung der räumlichen Abbildungen aus. Beim Versuch, das auditorische Sen-

sormodell auf Basis eines Histogramms der binauralen Kreuzkorrelation oder ihrer WTA-gefilterten Repräsentation zu erstellen, waren ähnliche Werteverteilungen wie in Abbildung 4.10a) zu beobachten. Allerdings führen Pausen im Audiosignal zu einer noch stärkeren Betonung sehr kleiner Werte, und eine plausible Definition der An- oder Abwesenheit auditorischer Ziele ist kaum möglich. Exemplarisch soll daher, in Analogie zu [APBB00], ein multisensorisches Bayesfilter mit identischen Sensor- und Bewegungsmodellen in den beteiligten Modalitäten untersucht werden. Korrespondierend zu Gleichung 4.6 und in Einklang mit den zitierten audio-visuellen Partikelfiltern erfolgt die Fusion der beiden unisensorischen Repräsentationen auf multiplikative Weise bei der Anwendung des Sensormodells:  $P(V, A|T) = P(V|T) \cdot P(A|T)$ .

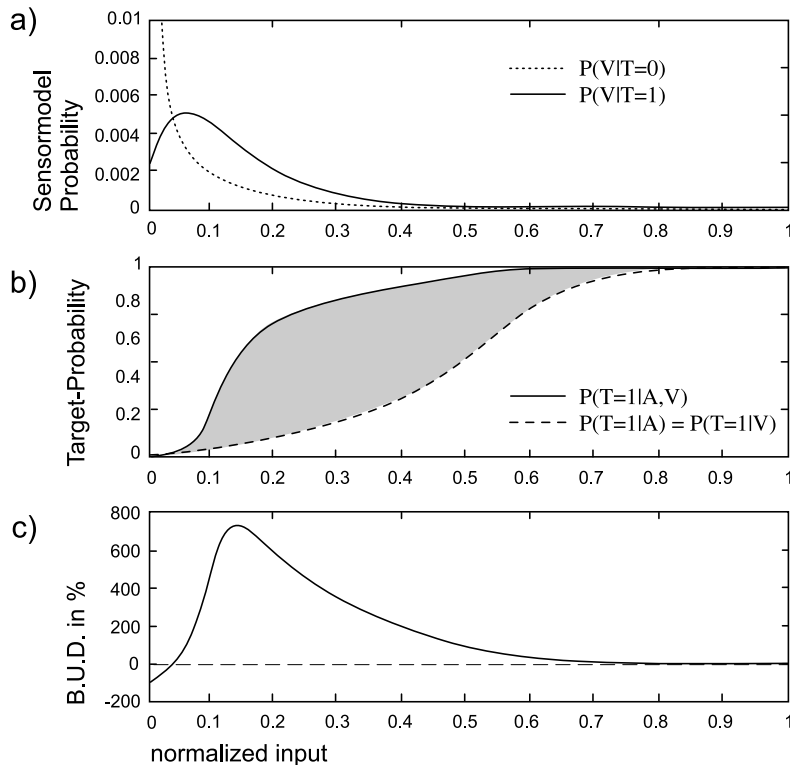


Abbildung 4.11: Visualisierung des rekursiven Bayesschen Filters bei Anwendung der Sensor- und Bewegungsmodelle aus Abbildung 4.10 (im Sensormodell  $P(V|T)$  wurden die experimentell gewonnenen Histogramme durch Exponentialfunktionen approximiert). Die Kurven zeigen einen Iterationsverlauf mit kleiner Schrittweite und geben recht genau die statischen Verhältnisse wider. Alle Winkelintervalle wurden mit den gleichen, linear wachsenden Eingangswerten  $V=A$  beaufschlagt.

Tatsächlich ist es möglich, bereits mit den bisherigen heuristischen Annahmen über Sensor- und Bewegungsmodell ein Bayessches Filter zu simulieren, das eine augenscheinlich sehr viel günstigere multisensorische Charakteristik aufweist als die hypothetischen Berechnungen von ANASTASIO. Greift man dessen Darstellungsweise



von Kodierungs-PDF, Target-Wahrscheinlichkeit und bimodal-unimodaler Differenz (BUD) auf, lassen sich die iterativen Filterergebnisse direkt mit den statischen Kurven in Abbildung 4.9 vergleichen. Abbildung 4.11 verdeutlicht, dass mit Hilfe des experimentell bestimmten Sensormodells weitreichende multisensorische Enhancement-Effekte realisiert werden. Die bimodale Target-Wahrscheinlichkeit steigt schon bei niedrigen Eingangsaktivierungen schnell an und verursacht BUD-Werte von mehreren hundert Prozent. Bemerkenswert ist vor allem, dass die BUD-Kurve trotz des erwünschten, zeitigen und hohen Maximums nur langsam abfällt. Offensichtlich lassen sich die Befunde einer maximalen Response Enhancement um 1000 Prozent und die umgekehrte Proportionalität von unimodaler Aktivierung und multimodaler Verstärkung mit den Target-Wahrscheinlichkeiten des neu entworfenen Filters wesentlich besser modellieren.

Um schließlich den Einfluss der lateralen Abhängigkeit im Bewegungsmodell und das Verhalten des Filters bei der Verarbeitung realer Daten zu demonstrieren, wurde die bereits aus Abschnitt 4.3.2 bekannte audio-visuelle Szene herangezogen. In ihr konkurrieren eine uni- und eine multimodale Geste (Winken und Händeklatschen) um den Aufmerksamkeitsfokus der simulierten Modelle. Am Diagramm der visuellen Target-Wahrscheinlichkeit  $P(T|V)$  in Abbildung 4.12 fällt zunächst auf, dass im Bayesschen Filter infolge der nur lokalen nachbarschaftlichen Korrespondenzen im Bewegungsmodell keine globale räumliche Selektion erzielt wird. Anders als die Aktivierung der simulierten Winner-Take-All Architektur (vergl. Abbildung 4.7) repräsentieren die Wahrscheinlichkeitswerte des Bayesfilters beide simultanen visuellen Ereignisse unabhängig und gleichberechtigt. Positiv zu bewerten sind die nichtlinearen multimodalen Effekte, die laut Diagramm der multisensorischen Wahrscheinlichkeit  $P(T|V, A)$  auch bei realen Eingangsmustern zum tragen kommen. Eine genauere Betrachtung der Werteverläufe im Winkelintervall der audio-visuellen Zielposition bestätigt, dass die Differenz der bimodalen und unimodalen Wahrscheinlichkeiten tatsächlich genau dann hohe Werte annimmt, wenn weder die auditorische noch die visuelle Einschätzung der Szene eine sichere Aussage über die Anwesenheit eines Zieles erlauben. Bei der exemplarischen Parametrisierung und Wichtung der sensorischen Eingänge reagiert die multisensorische Target-Wahrscheinlichkeit überhaupt nur auf korrelierte multimodale oder sehr eindeutige unimodale Ereignisse. Bei mehr oder weniger unsicheren unimodalen Stimuli bleiben die Werte im  $P(T|V, A)$ -Diagramm fast gleich Null und korrespondieren mit einer pauschalen neuronalen Response Depression nahe 100%. Um mit dem vorliegenden Filter ein robustes Orientierungsverhalten zu modellieren, müssen demnach immer alle drei Wahrscheinlichkeitskarten ausgewertet werden.

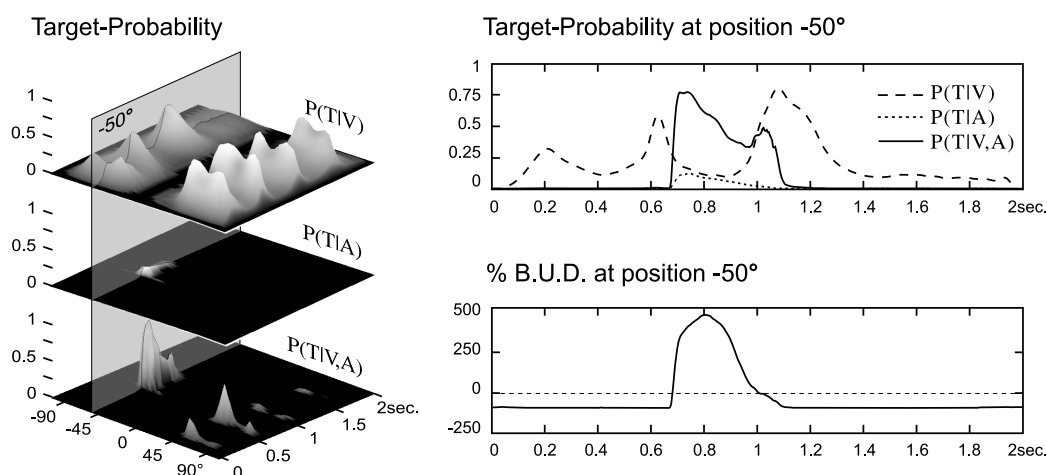


Abbildung 4.12: Visuell, auditorisch und multisensorisch basierte Target-Wahrscheinlichkeiten als Bayessche Filterantwort auf die bereits in Abbildung 4.7 vorgestellte Szene. Für das Winkelintervall der Zielposition wurden in den rechten Diagrammen die Wahrscheinlichkeitswerte einzeln dargestellt und ihre bimodal-unimodale Differenz (BUD) berechnet. Um hohe BUD-Werte zu demonstrieren, wurde bewusst ein niedriger Audiopegel eingestellt.

#### 4.3.4 Gegenüberstellung der Ansätze

Spätestens mit dem Trend, Bayesfilter für audio-visuelle Problemstellungen anzuwenden, stellt sich die Frage nach einem Vergleich zwischen den Modellvarianten mit künstlichen neuronalen Netzen und solchen auf Grundlage probabilistischer Algorithmen. So wurde schon versucht, des Verhaltens eines sogenannten Basis-Function-Networks mit Hilfe von ANASTASIOS Beschreibungen eines multimodalen Bayesfilters zu interpretieren [DP04]. Die dabei vorgenommene Anwendung des probabilistischen Ansatzes zur Evaluierung der topographischen Eigenschaften eines neuronalen Netzes bleibt jedoch vage und wird als allgemeines Modell verschiedener subkortikaler und kortikaler Effekte diskutiert. Nachdem in den beiden vorangegangenen Abschnitten aber zwei eigenständige Implementierungen mit entweder neuroanatomischer oder probabilistischer Motivation entworfen wurden, sollen diese nun auch anhand der konkreten Befunde zum Superior Colliculus verglichen werden. Es liegt nahe, einen solchen Vergleich unter Bezugnahme auf einige der in Abschnitt 4.3.1 formulierten Modellanforderungen zu führen:

##### *Response Enhancement*

Der von STEIN und MEREDITH geprägte Begriff der Response Enhancement bezieht sich auf mittelfristige neuronale Aktivierungen in separaten uni- und multisensorischen Experimenten und ist auf die Simulation des Winner-Take-All

Netzwerkes (WTA) direkt anwendbar. In der vorgeschlagenen WTA-Struktur wird die multisensorische Verstärkung durch die sigmoide Ausgabefunktion der simulierten Neurone und durch die lokalen exzitatorischen Rückkopplungen der Aktivierung ermöglicht. Eine in den rekurrenten Verschaltungen begründete Hysterese der Aktivierung realisiert dabei ohne weiteres multisensorisch sensitive Zeitfenster von mehreren hundert Millisekunden. Anstelle der Response Enhancement beschreibt in probabilistischen Modellen die Differenz von uni- und bi-sensorischen Target-Wahrscheinlichkeiten eine erhöhte Sicherheit der Aussage über die Anwesenheit von multimodalen Zielen. Ihren Ursprung hat die nichtlineare Verstärkung der Wahrscheinlichkeitswerte in der multiplikativen Verknüpfung der asymmetrischen Sensormodelle für  $P(Z|T=1)$  und  $P(Z|T=0)$ . Multisensorische Zeitfenster, in denen über die unmittelbare Stimulusdauer hinaus multisensorische Enhancement-Effekte auftreten, lassen sich mit den beschriebenen Bayesfiltern allerdings nicht realisieren. Infolge ihrer schärferen Unterscheidung zwischen An- und Abwesenheit von multimodalen Zielen sinkt die Target-Wahrscheinlichkeit am Ende audio-visueller Ereignisse noch schneller als im unimodalen Fall.

#### *Umgekehrte Proportionalität von unimodaler Reaktion und multimodaler Verstärkung*

Der Befund, dass in den tiefen Schichten des Superior Colliculus die deutlichste multisensorische Verstärkung gerade durch die Kombination von unisensorisch ineffektiven Stimuli verursacht wird, kann mit beiden Simulationsmodellen nachvollzogen werden. Im WTA-Netz limitieren die sigmoide Ausgabe sowie der globale Inhibitionsmechanismus die Enhancement-Effekte bei höheren Eingangsaktivierungen. In einer ähnlichen Weise wird auch die bimodal-unimodale Differenz im Bayesschen Filter begrenzt, wenn bereits eine der beiden unisensorischen Target-Wahrscheinlichkeiten Werte nahe Eins annimmt. Die Sicherheit der unimodalen Aussage über die Anwesenheit eines Zieles kann dann durch die Kombination von auditorischer und visueller Information nur noch wenig erhöht werden.

#### *Response Depression*

Die Verminderung der Aktivierung infolge einer Kombination von dekorrelierten akustischen und optischen Reizen kann plausibel mit dem Selektionsverhalten der WTA-Architektur erklärt werden. Im simulierten dynamischen neuronalen Feld bewirkt die globale und von der Gesamtaktivierung abhängige Inhibition einen topographischen Wettstreit um imaginäre physiologische Ressourcen. Im ursprünglichen Bayesschen Modell des Superior Colliculus nach ANASTASIO sind topographische

Abhängigkeiten als Folge globaler oder begrenzt lateraler Mechanismen nicht vorgesehen. In der Praxis fällt die bimodal-unimodale Differenz unabhängig von anderen topologischen Bereichen schnell auf Werte von nahezu  $-100\%$ , wenn einer der beiden sensorischen Eingänge nicht aktiviert wird. Nur die in Partikelfiltern angewandte Normierung der Target-Wahrscheinlichkeit über die Winkelintervalle könnte im weitesten Sinne Depression-Effekte bei simultanen und räumlich getrennten Ereignissen realisieren. Sie entspricht in der Anwendung des Bayesschen Filters für Trackingaufgaben aber einer anderen Zielstellung und würde außerdem die erreichten multisensorischen Enhancement Effekte beeinträchtigen. Response Depression Befunde lassen sich demnach mit den vorgestellten probabilistischen Algorithmen nicht adäquat modellieren.

#### *Selektivität der Wahrnehmung*

Obwohl eine Selektivität innerhalb der sensorischen Repräsentation zu den wichtigsten Grundlagen einer Aufmerksamkeitssteuerung zählt, wird ein räumlicher oder zeitlicher Selektionsaspekt im Rahmen der neurobiologischen Untersuchungen selten explizit diskutiert. Interpretiert man die Befunde zu Response Enhancement und Depression als Ausdruck der Selektivität der Wahrnehmung, erschließt sich ein wesentlicher Unterschied zwischen neuronalen und probabilistischen Ansätzen. Zunächst sind sowohl WTA-Netz als auch Bayesscher Filter in der Lage, eindeutige unimodale oder unterschwellige, aber korrespondierende multimodale Reizmuster zu „selektieren“. Darüberhinaus lässt ein typischer Winner-Take-All Prozess im dynamischen neuronalen Feld neben einem dominanten topologischen Bereich keine weiteren Aktivierungen zu. Ein gleichwertiges Verhalten kann bei einer Bayesschen Filterung, selbst mit der Normierung der Target-Wahrscheinlichkeit in den audio-visuellen Partikelfiltern [ZDD02], nicht realisiert werden. Im simulierten neuronalen Netzwerk stellen die Begrenzung einer Ressource durch die globale Inhibition und die spezifischen lateralen Verschaltungen interne Eigenschaften des wahrnehmenden Systems dar. Probabilistische Modelle beschreiben jedoch nur Eigenschaften der Umwelt und der Kodierung der sensorischen Eingänge, ohne dabei einen globalen räumlichen Kontext herzustellen. Die notwendige und sinnvolle Annahme der statistischen Unabhängigkeit von räumlich getrennten Ereignissen verbietet im probabilistischen Modell geradezu einen selektiven Charakter der topographischen Repräsentation.

#### *Motorische Kodierung*

Während die Realisierung topographischer Register und die Adaption der auditorischen rezeptiven Felder nach Augenbewegungen noch nicht Gegenstand der

multisensorischen Modellierung sein sollten, ist die grundsätzliche Eignung der Simulationsergebnisse zur motorischen Kodierung ein nicht zu vernachlässigendes Kriterium für die Anwendbarkeit der Verfahren. In der WTA-Implementierung schlagen sich die Reaktionen auf uni- und multimodale Ereignisse in ein und derselben Aufmerksamkeitskarte nieder. Die Aktivierungen dieser Karte können beim Überschreiten einer arbiträren Schwelle entsprechend der zugehörigen Zielposition direkt als Motorkommando interpretiert werden. Die bisherigen Simulationsergebnisse des multisensorischen Bayesfilters lassen dagegen vermuten, dass zur Auswertung der Target-Wahrscheinlichkeiten ein zusätzliches Auswahlverfahren notwendig ist. So müssten für die visuelle, auditorische und multisensorische Wahrscheinlichkeitskarte separate Schwellwerte adaptiert werden, um bei zu geringen Werten der multimodalen Target-Wahrscheinlichkeit die Reaktion auf unimodale Reize zu garantieren.

Auch mit dem Entwurf des spezifischen rekursiven Bayesfilters bleiben einige grundsätzliche Unterschiede im Verhalten der probabilistischen und künstlichen neuronalen Simulationsmodelle bestehen. Eine wesentliche Hürde bei der vergleichenden Evaluierung beider Ansätze stellt die zeitliche Dynamik der multisensorischen Effekte dar. Die Zeitintervalle, in denen die Befunde zur Response Enhancement anhand elektrophysiologischer Ableitungen ermittelt werden, müssen mit etwa 500–1000ms sehr lang sein, um eine angemessene Repräsentation von Zielbewegungen zu erlauben [MNS87, PY02]. Die Beurteilung des WTA-Verhaltens über eine solche Zeitspanne hinweg ist unkritisch und sogar notwendig, damit die rekurrenten Strukturen im dynamischen neuronalen Feld wirken können. Bei der Simulation des Bayesfilters und der Berechnung seiner bimodal-unimodalen Wahrscheinlichkeitsdifferenz kann eine längere zeitliche Mittelung jedoch zu Fehleinschätzungen führen. Sowohl in den Testsignalen der neurologischen Untersuchungen als auch in realen, alltäglichen Situationen werden akustische und optische Reize selten identisch lange andauern (vergl. Abbildungen 4.4, 4.5 und 4.12). Da die multisensorischen Enhancement-Effekte des Bayesschen Filters aber auf die unmittelbare Stimulusdauer begrenzt sind, vereiteln die ansonsten auftretenden, negativen BUD-Werte eine objektive Einschätzung der Simulation. Beispielsweise würde die zeitliche Mittelung der exemplarischen Target-Wahrscheinlichkeiten aus Abbildung 4.12 nicht eine multisensorische Verstärkung von fast 500%, sondern sogar eine scheinbare Verminderung der Werte im multimodalen Fall ergeben. Ohne willkürliche Einschränkungen in der Gestalt der audio-visuellen Testsignale lassen sich probabilistische Enhancement-Effekte nicht im zeitlichen Mittel, sondern nur anhand der detektierten Maximalwerte oder der Summation der positiven BUD-Intervalle

charakterisieren. Abgesehen von ihrer unterschiedlichen Bedeutung und Berechnungsgrundlage können neuronale Response Enhancement und bimodal-unimodale Wahrscheinlichkeitsdifferenz also auch im praktischen Experiment nicht direkt miteinander verglichen werden und bedürfen stattdessen einer spezifischen Evaluierung.

Dennoch sollen im folgenden Kapitel 5 umfangreichere vergleichende Experimente realisiert werden, um die bislang gezeigten, beispielhaften Simulationsergebnisse und deren Interpretation zu validieren. Neben Vor- und Nachteilen der Verfahren für die praktische Anwendung sind dabei insbesondere Fragen zu Robustheit, Stabilität und Parameteroptimierung von Interesse, die anhand einzelner multisensorischer Versuche noch nicht zu klären waren.

# Kapitel 5

## Experimentelle Untersuchungen

### 5.1 Motivation des experimentellen Ansatzes

#### Analytische Beschreibung vs. experimentelle Evaluation

Beim Entwurf der visuellen, auditorischen und multisensorischen Modellkomponenten in den Kapiteln 2, 3 und 4 wurde ein Sammelsurium bekannter oder modifizierter Algorithmen herangezogen, aus dem sich letztendlich die sinnvollen Varianten eines kompletten Simulationssystems herauskristallisierten. Zu den favorisierten Optionen zählen die binaurale Kreuzkorrelation und anschließende WTA-Filterung im auditorischen Modell, die einfache, räumlich-zeitliche Bewegungskodierung in Bildfolgen sowie die Realisierung der auditorisch-visuellen Integration mittels WTA-Netz oder Bayesfilter. Mit dem Anspruch einer vergleichenden Untersuchung der neurobiologisch oder mathematisch-technisch motivierten Ansätze könnte nun nach einer möglichst weitgehenden analytischen Beschreibung aller Modellkomponenten gefragt werden. Allerdings zeigte bereits die Diskussion der konzeptionellen Unterschiede zwischen Response Enhancement und bimodal-unimodaler Differenz (Vergl. Abschnitt 4.3.4), dass die mathematischen Notationen für neuronale Aktivierungen und probabilistische Zustandsgrößen unterschiedlich zu interpretieren sind. Kommen zu den Schwierigkeiten, die sich aus den anspruchsvollen und heterogenen Formelapparaten ergeben, noch einige unumgängliche Einschränkungen und Vereinfachungen bei der analytischen Evaluation und Optimierung hinzu, dürften die resultierenden Parametrisierungen und Arbeitsbereiche für einen funktionellen Vergleich der Algorithmen oder deren praktische Anwendung kaum relevant sein. Beispielsweise widmen sich zahlreiche Untersuchungen dem Konvergenzverhalten und der Stabilität typischer spike- und ratenkodierter WTA-Strukturen [WW93, YG95, FCK96, Maa00b, Maa00a] und nennen dabei wie-

derholt Anwendungen wie die VLSI-Implementierung [LRMM89, MH94] als praktische Aspekte und Motivation. Eigene Experimente und analoge Hardwarerealisierungen des hier vorgestellten auditorischen Teilmodells [ISP99] ließen aber erkennen, dass die einschränkenden Randbedingungen einer analytischen Beschreibung (konstante Inputs, einheitliche Wichtungen etc.) tatsächlich nur schwer mit einer realen Umsetzung und Anwendung der Modelle vereinbar sind.

Um im multisensorischen System spezifische Response-Eigenschaften des Superior Colliculus nachzubilden, könnten schließlich auch unkonventionelle Arbeitsbereiche des dynamischen neuronalen Feldes geeignet sein, die von einem analytisch optimalen WTA-Prozess abweichen. Gerade für die kurzen Zeitintervalle, in denen eine initiale Kodierung neuer Reizausprägungen stattfindet, sind detaillierte Stabilitätsbetrachtungen oder die Forderung nach einem langfristig starken WTA-Verhalten mit der Fähigkeit der fortlaufenden Aktivierungsverlagerung weniger wichtig. In einem einfachen sensomotorischen Szenario provozieren einsetzende Geräusche und plötzliche Bewegungen umgehend motorische Reaktionen, für deren Dauer irreführende visuelle Aktivierungen infolge der Eigenbewegung durch die nahezu vollständige Inhibition des sensorischen Inputs [RMKP91, WFG95] unterbunden werden. Da im SC aber kein Mechanismus existiert, der eine vorherige topographische Kodierung in die neue sensorische Geometrie nach einem ausgeführten Motorkommando transformiert, sollte die multisensorische Abbildung im technischen System nach jeder Bewegung von Kamera und Mikrofonen neu initialisiert werden. Im Simulationsmodell des WTA-Netzes korrespondiert dies mit der ursprünglich schon von AMARI und KOHONEN empfohlenen Erweiterung ihrer Architekturen durch geeignete Resetmechanismen.

Für die Diskussion über Konvergenz und Stabilität von starken oder schwachen WTA-Prozessen gibt es im Bereich der probabilistischen Verfahren keine direkte Entsprechung. Oft unterstellen die Autoren ihren Simulationsmodellen pauschal ein sehr robustes und adaptives Verhalten, wodurch eine günstige Initialisierung oder die separate Optimierung von Parametern überflüssig würden [BJA03, ZDD02]. Quantitative Eigenschaften wie die bimodal-unimodale Differenz hängen aber sehr wohl von kritischen Modellparametern ab. Deutlich wurde dies an den in Abschnitt 4.3.3 noch ungeklärten Details von Sensor- und Bewegungsmodell, die allein durch theoretische Überlegungen bei der Modellierung oder anhand statistischer Aussagen über exemplarische, sensorische Inputs noch nicht optimal zu gestalten waren.

Wie bereits bei der Evaluation der auditorischen Modellvarianten im Abschnitt 2.3 soll daher auch zum weiteren Vergleich von WTA- und Bayes-Konzept ein pragmatischer, experimenteller Ansatz verfolgt werden. Die Annahmen über die typischen



funktionellen Eigenschaften der Simulationsmodelle, die im vorangegangenen Kapitel aus den multisensorischen Response-Eigenschaften des Superior Colliculus abgeleitet wurden, stellen dabei eine gut strukturierte Heuristik als Ausgangspunkt dar. Bei der Gestaltung geeigneter experimenteller Prozeduren gilt es nun einmal mehr, vielfältige und universelle Stimulusformen zu finden, denn die analytisch nicht beschreibbaren räumlich-zeitlichen Merkmale von natürlichen optischen oder akustischen Ereignissen und deren unterschiedliche Korreliertheit sind wesentliche Faktoren bei der Ausprägung der multisensorischen Effekte [MS86, SM93].

### **Evaluation früher Wahrnehmungsleistungen**

Im Rahmen der auditorischen, laufzeitbasierten Modellierung in Kapitel 2 war es möglich, die angestrebte Funktionalität als Lokalisationsaufgabe zu formulieren und in entsprechenden Benchmarks die Ortung einzelner Quellen als erfolgreich oder fehlerhaft zu bewerten. Ein Fehlerpotential war insofern gegeben, da die interaurale Laufzeit als berechnetes Richtungsmerkmal zum Teil deutlich durch die Raumakustik beeinträchtigt wird. Es galt zu klären, ob sich die Lokalisationsleistung bei schlechter Akustik durch verschiedene sensorische Kodierungsformen, binaurale Korrelationsvarianten oder die WTA-Filterung der Richtungsabbildung verbessern lässt. Eine solche Evaluation im topographischen Ortscode ist für die projizierte Bewegungskarte des visuellen Teilmodells nicht erforderlich. Die in Kapitel 3 diskutierten Filterungen und Transformationen zur Bewegungskodierung können lediglich die Intensität, die Schärfe oder den zeitlichen Verlauf der visuellen Repräsentation beeinflussen, nicht jedoch die Position einer Aktivierung. Zwar könnte mit der Nutzung der Bewegung als einzigem Merkmal der aus dem auditorischen System bekannte Zusammenhang zwischen dem Auftreten eines lokalisierten Stimulus und dem Vorhandensein eines Objektes hergestellt und auch das Verhalten des visuellen Modells als Ortung interpretiert werden. Es ist aber offensichtlich, dass die Richtung der stärksten Intensitätsänderung in einer Bildfolge auch schon die Position des Maximums in der Bewegungskarte bestimmt und die bloße topographische Projektion keine zu evaluierende Lokalisationsleistung darstellt.

Zur Beschreibung der frühen und objektunspezifischen, visuellen oder multisensorischen Mechanismen ist die Frage nach korrekten oder falschen Ortungsergebnissen demnach ungeeignet. Sinnvoller erscheint es, in Analogie zu den zitierten Wahrnehmungsexperimenten und elektrophysiologischen Ableitungen auch den Simulationsmodellen mehr oder weniger korrelierte audio-visuelle Stimuli zu präsentieren und die künstliche sensorische Abbildung an der vorgegebenen Zielposition auszuwerten.

## 5.2 Szenarien, Experimente und Benchmarks

### 5.2.1 Datenbank für audio–visuelle Szenarien

Eine erste grundlegende Frage zur Bestimmung einer adäquaten experimentellen Strategie betrifft die Entscheidung, ob Echtzeit– oder offline Versuche realisiert werden sollen. Sowohl für den angestrebten Vergleich zwischen multisensorischem WTA–Netz und Bayesfilter als auch zur Evaluierung verschiedener Parametrisierungen der implementierten Modelle müssen die sensorischen Inputs reproduzierbar sein. Für diesen Zweck erschienen offline Simulationen auf Basis von Audio– und Videoaufnahmen besser geeignet als reale Einzelexperimente mit einem technischen Demonstrator (vergl. Abschnitt 5.3). Ein weiteres typisches Problem, das es unabhängig vom Klassifikations– oder Aufmerksamkeitskontext des untersuchten Modellverhaltens zu lösen gilt, stellt die Wahl zwischen realen oder künstlichen Stimuli dar. Einerseits sind die Variationsmöglichkeiten und die Reproduzierbarkeit realer sensorischer Inputs begrenzt – andererseits geben simulierte, audio–visuelle Szenen die räumlich–zeitliche Komplexität realer Reize nicht wieder. In Anbetracht dieses Dilemmas musste ein angemessener Kompromiss gefunden werden.

Als universelles Instrument zur Bewertung der Modellvarianten wurde schließlich ein eigener experimenteller Ansatz entwickelt, der die Vorteile realer audio–visueller Szenen mit den Gestaltungsmöglichkeiten virtueller Experimente verbindet. Die Grundlage der folgenden Versuche bilden separate Audio– und Videoaufnahmen von elementaren akustischen und optischen Stimuli, die in verschiedenen Variationen dargeboten wurden. Um diversen akustischen Effekten Rechnung zu tragen, die zur Beeinflussung der laufzeitbasierten, berechneten Abbildung im auditorischen System führen, erfolgte die Beschallung der Stereomikrofonanordnung entsprechend der topographischen Konfiguration der Modelle aus mehreren Richtungen. Auf diese Weise entstand mit einem überschaubaren Aufwand eine Datenbank mit kurzen Bildfolgen und Audiosequenzen, die nach der Art der visuellen oder akustischen Ereignisse und nach Instanzen der variierten Repräsentation geordnet vorliegen. Unter den Audioaufnahmen werden außerdem Instanzen der Geräusche für verschiedene Richtungen unterschieden.

Nach vorzugebenden Regeln lassen sich die elementaren akustischen und optischen Ereignisse zu sinnvollen, einfachen oder komplexen Szenen mit korrelierten oder konkurrierenden Stimuluskombinationen zusammenfügen. Die Komposition audio–visueller Szenen kann aber nicht nur für einzelne Experimente erfolgen, sondern in

einer automatischen und scriptbasierten Form auch umfangreiche Benchmarks mit reproduzierbaren offline Simulationen realisieren. In einem solchen Benchmark-Konzept sollen für die weitere Diskussion folgende Begriffe Verwendung finden:

- Das **Szenario** beschreibt die Gesamtheit der elementaren akustischen und optischen Ereignisse der Datenbank und die Regeln zu deren Kombination.
- Als **Szene** wird eine konkrete, audio-visuelle Realisierung bezeichnet, die ein oder mehrere Ereignisse beinhalten kann.
- Unter dem Begriff **Benchmark** ist schließlich eine repräsentative und reproduzierbare Folge von Szenen zur Evaluierung einer Modellvariante oder einer Parametrisierung zu verstehen.

In einer einfachen, scriptbasierten Implementierung prüft ein Szenengenerator anhand der Regeln, mit denen ein Szenario definiert wurde, die Gültigkeit eines Szenenkommandos, löst bei Bedarf Joker-Parameter für Instanzen, Winkel und Zeitangaben auf und erzeugt die Bildfolge und das Stereosignal für eine Szene (s. Anhang C).

Das Konzept von virtuellen Experimenten auf der Grundlage gespeicherter, realer akustischer und optischer Reize lässt einen großen Spielraum bei der Gestaltung der Aufnahmen. Die elementaren Stimulusformen eines Szenarios können beliebig einfach oder komplex sein – eine Orientierung an den zitierten neurobiologischen Experimenten ist ebenso vorstellbar wie die Simulation potentieller Situationen beim Einsatz eines technischen Demonstrators. Allein aus dem Verzicht auf die Kodierung objektspezifischer, sensorischer Merkmale können zunächst kaum Anhaltspunkte für ein konkretes audio-visuelles Szenario abgeleitet werden. In Bezug auf die bereits in Abschnitt 3.4 hervorgehobenen Invarianzen bei der Bewegungskodierung kann lediglich an die Unabhängigkeit des visuellen Simulationsmodells gegenüber Farbe, Beleuchtung und Hintergrundstruktur erinnert werden. Um pauschale Fehleinschätzungen der multisensorischen Mechanismen zu vermeiden, sollten darüber hinaus die akustischen Bedingungen so gewählt werden, dass eine sichere und topographisch korrekte Abbildung der Audiokomponente multimodaler Ereignisse garantiert wird.

An dieser Stelle kann nach den experimentellen Szenarien der in Kapitel 4 zitierten audio-visuellen Ansätze gefragt werden. Während die Videosequenzen zur Demonstration der verschiedenen Bayesfilter aufgrund ihrer Orientierung an Klassifikations- und Trackingproblemen nicht in Betracht kommen, erscheint es interessanter, mit welchen sensorischen Daten die audio-visuelle WTA-Architektur RUCcis getestet wurde. In

[RWE00] schildert er einen pragmatisch gestalteten Versuchsaufbau mit einer artifizialen Lampen- und Lautsprecherkonfiguration, die sich in einem sonst dunklen und hallarmen Raum befindet. Wie die zuvor beschriebene Szenendatenbank bietet auch RUCCIS technische Anordnung den Vorteil der leichten Reproduzierbarkeit von akustischen und optischen Ereignissen. Als Schwachpunkt ist allerdings die Erzeugung punktförmiger Lichtreize anzusehen. Zwar könnte man die Intensitätsänderung durch die Bewegung eines Zieles mit dem Aufleuchten oder Erlöschen einer Lampe vergleichen, die frühe auditorisch-visuelle Fusion basiert jedoch auf einer nur unscharfen topographischen Kodierung, in der visuelle Stimuli eine gewisse räumliche Ausdehnung besitzen müssen, um berücksichtigt zu werden. Diesem Umstand tragen auch viele Wahrnehmungsexperimente Rechnung, indem anstelle punktförmiger Reize streifenartige Stimuli bewegt werden. Allerdings sind die im Grunde aufschlussreichen Informationen zu den experimentellen Setups von MEREDITH, STEIN oder WALLACE wiederum nicht detailliert genug, um etwa deren konkrete Versuche und Befunde akribisch und quantitativ exakt nachzuvollziehen (vergl. [MS86, SM93, WWS96]). Anders als bei RUCCIS Arbeiten an einer „*Robotic Barn Owl*“ sollen hier ja auch nicht die Wahrnehmungsleistung eines spezifischen Versuchstieres simuliert werden.

Vor dem Hintergrund der möglichst allgemeinen Anwendbarkeit der Ergebnisse muss letztendlich auch angezweifelt werden, ob ein stark vereinfachtes und artifizielles Setup die räumliche und zeitliche Dynamik realer Szenen abbilden kann. Die prinzipiellen Befunde zur audio-visuellen Fusion im SC, die als Vorbild und Motivation des gesamten Modellkonzeptes angeführt wurden, spiegeln sich nachweislich auch in den motorischen Reaktionen auf komplexe sensorische Situationen wider [CVWMVO02]. Es ist daher legitim, die Geräusche und elementaren, visuellen Bewegungsmuster einer Szenariendatenbank schon in Hinblick auf eine potentielle Anwendung auszuwählen. Für die folgenden Untersuchungen wurde ein hypothetisches Szenario der Mensch-Maschine Interaktion mit visuellen Gesten, Geräuschen und Lautäußerungen von potentiellen Nutzern eines technischen Systems in einer Datenbasis repräsentiert. Um den Aufwand bei der Erstellung der Datenbank zu begrenzen, ließ sich das Repertoire an akustischen und optischen Ereignissen durch plausible Randbedingungen des Szenarios konkretisieren. So sollten die sensorischen Ereignisse von Personen verursacht werden, die sich etwa in derselben horizontalen Ebene mit der Aufnahmeeinrichtung befinden und höchstens einige Meter weit entfernt sind. Bei den visuellen Bewegungsmustern lassen sich eher statische Gesten von stehenden Personen und eher dynamische Teilszenen unterscheiden, in denen Personen an der Kamera vorübergehen, sich auf sie zu oder von ihr weg bewegen. Zu den akustischen Ereignissen zählen in erster Linie kur-

ze Sprachsignale, aber auch andere Geräusche wie beispielsweise ein Händeklatschen. Die Realisierung relativ natürlicher Gesten und Bewegungsabläufe stellt weiterhin eine gewisse Variabilität in der Art der räumlichen und zeitlichen Korreliertheit der audio-visuellen Szenen sicher. Je nachdem wie ausladend eine Geste von der einen oder anderen Person ausgeführt wird, umso mehr oder weniger gut werden die topographischen Kodierungen der Bewegungskarte und der Schallortung korrespondieren. Bewegungen dauern oft noch an, obwohl die kurzen Sprachsignale schon verstummt sind. Im Gegensatz dazu treten bei einem Händeklatschen die visuellen Bewegungsmerkmale schon vor dem akustischen Ereignis auf. In Anlehnung an die Befunde über multisensorisch relevante Zeitfenster von bis zu 500ms und die Versuchsdauer der zitierten neurophysiologischen Ableitungen zur Quantifizierung der neuronalen Response Enhancement betrug die Dauer der hier realisierten, elementaren Teilszenen generell ein bis zwei Sekunden. Die exemplarische Datenbasis beinhaltet Aufnahmen von fünf Personen, die nach eigenem Ermessen jeweils 14 Gesten und Bewegungen und 11 Geräusche (gesprochene Wörter und Händeklatschen) darboten (vergl. Anhang C). Es soll davon ausgegangen werden, dass auf Grundlage dieses sensorischen Datenmaterials belastbare statistische Aussagen möglich sind.



Abbildung 5.1: Veranschaulichung einer exemplarischen Szene mit zwei audio-visuellen und einem nur visuellen Ereignis. Die Details zum Aufnahme-Setup, zur Szenariodefinition und zur scriptbasierten Generierung konkreter Szenen werden in Anhang C beschrieben.

## 5.2.2 Kriterien zur Evaluierung

Die beschriebene Szenariendatenbank kann als erster von zwei Aspekten des Benchmarkkonzeptes angesehen werden. Gelingt es, die geometrischen und dynamischen Merkmale eines Szenarios in der Datenbasis angemessen zu repräsentieren, kann prognostiziert werden, dass sich eine konkret parametrisierte Modellvariante mit guten Simulationsergebnissen im Benchmark auch im realen Einsatz bewährt. Nach der Realisierung der Datenbank müssen nun die Kriterien zur Bewertung der Simulationsergebnisse als zweite Komponente des experimentellen Ansatzes bestimmt werden. Mit dem Verzicht auf die Modellierung von höheren kognitiven Wahrnehmungsleistungen

wurde bereits im vorangegangenen Abschnitt 5.1 begründet, warum das Modellverhalten nicht im Tracking- oder Ortungskontext zu interpretieren ist. Stattdessen gilt es, die Befunde über multisensorische Mechanismen im Superior Colliculus in einen quantifizierbaren Bewertungsapparat zu überführen.

Konkrete Formeln zur Berechnung der neuronalen Response Enhancement oder die unter Vorbehalten definierte bimodal-unimodale Differenz der Targetwahrscheinlichkeiten können zu diesem Zweck unmittelbar auf einzelne Simulationsergebnisse angewendet und für umfangreiche Benchmarks gemittelt werden. Der qualitative Zusammenhang zwischen einer geringen Effektivität unimodaler Stimuli und dem korrespondierenden Auftreten einer maximalen multisensorischen Verstärkung lässt sich jedoch weder an einzelnen Experimenten demonstrieren, noch existieren bislang Berechnungsvorschriften zur Quantifizierung des Effektes in einer Folge von Versuchen. Um neben der direkten multisensorischen Verstärkung auch solche abgeleiteten Modelleigenschaften zu bewerten, soll zunächst eine geeignete Darstellung der Simulationsergebnisse eines gesamten Benchmarks gefunden werden. Zur Veranschaulichung des Zusammenhangs zwischen unimodaler Effektivität und multimodaler Verstärkung muss die an einer Zielposition beobachtete Verstärkung im multisensorischen Experiment in Bezug zu den jeweils rein visuellen und rein auditorischen Modellantworten - beispielsweise zu deren Summe - gesetzt werden. In einem entsprechenden Diagramm korrespondieren die Beobachtungen für ein visuell, auditorisch und multisensorisch simuliertes Experiment mit einem einzelnen Punkt, während ein kompletter Benchmark durch einen Scatterplot (eine Punktwolke) repräsentiert wird. Abbildung 5.2 zeigt, wie ein Benchmark-Scatterplot sowohl für die neuronale Response Enhancement (RE) des WTA-Modells als auch für die bimodal-unimodale Differenz (BUD) der Targetwahrscheinlichkeit eines Bayesfilters angegeben werden kann.

Aus der gemeinsamen Berechnungsvorschrift für die RE- und BUD-Werte resultiert ein theoretisch möglicher Bereich, den die Punktwolke eines Benchmarks in den beschriebenen Diagrammen okkupiert. Die Variabilität, die ein Scatterplot innerhalb dieses Bereiches aufweist, wird durch die unterschiedliche Amplitude und die verschieden ausgeprägte räumlich-zeitliche Korreliertheit der auditorischen und visuellen Stimuli verursacht. Außerdem hängen Form und Ausrichtung der erzeugten Punktwolke aber auch vom Algorithmus und von der Parametrisierung des zugrundeliegenden Simulationsmodells ab. Die gesamte experimentelle Evaluierung der Modellvarianten stützt sich auf die These, dass der Scatterplot eines Benchmarks alle angestrebten multisensorischen Eigenschaften repräsentiert [SG03, SG04]. Mit der Darstellung der Benchmarks als Punktwolken stehen schließlich auch diverse mathematische Werk-

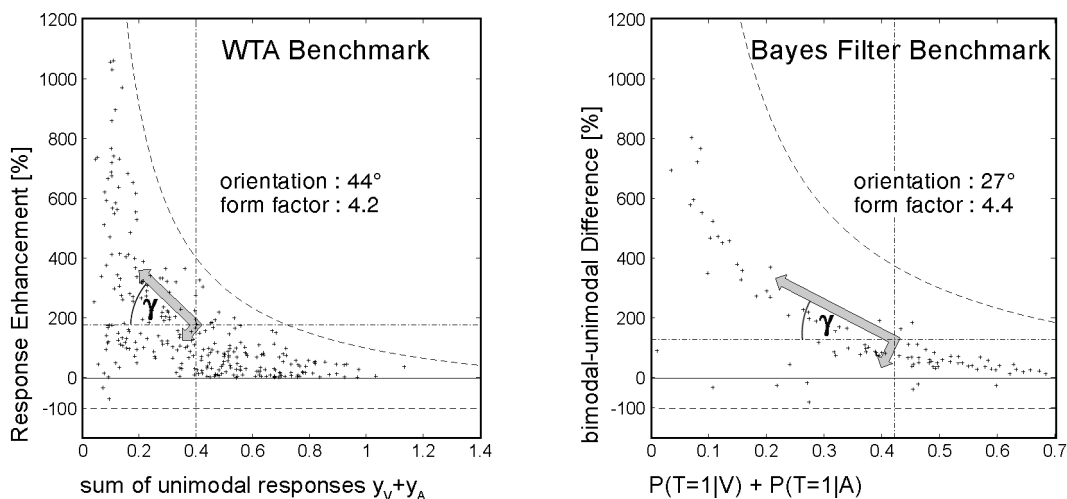


Abbildung 5.2: Exemplarische Benchmarks von WTA-Netz und Bayesfilter in der gleichen formalen Darstellung. Jeder Punkt in den Diagrammen korrespondiert mit einem konkreten Experiment, für das jeweils die Verstärkung bei multisensorischen Stimuli in Bezug zur Summe von normierter auditorischer und visueller Modellantwort gesetzt wurde. Als einziger Unterschied in den Berechnungsgrundlagen wurde die Ausgabe des Bayesfilters nicht wie die des WTA-Netzes für die gesamte Dauer der Szene gemessen, sondern nur in den Zeitabschnitten, in denen tatsächlich multisensorischer Input auftrat (vergl. Abschnitt 4.3.4). Die Darstellung ist im Intervall (0..2) definiert. Gestrichelte Linien markieren die obere und untere Schranke der multisensorischen Verstärkung. Die eingezeichneten Pfeile stellen in Orientierung (Winkel  $\gamma$ ) und Längenverhältnis die Hauptkomponenten der Punktwolken dar.

Links: Summe der normierten unisensorischen Aktivierungen  $y_V + y_A$  und Response Enhancement des WTA-Netzes.

Rechts: Summe der unisensorischen Targetwahrscheinlichkeiten  $P(T=1|V) + P(T=1|A)$  und bimodal-unimodale Differenz des Bayesfilters.

zeuge bereit, um die qualitativen Vorgaben für die Ausprägung der multisensorischen Effekte zu quantifizieren. Dank der identischen Darstellungsweise von RE- und BUD-Eigenschaften können die folgenden Kriterien gleichermaßen zur Evaluierung von WTA-Netz und Bayesfilter angewandt werden:

- Ein **Maximumkriterium** definiert ein plausibles Intervall, in dem die maximalen RE- bzw. BUD-Werte eines Benchmarks liegen sollten.
- Mit Hilfe eines **Mittelwertkriteriums** wird ein Mindestniveau für die durchschnittliche multisensorische Verstärkung aller Experimente festgelegt. Gemeinsam mit dem Maximumkriterium kann damit eingeschätzt werden, ob multimodale Enhancement-Effekte typisch für das Modellverhalten sind, ohne dass andererseits schwer zu motivierende Extremwerte auftreten.

- Aussagen über den Zusammenhang zwischen unisensorischer Modellantwort und multisensorischer Verstärkung sind anhand der Form und der Ausrichtung der Punktwolke eines Benchmarks möglich. Dazu können beispielsweise eine Hauptkomponentenanalyse (PCA) des Scatterplots durchgeführt und dabei die Eigenvektoren der Kovarianzmatrix der Punktwolke bestimmt werden. Ein **Orientierungskriterium** ist dann sinnvoll für den Winkel der Hauptkomponente (korrespondierend zum betragsgrößten Eigenvektor) im Benchmarkdiagramm zu definieren. Abbildung 5.2 verdeutlicht, dass Ausrichtungen zwischen 0 und 90 Grad die geforderte, umgekehrte Proportionalität von unisensorischer Effektivität und multisensorischer Verstärkung beschreiben.
- Neben der richtigen Ausrichtung der Hauptkomponente ist es außerdem wichtig, dass diese die Varianz im Scatterplot tatsächlich auch dominiert. So wäre eine eher runde Punktwolke mit annähernd gleich starken Eigenvektoren bezüglich der zu bewertenden Modelleigenschaften kaum aussagekräftig. Ein **Formkriterium** prüft deshalb auch das Verhältnis der Beträge der Eigenvektoren.

In einem Benchmark mit korrelierten, audio-visuellen Szenen sollte der für die Zielposition ermittelte Scatterplot eines adäquaten multisensorischen Modells eine langgezogene Punktwolke aufweisen, die sich diagonal im positiven Teil des theoretischen Verstärkungsbereiches entfaltet (vergl. Abbildung 5.2). Mit den vier genannten Kriterien für maximale und durchschnittliche Verstärkung sowie Form und Ausrichtung des Scatterplots lässt sich diese verbale Forderung plausibel und objektiv bewerten.

Weitaus problematischer ist die Beurteilung der multisensorischen Response Depression, die in Szenen zu beobachten sein sollte, in denen konkurrierende Stimuluskombinationen auftreten (vergl. Abbildung 4.7). In exemplarischen Versuchsreihen mit dem WTA-Modell konnten multisensorische Depression-Effekte problemlos demonstriert werden [SG04]. In Anbetracht der allgemein formulierten Depression-Befunde [MS86] ist eine konkrete Gestaltung der räumlichen und zeitlichen Merkmale von de-korrelierten Stimuli aber ebenso schwierig wie die Formulierung von quantifizierbaren Kriterien zur Bewertung der Simulationsergebnisse. Weiterführenden Depression-Benchmarks wurden deshalb bislang noch nicht realisiert. Für die Gegenüberstellung von WTA-Netz und Bayesfilter könnten Depression-Kriterien ohnehin keinen Beitrag leisten, da die Auswertung weitreichender topographischer oder zeitlicher Zusammenhänge mit dem vorliegenden probabilistischen Formelapparat nicht möglich ist.

Die angestrebten multisensorischen Merkmale beschreiben das Verhalten der Modellvarianten unter dem Gesichtspunkt der neurobiologischen Motivation. Einige tech-



nische Randbedingungen und spezifische Eigenschaften der WTA-Architektur und des Bayesschen Algorithmus bleiben in den bislang definierten Kriterien jedoch unberücksichtigt. Beispielsweise wäre es interessant, in welchen typischen Arbeitspunkten die simulierten Neurone in den WTA-Prozessen betrieben werden oder wie sicher die Schätzung des Bayesfilters über die Anwesenheit eines sensorischen Zieles ist. Im benchmarkbasierten experimentellen Ansatz liegt es nahe, auch solche modellbezogenen Fragestellungen zu untersuchen. Gerade angesichts der schwierigen analytischen Behandlung der vorgeschlagenen Algorithmen bieten zwei weitere spezifische Kriterien eine gute Alternative, um die reguläre Arbeitsweise von WTA-Netz und Bayesfilter zu überprüfen.

- **WTA-Kriterium:** Die Beurteilung der WTA-Prozesse kann schon ohne die Auswertung spezieller Diagramme erfolgen. Bereits ein einfaches Mittelwertkriterium für die multimodale Aktivierung (VA-Response) hilft beim Auffinden geeigneter Parametrisierungen. Mit einer oberen und unteren Schranke für zulässige Aktivierungen der Gewinnerregion werden zu stark gesättigte oder zu unterschwellige WTA-Zustände unterbunden.
- **Bayes-Kriterium:** Anstelle der multimodalen Aktivierung ist für das Verhalten des probabilistischen Modells die bedingte multisensorische Targetwahrscheinlichkeit  $P(T=1|V,A)$  ausschlaggebend. Das Bayesfilter sollte die Anwesenheit eines Zieles durch möglichst hohe Wahrscheinlichkeitswerte anzeigen, weshalb nicht ein Intervall, sondern ein unterer Grenzwert für  $P(T=1|V,A)$  vorgeschlagen wird.

### 5.2.3 Benchmarkbasierte Optimierung

Bevor ein aussagekräftiger, quantitativer Vergleich zwischen WTA-Netz und probabilistischer Modellvariante vorgenommen werden kann, gilt es zunächst, geeignete Wertebereiche für eine Vielzahl von Modellparametern zu finden. Um den Freiheitsgrad für denkbare Parametervariationen zu begrenzen, können in beiden Modellansätzen einige Annahmen über angemessene Randbedingungen getroffen werden. Im WTA-Netz lassen sich Breite und Gestalt der Feedback-Verschaltung an die Geometrie der rezeptiven Felder und die zu erwartenden Stimulusformen anpassen. Ebenso erscheint es sinnvoll, die Zeitkonstante der dynamischen Amari-Neurone entsprechend der angestrebten, multisensorisch sensitiven Zeitfenster vorzugeben. Die pauschale Wichtung der sensorischen Eingänge ist unkritisch so zu wählen, dass typische Bewegungen und Geräusche

sich ähnlich stark in den topographischen Karten abbilden und tatsächlich auch WTA-Prozesse in Gang setzen (vergl. Abbildung 4.8). Für eine Reihe weiterer Parameter wie die Wichtung der global inhibitorischen und lokal exzitatorischen Rekurrenzen oder den Anstieg der Ausgangssigmoide der Neurone sind kaum plausible Vorgaben möglich. Sie bestimmen die Dimension eines zu optimierenden Parameterraumes und somit die Zahl der notwendigen Variationen und Benchmarks. Eine vergleichbare, experimentelle Strategie wird zur Bewertung und Optimierung des multisensorischen Bayesfilters verfolgt: während die statistischen Eigenschaften der audio-visuellen Stimuli ein adäquates Sensormodell vorgeben, sollen die weniger leicht messbaren Parameter des Bewegungsmodells im Experiment bestimmt werden. Im Rahmen eines exemplarischen Vergleichs wurden WTA-Netz und Bayesfilter nach sorgfältiger Abschätzung des Berechnungsaufwandes in einem jeweils dreidimensionalen Parameterraum der folgenden multikriteriellen Optimierung unterzogen:

- WTA-Parameter
  - globale Inhibition: [0.1, 0.125, 0.15, 0.175, 0.2, 0.25, 0.3]
  - Wichtung des lokalen Feedbacks: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]
  - Sigmaparameter der neuronalen Ausgabefunktion: [4, 5, 6, 7, 8]
- Bayes-Parameter
  - Grundbetrag in der Wahrscheinlichkeitstabelle des Bewegungsmodells für das Erscheinen eines Zieles:  
 $P_0(T_t = 1|T_{t-1}) = [0.001, 0.002, 0.005, 0.01, 0.02, 0.04, 0.1]$
  - entsprechender Grundbetrag für das Verschwinden eines Zieles:  
 $P_0(T_t = 0|T_{t-1}) = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$
  - Sigmaparameter des Bewegungsmodells: [0.1, 0.5, 1, 2, 3]
- Multisensorische Kriterien
  - Maximumkriterium:  $500\% < RE_{max}, BUD_{max} < 1500\%$
  - Mittelwertkriterium:  $50\% < RE_{mean}, BUD_{mean}$
  - Orientierung der Hauptkomponente:  $15^\circ < \gamma < 75^\circ$
  - Formkriterium (Beträge der Eigenvektoren):  $Faktor > 2$
- Modellspezifische Kriterien
  - WTA-Kriterium:  $0.1 < \text{mittlere VA-Response} < 0.5$
  - Bayes-Kriterium:  $P(T = 1|V, A) > 0.5$

Für die geschilderte Prozedur mussten beide Modellvarianten mit  $7 \times 7 \times 5 = 245$  verschiedenen Parametrisierungen jeweils einen Benchmark mit einer vorgegebenen Folge von Szenen durchlaufen. Das dabei repräsentierte Szenario beschränkte sich auf etwa 100 lokale sensorische Ereignisse mit unterschiedlich stark korrelierten audio-visuellen Stimuluskombinationen (s. Anhang D). Unter Berücksichtigung der zu erwartenden Streuung in den multisensorischen Eigenschaften (vergl. Abbildung 5.2) sollten in den WTA-Benchmarks weitere Variationen der Stimulusintensität (Lautstärke und Helligkeit) die statistische Aussagekraft erhöhen. Jeder Benchmark eines konkreten Parametersets wurde mit Hilfe der vier allgemeinen, multisensorischen und des zusätzlichen, modellspezifischen Kriteriums bewertet. Parametrisierungen, deren Benchmarkergeb-

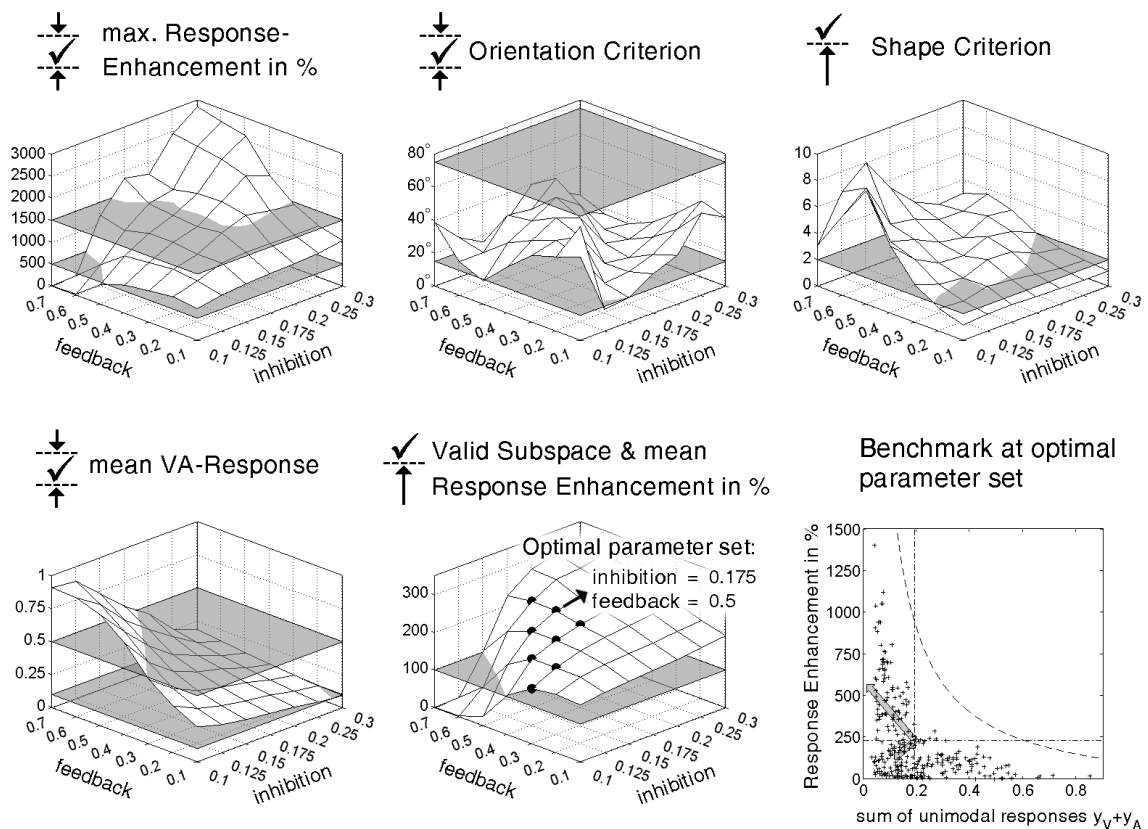


Abbildung 5.3: Ausgewählte Benchmarkergebnisse des WTA-Netzes für Parametervariationen der inhibitorischen und exzitatorischen Feedback-Wichtung. Im Diagramm der durchschnittlichen Response Enhancement (unten, Mitte) wurde der gültige Parameter-Unterraum markiert und das Parameterset [Inhibition=0.175, Feedback=0.6] aufgrund seines hohen, durchschnittlichen RE-Wertes als optimal eingestuft. Mit diesen Einstellungen zeigte das multisensorische WTA-Netz entsprechend der fünf Kriterien zur Evaluierung folgende Eigenschaften:  $RE_{max} = 1402\%$ ,  $RE_{mean} = 233\%$ , *mittlere VA-Response* = 0.33, *Orientierung* =  $48^\circ$ , *Formfaktor* = 2.6.

nisse allen fünf Kriterien genügen, bilden eine gültige Region innerhalb des diskreten Parameterraumes. Unter allen gültigen Parametersets kann pragmatisch dasjenige mit dem höchsten Mittelwert der multisensorischen Verstärkung als optimal angesehen werden. Zur übersichtlicheren Darstellung zeigen die Abbildungen 5.3 und 5.4 zweidimensionale Ausschnitte der Parameterräume von WTA-Netz und Bayesfilter sowie deren quantitative Bewertung. Abbildung 5.5 stellt die ermittelten, gültigen Parameterräume des WTA-basierten und des probabilistischen Simulationsmodells gegenüber. Die vollständige Dokumentation einer dreidimensionalen Parameteroptimierung und weiterer Benchmarks mit variierenden sensorischen Kodierungen bei der visuellen Vorverarbeitung erfolgt in Anhang D.

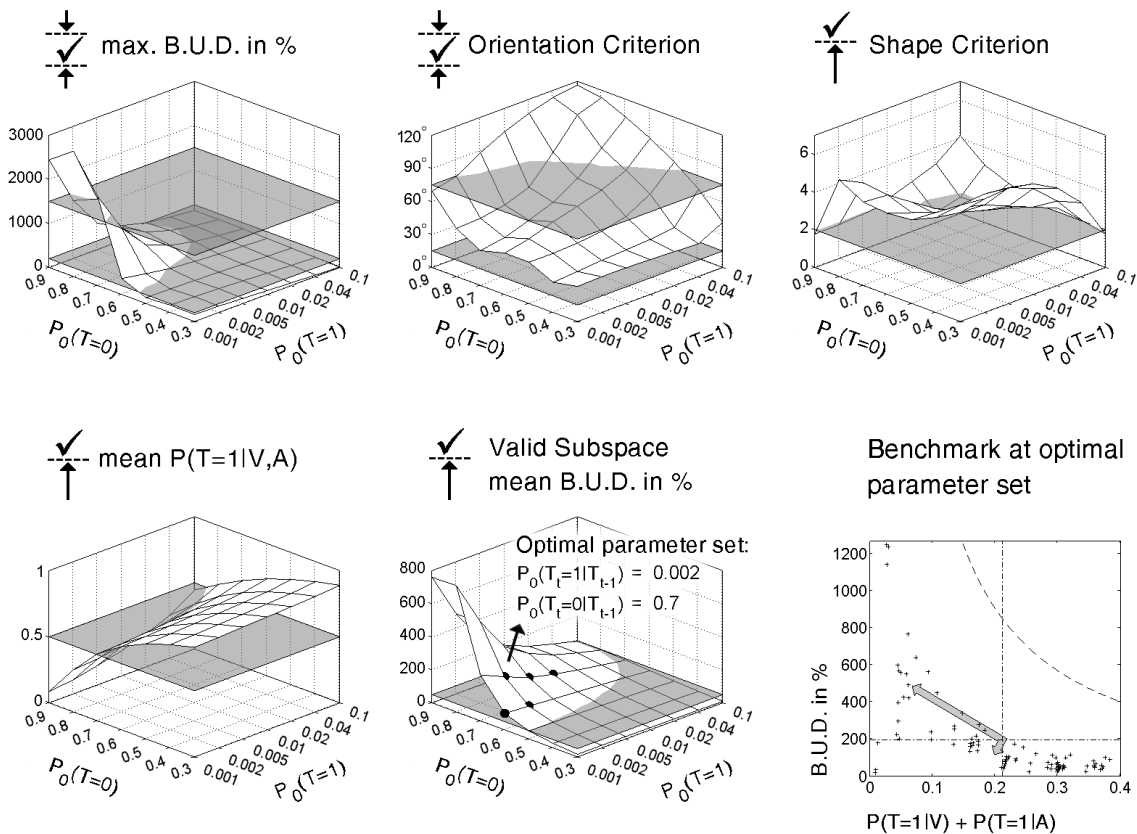


Abbildung 5.4: Benchmarkergebnisse des Bayesfilters für Variationen der Grundbeträge in den Wahrscheinlichkeitstabellen des Bewegungsmodells. Im Diagramm der mittleren bimodal-unimodalen Wahrscheinlichkeitsdifferenz (unten, Mitte) wurde der gültige Parameter-Unterraum markiert und das Parameterset  $[P_0(T_t=1|T_{t-1})=0.002, P_0(T_t=0|T_{t-1})=0.7]$  aufgrund seiner hohen durchschnittlichen BUD-Werte als optimal eingestuft. Mit diesen Einstellungen zeigte das Bayesfilter entsprechend der fünf Kriterien zur Evaluierung die Eigenschaften:  $BUD_{max}=1246\%$ ,  $BUD_{mean}=196\%$ , *mittlere bedingte Targetwahrscheinlichkeit*  $P(T=1|V, A)=0.52$ , *Orientierung* $=35^\circ$ , *Formfaktor* $=6.4$ .

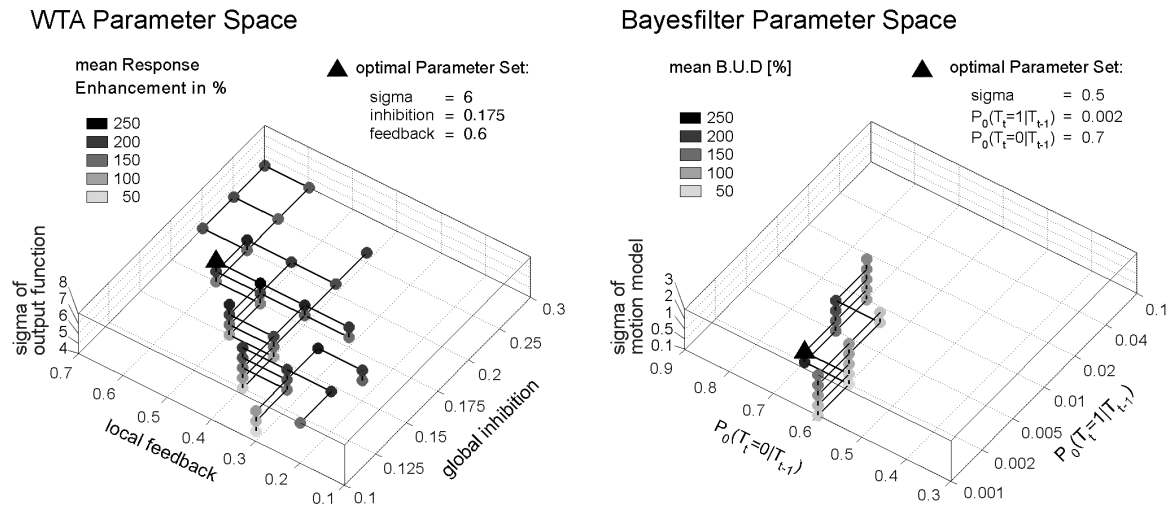


Abbildung 5.5: Gültige Cluster in den diskreten dreidimensionalen Parameterräumen, die bei der Optimierung von WTA-Netz und Bayesfilter ermittelt wurden. Die optimalen Parameter sets wurden innerhalb der gültigen Unterräume anhand der mittleren Response Enhancement bzw. der durchschnittlichen bimodal-unimodalen Wahrscheinlichkeitsdifferenz bestimmt.

## 5.2.4 Interpretation der Simulationsergebnisse

Die in den Abbildungen 5.4 und 5.3 zusammengefassten Benchmarkergebnisse lassen erkennen, dass mit den exemplarischen Versuchen aus Kapitel 4 keine zufälligen Effekte, sondern tatsächlich typische Modelleigenschaften demonstriert wurden. Unter dem Vorbehalt der unterschiedlichen Interpretation von Response Enhancement und bimodal-unimodaler Wahrscheinlichkeitsdifferenz können die grundlegenden Befunde und vermuteten Gesetzmäßigkeiten der multisensorischen Integration mit beiden untersuchten Modellvarianten nachvollzogen und durch belastbare statistische Aussagen untermauert werden. Der benchmarkbasierte, experimentelle Ansatz ermöglichte es insbesondere, plausible Maxima und Durchschnittswerte der multisensorischen Verstärkung sowie die umgekehrte Proportionalität von bimodaler Verstärkung und unimodaler Effektivität der Stimuli nachzuweisen. Die Diagramme der zu diesem Zweck eingeführten Kriterien zeigen eine über große Parameterbereiche hinweg homogene und stetige Bewertung der Simulationen, woraus auf eine prinzipielle Robustheit und Praxistauglichkeit beider Verfahren geschlossen werden kann.

Die ermittelten, gültigen Parameterräume, in denen alle multisensorischen und modellspezifischen Anforderungen erfüllt werden, bilden für sämtliche untersuchten Modellvarianten mehr oder weniger kompakte Cluster (s. Abbildung 5.5 und Anhang D). Die zusammenhängenden, gültigen Bereiche in den Parameterräumen und

die stetige Entwicklung der durchschnittlichen multisensorischen Verstärkung als Auswahlkriterium für ein „bestes“ Parameterset belegen die Zuverlässigkeit und Aussagekraft des Optimierungsverfahrens.

#### *WTA Modell*

Wie nicht anders zu erwarten war, nimmt die Ausprägung der WTA-Prozesse und deren Einfluss auf das Modellverhalten mit wachsenden Wichtungen der inhibitorischen und exzitatorischen Rückkopplungen zu (s. Abbildung 5.3). Besonders günstige, multisensorische Eigenschaften werden erzielt, wenn lokales Feedback und globale Inhibition in einem ausgewogenen Verhältnis stehen. Das typische Cluster der gültigen WTA-Parametrisierungen liegt deshalb diagonal im getesteten Parameterraum (s. Abbildung 5.5 und Anhang D). Ferner bewirkt auch eine steilere sigmoide Ausgabefunktion der WTA-Zellen eine zunehmend nichtlineare Übertragung und höhere Enhancement-Werte.

Die großen Intervalle im Parameterraum, in dem die angestrebten multisensorischen Eigenschaften zu beobachten sind, werden auch durch das WTA-spezifische Kriterium zur Sicherstellung regulärer Arbeitsbereiche kaum weiter eingeschränkt. Lediglich sehr hohe Feedback-Wichtungen führen zur Sättigung der Netzaktivierung, wenn gleichzeitig die Inhibition als Antagonist im WTA-Prozess zu schwach ist. In den durchgeführten Versuchsreihen verhielt sich das WTA-Modell aufgrund seiner massiv rekurrenten Komponenten weniger deterministisch als der simulierte Bayesfilter und zeigte eine stärkere Streuung in den Benchmark-Scatterplots. Dies stellt allerdings nicht zwingend einen Nachteil dar, denn die Streuung der Punktwolke einer Versuchsreihe ist Voraussetzung zur Evaluierung und sinnvollen Anwendung der beiden PCA-basierten Kriterien.

#### *Probabilistisches Modell*

Zur Beurteilung des Bayesfilters erwies sich eine Auswertung der Modellantwort über die komplette Szenendauer von ein bis zwei Sekunden als ungeeignet. Die pauschale bimodal-unimodale Differenz (BUD) von nahezu minus 100 Prozent, mit der das Simulationsmodell auf sehr geringe Geräuschpegel oder ausbleibende visuelle Bewegungsmuster reagiert, würde die verhältnismäßig kurzen Enhancement-Effekte im zeitlichen Mittel überdecken. In die Ermittlung des Modellausganges für ein Experiment gingen deshalb nur Zeitintervalle ein, in denen sowohl der Audiopegel als auch die Amplitude der visuellen Bewegungskodierung erkennbar über dem Niveau des Sensorauschens lagen.

Unter dieser Voraussetzung kann WTA-Netz und Bayesfilter ein durchaus ähnliches Modellverhalten bescheinigt werden. Korrespondierend zur Wirkung von Feedback und Inhibition des künstlichen neuronalen Netzes können die Parameter des probabilistischen Bewegungsmodells  $P_0(T=0)$  und  $P_0(T=1)$  interpretiert werden. Die deutlichste Verstärkung der multisensorischen Targetwahrscheinlichkeit zeigt das Bayesfilter, wenn es von einer eher pessimistischen Annahme über das Auftreten von Zielen ausgeht. Werden geringe Wahrscheinlichkeiten für das Einsetzen neuer Stimuli und gleichzeitig hohe Werte für das Verschwinden von Zielen vorgegeben, kann sich der Informationsgewinn durch die Kombination von visueller und auditorischer Umweltrepräsentation nachhaltiger im Bayesschen Wahrscheinlichkeits-Update niederschlagen. In sämtlichen Benchmarks traten bei gegebenem Sensormodell überhaupt nur positive BUD-Werte auf, wenn die Grundbeträge in der Wahrscheinlichkeitstabelle des Bewegungsmodells  $P_0(T_t=1|T_{t-1}) \leq 0.02$  und  $P_0(T_t=0|T_{t-1}) \geq 0.6$  betragen. Während derartige Parametrisierungen gute multisensorische Eigenschaften zeigen, offenbart die Bewertung durch das spezifische Bayeskriterium ein erstes Problem. Gerade wenn die stärksten Enhancement-Effekte zu beobachten waren, blieb die resultierende multisensorische Targetwahrscheinlichkeit oft auf einem sehr niedrigen Niveau (s. Abbildung 5.4). Da die Aussagekraft von unsicheren Bayesschen Schätzungen gerade vor dem Hintergrund einer abstrakten, prämotorischen Kodierung kritisch zu bewerten ist, wurden solche Bereiche durch die Bedingung  $P(T=1|V, A) > 0.5$  ausgeschlossen.

Der Sigmaparameter des Bewegungsmodells, durch den die lateralen Abhängigkeiten in der Bewegungskodierung mehr oder weniger stark in die Berechnung einbezogen werden sollen, wirkt sich in den Benchmarks sichtbar, jedoch nicht dramatisch aus. Bei konstanten übrigen Parametern erzielte die breiteste nachbarschaftliche Wichtung (mit den höchsten  $\sigma$ -Werten) stets die besten multisensorischen Eigenschaften. Der lokale räumliche Ausschnitt, der im Bewegungsmodell berücksichtigt werden kann, ist aber durch den mit dem Radius exponentiell steigenden numerischen Aufwand bei der Simulation begrenzt. In Verbindung mit der unscharfen topographischen Abbildung infolge der notwendigerweise großen rezeptiven Felder kommen die Vorteile des komplizierten Bewegungsmodells möglicherweise noch nicht voll zur Geltung.

#### *Einfluss der visuellen Vorverarbeitung*

Im Anhang D sind die Benchmarkergebnisse von WTA-Netz und Bayesfilter für jeweils zwei Varianten der visuellen sensorischen Kodierung dargestellt. Zur Vorverarbeitung der Bildfolgen sind die primäre Bewegungskodierung durch die Spaltensummen der Differenzbilder und die topographische Transformation entsprechend der

Geometrie der rezeptiven Felder unumgänglich. Über diese minimale Konfiguration hinaus wurden in Kapitel 3 eine zeitliche Filterfunktion und die sigmoide Limitierung der Stimulusintensität als optionale Verarbeitungsschritte vorgeschlagen. In den bisherigen Versuchsreihen mit räumlich und zeitlich mehr oder weniger stark korrelierten audio-visuellen Stimuluskombinationen führte die einfache topographische Bewegungskodierung ebenso wie die zeitlich und dynamisch gefilterte Variante zu plausiblen Resultaten.

Prinzipiell muss im Falle einer aufwendigeren visuellen Vorverarbeitung beachtet werden, dass eine zu strikte Limitierung oder Normierung der Stimulusamplitude, etwa durch eine sigmoide Übertragungsfunktion, den Spielraum für die multisensorische Verstärkung einschränkt. Um die gewünschten Gesetzmäßigkeiten der auditorisch-

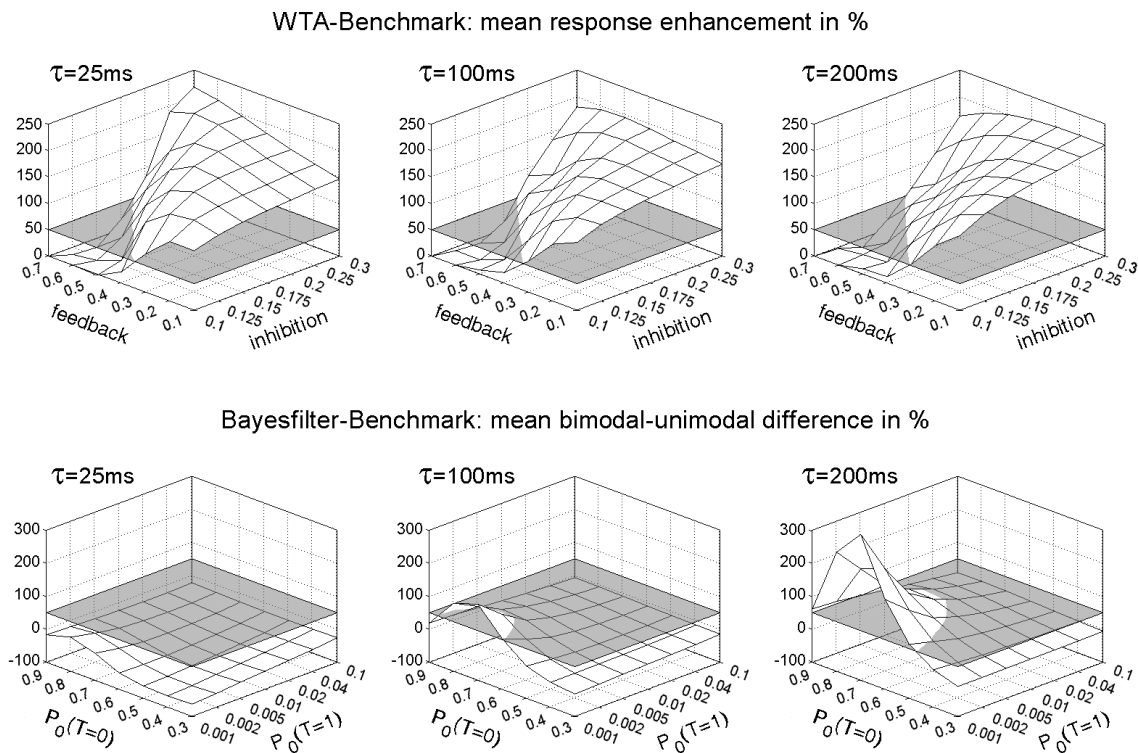


Abbildung 5.6: Benchmarkergebnisse von WTA-Netz und Bayesfilter für die gleichen Parametervariationen wie in den Abbildungen 5.3 und 5.4, jedoch mit unterschiedlichen Zeitkonstanten der Bewegungskodierung (vergl. Abschnitt 3.4.2). Um den Einfluss der visuellen Vorverarbeitung auf die Realisierung multisensorisch sensitiver Zeitfenster hervorzuheben, setzte in den Szenen der Benchmarks das akustische Ereignis erst nach dem Ende der visuellen Gesten mit einer Verzögerung von maximal 300ms ein. Die Ausgänge beider Modellvarianten wurden für die Dauer der Geräusche bewertet. Anders als das WTA-Netz (oben) kann das für Enhancement-Effekte optimierte Bayesfilter (unten) ohne eine geeignete Kodierung des visuellen Eingangs keine eigenen multisensorischen Zeitfenster generieren.



visuellen Integration demonstrieren zu können, müssen WTA-Netz und Bayesfilter gleichermaßen sensorische Inputs erhalten, die eine gewisse Variabilität und Unsicherheit aufweisen. Gerade angesichts des eher binären und selektiven Charakters der vorausgegangenen WTA-Filterung bei der auditorischen Kodierung sollte eine sigmoide Ausgabefunktion im visuellen Teilmodell nicht zu steil eingestellt werden.

Ob die in Abschnitt 3.4.2 vorgeschlagene zeitliche Kodierung der visuellen Bewegungsmuster die Realisierung multisensorischer Zeitfenster begünstigt, lässt sich schließlich nur in einem spezifischen Szenario mit angemessenen Pausen zwischen visuellen und akustischen Ereignissen zeigen. Abbildung 5.6 stellt abschließend entsprechende Benchmarks des WTA-Netzes und des Bayesfilters gegenüber, in denen die Geräusche erst am Ende oder kurz nach den visuellen Gesten zu hören waren. Neben den schon beschriebenen Parametervariationen in den multisensorischen Modellen wurde der visuelle Input mit einer zeitlichen Filterfunktion  $\frac{t}{\tau}e^{1-\frac{t}{\tau}}$  und verschiedenen Parametern  $\tau$  beaufschlagt. Offensichtlich können die maximal 300ms langen Pausen zwischen visuellen und auditorischen Stimuli im rekurrenten WTA-Netz mit Hilfe der Hysterese in der topographischen Abbildung unabhängig von der zeitlichen Dynamik der visuellen Kodierung überbrückt werden. Anders reagiert der für Enhancement-Effekte optimierte Bayesfilter. Seine Targetwahrscheinlichkeit sinkt am Stimulusende sehr schnell und verhindert, dass längere zeitliche Zusammenhänge ausgewertet werden. Positive BUD-Werte sind demnach nur möglich, wenn eine hinreichend träge visuelle Vorverarbeitung die Präsenz multisensorischer Inputs künstlich verlängert.

### 5.3 Hardware Realtime–Demonstrator

#### Interpretation der multisensorischen Karte als prämotorische Kodierung

Nachdem die Topographie, die multisensorischen Merkmale und die zeitliche Dynamik der simulierten Modellvarianten untersucht wurden, soll abschließend auf einige offene Fragen zur Interpretation der Modellantworten als prämotorische Kodierung eingegangen werden. Während die Befunde über die sensorischen Repräsentationen im Superior Colliculus klare funktionelle Vorgaben und plausible Evaluierungskriterien lieferten, bleiben einige Aspekte bei der Integration von Motorkommandos in ein neurobiologisch plausibles Modellkonzept problematisch. Allein die Größe der visuellen und auditorischen rezeptiven Felder und die resultierende Unschärfe der topographischen Abbildung lassen Raum für Spekulationen. Im einfachsten Fall kann pragmatisch ein Schwellwert definiert werden, bei dessen Überschreitung die Position des Maximums in der Modellantwort das motorische Ziel bestimmt. Ob die Generierung von Zielpositionen im Superior Colliculus aber tatsächlich nur auf solchen lokalen Mechanismen beruht oder ob darüber hinaus eine verteilte Kodierung ausgewertet wird, ist bislang unklar. Selbst eine räumlich–zeitliche Analyse der multisensorischen Reizmuster könnte von Nutzen sein, um durch eine Extrapolation des motorischen Zieles in die Bewegungsrichtung des Stimulus angemessen auf schnelle und andauernde Bewegungen zu reagieren<sup>1</sup>.

Abgesehen von der soeben beschriebenen, trivial oder aufwendig zu gestaltenden Ermittlung der motorischen Zielposition gibt es schließlich auch einige diskussionswürdige und grundlegende Unterschiede in der Art und Bedeutung verschiedener Motorkommandos, auf die selbst im Rahmen der neurologischen und wahrnehmungspsychologischen Untersuchungen selten eingegangen wird. Wie die meisten ihrer Kollegen sprechen die Biologen STEIN und MEREDITH in ihren Veröffentlichungen meist pauschal von einem Orientierungs– oder Zuwendungsverhalten, das durch die Mechanismen im Superior Colliculus unterstützt werde [MS86, WS94, Ste98]. Angesichts der Charakterisierung der frühen Wahrnehmungsleistungen als schnelle und unbewusste Verarbeitung ohne den Umweg der kortikalen, symbolischen Interpretation sollten neben der Zuwendung zum Stimulus aber unbedingt auch Fluchtreflexe zu einem elementaren sensomotorischen Repertoire gehören. Dieser naheliegenden Hypothese wurde bisher

---

<sup>1</sup>Ebenso ist es vorstellbar, Bewegungsrichtungen nicht erst in der multisensorischen Karte auszuwerten, sondern bereits bei der primären visuellen und auditorischen Kodierung. Anhaltspunkte für eine entsprechende Erweiterung der in den Kapiteln 2 und 3 beschriebenen Modelle geben beispielsweise [RP95, WWB<sup>+</sup>04] und [WO98, IHM01].

nur in wenigen Arbeiten nachgegangen. Aktivierungen des SC bzw. des Tectum Opticum konnten in Verbindung mit beiden Arten von motorischen Reaktionen unter anderem bei Fischen [HRST98] und beim Hamster [NLS88] nachgewiesen werden. Ein neuronaler Mechanismus, der zwischen Zuwenden oder Flucht entscheidet, wird wahrscheinlich einmal mehr spezifisch für Lebensraum und Lebensweise ausgeprägt sein und außerdem geeignete Stimuluseigenschaften berücksichtigen. Untersuchungen am SC der Ratte zeigten, dass Stimuluspositionen am Boden, die häufiger von Beutetieren hervorgerufen werden, eher eine Orientierung zum Stimulus hin bewirken, während von oben eintreffende Reize meist einen Fluchtreflex vor potentiellen Bedrohungen auslösen [WKR<sup>+</sup>90, SDR86]. Die Demonstration beider sensomotorischer Reaktionen in einem gemeinsamen Simulationsmodell ist unter Umständen schon durch eine einfache Intensitätskodierung möglich: Auf geringe Stimulusamplituden könnte mit einem Zuwenden reagiert werden, auf laute Geräusche und schnelle Bewegungen dagegen mit einem Fluchtreflex. Letztendlich ist auch nicht auszuschließen, dass die Ausgestaltung der prämotorischen Steuerung selbst auf dem frühen neuronalen Niveau des Superior Colliculus über dessen Beeinflussung durch die assoziativen, sensorischen Kortexareale AES und LS an die aktuelle Situation und das momentane Verhalten eines Individuums angepasst wird.

Ein weiterer Problemkreis, der sich als Gegenstand der sensomotorischen Modellierung anbietet, ist schließlich die Koordination von Kopfbewegungen und Augensakkaden unter Aufrechterhaltung der auditorisch-visuellen, topographischen Register. Auch dabei können komplexe Aufgaben wie die Kalibrierung und Adaption von Änderungen in der sensorischen Geometrie formuliert werden [RWE00]. Sind die geometrischen Eigenschaften einer technischen sensorischen Konfiguration bekannt und unveränderlich, wäre jedoch ebenso eine pragmatische Lösung wie die in Abschnitt 4.3.1 vorgeschlagene topographische Schablone für die IC-SC Projektion angemessen.

In der vorliegenden Arbeit wurde der Schwerpunkt bewusst auf die Modellierung der sensorischen Repräsentationen gelegt und auf eine konkrete Integration der simulierten auditorisch-visuellen Aufmerksamkeitskarte in komplexere kognitive und sensomotorische Systeme verzichtet. Allein die Realisierung eines umfangreichen Repertoires an motorischen Reaktionen würde ein neues Kapitel der Recherche neurobiologischer Befunde und der Diskussion von technischen Randbedingungen aufschlagen. Um den abstrakten Benchmarks aus dem vorangegangenen Abschnitt 5.2.3 eine anschaulichere Demonstration der Simulationsergebnisse gegenüberzustellen, war es dennoch wünschenswert, zumindest eine einfache technische Anwendung zu realisieren, die unmittelbar auf akustische und optische Ereignisse in ihrer Umgebung reagiert.

## Implementierung als Roboter-Demonstrator

Im Gegensatz zu den virtuellen offline Experimenten zur Optimierung und vergleichenden Untersuchung der sensorischen Modellvarianten hatte die ingenieurtechnische Umsetzung eines einfachen motorischen Regimes eine echtzeitfähige Hardwarelösung zum Ziel. Dabei sollten keine methodischen Aufgabenstellungen, sondern vielmehr die Frage der Praxistauglichkeit von ausgewählten Algorithmen im Mittelpunkt stehen. Immerhin bedeutet die anwendungsgerechte Implementierung wesentlicher Komponenten des offline getesteten Simulationsmodells nicht weniger als eine effiziente Sequentialisierung des von Natur aus parallelen Konzeptes früher Wahrnehmungsleistungen. Als anspruchsvolle Randbedingungen wurden zum einen die Verwendung einer einfachen Standard-Hardware und andererseits eine echte online Verarbeitung des sensorischen Datenstroms ohne vorgeschaltete Onset-Filter angestrebt.

In der Audiokomponente des Systems kam die aus Kapitel 2 bekannte sensorische Geometrie mit einem Mikrofonabstand von 15cm und einer Abtastfrequenz von 44,1kHz zum Einsatz. Die auditorische, laufzeitbasierte Richtungsabbildung wurde durch eine Kreuzkorrelation der ungefilterten Stereosignale und eine anschließende WTA-Filterung erzeugt. Die Videoverarbeitung basierte bei einer Frame-Rate von

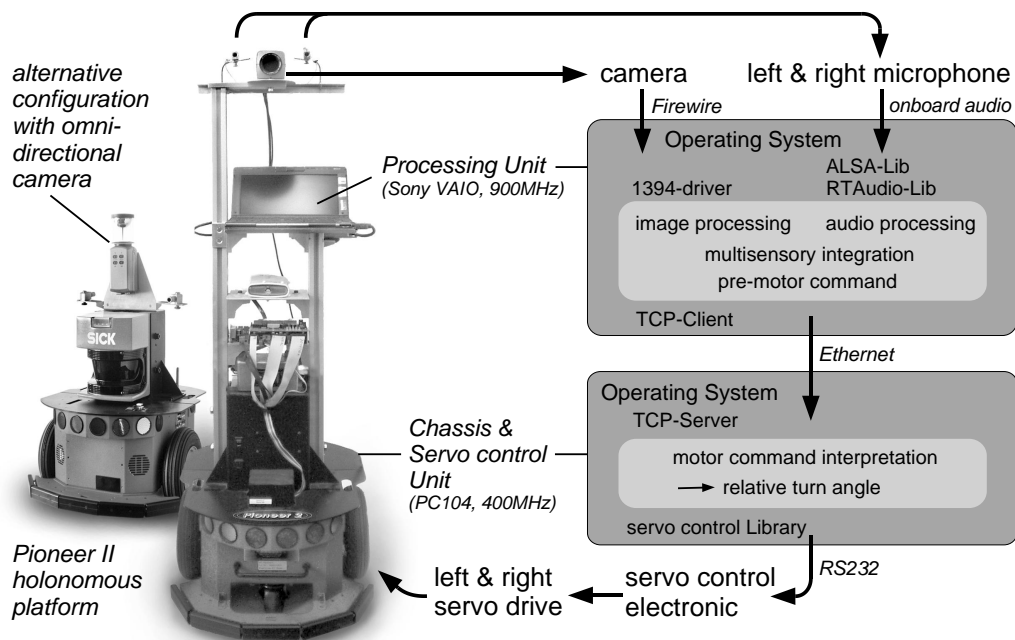


Abbildung 5.7: Technische Realisierung des Demonstrators mit einem Subnotebook-Computer zur Simulation der sensorischen Modelle und einer mobilen Roboterplattform *Pioneer II* zur Ausführung von Drehbewegungen als holonome motorische Reaktion.

30Hz auf 8bit-Grauwertbildern mit einer Auflösung von 160x120 Pixeln. Eine leichte Weitwinklereinstellung der verwendeten Kamera lieferte einen exemplarischen horizontalen Bildausschnitt von ca. 45 Grad. Die primäre visuelle Bewegungskodierung in einer eindimensionalen Richtungskarte erfolgte durch die spaltenweise Summation der Pixel-Differenzen der Bildfolge. Zur multisensorischen Integration diente schließlich eine WTA-basierte Implementierung die mit Iterationsraten von bis zu 645Hz betrieben werden konnte und damit die adäquate Verarbeitung der zeitlichen Dynamik in den visuellen und akustischen Sensordaten garantierte. Die motorische Kodierung wurde in einfacher Weise durch den Vergleich der maximalen WTA-Aktivierung mit einer justierbaren Schwelle realisiert. Ein Überschreiten der Schwelle löste unmittelbar ein Motorkommando mit der zugehörigen topographischen Zielposition aus. Zur Ausführung des Motorkommandos wurde die auf einer mobilen Plattform installierte audio-visuelle Sensorik gemeinsam gedreht. In Anlehnung an die strikte Inhibition des sensorischen Inputs bei motorischen Reaktionen des biologischen Vorbilds wurde das Simulationssystem nach jeder Ausführung einer eigenen Bewegung neu initialisiert. Abbildung 5.7 gibt eine Übersicht zur technischen Realisierung des Demonstrators.

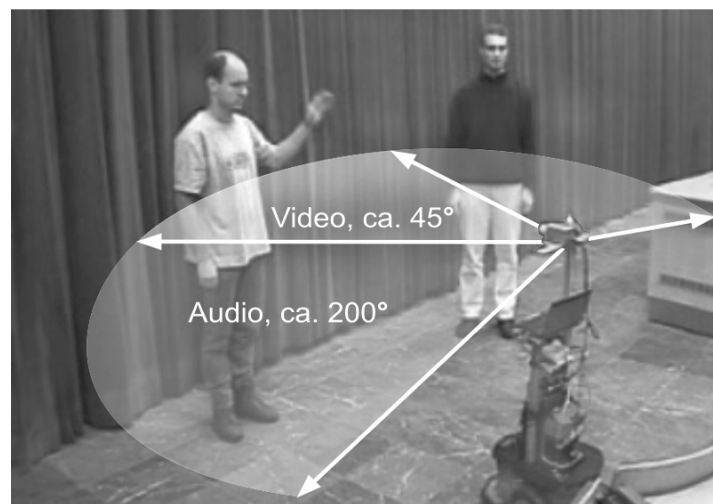


Abbildung 5.8: Veranschaulichung des Demonstrator-Szenarios. In der exemplarischen sensorischen Konfiguration führten Geräuschrictungen, die außerhalb des Bildwinkels der Kamera lagen, zu einer adäquaten Drehbewegung und zur zuverlässigen visuellen Erfassung der Stimulusposition.

In realen Situationen, die dem geschilderten Benchmarkszenario nachempfunden waren (s. Abbildung 5.8), konnte neben der Echtzeitfähigkeit der sensorischen Verarbeitung auch ein augenscheinlich plausibles Modellverhalten nachgewiesen werden. Insbesondere reagierte der Demonstrator mit derselben universellen Parametrisierung

sowohl auf rein akustische und optische, als auch auf multimodale Ereignisse. Dabei schien die bewusst unscharf gehaltene topographische Repräsentation, mit der die biologisch motivierten, multisensorischen Response-Eigenschaften garantiert werden sollten, die Genauigkeit der motorischen Kodierung kaum zu beeinträchtigen. Ohne dass zu dieser Problematik umfangreiche und systematische Untersuchungen vorgesehen waren, konnte beispielsweise demonstriert werden, wie verschiedene akustische Ereignisse, deren Richtung außerhalb des Bildwinkels der Kamera lagen, zu einer angemessenen Drehbewegung und zur zuverlässigen visuellen Erfassung der entsprechenden Position führten.

# Kapitel 6

## Diskussion und Ausblick

### Interdisziplinärer Anspruch

Der enorme Wissenszuwachs, mit dem die Hirnforschung unser Verständnis vieler Wahrnehmungsleistungen erweitert und das wachsende Interesse der Ingenieurwissenschaften an leistungsfähigen, künstlichen sensorischen Systemen gaben den weiten inhaltlichen und methodischen Rahmen dieser Arbeit vor. An der konkreten Aufgabenstellung einer frühestmöglichen, multisensorischen Integration von Sehen und räumlichem Hören konnte beispielhaft demonstriert werden, welche Defizite das in der Informatik etablierte Gleichsetzen von Wahrnehmung und Objekterkennung birgt. Mit dem Anspruch der kognitiven Robotik, immer universellere und natürlicher agierende Systeme zu realisieren, muss die konventionelle Klassifikationsproblematik der Bild- und Audioverarbeitung zwingend um unbewusste und subsymbolische Leistungen erweitert werden. In Bezug auf die räumlichen Aspekte der Wahrnehmung bedeutet der Verzicht auf objekt- oder kontextspezifische sensorische Informationen, dass ein simuliertes Modellverhalten nicht mehr im herkömmlichen Sinne als Ortung zu bewerten ist. Für frühe und unbewusste sensorische Mechanismen, wie das Auslösen motorischer Reflexe oder die primäre Aufmerksamkeitssteuerung, können angemessene Modellvorgaben folglich kaum aus technischen Anwendungsszenarien abgeleitet werden.

Vor diesem Hintergrund sollten die biologischen Befunde nicht nur als allgemeine Motivation für die Aufgabenstellungen der Neuroinformatik verstanden werden. Als qualitativ und quantitativ optimale, evolutionäre Lösungen für elementare Wahrnehmungsprobleme können die natürlichen sensorischen Systeme funktionale und algorithmische Modellkonzepte und gleichzeitig Kriterien zur Evaluierung der simulierten, neuronalen Antwortmuster vorgeben. In einer ingenieurwissenschaftlichen Arbeit erfordert dieser alternative Ansatz eine ungewöhnlich intensive Auseinandersetzung mit

neurologischen und wahrnehmungspsychologischen Befunden. Dabei geht eine Problembeschreibung, die den spezifischen Sichtweisen der Biologie, Informatik und Mathematik gleichzeitig gerecht werden will, unvermeidlich Kompromisse ein: biologische Befunde liefern ein nur lückenhaftes Bild der sensorischen Mechanismen, die künstlichen neuronalen Netze der Neuroinformatik sind analytisch schwer zu beherrschen, und abstrakte mathematische und probabilistische Beschreibungen spiegeln wichtige Aspekte der natürlichen Wahrnehmung nicht wider. Andererseits werden viele Stärken und Schwächen der entweder biologienahen oder mathematisch-technisch orientierten Modelle erst in einem gemeinsamen Kontext deutlich und ermöglichen so die Modifikation und Erweiterung bestehender Ansätze. Beispielhafte Ergebnisse einer interdisziplinären Diskussion von Wahrnehmungsleistungen sind die kritische Unterscheidung zwischen unspezifischen oder objektbezogenen Merkmalen bei der unisensorischen Vorverarbeitung, die Evaluierung und Optimierung der Simulationsmodelle oder die neurobiologisch motivierte Erweiterung des multisensorischen Bayesfilters.

### **Methodik und wissenschaftlicher Beitrag**

Das etablierte Verständnis der Wahrnehmung hat seinen Ursprung im Wesentlichen in der Untersuchung visueller Leistungen und kortikaler Hirnfunktionen. Deutlich wurde dies auch an den im ersten Kapitel zitierten Aufmerksamkeitsmodellen der parallelen und seriellen Suche, der selektiven Bahnung oder der Theorien des Wettbewerbs um neuronale Ressourcen. Beim Versuch, die Ergebnisse einer vorausgegangenen Untersuchung von Schallortungsmodellen [SP99] in komplexere und multisensorische Systeme zu integrieren, warf das von der visuellen Verarbeitung grundverschiedene räumlich-zeitliche Konzept des räumlichen Hörens interessante Fragen auf. Während die retinotopische Umweltrepräsentation beim Sehen sowohl für die räumliche Orientierung als auch für alle formbasierten Klassifikationsaufgaben unabdingbar ist, dient die topographische Abbildung im auditorischen System ausschließlich zur Ortung oder zur Separation von Schallquellen, nicht jedoch zu deren Klassifikation. Die aufwendige Realisierung einer berechneten Topographie schon auf dem Niveau des auditorischen Hirnstammes und die bereits im Mittelhirn zu beobachtende, auditorisch-visuelle Integration sprechen für die große Bedeutung der primären multisensorischen Mechanismen bei der räumlichen Wahrnehmung. Gleichzeitig finden sich solche subkortikalen Leistungen in den bisher meist an der visuellen und symbolischen Verarbeitung orientierten Aufmerksamkeitsmodellen nur sehr bedingt wieder. Den Ausgangspunkt dieser Arbeit bildeten deshalb Überlegungen zum Verständnis von Aufmerksamkeit und Objektbe-



---

griff und der Vorschlag einer konzeptionellen Trennung zwischen unspezifischen und klassifikationsrelevanten sensorischen Informationen. Mit der Konzentration auf frühe neuronale Mechanismen und unspezifische sensorische Merkmale war die grundsätzliche Entscheidung verbunden, auf anwendungsbezogene Modellvorgaben zu verzichten und stattdessen die Simulation biologischer Befunde als Zielstellung und Bewertungskriterium zu verstehen.

In diesem methodischen Rahmen erfolgte im Kapitel 2 zunächst eine Beschreibung der berechneten auditorischen Topographie und eine Bestandsaufnahme binauraler Schallortungsalgorithmen. Im Ergebnis der Diskussion von laufzeit- und intensitätsbasierten Lokalisationsmechanismen und nach einer kritischen Bewertung der neurobiologisch motivierten, sensorischen Kodierungen standen sich ein eigener Korrelationsansatz mit anschließender WTA-Filterung und eine etwa ebenbürtige probabilistische Laufzeitschätzung gegenüber.

Anders als im Falle der auditorischen Modellierung konnte beim folgenden Entwurf einer adäquaten visuellen Vorverarbeitung in Kapitel 3 nicht auf eine Vielzahl etablierter Modellvarianten zurückgegriffen werden. Hier lieferte die Forderung nach einer Kodierung universeller und objektunabhängiger sensorischer Merkmale wichtige Anhaltspunkte für eine Modellierung, die wenig mit der konventionellen Bildverarbeitung gemein hat. Es konnte nachgewiesen werden, dass die angestrebte Unabhängigkeit von Farbe und detaillierten Forminformationen sehr genau mit der unscharfen Bewegungsabbildung der visuellen Afferenzen des Superior Colliculus – dem multisensorischen Integrationsort im Mittelhirn – korrespondiert. Als zusätzliche Modellierungsoptionen wurden außerdem eine sigmoide Limitierung der Reizintensität und eine an Latenzmessungen orientierte, zeitliche Dynamik der Kodierung vorgeschlagen.

Auf Grundlage der topographisch kompatiblen, unisensorischen Repräsentationen im Ergebnis der visuellen und auditorischen Vorverarbeitung stand im 4. Kapitel die frühe multisensorische Integration als Kern des vorgestellten Wahrnehmungsmodells im Mittelpunkt. Auch hier ließen die wenigen bislang bekannten Modellansätze mit künstlichen neuronalen Netzen oder abstrakten probabilistischen Beschreibungen zwei grundsätzliche Herangehensweisen erkennen. Allerdings wiesen sowohl das auf topographische und motorische Aspekte spezialisierte, WTA-verwandte Simulationssystem RUCCIS [REW99, RWE00] als auch der Grundsatzartikel zum neurobiologisch motivierten auditorisch-visuellen Bayesfilter von ANASTASIO [APBB00] diverse Schwächen auf. Mit dem interdisziplinären Anspruch der vorliegenden Arbeit wurde schließlich eine Weiterentwicklung und vergleichende Untersuchung der beiden grundlegenden Paradigmen versucht, die sich einerseits an neurobiologischen Befunden als Modellvorgabe

orientiert und gleichzeitig ingenieurwissenschaftliche Aspekte und Randbedingungen der Implementierung und praktischen Anwendung berücksichtigt:

- Das eigene, WTA-basierte Fusionsmodell wurde konsequent am Vorbild neuroanatomischer Befunde und multisensorischer Response-Eigenschaften motiviert und gestaltet.
- Als probabilistisches Referenzverfahren konnten die theoretischen Betrachtungen zur Bayesschen Filterfunktion des Superior Colliculus in ein ingenieurtechnisch und neurobiologisch plausibel erweitertes Simulationsmodell überführt werden.
- Zur Evaluierung des Modellverhaltens anhand der Gesetzmäßigkeiten der frühen auditorisch-visuellen Integration wurden aus den multisensorischen Response-Eigenschaften des Superior Colliculus quantifizierbare Bewertungskriterien abgeleitet.
- Unter ausdrücklicher Abgrenzung zu Klassifikations- und Trackingaufgaben gelang es, die simulierten, unbewussten Wahrnehmungsleistungen in offline Versuchsreihen noch ohne einen Handlungskontext, aber schon in Bezug zu realen Anwendungsszenarien, zu untersuchen. Auf Basis der neurobiologisch motivierten Bewertungsregeln konnte in Verbindung mit einer eigens entworfenen audiovisuellen Szenen-Datenbank eine vergleichende, multikriterielle Optimierung von WTA-Netz und Bayesfilter realisiert werden.
- Um die praktische Anwendbarkeit der Simulationsverfahren zu demonstrieren, wurden wesentliche Komponenten des WTA-basierten multisensorischen Systems in einem echtzeitfähigen Demonstrator implementiert.

### **Wertung der Ergebnisse**

Die im Grunde einfach formulierte Aufgabenstellung der objektunabhängigen, räumlichen Aufmerksamkeitssteuerung hält eine Vielzahl methodischer Herausforderungen bereit und kann am natürlichen Vorbild und an künstlichen sensorischen Systemen diskutiert werden. Ein erstes Anliegen dieser Arbeit ist es, die neurobiologischen und wahrnehmungspsychologischen Befunde und die ingenieurwissenschaftlichen Fragen der Simulationsmodelle in einem interdisziplinären Kontext für Informatiker und Biologen gleichermaßen lesbar zu machen. Dazu wurde versucht, ein Niveau der biologischen und mathematischen Beschreibungen zu finden, das möglichst exakt und dabei für einen großen Leserkreis nachvollziehbar ist.

Zur Veranschaulichung der beschriebenen sensorischen Mechanismen dokumentierten durchgehend eigene Simulationsergebnisse die Stärken und Schwächen der Modellansätze. Grundlegende Vorgaben wie die multisensorische Response Enhancement für räumlich und zeitlich korrelierte Stimuli oder die umgekehrte Proportionalität von unimodaler Effektivität und multisensorischer Verstärkung konnten sowohl mit dem künstlichen neuronalen Netz als auch mittels Bayesfilter demonstriert werden. Im Vergleich zum probabilistischen Verfahren erwies sich die algorithmisch einfachere WTA-Variante in der Simulation als effektive und leichter zu parametrisierende Lösung. Durch die Kombination von günstigen multisensorischen Response-Eigenschaften mit einem typischen Selektionsverhalten hat das WTA-Konzept gleichzeitig ein Potential, das über die Möglichkeiten der bislang implementierten Bayesfilter hinausgeht. Die Amari-Felddynamik vermag schnell auf neue Stimuli zu reagieren und realisiert gleichzeitig durch ihre inherente Hysterese große, multisensorisch sensitive Zeitfenster.

Die topographischen Repräsentationen der beiden untersuchten Simulationsmodelle können einerseits als räumliche Aufmerksamkeitskarte zur Unterstützung von verdeckten Verhaltenskomponenten wie der visuellen Suche interpretiert werden. Ebenso ist auf ihrer Grundlage eine direkte motorische Kodierung von reflexhaften Bewegungen denkbar. Anhand des WTA-basierten Hardware demonstrators wurden Echtzeitfähigkeit und Robustheit eines einfachen Orientierungsverhaltens für reale visuelle, akustische und multimodale Situationen nachgewiesen.

Als Resümee zur multimodalen Modellierung können einige grundsätzliche Erkenntnisse hervorgehoben werden. Im Rahmen früher sensomotorischer Mechanismen, die ohne kortikale Analyse bereits im Mittelhirn gesteuert werden, scheinen schnelle und zuverlässige Reaktion wichtiger zu sein als eine hohe Genauigkeit der motorischen Kodierung. Damit multisensorische Effekte auf einem subsymbolischen Niveau zum tragen kommen, müssen die sensorischen Repräsentationen aller beteiligten Modalitäten notwendigerweise große rezeptive Felder besitzen und unspezifische Informationen wie Geräuschrichtungen oder visuelle Bewegungen übertragen. In der Vergangenheit wurde wiederholt versucht, Gemeinsamkeiten zwischen visueller und auditorischer Verarbeitung zu konstruieren und etwa die Repräsentation visuell erfasster Richtungen und akustischer Frequenzen gleichzusetzen. Dass aus solchen Parallelen keinesfalls auf einen gemeinsamen Charakter oder eine Redundanz der Wahrnehmung in verschiedenen Modalitäten zu schließen ist, wird schon daran deutlich, dass sich im Laufe der Evolution kaum Spezialisierungen auf einzelne Modalitäten, sondern die vorherrschenden multisensorischen Strategien entwickelt haben. Abgesehen von der Sicherstellung kompatibler Topographien ist es im Rahmen einer frühen, sensorischen Verarbeitung

nicht sinnvoll, nach Gemeinsamkeiten von auditorischen und visuellen Reizmustern zu suchen. Vielmehr erhöhen gerade die unterschiedlichen räumlich–zeitlichen Konzepte und spezifischen Stärken der einzelnen Modalitäten den Nutzen bei ihrer multisensorischen Integration. Auch der unterschiedlich hohe Aufwand der unisensorischen Vorverarbeitung ist unter diesem Gesichtspunkt nicht als Manko, sondern als adäquate Simulation der berechneten, auditorischen und der lediglich projizierten, visuellen Topographie zu verstehen.

### **Weitere Entwicklung der Modelle und Simulationen**

Die exemplarische, auditorisch–visuelle Integration auf Basis einer Azimutwinkelkarte kann als kleinster gemeinsamer Nenner bei der Modellierung der vielfältigen neuronalen Mechanismen und zitierten Befunde verstanden werden. Die Erweiterung zu einer zweidimensionalen Topographie ist unter Beibehaltung der multisensorischen Mechanismen ebenso vorstellbar wie der Einsatz anderer Kodierungs- und Implementierungsoptionen in bestimmten Teilmodellen oder die Anpassung der Bewertungskriterien an konkrete Anwendungsszenarien.

Im visuellen Teilmodell gibt es zunächst keine prinzipielle Alternative zur Bewegungskodierung durch Intensitätsunterschiede. Zur Vorbereitung sensomotorischer Leistungen kann allerdings eine frühe Kodierung der Bewegungsrichtung visueller Stimuli erwogen werden. Gestaltungsspielraum bietet die Geometrie der rezeptiven Felder, die leicht an typische Szenarien anzupassen oder durch eine Elevationsdarstellung zu ergänzen ist. Desweiteren sollte sich die Dynamik von Intensität und zeitlicher Kodierung an der nachfolgenden Verarbeitung orientieren, um beispielsweise multisensorische Zeitfenster für probabilistische Fusionsmodelle zu ermöglichen.

Auch im auditorischen Modell ließe sich auf Basis einer geeigneten Mikrofonkonfiguration eine vertikale Richtungskarte bewerkstelligen. In einem am biologischen Vorbild orientierten Modell wären dazu die Erzeugung, Kodierung und Auswertung von richtungsspezifischen, akustischen Übertragungsfunktionen erforderlich. Am leichtesten wären solche äußerst komplexen Verarbeitungsleistungen vermutlich in separaten Timing– und Intensity–Pfadern realisierbar, in denen die Stereolaufzeit in bekannter Weise den Azimutwinkel und intensitätsbezogene Informationen die Elevation kodieren. In einem typischen Szenario, in dem sich die meisten akustischen und optischen Ereignisse etwa in derselben horizontalen Ebene mit der sensorischen Einrichtung befinden, ist der Nutzen einer zweidimensionalen Topographie für eine primäre, räumliche Aufmerksamkeitskarte gemessen am hohen Aufwand eher gering. Ähnliches gilt

für die Implementierungsoptionen des binauralen Kreuzkorrelators. Theoretisch kann eine tonotope, sensorische Repräsentation in parallelen Frequenzbändern und die detailliertere Simulation der Onset- und Phaselocked-Kodierungen im Cochlearen Kern Hall- und Echos unterdrücken und so die Schallortung unter ungünstigen akustischen Bedingungen verbessern. Gerade angesichts der schon guten Ergebnisse der hier favorisierten Korrelation der Mikrofon-signale mit anschließender WTA-Filterung erscheint der enorme zusätzliche Berechnungsaufwand einer tonotopen Variante für die praktische Anwendung aber kaum gerechtfertigt.

Im Gegensatz zu unserem fundierten Kenntnisstand zur visuellen und auditorischen Verarbeitung stützen sich die wenigen Modelle der frühen multisensorischen Integration auf jüngere Befunde und betreffen neuronale Mechanismen, die Gegenstand aktueller und grundsätzlicher Diskussionen sind [Kat96]. Deutlich wurde dies insbesondere am Bayesschen Modellansatz, der zunächst eine elegante mathematische Abstraktionsmöglichkeit verspricht, bei genauerer Betrachtung jedoch eher für pixelbasierte, audiovisuelle Anwendungen und weniger für biologisch motivierte, sensorische Topographien geeignet erscheint. Zwar konnten bei der Modifikation und Erweiterung des in [APBB00] beschriebenen Bayesfilters durch realistische Sensor- und Bewegungsmodelle plausible multisensorische Verstärkungseffekte und lokale topographische Abhängigkeiten realisiert werden. Die Auswertung größerer räumlicher Zusammenhänge oder die Erzeugung eigener, multisensorisch sensitiver Zeitfenster sind im probabilistischen Modell jedoch offene Fragen für zukünftige Untersuchungen.

Eingebettet in komplexere Wahrnehmungssysteme können die vorgestellten Simulationsmodelle schließlich zwei grundlegende Funktionen erfüllen. Als multisensorische Repräsentation im afferenten Projektionsweg liefern sie neue Aufmerksamkeitsbereiche für verdeckte, höhere Mechanismen wie die visuelle Suche. Gleichzeitig motivieren die Befunde über integrierte sensomotorische Karten im Superior Colliculus eine selbständige Kodierung von Motorkommandos, um schnell und reflexartig Reaktionen auszulösen. Angefangen von der lokalen oder verteilten Kodierung motorischer Ziele über die Entscheidung zwischen Zuwendung oder Fluchtreflex bis hin zur Koordinierung von Augen- und Kopfbewegungen unter Aufrechterhaltung der auditorisch-visuellen, topographischen Register halten die motorischen Aspekte bei der Anwendung der Modelle viele Herausforderungen bereit. Mit seinen frühen multisensorischen und motorischen Funktionen und seiner exklusiven Lage in afferenten und efferenten Projektionswegen stellt der Superior Colliculus sicher noch länger ein interessantes Forschungsobjekt dar. Für den Entwurf und die Anwendung komplexerer Simulationsmodelle lieferte die Diskussion seiner kortikalen Wechselwirkungen bereits viele Anhaltspunkte.



# Anhang A

## Neuronale Modelle

### Ausgewählte Gewichtsfunktionen und Differentialgleichungen

Zur Modellierung der neuronalen Potentialverläufe PSP und AHP, zur Realisierung der zeitlichen Dynamik in den WTA-Feldgleichungen und bei der zeitlichen Filterung der visuellen Bewegungskodierung wurden zwei universell anwendbare Exponentialfunktionen  $\frac{t}{\tau}e^{1-\frac{t}{\tau}}$  und  $e^{-\frac{t}{\tau}}$  herangezogen. Mit entsprechenden Parametern  $\tau$  können diese als vorgegebene Impulsantwort das typische Übertragungsverhalten der elementaren Modellkomponenten beschreiben. Für die verschiedenen Notationsmöglichkeiten als Zeitfunktion, Übertragungsfunktion oder Differentialgleichung (DGL) soll hier eine schlüssige Herleitung gegeben werden, die außerdem parametrisierbare zeitdiskrete Filter in einer unmittelbar zu implementierenden Form bereitstellt.

Eine erste exponentielle Gewichtsfunktion hat sich zur Beschreibung postsynaptischer Potentiale in Spikeresponse Neuronenmodellen etabliert (s. Abschnitt 2.2.6, Gleichung 2.3) und wurde in Abschnitt 3.4 pragmatisch zur optionalen zeitlichen Filterung der visuellen Bewegungskarte vorgeschlagen. Sie wird oft als  $\alpha$ -Funktion bezeichnet und ist auf das Maximum 1 normiert, welches sie zum Zeitpunkt  $\tau$  erreicht, um danach auf 0 abzuklingen. Ausgehend vom Übertragungsverhalten im Zeitbereich:

$$g(t) = \frac{t}{\tau}e^{1-\frac{t}{\tau}} \tag{A.1}$$

findet man mit Hilfe der Laplace-Transformation:

$$\begin{aligned} \mathcal{L}\left\{\frac{t}{\tau}e^{1-\frac{t}{\tau}}\right\} &= \mathcal{L}\left\{\frac{e}{\tau}te^{-\frac{1}{\tau}t}\right\} \\ &= \frac{e}{\tau} \frac{1}{(s + \frac{1}{\tau})^2} \\ &= \frac{e}{\tau s^2 + 2s + \frac{1}{\tau}} \end{aligned}$$

die Übertragungsfunktion der Synapse im Bildbereich:

$$G(s) = \frac{Y(s)}{X(s)} = \frac{e}{\tau s^2 + 2s + \frac{1}{\tau}} \quad (\text{A.2})$$

Die Rücktransformation liefert eine Differentialgleichung zur Beschreibung des synaptischen Übertragungsverhaltens.

$$\begin{aligned} \tau \ddot{y} + 2\dot{y} + \frac{1}{\tau}y &= e x \\ \ddot{y} &= -\frac{2}{\tau}\dot{y} - \frac{1}{\tau^2}y + \frac{e}{\tau}x \end{aligned} \quad (\text{A.3})$$

Desweiteren veranschaulicht Tabelle A.1 mit einem Struktogramm die Umsetzung der DGL. In gleicher Weise lässt sich auch die Differentialgleichung herleiten, mit der das einfache Abklingverhalten des refraktären AHP oder die Gewichtsfunktion der ratenkodierten WTA-Neurone modelliert wurde:

$$g(t) = e^{-\frac{t}{\tau}} \quad (\text{A.4})$$

$$\mathcal{L}\left\{e^{-\frac{t}{\tau}}\right\} = \frac{1}{s + \frac{1}{\tau}}$$

$$G(s) = \frac{Y(s)}{X(s)} = \frac{1}{s + \frac{1}{\tau}} \quad (\text{A.5})$$

$$\dot{y} = x - \frac{1}{\tau}y \quad (\text{A.6})$$

### Simulation des Übertragungsverhaltens mit zeitdiskreten Filtern

Für die getaktete, iterative Simulation ist es notwendig, die vorgegebene Impulsantwort zeitdiskret zu beschreiben. Am Beispiel der  $\alpha$ -Funktion soll zu diesem Zweck eine Differenzengleichung konstruiert werden, die auf die Eingangsfolge  $\{1, 0, 0, \dots\}$  mit der Ausgabe  $z_n = \frac{n}{\tau}e^{1-\frac{n}{\tau}}$  reagiert. Als Ausgangspunkt liefert die Z-Transformation:

$$\mathcal{Z}(z_n) = \mathcal{Z}\left(\frac{n}{\tau}e^{1-\frac{n}{\tau}}\right)$$

$$\mathcal{Z}(e^{\alpha n}) = \frac{z}{z - e^\alpha}$$

Mit  $\alpha = -\frac{1}{\tau} \Rightarrow \frac{n}{\tau}e^{1-\frac{n}{\tau}}$  erhält man:

$$f(n) = -\alpha e \cdot n \cdot e^{\alpha n}$$

Bei der Berechnung der Z-Transformierten von  $f(n)$  gilt:

a) wegen der Linearität der Transformation:

$$\begin{aligned} \mathcal{Z}(\alpha f_n) &= \alpha \mathcal{Z}(f_n) \\ \Rightarrow \mathcal{Z}(-\alpha e \cdot n e^{\alpha n}) &= -\alpha e \cdot \mathcal{Z}(n e^{\alpha n}) \end{aligned}$$



b) aufgrund des Multiplikationssatzes:

$$\begin{aligned}\mathcal{Z}(nf_n) &= -z [\mathcal{Z}(f_n)]' \\ \Rightarrow -\alpha e \cdot \mathcal{Z}(ne^{\alpha n}) &= \alpha e \cdot z \frac{d}{dz} \left( \frac{z}{z - e^\alpha} \right) \\ &= \frac{-\alpha z \cdot e^{1+\alpha}}{(z - e^\alpha)^2}\end{aligned}$$

Die Differenzgleichung lässt sich nun wie folgt herleiten:

$$\begin{aligned}\mathcal{Z}(f_n) &= -\frac{\alpha z \cdot e^{1+\alpha}}{(z - e^\alpha)^2} \\ (z^2 - 2ze^\alpha + e^{2\alpha}) \mathcal{Z}(f_n) &= -\alpha z \cdot e^{1+\alpha} \\ -\frac{1}{\alpha z \cdot e^{1+\alpha}} (z^2 - 2ze^\alpha + e^{2\alpha}) \mathcal{Z}(f_n) &= 1 = \mathcal{Z}(\{1, 0, 0, \dots\}) \\ -\frac{1}{\alpha e^{1+\alpha}} \left( z - 2e^\alpha + \frac{e^{2\alpha}}{z} \right) \mathcal{Z}(f_n) &= 1\end{aligned}$$

Die Rücktransformation führt zu:

$$-\frac{1}{\alpha e^{1+\alpha}} (z_{n+1} - 2e^\alpha z_n + e^{2\alpha} z_{n-1}) = u_n$$

Die Werte der Folge  $\{z_n\}; n = 0, 1, 2, \dots$  können rekursiv berechnet werden:

$$\begin{aligned}z_{n+1} - 2e^\alpha z_n + e^{2\alpha} z_{n-1} &= -\alpha e^{1+\alpha} u_n \\ z_{n+1} &= -\alpha e^{1+\alpha} u_n + 2e^\alpha z_n - e^{2\alpha} z_{n-1}\end{aligned}$$

Durch Einsetzen von  $\alpha = -\frac{1}{\tau}$  erhält man schließlich die Bemessungsvorschrift für ein Filter zweiter Ordnung, das die geforderte Impulsantwort zeigt:

$$\left. \begin{aligned}a &= \frac{1}{\tau} e^{1-\frac{1}{\tau}} \\ b_1 &= 2e^{-\frac{1}{\tau}} \\ b_2 &= -e^{-\frac{2}{\tau}}\end{aligned} \right\} \Rightarrow z_{n+1} = a u_n + b_1 z_n + b_2 z_{n-1} \quad (\text{A.7})$$

Unter Verzicht auf die entsprechend zu führende Herleitung kann das Filter erster Ordnung zur Modellierung eines einfachen Abklingverhaltens angegeben werden:

$$z_{n+1} = u_n + e^{-\frac{1}{\tau}} z_n \quad (\text{A.8})$$

Der Parameter  $\tau$  ist hier ohne Einheit und an den Systemtakt gebunden. In der Simulation erfolgt die iterative Berechnung unmittelbar an den diskreten Datenstrom  $u$  gekoppelt, wobei die Schrittweite klein bezüglich der zeitlichen Dynamik in den Eingangsdaten zu wählen ist. Tabelle A.1 stellt die beiden grundlegenden Übertragungsverhalten und ihre zeitdiskrete Umsetzung zusammenfassend gegenüber.

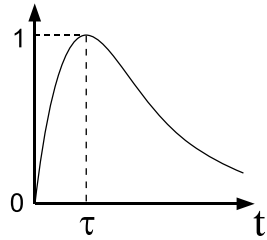
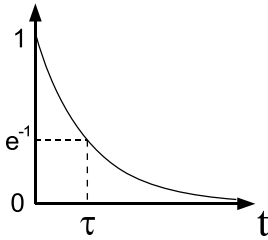
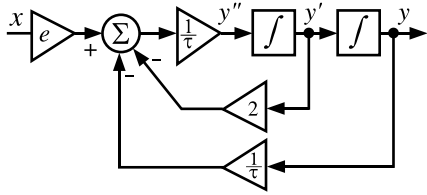
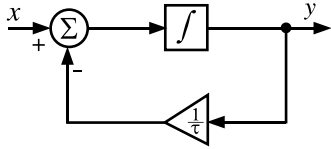
Funktion	$f_\alpha(t) = \frac{t}{\tau} e^{1-\frac{t}{\tau}}$	$f_\beta(t) = e^{-\frac{t}{\tau}}$
Kurve und Parameter $\tau$		
Übertragungsfunktion im Laplacebereich	$G(s) = \frac{Y(s)}{X(s)} = \frac{e}{\tau s^2 + 2s + \frac{1}{\tau}}$	$G(s) = \frac{Y(s)}{X(s)} = \frac{1}{s + \frac{1}{\tau}}$
Differentialgleichung	$\ddot{y} = -\frac{2}{\tau}\dot{y} - \frac{1}{\tau^2}y + \frac{e}{\tau}x$	$\dot{y} = x - \frac{1}{\tau}y$
Struktogramm		
Zeitdiskreter Iterationsfilter	$\left. \begin{array}{l} a = \frac{1}{\tau} e^{1-\frac{1}{\tau}} \\ b_1 = 2e^{-\frac{1}{\tau}} \\ b_2 = -e^{-\frac{2}{\tau}} \end{array} \right\} \Rightarrow \begin{array}{l} z_{n+1} = a u_n \\ \quad + b_1 z_n \\ \quad + b_2 z_{n-1} \end{array}$	$z_{n+1} = u_n + e^{-\frac{1}{\tau}} z_n$
Anwendung	<ul style="list-style-type: none"> <li>• Postsynaptisches Potential (PSP) in spikebasierten Neuronenmodellen</li> <li>• zeitliche Filterung im Model der visuellen Bewegungskodierung</li> </ul>	<ul style="list-style-type: none"> <li>• Refraktionspotential (AHP) in spikebasierten Neuronenmodellen</li> <li>• Modellierung der zeitlichen Dynamik in den neuronalen Feldgleichungen von WTA-Netzen</li> </ul>

Tabelle A.1: Alpha-Gewichtsfunktion und einfache, exponentielle Abklingfunktion in verschiedenen Darstellungsweisen, deren zeitdiskrete Notation für die Implementierung mit fester Iterationsschrittweite sowie ihre Verwendung in den simulierten Modellkomponenten.

# Anhang B

## Probabilistische Modelle

### B.1 Informationstheoretischer Delay-Schätzer

Die Aufgabe der Laufzeitbestimmung in Stereosignalen wird im technischen Umfeld wie auch bei der Interpretation der neuronalen Grundlagen des räumlichen Hörens meist als Korrelationsproblem beschrieben. Entsprechenden Methoden, die auf eine Kreuzkorrelation zurückzuführen sind, ist gemein, dass sie nach einer größtmöglichen Korrespondenz zwischen Wertepaaren  $[\underline{x}(t), \underline{y}(t + \tau)]$  suchen und sonst keine weiteren Abhängigkeiten zwischen den Signalvektoren auswerten. Bei der Frage nach einem Referenzverfahren zur Laufzeitbestimmung wurde bereits in Abschnitt 2.3 auf die Möglichkeit hingewiesen, mit einer geeignet gestalteten Maximum-Likelihood Schätzung (MLE) auch höhere statistische Abhängigkeiten als die Korreliertheit der Signale auszuwerten. Als bekannte MLE-Variante wird häufig die generalisierte Kreuzkorrelation von KNAPP und CARTER zitiert, die mit der Motivation einer spektralen Störschallunterdrückung im Frequenzbereich arbeitet [KC76]. Die hier gewählte Referenzmethode des informationstheoretischen Delay-Schätzers von MODDEMEIJER entspricht nach [Mod88] einer Realisierung des Ansatzes von KNAPP und CARTER im Zeitbereich und stellt ein allgemeines Verfahren zur Laufzeitermittlung ohne a priori Annahmen über die Signaleigenschaften bereit. Unter methodischen Gesichtspunkten ist eine MLE-basierte Laufzeitschätzung im Rahmen dieser Arbeit schon deshalb interessant, da mit ihrer Hilfe die Idee einer vergleichenden Untersuchung von neuronalen und probabilistischen Verfahren nicht auf das multisensorische Integrationsmodell beschränkt bleibt. Nachdem sich die von MODDEMEIJER zur Verfügung gestellte Implementierung des noch wenig etablierten Verfahrens als effizient und robust erwiesen hat, sollen die in [Mod88] formulierten Grundgedanken an dieser Stelle wiedergegeben werden.

MODDEMEIJER beginnt seine Überlegungen zur Definition einer informationstheoretischen Laufzeit mit der Betrachtung zweier Signalvektoren  $\underline{x}_n$  und  $\underline{y}_n$  als zeitdiskrete Realisierungen von stationären stochastischen Prozessen. In der Umgebung eines Samples  $m$  unterteilt er ein Intervall der Länge  $2M$  in einen Past- und einen Future-Vektor :

$$\begin{aligned}\underline{P}_x(m; M) &= (\underline{x}_{m-M}, \dots, \underline{x}_{m-2}, \underline{x}_{m-1})^T \\ \underline{F}_x(m; M) &= (\underline{x}_m, \underline{x}_{m+1}, \dots, \underline{x}_{m+M-1})^T\end{aligned}$$

Den nicht redundanten Informationsgehalt des Signals, der durch eine Verknüpfung von Past- und Future-Vektor kodiert wird, nennt MODDEMEIJER die mutuelle Information und notiert diese zunächst in einer abstrakten Form:

$$I\{\underline{P}_x(m); \underline{F}_x(m)\} = \lim_{M \rightarrow \infty} I\{\underline{P}_x(m; M); \underline{F}_x(m; M)\} \quad (\text{B.1})$$

Eine gleichlautende Deklaration gilt für den Signalvektor  $\underline{y}_n$ . Mögliche Zeitverzögerungen können nun als Verschiebung des Signals  $y$  um  $j$  Samples relativ zum Signal  $x$  definiert werden. Die Verknüpfung der beiden Signale durch den Verschiebeparameter  $j$  kann in der Notation von joint Past- und joint Future-Vektoren berücksichtigt werden:

$$\underline{P}(m; j; M) = \left( \underline{P}_x^T(m; M), \underline{P}_y^T(m + j; M) \right)^T \quad (\text{entsprechend für } \underline{F})$$

Wendet man das Konzept der mutualen Information auf die joint Past- und joint Future-Vektoren an, kann gefragt werden, für welche Verschiebung  $j$  der gemeinsam kodierte Informationsgehalt minimal ist. Dieser Fall stelle eine optimale Trennung zwischen Past- und Future-Intervall in dem Sinne dar, dass der geringste Informationsfluss zwischen beiden stattfindet. Die Verschiebung  $j$  entspräche dann der tatsächlichen Laufzeit zwischen den Signalvektoren  $x$  und  $y$ , die im Ergebnis der informationstheoretischen Delay-Schätzung synchronisiert werden und eine maximale Abhängigkeit aufweisen. Die Notation der mutualen Information von joint Past- und joint Future-Vektor und das geforderte Minimum in Abhängigkeit von  $j$  lauten in Analogie zur Gleichung B.1:

$$\begin{aligned}I\{\underline{P}(m, j); \underline{F}(m, j)\} &= \lim_{M \rightarrow \infty} I\{\underline{P}(m, j; M); \underline{F}(m, j; M)\} \\ I\{\underline{P}(m, j); \underline{F}(m, j)\} &\leq I\{\underline{P}(m, i); \underline{F}(m, i)\} \quad \text{für } i \in \mathcal{Z}\end{aligned} \quad (\text{B.2})$$

Anschließend demonstriert MODDEMEIJER, wie sich die abstrakte Darstellung der mutualen Information in eine adäquate Likelihood-Funktion überführen lässt. Er

nimmt an, dass die Signale normalverteilt mit den Erwartungswerten  $E\{\underline{x}\}=E\{\underline{y}\}=0$  sind:

$$f\{\underline{F}(m, j; M)\} = \frac{1}{(2\pi)^M \sqrt{\det C(j)}} e^{-\frac{1}{2} \underline{F}(m, j; M)^T C(j)^{-1} \underline{F}(m, j; M)} \quad (\text{B.3})$$

wobei  $C(j)$  die Covarianzmatrix beschreibe:

$$E \left\{ \begin{bmatrix} \underline{F}_x(m; M) \\ \underline{F}_y(m+j; M) \end{bmatrix} \cdot \begin{bmatrix} \underline{F}_x(m; M) \\ \underline{F}_y(m+j; M) \end{bmatrix}^T \right\} = \begin{bmatrix} C_{xx} & C_{xy}(j) \\ C_{yx}(j) & C_{yy} \end{bmatrix} \quad (\text{B.4})$$

Nach [Sha48] kann die Entropie der joint Vektoren bestimmt:

$$H\{\underline{F}(m, j; M)\} = M \log 2\pi + \frac{1}{2} \log \det C(j) + M \quad (\text{B.5})$$

und letztendlich die Berechnung der mutuellen Information konkretisiert werden:

$$I\{\underline{F}_x(m; M); \underline{F}_y(m+j; M)\} = -\frac{1}{2} \log \frac{\det C(j)}{\det C_{xx} \cdot \det C_{yy}} \quad (\text{B.6})$$

Schließlich habe eine Maximierung der mutuellen Information der joint Future-Vektoren  $I\{\underline{F}_x(m; M); \underline{F}_y(m+j; M)\}$  die gleiche Funktion wie die Minimierung von  $I\{\underline{P}(m, j); \underline{F}(m, j)\}$ , sei dabei aber numerisch effizienter, da in Gleichung B.6 die Division durch  $\det C_{xx} \cdot \det C_{yy}$  eine verschiebungsinvariante Normierung darstelle und zur Extremwertbestimmung lediglich  $\det C(j)$  betrachtet werden müsse.

Der Parameter  $M$  bestimmt die Ordnung der berücksichtigten, statistischen Momente und nimmt für endliche Signalvektoren gezwungenermaßen finite Werte an. In der praktischen Anwendung zur Schätzung der Stereolauftzeit liefern bereits kleine Werte für  $M$  sehr gute Ergebnisse. In den Benchmarks zur vergleichenden Untersuchung der Schallortungsvarianten in Abschnitt 2.3 wurde beispielsweise ein informationstheoretischer Delay-Schätzer vierter Ordnung benutzt. Mit  $M=1$  verhält sich das Verfahren äquivalent zur Kreuzkorrelationsfunktion.

In Abbildung B.1 wird deutlich, dass die von MODDEMEIJER beabsichtigte Einbeziehung untergeordneter Wertepaare in einem Bereich  $[m-M \dots m+M-1]$  eine brisante Frage bezüglich der funktionellen Interpretation von neuronalen Strukturen zur interauralen Laufzeitkodierung aufwirft. Die etablierten Koinzidenzmodelle der binauralen Schallortung deuten die Verarbeitungsleistung bei der Realisierung einer berechneten, ortskodierten Laufzeitkarte nach wie vor als Korrelationsanalyse [JSY98, FKB00]. MODDEMEIJERS bislang wenig beachteter MLE-Ansatz im Zeitbereich legt aber nahe, dass die prinzipielle Auswertung höherer statistischer Zusammenhänge auch im

Verbund der Delaylines möglich sein sollte. Die massiv divergenten axonalen Verzweigungen, die im laufzeitsensitiven Nucleus laminaris der Eule nachgewiesen wurden, könnten vor diesem Hintergrund nicht nur einer topographischen Kalibrierung des ITD–Ortscodes dienen, sondern auch mehrere Instanzen der verzögerten Reizmuster für eine über die Korrelation hinausgehende Verknüpfung bereitstellen. Es erscheint kaum plausibel, dass der evolutionäre Optimierungsprozess bei der neuronalen ITD–Detektion nach der Realisierung der Korrelationsfunktion halt macht, wenn ganz ähnliche Strukturen bessere und zuverlässigere Ortungsergebnisse erzielen würden.

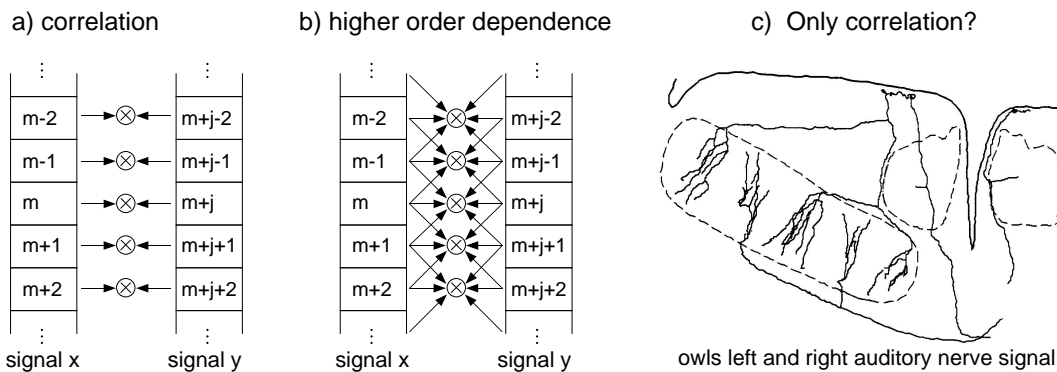


Abbildung B.1: Veranschaulichung der Ordnung des informationstheoretischen Delay–Schätzers nach MODDEMEIJER, der sich a) im Falle  $M=1$  äquivalent zur Kreuzkorrelationsfunktion verhält und b) mit  $M>1$  auch komplexere Zusammenhänge als die Korreliertheit der Signale auswerten kann. c) Die etablierte, funktionelle Interpretation neuronaler Architekturen zur Laufzeitkodierung beschränkt sich auf die Kreuzkorrelation von rechtem und linkem Hörnervmuster (vergl. [Kon93] und Abbildung 2.10).

Letztendlich bieten sowohl das Verfahren von KNAPP und CARTER [KC76] als auch der alternative Delay–Schätzer MODDEMEIJERS [Mod88] Ansatzpunkte für eine kritische Diskussion des über fünfzig Jahre alten Koinzidenzmodells nach JEFFRESS [Jef48]. Die Grundzüge der im Frequenzbereich arbeitenden generalisierten Kreuzkorrelation vorgefilterter Signale [KC76] finden sich im neurologischen Tonotopieprinzip, in den spezifischen Kodierungsformen der cochlearen Kerne und in den periodischen Phasenartefakten der Koinzidenzmuster wieder. MODDEMEIJERS Delay–Schätzer und die spezifische neuroanatomische Divergenz in den biologischen Delayline–Strukturen sprechen gleichzeitig für eine über die Korrelation hinausreichende Strategie im Zeitbereich. Die ausführliche informationstheoretische Diskussion der beiden probabilistischen Paradigmen und des eigenen Ansatzes einer WTA–gefilterten Kreuzkorrelation liegt nicht mehr im Fokus der vorliegenden Arbeit, sollte als interessante und offene Fragestellung aber erwähnt werden.

## B.2 Sensorisches Bayesfilter

### Iterative Bayessche Berechnung

In Gleichung 4.3 wurde die Bayessche Formel für visuell geschätzte Targetwahrscheinlichkeiten nach [APBB00] angegeben. Diese lässt sich zunächst in eine Iterationsvorschrift umformen, indem der zeitliche Verlauf der Beobachtungen  $V_t, V_{t-1}, \dots, V_0$  betrachtet:

$$P(T_t|V_t, V_{t-1}, \dots, V_0) = \frac{P(V_t, V_{t-1}, \dots, V_0|T_t) \cdot P(T_t)}{P(V_t, V_{t-1}, \dots, V_0)}$$

und die Multiplikationsregel für bedingte Wahrscheinlichkeiten auf die Terme in Zähler und Nenner angewandt wird:

$$\begin{aligned} P(V_t, V_{t-1}, \dots, V_0|T_t) \cdot P(T_t) &= P(T_t, V_t, V_{t-1}, \dots, V_0) \\ &= P(T_t, V_t|V_{t-1}, \dots, V_0) \cdot P(V_{t-1}, \dots, V_0) \end{aligned}$$

$$P(V_t, V_{t-1}, \dots, V_0) = P(V_t|V_{t-1}, \dots, V_0) \cdot P(V_{t-1}, \dots, V_0)$$

$$P(T_t|V_t, V_{t-1}, \dots, V_0) = \frac{P(T_t, V_t|V_{t-1}, \dots, V_0)}{P(V_t|V_{t-1}, \dots, V_0)}$$

Fasst man die von T unabhängige Normierung durch  $P(V_t|V_{t-1} \dots V_0)$  in einem Faktor  $\alpha$  zusammen, kann nach erneuter Anwendung der Multiplikationsregel und unter der Annahme von Markov-Prozessen der iterative Charakter der Berechnung deutlich gemacht werden:

$$\begin{aligned} P(T_t|V_t, V_{t-1}, \dots, V_0) &= \alpha \underbrace{P(V_t|T_t, V_{t-1}, \dots, V_0)}_{\stackrel{!}{=}P(V_t|T_t) \text{ (lt. Markovbed.)}} \cdot P(T_t|V_{t-1}, \dots, V_0) \\ &= \alpha \underbrace{P(V_t|T_t)}_{\text{Sensormodell}} \int \underbrace{P(T_{t-1}|V_{t-1}, \dots, V_0)}_{\text{alter Belief}} \cdot \underbrace{P(T_t|T_{t-1})}_{\text{Bewegungsmodell}} dT_{t-1} \end{aligned}$$

Als bisherige Schätzung der bedingten Targetwahrscheinlichkeiten geht  $P(T_{t-1}|V_{t-1}, \dots, V_0)$  in die Aktualisierung des *Beliefs* des Filters ein. Außerdem bewertet ein Sensormodell die momentane sensorische Kodierung, während das Bewegungsmodell eine Anpassung des Bayesschen Filters an die räumlich-zeitliche Dynamik der Daten erlaubt (vergl. Abschnitt 4.3.3).

### Einbeziehung lateraler Abhängigkeiten

Um beim Auftreten von Zielen Abhängigkeiten im topographischen Verbund zu berücksichtigen, können in die Notation der Targetwahrscheinlichkeiten und sensorischen Kodierungen formal die Winkelintervalle  $\phi = [\varphi_1, \dots, \varphi_n]$  eingeführt werden. Zur übersichtlicheren Darstellung soll eine verkürzte Schreibweise Anwendung finden:

$$T_{t,\phi} = T_{t,\varphi_1}, T_{t,\varphi_2}, \dots, T_{t,\varphi_n}$$

$$T_{t,\phi}, \dots, T_{0,\phi} = \begin{cases} T_{t,\varphi_1}, & T_{t,\varphi_2}, & \dots, & T_{t,\varphi_n}, \\ T_{t-1,\varphi_1}, & T_{t-1,\varphi_2}, & \dots, & T_{t-1,\varphi_n}, \\ \vdots & & & \\ T_{0,\varphi_1}, & T_{0,\varphi_2}, & \dots, & T_{0,\varphi_n}, \end{cases} \quad \begin{array}{l} \text{(entsprechende Notation} \\ \text{für } V_{t,\phi} \text{ und } V_{t,\phi}, \dots, V_{0,\phi}) \end{array}$$

Der Belief des Bayesschen Filters mit lateralen Verknüpfungen nimmt die in Gleichung 4.8 angegebene Form  $P(T_{t,\phi}|V_{t,\phi}, \dots, V_{0,\phi})$  an:

$$P(T_{t,\phi}|V_{t,\phi}, \dots, V_{0,\phi}) = \alpha P(V_{t,\phi}|T_{t,\phi}) \int P(T_{t-1,\phi}|V_{t-1,\phi}, \dots, V_{0,\phi}) \cdot P(T_{t,\phi}|T_{t-1,\phi}) dT_{t-1,\phi}$$

Die  $2^{|\phi|}$  Belegungen der Integrationsvariable mit  $T \in \{1, 0\}$  machen deutlich, dass eine konkrete Implementierung weitere Vereinfachungen und Einschränkungen erfordert. Im folgenden Lösungsansatz sollen laterale Abhängigkeiten ausschließlich im Bewegungsmodell realisiert werden. Die Eigenschaften des Sensormodells können insbesondere für die hier angestrebte grobe räumliche Auflösung auch ohne lateralen Kontext formuliert werden. Betrachtet man zum Zeitpunkt  $t$  unabhängige Winkelintervalle und fordert entsprechend:

$$P(T_{t,\phi}|V_{t,\phi}) \stackrel{!}{=} \prod_{i=1}^n P(T_{t,\varphi_i}|V_{t,\phi})$$

so gilt für jeden Winkel:

$$P(T_{t,\varphi_i}|V_{t,\phi}, \dots, V_{0,\phi}) = \alpha \underbrace{P(V_{t,\phi}|T_{t,\varphi_i}, V_{t-1,\phi}, \dots, V_{0,\phi})}_a \cdot \underbrace{P(T_{t,\varphi_i}|V_{t-1,\phi}, \dots, V_{0,\phi})}_b$$

Nimmt man an, dass der Teilausdruck (a) gleichverteilt in allen  $V_{t,\varphi_m}$  mit  $m \neq i$  ist, kann ein lateral unabhängiges Sensormodell für beliebige, einzelne Winkel  $i$  unter erneuter Zuhilfenahme der Markovbedingung definiert werden:

$$a) \quad P(V_{t,\phi}|T_{t,\varphi_i}, V_{t-1,\phi}, \dots, V_{0,\phi}) = \beta \underbrace{P(V_{t,\varphi_i}|T_{t,\varphi_i}, V_{t-1,\phi}, \dots, V_{0,\phi})}_{\stackrel{!}{=} P(V_{t,\varphi_i}|T_{t,\varphi_i}) \text{ (lt. Markovbed.)}} = \beta \underbrace{P(V_{t,\varphi_i}|T_{t,\varphi_i})}_{\text{Sensormodell}}$$



Ausdruck (b) berücksichtigt nach wie vor die Abhängigkeiten zwischen verschiedenen Winkeln und lässt sich wiederum in die Notationen für ein Bewegungsmodell und für die vorausgegangene Schätzung der bedingten Targetwahrscheinlichkeit, den alten Belief, zerlegen:

$$\begin{aligned} \text{b) } P(T_{t,\varphi_i} | V_{t-1,\phi}, \dots, V_{0,\phi}) &= \int P(T_{t,\varphi_i} | T_{t-1,\phi}) \cdot P(T_{t-1,\phi} | V_{t-1,\phi}, \dots, V_{0,\phi}) dT_{t-1,\phi} \\ &= \int \underbrace{P(T_{t,\varphi_i} | T_{t-1,\phi})}_{\text{Bewegungsmodell}} \underbrace{\prod_{m=1}^n P(T_{t-1,\varphi_m} | V_{t-1,\phi}, \dots, V_{0,\phi})}_{\text{alter Belief}} dT_{t-1,\phi} \end{aligned}$$

Angesichts einer an der auditorischen Richtungsauflösung orientierten Topographie mit relativ großen Winkelintervallen und der andererseits beliebig klein wählbaren Iterationsschrittweite muss das Bewegungsmodell nicht den gesamten, sensorisch erfassten Bereich gleichzeitig berücksichtigen. Unter der Annahme einer gewissen physikalischen Trägheit der Ziele wurde in den Benchmarks zur Evaluierung des Bayesfilters exemplarisch ein lokaler Winkelbereich  $\phi = [\varphi_{i-2}, \dots, \varphi_{i+2}]$  betrachtet. In einem solchen Intervall lassen sich die resultierenden  $2^5$  Werte leicht tabellieren und in die Berechnung einbeziehen.

Anders als die Gestalt des Sensormodells, die in Abschnitt 4.3.3 aus einem Histogramm der sensorischen Kodierung abgeleitet wurde, müssen für die konkrete Form des Bewegungsmodells einige arbiträre Annahmen getroffen werden. Als naheliegende Lösung kann eine hypothetische Verteilung der Geschwindigkeiten von Zielen mit den möglichen Belegungen der binären Targetvariable ( $T_\phi = [0, 0, 0, 0, 0], \dots, T_\phi = [1, 1, 1, 1, 1]$ ) gefaltet werden. Die Berechnungsvorschrift des hier gewählten, normalverteilten Bewegungsmodells lautet demnach:

$$P_{move}(T_{t,\varphi_i} = 1) = \sum_{k=-2}^2 \left( \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{k^2}{\sigma^2}} \right) \cdot T_{t-1,\varphi_{i+k}}$$

Zusätzliche Basiswahrscheinlichkeiten für das Auftreten neuer Ziele im Falle  $T_{t-1,\varphi_i} = 0$  oder für das Verschwinden von Zielen bei  $T_{t-1,\varphi_i} = 1$  vervollständigen das Bewegungsmodell. Zur Einbeziehung der Grundbeträge  $P_0(T=1)$  für das Erscheinen sowie  $P_0(T=0)$  für das Ausbleiben von Zielen soll die Tabellierung wie folgt realisiert werden:

$$\begin{aligned} P(T_{t,\varphi_i} = 1 | T_{t-1,\phi}) &= P_0(T=1) + (1 - P_0(T=1) - P_0(T=0)) \cdot P_{move}(T_{t,\varphi_i} = 1) \\ P(T_{t,\varphi_i} = 0 | T_{t-1,\phi}) &= 1 - P(T_{t,\varphi_i} = 1 | T_{t-1,\phi}) \\ &\text{mit } P_0(T=1) + P_0(T=0) \leq 1 \end{aligned}$$

Abbildung B.2 veranschaulicht die Zusammenhänge zwischen den Basiswahrscheinlichkeiten und dem Einfluss der angenommenen Normalverteilung der Zielbewegungen. Mit  $P_0(T=1)$ ,  $P_0(T=0)$  und dem  $\sigma$ -Parameter der Normalverteilung der Zielbewegungen kann der optimistische oder pessimistische Charakter des Bewegungsmodells und dessen Anpassung an eher statische oder dynamische Szenarien gesteuert werden. Im Rahmen der benchmarkbasierten Evaluation des Bayesschen Filters in Kapitel 5 gaben diese drei Werte folglich den Parameterraum der Optimierungsaufgabe vor.

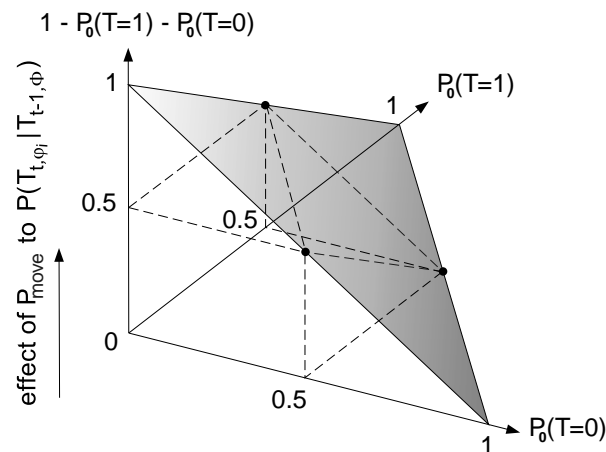


Abbildung B.2: Parametrisierung des Bewegungsmodells durch die Basiswahrscheinlichkeiten  $P_0(T=1)$  und  $P_0(T=0)$ . Hohe Werte von  $P_0(T=1)$  führen zu einer optimistischen a priori Einschätzung der Targetwahrscheinlichkeit, große  $P_0(T=0)$  stellen hingegen eine pessimistische Annahme über das Auftreten von Zielen dar. Je kleiner die Summe beider Parameter ist, um so deterministischer wird der Einfluss des Bewegungsmodells durch die über  $P_{move}$  verknüpften, bisherigen Wahrscheinlichkeiten der Targetvariable im lateralen Intervall  $\phi$ .

In Analogie zum theoretischen, nichtiterativen Ansatz ANASTASIOS [APBB00] wird die posteriore Targetwahrscheinlichkeit  $P(T_{t,\phi}|V_{t,\phi}, \dots, V_{0,\phi})$  für die Hypothesen  $T=1$  und  $T=0$  berechnet und entsprechend der Gesamtwahrscheinlichkeit  $P(T=1) + P(T=0) = 1$  normiert. Exemplarisch können Sensor- und Bewegungsmodelle in identischer Form zur Verarbeitung der auditorischen, topographischen Reizmuster dienen. Die Fusion der beiden unisensorischen Repräsentationen erfolgt auf multiplikative Weise bei der Anwendung des Sensormodells:  $P(V, A|T) = P(V|T) \cdot P(A|T)$ .

# Anhang C

## Aufnahme-Setup und Datenbank

Für die datenbankbasierten Offline-Experimente wurden visuelle und akustische Umweltsituationen in getrennten Teilszenen gespeichert und später zu variierenden, audiovisuellen Reizkombinationen zusammengefügt. Um die Demonstration von Anwendungsszenarien zu ermöglichen, sollten Gesten und Lautäußerungen realer Personen als elementare optische und akustische Ereignisse dienen. Abgesehen davon, dass in den Experimenten ein inhaltlicher oder semantischer Kontext der Szenen keine Rolle spielt, ist eine solche Herangehensweise auch angesichts der geringen Ortsauflösung der modellierten Topographien gerechtfertigt. Infolge der großen rezeptiven Felder werden praktisch keine Details von Gesichtern und somit auch keine spezifischen visuellen Bewegungsmuster beim Erzeugen von Sprachsignalen kodiert.

### Visuelle Szenen

Die Aufnahme der elementaren visuellen Teilszenen erfolgte mit einer digitalen Videokamera (Sony DFW-VL500) in einer diffusen Kunstlichtsituation vor einem neutralen Hintergrund bei einer Bildrate von 30Hz und einer Auflösung von 320x240 Bildpunkten. Ein Set von 14 Gesten (s. Abbildung C.3) wurde nach eigenem Ermessen von 5 Personen dargeboten, die sich dabei in einer Entfernung von maximal 3 Metern vor der Kamera bewegten. Um die spätere Komposition von komplexeren Szenen zu vereinfachen, kam ein omnidirektionales Spiegelobjektiv zum Einsatz, mit dem Azimutabhängige Verzerrungen vermieden wurden. Die elementaren Szenen füllten horizontale Bildausschnitte von 40–90 Grad und wurden nach einer kartesischen Transformation als Einzelbildfolge gespeichert (Abbildung C.1). Die Auflösung der transformierten Bilder betrug 110 Pixel in der Bildhöhe und je nach Art der Bewegung zwischen 80 und 180 Pixeln in der Breite des Bildes. Obwohl die vorgestellten Modelle nur farbun-

abhängige Intensitätsinformationen auswerten, wurden in der Videodatenbank vorerst 24Bit-*RGB*-Dateien gespeichert. Auch die visuellen Szenen in den Abbildungen 3.4a, 4.7 und 5.1 entstanden durch die Kombination einzelner Gesten dieser Datenbank.

a) person 4, "wave with left hand"



b) person 1, "get closer turn"



c) person 2, "walk 40 degree from left to right"

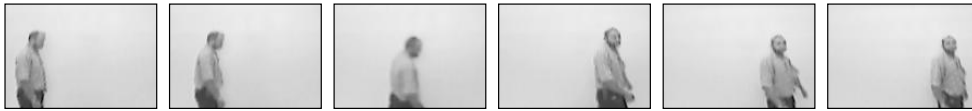


Abbildung C.1: Drei visuelle Teilszenen der Videodatenbank. Bei einer Bewegungsdauer zwischen 1.3 und 2.3 Sekunden enthalten die visuellen Sequenzen 39–69 Einzelbilder.

## Akustische Szenen

Getrennt von den Aufnahmen für den visuellen Teil der Datenbank gaben dieselben 5 Personen ein- und zweisilbige Kommandowörter ("Achtung", "Hallo", "Hier", "Stop") wieder. Desweiteren gehörten Zahlworte ("One", "Two", "Three", "Four"), ein Händeklatschen und ferner auch ein Klick-Geräusch sowie ein Noise-Burst zum Set der akustischen Ereignisse. Alle Sprachsignale und anderen Geräusche wurden in einer hallarmen Umgebung aufgenommen und über eine hochwertige Wiedergabeeinrichtung in

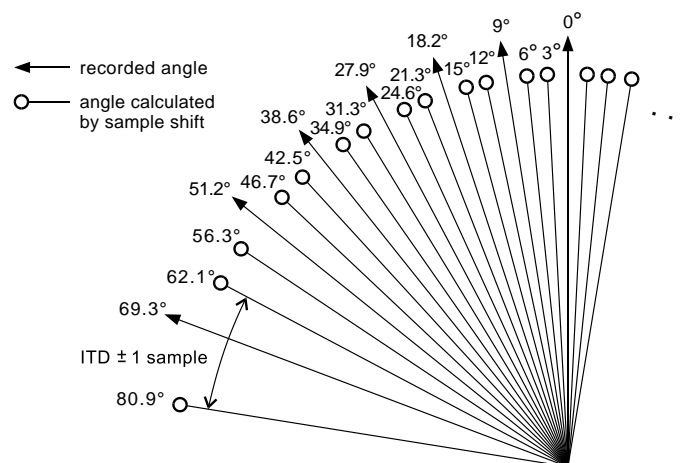


Abbildung C.2: Aufgenommene und berechnete Winkel der Audiodatenbank. Rechte und linke Instanzen der Schallrichtung entstanden durch Vertauschen der Stereokanäle.

einem Hörsaal reproduziert. Dabei erfolgte die Beschallung einer Konfiguration von zwei Kondensatormikrofonen mit Kugelcharakteristik und einem Basisabstand von 15cm aus 3 Metern Entfernung sowie aus verschiedenen Winkeln. Die gewählten Winkel korrespondieren mit den diskreten Stereolaufzeiten, die bei einer seitlichen Auslenkung der Schallquelle in den mit 44.1kHz digitalisierten Audiosignalen entstehen (vergl. Abbildungen 2.12). Um einerseits realistische akustische Bedingungen zu garantieren und gleichzeitig den Aufwand bei der Erstellung der Datenbank zu begrenzen, wurde nur jede dritte mögliche Winkelstufe aufgenommen und die benachbarten Richtungen durch eine Verschiebung der Signale um  $\pm 1$  Sample künstlich erzeugt (Abbildung C.2).

### Datenbank-Konzept

Im Ergebnis der Video- und Audioaufnahmen lagen Gesten und Geräusche geordnet nach Art des optischen oder akustischen Ereignisses und geordnet nach darbietenden Personen vor. Unter den Audioaufnahmen gibt es weiterhin Instanzen für 39 verschiedene Winkel. Zur Deklaration eines Szenarios sind nun Regeln vorzugeben, welche akustischen Ereignisse mit einer bestimmten Geste verknüpft werden dürfen. Zusätzlich wird unter den visuellen Teilszenen eine Unterscheidung in zwei Kategorien für entweder lokale Gesten oder das Vorbeilaufen von Personen eingeführt. Abbildung C.3 veranschaulicht die Organisationsstruktur der Bild und Audiodaten und die Syntax bei der Szenario-Deklaration.

Um auf Grundlage des gespeicherten Bild- und Tonmaterials konkrete Szenen zu erzeugen, stehen scriptbasierte Hilfsmittel zur Verfügung. Zunächst kann ein Parser prüfen, ob die Deklaration eines Szenarios konsistent mit der Verzeichnisstruktur und den vorhandenen Dateien ist. Anschließend ist die Generierung einzelner audio-visueller Sequenzen entsprechend eines Szenenkommandos möglich:

```
scene(Label1, Take1, Position1, Time1, Audio1, Mode1, Label2, ...)
```

Eine Teilszene wird durch jeweils sechs Parameter beschrieben. Bis zu drei, räumlich nicht überlappende Teilszenen lassen sich zu einer audio-visuellen Sequenz kombinieren. Die einzelnen Parameter haben die folgenden Funktionen:

**Label:** bezeichnet die Art der Bewegung (z.B. 'wave\_1' oder 'getcloser\_turn'). Anstelle konkreter visueller Ereignisse ist auch die Angabe einer Kategorie ('local', 'trans') oder eines Jokerzeichens '\*' möglich (vergl. Abbildung C.3).

**Take:** Angabe der Person, die eine Geste oder Bewegung ausführt. Alternativ kann das Jokerzeichen '\*' benutzt werden.

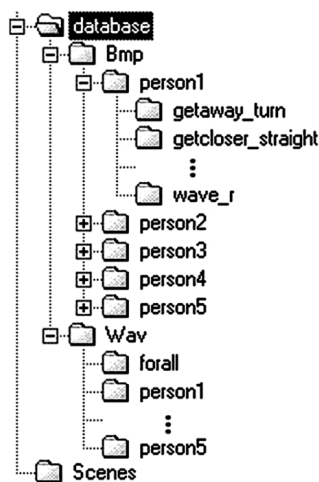
Position: gibt bezogen auf die Mitte der Teilszene in einem Bereich von maximal  $\pm 100$  Grad die Richtung an, in der eine Geste und die dazugehörigen akustischen Ereignisse platziert werden sollen. Alternativ kann ein Intervall in der Form  $[\varphi_l, \varphi_r]$  notiert werden.

Time: bestimmt die Startzeit einer Geste. Auch die Startzeit kann variabel in einem Bereich  $[t_{min}, t_{max}]$  definiert werden.

Audio: bezeichnet das akustische Ereignis, das aus derselben Richtung und zur gleichen Startzeit wie die visuelle Bewegung zu hören sein soll. Erlaubt sind konkrete Angaben ('Achtung', 'one,two,three') oder das Jokerzeichen '\*'.

Mode: entscheidet, ob ein rein visuelles, akustisches oder multimodales Ereignis generiert wird. Die entsprechenden Notationen lauten 'V', 'A' und 'M'.

Ein Szenengenerator prüft die Gültigkeit eines Szenenkommandos, löst Kategorien, Intervalle und Jokerzeichen auf und erzeugt Bildfolge und Audiosignal der Szene. Die beschriebene audio-visuelle Datenbank wurde in [SG04] vorgestellt und steht seither öffentlich zugänglich auf der Webpräsenz des Fachgebiets für Neuroinformatik und Kognitive Robotik der Technischen Universität Ilmenau zusammen mit MATLAB-Scripten zur Generierung von Szenen zum Download bereit.



```
Sourcedirectory='G:/matlab/multimod/database';
Workingdirectory='G:/matlab/multimod/scene';
```

LABEL	CATEGORY	LINKED AUDIO-EVENTS
wave_r,	local,	[ hallo   achtung   hier ]
wave_l,	local,	[ hallo   achtung   hier ]
wave_b,	local,	[ hallo   achtung   hier ]
handup_r,	local,	[ hallo   achtung   hier   stop ]
handup_l,	local,	[ hallo   achtung   hier   stop ]
handup_b,	local,	[ hallo   achtung   hier   stop ]
oneclap,	local,	[ clap ]
getaway_turn,	local,	[ one   two   three   four ]
getcloser_straight,	local,	[ hallo   achtung   hier   stop ]
getcloser_turn,	local,	[ hallo   achtung   hier   stop ]
walk40_lr,	trans,	[ one,two,three ]
walk40_rl,	trans,	[ one,two,three ]
walk60_lr,	trans,	[ one,two,three   one,two,three,four ]
walk60_rl,	trans,	[ one,two,three   one,two,three,four ]

Abbildung C.3: Verzeichnisstruktur der Bild und Audiodaten (links) und exemplarische Deklaration eines Szenarios (rechts). Senkrechte Striche trennen einzelne Geräusche innerhalb der Menge von jeweils zulässigen akustischen Ereignissen (z.B. [ hallo | achtung | hier ]). Durch Komma getrennte Aufzählungen beschreiben hingegen Geräusche, die in einer Szene nacheinander zu hören sein sollen. So kann beispielsweise anhand der Schallrichtung die Positionsänderung einer Person nachvollzogen werden, die an der Kamera vorbeiläuft und dabei zählt (one,two,three,... ).

# Anhang D

## Multisensorische Benchmarks

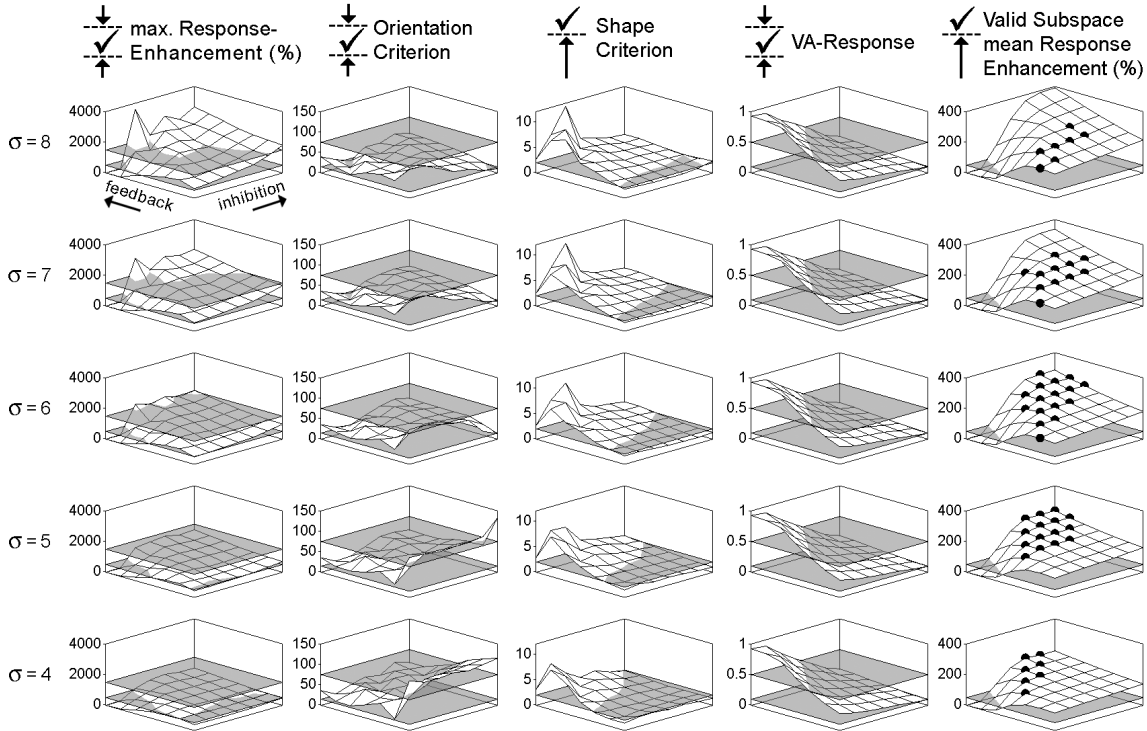
Im Rahmen der multisensorischen Experimente wurden noch nicht alle Möglichkeiten des audio-visuellen Datenbankkonzeptes ausgeschöpft. Beispielsweise ist die Demonstration einer multisensorischen Response Depression bei räumlich dekorrelierten Stimuluskombinationen in Analogie zu den experimentellen Setups von STEIN und MEREDITH [MS86] in einzelnen audio-visuellen Sequenzen durchaus möglich. Konkrete Vorgaben, wie die Geometrie eines entsprechenden Szenarios auf die Topographie und die rezeptiven Felder des SC-Modells anzupassen wäre gibt es bislang jedoch ebenso wenig wie ein plausibles Konzept zur Beschreibung der zeitlichen auditorisch-visuellen Korrespondenz und der großen multisensorisch sensitiven Zeitfenster. Das Szenario der folgenden multisensorischen Benchmarks war daher ausschließlich auf die Kriterien zur Evaluierung der Response Enhancement Effekte abgestimmt. In jeder der vier Versuchsreihen (Abbildungen D.1–D.4) wurde dasselbe Set von 100 lokalen Gesten und dazu räumlich und zeitlich korrelierten Audiosignalen verwendet (vergl. Abbildung C.3).

Die Vorverarbeitung der auditorischen ITD-basierten Azimutwinkelkarte beinhaltete in allen Benchmarks die Kreuzkorrelation der Mikrofonsignale und eine anschließende WTA-Filterung im Ortscode mit 500Hz Iterationsrate und folgenden Parametern:  $\tau = 50ms$ ,  $\sigma_{out} = 1$ ,  $Feedback = [0.1, 0.2, 0.4, 0.2, 0.1]$ ,  $Inhibition = 0.25$ .

Zur Erzeugung der visuellen Bewegungskarte wurden die Spaltensummen der Grauwert-Differenzbilder berechnet und die in Abbildung 3.3 dargestellten rezeptiven Felder simuliert. In den Benchmarks D.2 und D.4 kamen zusätzlich die in Abschnitt 3.4.2 vorgeschlagene sigmoide Dynamikbegrenzung ( $\sigma_{out} = 5$ ) sowie eine moderate zeitliche Glättung ( $\tau = 33ms$ ) zur Anwendung.

### WTA-Benchmark I

#### Whole Benchmark & Criteria



#### Valid Parameter Subspace & Benchmark-data for the optimal Parameter Set

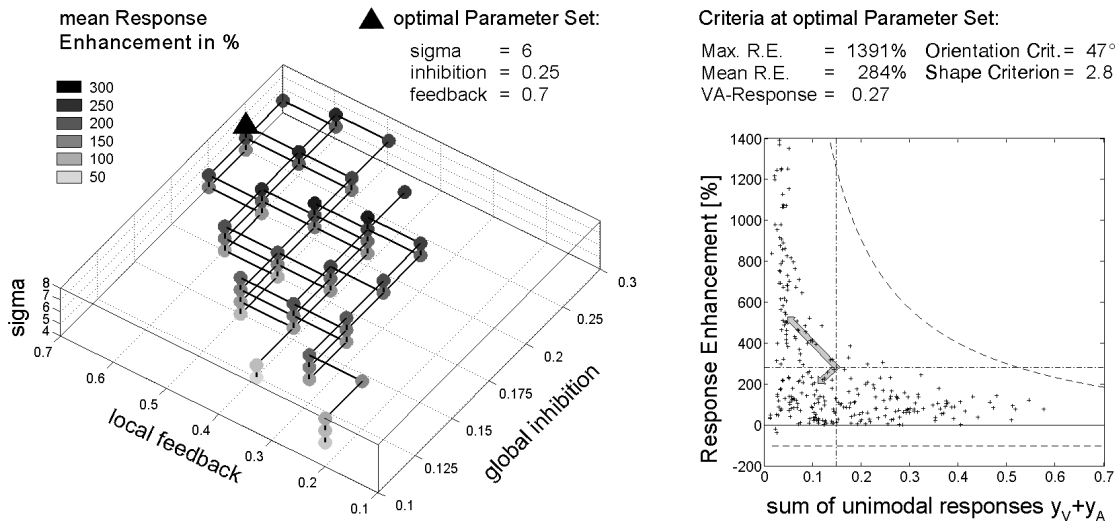
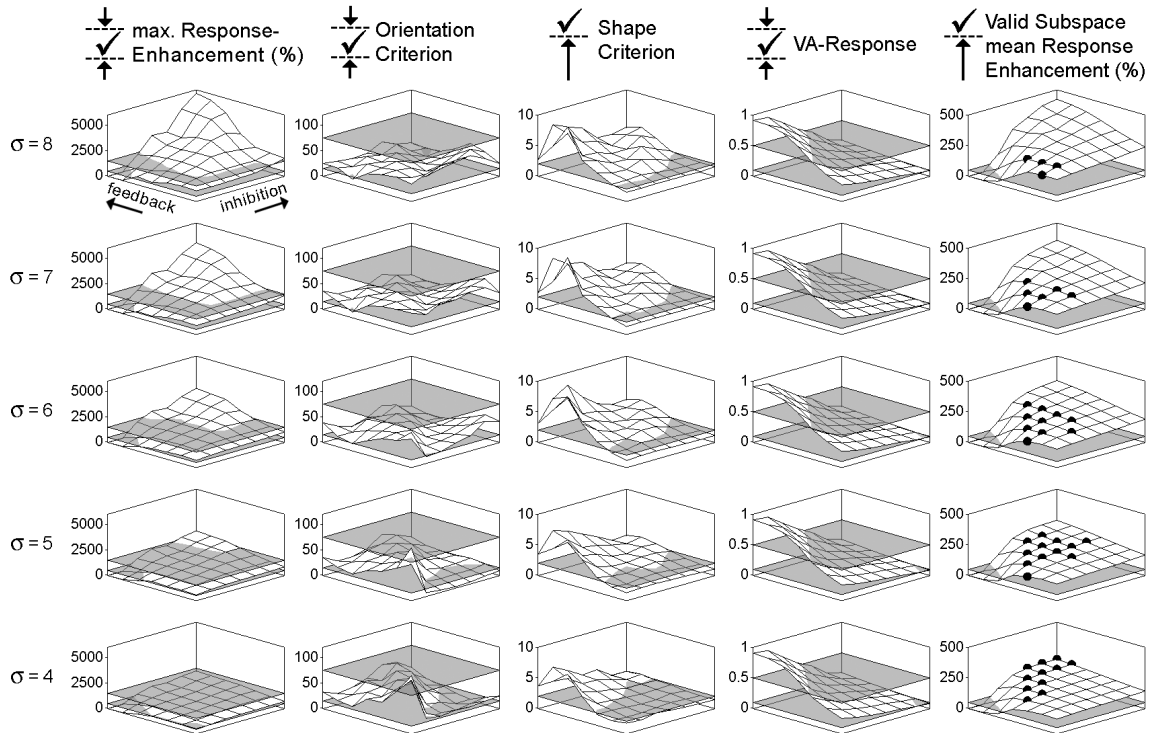


Abbildung D.1: Benchmarkergebnisse des WTA-Netzes **ohne** zeitliche Filterung und sigmoide Ausgabefunktion im Modell der visuellen sensorischen Kodierung. Das Set von 100 Test-szenen wurde in jeweils drei Variationen der Stimulusintensität dargeboten. Kriterium zur Bestimmung eines optimalen Parametersets im gültigen Parameterraum war die zu maximierende durchschnittliche Response Enhancement. Weitere, unveränderliche WTA-Parameter betragen: Iterationsrate=250Hz, Radius des Feedbackvektors=10 und  $\tau = 250ms$ .



## WTA-Benchmark II

### Whole Benchmark & Criteria



### Valid Parameter Subspace & Benchmark-data for the optimal Parameter Set

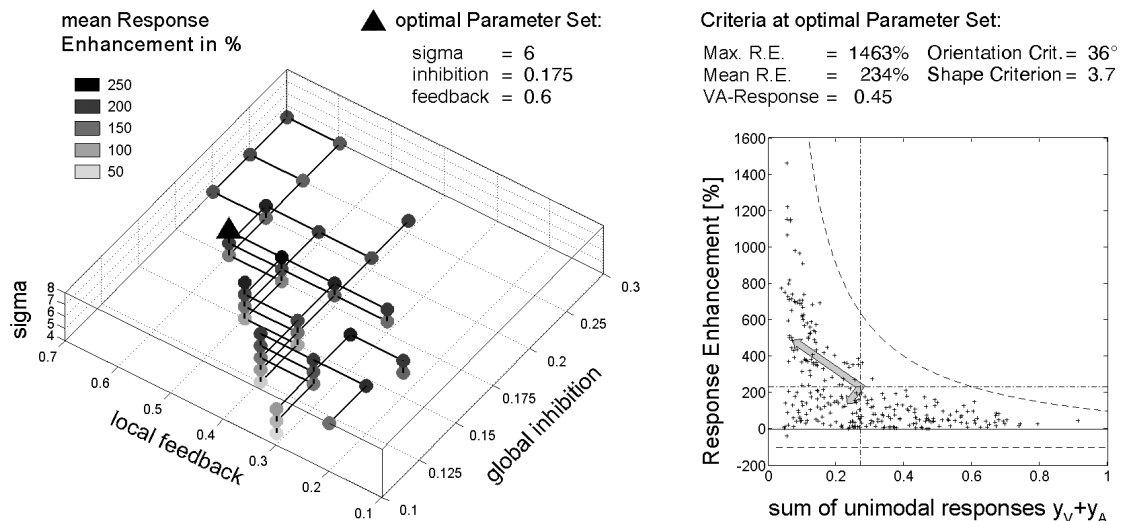
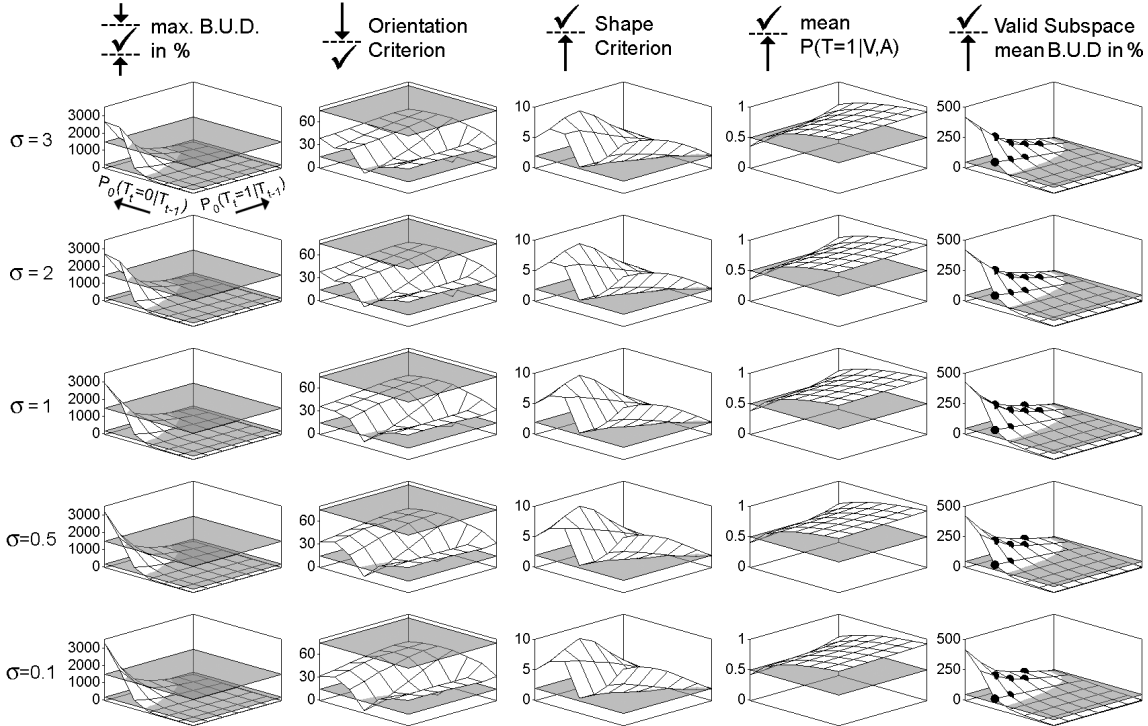


Abbildung D.2: Benchmarkergebnisse des WTA-Netzes **mit** zeitlicher Filterung und sigmoider Ausgabefunktion im Modell der visuellen sensorischen Kodierung. Das Set von 100 Test-szenen wurde in jeweils drei Variationen der Stimulusintensität dargeboten. Kriterium zur Bestimmung eines optimalen Parametersets im gültigen Parameterraum war die zu maximierende durchschnittliche Response Enhancement. Weitere, unveränderliche WTA-Parameter betragen: Iterationsrate=250Hz, Radius des Feedbackvektors=10 und  $\tau = 250ms$ .

### Bayes-Benchmark I

#### Whole Benchmark & Criteria



#### Valid Parameter Subspace & Benchmark-data for the optimal Parameter Set

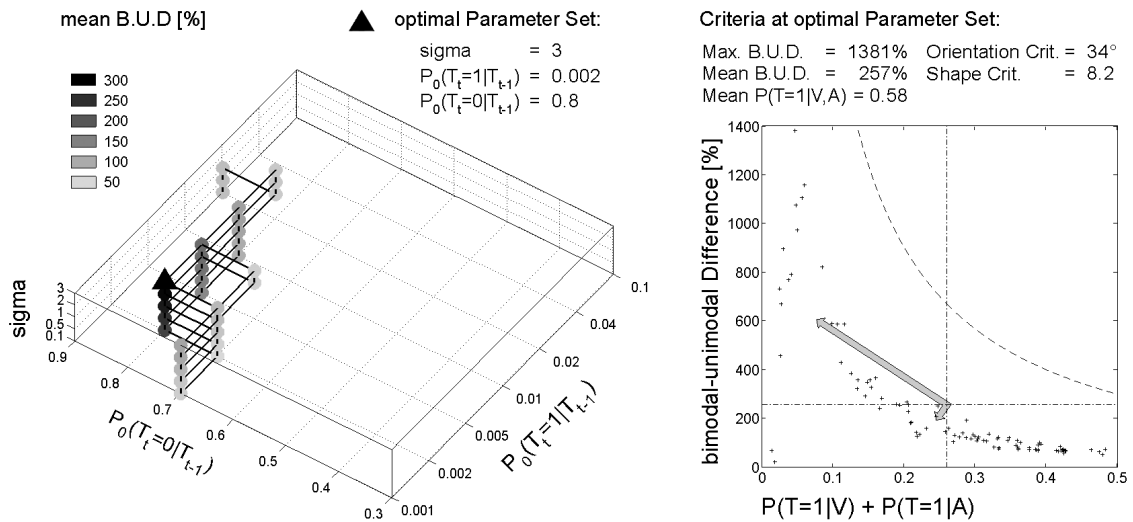
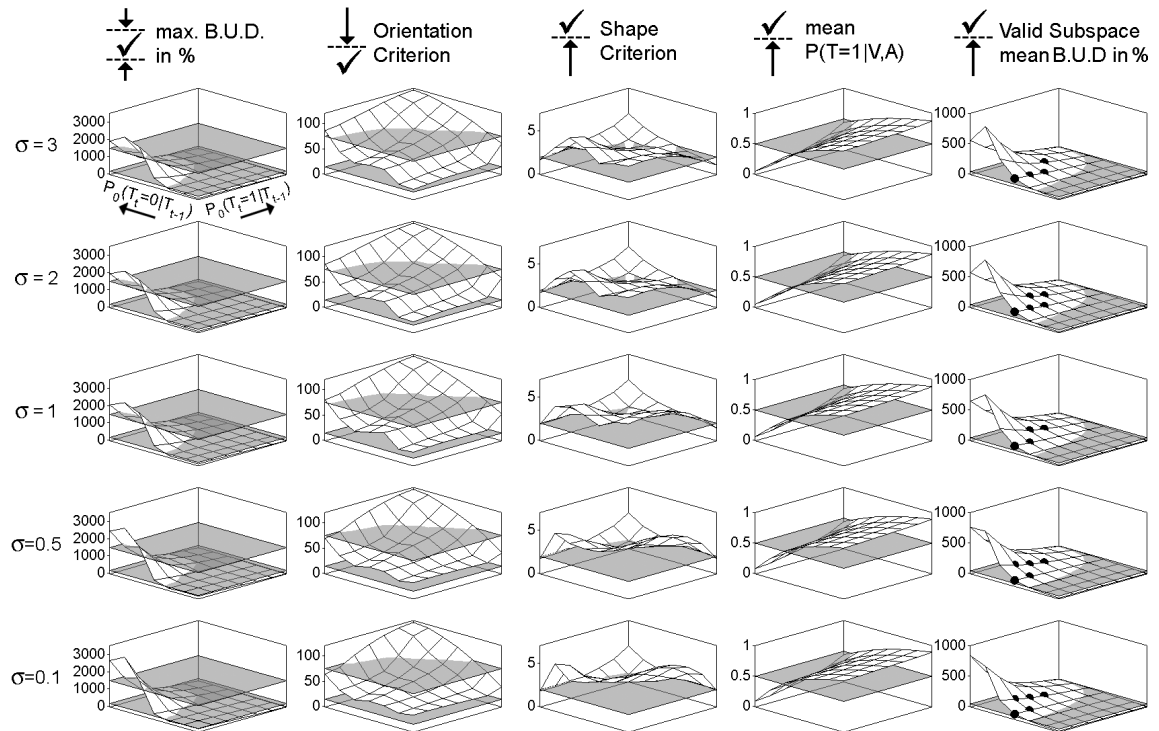


Abbildung D.3: Benchmarkergebnisse des Bayesfilters **ohne** zeitliche Filterung und sigmoide Ausgabefunktion im Modell der visuellen sensorischen Kodierung. Kriterium zur Bestimmung eines optimalen Parametersets im gültigen Parameterraum war die zu maximierende durchschnittliche bimodal-unimodale Differenz. Die Iterationsrate der Simulation betrug 120Hz.

## Bayes-Benchmark II

### Whole Benchmark & Criteria



### Valid Parameter Subspace & Benchmark-data for the optimal Parameter Set

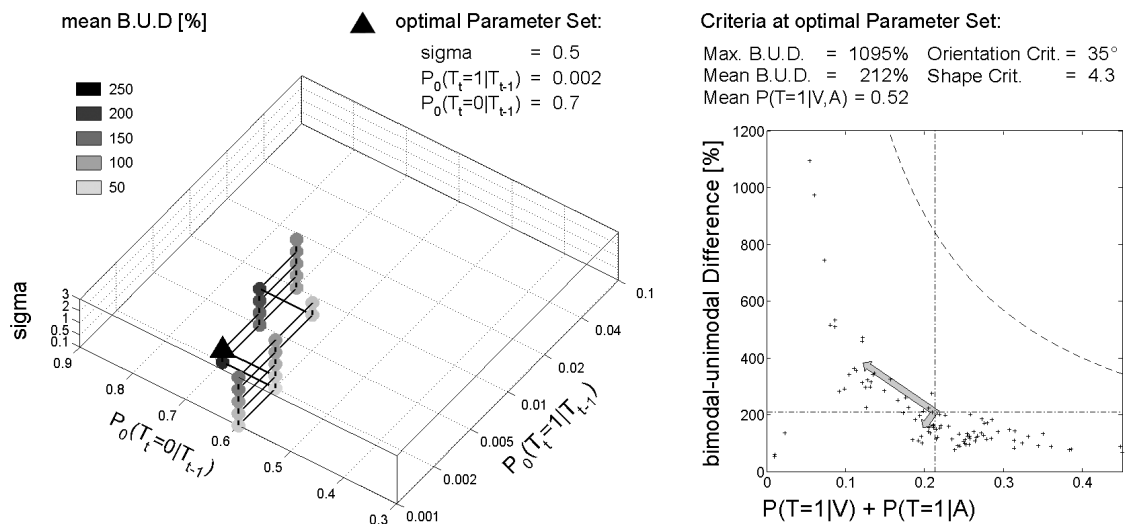


Abbildung D.4: Benchmarkergebnisse des Bayesfilters mit zeitlicher Filterung und sigmoider Ausgabefunktion im Modell der visuellen sensorischen Kodierung. Kriterium zur Bestimmung eines optimalen Parametersets im gültigen Parameterraum war die zu maximierende durchschnittliche bimodal-unimodale Differenz. Die Iterationsrate der Simulation betrug 120Hz.



# Glossar

Die verwendeten Abkürzungen korrespondieren nicht immer mit den im Text ausformulierten Begriffen. Der Grund dafür ist die Absicht, größtmögliche Konsistenz mit der angegebenen, meist englischsprachigen Literatur zu wahren und die Orientierung beim Nachschlagen zu erleichtern.

- |          |  |           |   |
|----------|--|-----------|---|
| $\Theta$ | Reizschwelle eines Neuronenmodells, die zum Auslösen eines Aktionspotentials (Spike) überschritten werden muss.  | afferent  | aufsteigend; in Richtung höherer neuronaler Strukturen  |
| $\tau$   | Zeitkonstante in Exponentialfunktionen zur Beschreibung von Potentialverläufen; Verzögerungszeit.  | AHP       | After hyperpolarization potential. Refraktionspotential einer Nervenzelle.  |
| $\phi$   | Winkelbereich (Azimut), hier in probabilistischen Modellen. Entspricht den möglichen, horizontalen Richtungen auditorischer und visueller Ereignisse.  | AN        | Hörnerv (auditory nerve)  |
| A        | Auditorische, sensorische Kodierung. Entspricht hier der Repräsentation horizontaler Richtungen durch die binaurale Laufzeit $\uparrow$ ITD. Sensorische Kodierungen können Spikefolgen, Raten oder abstrakte mathematische Größen wie die Kreuzkorrelation der Mikrofon-signale sein. | anterior  | vorderer bzw. zuerst innervierter Bereich.  |
| AES      | Anterior ectosylvischer Sulcus. Multisensorischer Grenzbereich zwischen visuellen, auditorischen und somatosensorischen Arealen des assoziativen sensorischen Kortex. Projiziert u.a. unisensorisch in den $\uparrow$ SCd.   | APGF      | all-pole-gammatone filter   |
| AEV      | Visuelles Subareal des $\uparrow$ AES.   | AVCN      | Anteroventraler Teil des $\uparrow$ NC. Von Bedeutung für die phasengenaue Kodierung auditorischer Reizmuster.  |
|          |  | bb        | auf Farbinformationen bezogen breitbandig (im Gegensatz zu $\uparrow$ co)   |
|          |  | belief    | posteriorer Wahrscheinlichkeits-Update eines Bayesfilters für $\uparrow P(T A, V)$ .  |
|          |  | Benchmark | Bewertung einer reproduzierbaren Folge von audio-visuellen $\uparrow$ Szenen, um die Eignung der multisensorischen Modellvarianten in einem $\uparrow$ Szenario zu beschreiben. |
|          |  | BG        | Basalganglien   |
|          |  | BS        | Hirnstamm (Brainstem)   |
|          |  | BMF       | Best modulation frequency   |

caudal	zum Körperende hin; unten	Hallradius	Abstand von einer Schallquelle, in dem die Schalldruckpegel von direktem und diffusem Schallfeld gleich hoch sind. Außerhalb des Hallradius dominieren diffuse Schallanteile (Echos, Raumresonanzen) und verhindern z.B. eine sichere Detektion der Stereophasenlage des Direktschalls.
CD	Characteristic delay	HRTF	Richtungsabhängige Übertragung akustischer Frequenzen am Kopf (head-related transfer function).
CF	Characteristic frequency	I&F	integrate and fire (Neuronenmodell).
CN	Nucleus cochlearis. Erster Kern der zentralen Hörbahn.	ICc/d/x	zentraler, dorsaler und externer Inferior Colliculus. Auditorischer Teil des ↑Tectum.
co	Color-opponent, auf einen Teil des visuellen Spektrums beschränkte Kodierung (im Gegensatz zu ↑bb).	IE-Unit	Kontralateral inhibiertes – ipsilateral exzitatorisches ↑rezeptives Feld.
DCN	Dorsaler Teil des ↑CN.	IID	Interaural intensity difference
DNLL	Dorsaler Teil der ↑NLL.	INLL	Intermediärer Teil der ↑NLL.
dorsal	1. rücklings, hinten liegend. 2. idealisierte Zusammenfassung kortikaler Orientierungsleistungen (dorsaler „Wo?“-Pfad).	IPD	Interaural phase differences. Mehrdeutiges binaurales Merkmal der zeitlichen Kodierung akustischer Ereignisse, aus dem die ↑ITD abgeleitet wird.
EE-Unit	kontralateral- und ipsilateral exzitatorisches Feld.	IPS	Intraparietaler Sulcus; Region des ↑PP, die an der visuellen Suche beteiligt ist (vergl. ↑FEF), in der aber auch multisensorische Effekte nachgewiesen wurden.
efferent	absteigend, vom Zentralnervensystem wegführend	ipsilateral	auf/von derselben Seite
FEF	Frontales/primäres Augenfeld (frontal eye field); im ventromedialen Frontalkortex gelegen und an der Steuerung willkürlicher Augenbewegungen beteiligt.	IT	Inferior temporaler Kortex; speichert positions-, skalierungs- und orientierungs-invariante Prototypen zur Objekterkennung. Zielgebiet im Modell des ↑ventralen Pfades.
Field AES	Auditorisches Subareal des ↑AES.		
fMRI	Funktionelle Magnet-Resonanz Tomographie (auch fMRT). Auch wenn mit tomographischen Verfahren im Vergleich zu invasiven Techniken weniger detaillierte Ergebnisse erzielt werden, sind fMRI-basierte Befunde aus ethischen Gründen bevorzugt zu zitieren.		
Fovea	Bereich der höchsten Ortsauflösung im zentralen Teil des Gesichtsfeldes.		

ITD	Interaural time difference. Entspricht im Rahmen der Schallortung der Stereolaufzeit und kodiert nichtlinear den horizontalen Einfallswinkel von Geräuschen.	MGBd/m/v	Medial geniculate body – dorsaler, medialer oder ventraler Teil. Thalamischer Kern, über den der ↑IC mit dem auditorischen Kortex verbunden ist.
K-Ganglien	Koniozelluläre Sehnervenzellen der Primaten; kodieren Helligkeitsdifferenzen und visuelle Bewegungsrichtungen (bei Wirbeltieren allg. W-Zellen).	MLGN	Bereich des ↑LGN, der von ↑M-Ganglien innerviert wird.
KKF	Kreuzkorrelationsfunktion	MNTB	Medial nucleus of trapezoid body; auditorische Struktur innerhalb des ↑SOC.
kontralateral	auf/von der gegenüberliegenden Seite	MSO	Medial superior olive; binauraler Kern des ↑SOC, der vorrangig ↑ITDs kodiert.
lateral	seitlich gelegen	MT	Medial temporales Kortexareal (auch V5), beteiligt an der Kodierung von Bewegungen und motorischen Funktionen.
Lateralisation	Auditorische Lokalisationsleistung, bei der ausschließlich eine rechts/links-Staffelung von Geräuschen realisiert wird. Elevation, Entfernung oder vor/hinter-Unterscheidung bleiben, ähnlich der Stereophonie, unberücksichtigt.	NA	Nucleus angularis; auditorischer Kern zur Intensitätskodierung im Hirnstamm der Vögel (vergl. ↑PVCN der Säugetiere).
LGN	Corpus geniculatum laterale; verschaltet die Ganglien des Sehnervs mit ↑V1.	nasal	hier bezogen auf das mit der Gesichtsmitte korrespondierende Zentrum von Gesichtsfeld oder auditorischem Raum.
LS	Lateraler suprasylvischer Sulcus. Extraprimäres visuelles Kortexareal mit multisensorischen Eigenschaften. Projiziert u.a. unisensorisch in den ↑SCd.	NL	Nucleus laminaris; binauraler auditorischer Kern zur ↑ITD-Detektion im Hirnstamm der Vögel (vergl. ↑MSO der Säugetiere).
LSO	Lateral superior olive. Binauraler Kern des ↑SOC, der vorrangig ↑IIDs kodiert.	NLL	Nuclei lemnisci lateralis. Auditorische Struktur des Mittelhirns, die ↑efferent mit ↑CN und ↑SOC sowie ↑afferent mit dem kontralateralen NLL und dem ↑IC verschaltet ist.
M-Ganglien	Magnozellularer Sehnervenzellen der Primaten; kodieren vorrangig Helligkeits- und Kontrastinformationen der Stäbchen-Photorezeptoren (bei Wirbeltieren allg. Y-Zellen).	NM	Nucleus magnocellularis; auditorischer Kern zur Phasenkodierung im Hirnstamm der Vögel (vergl. ↑AVCN der Säugetiere).
medial	in der Mitte gelegen	NRT	Reticular nucleus of thalamus

- OT Tectum opticum, Entsprechung zum  $\uparrow$ SC der Säugetiere bei Fischen, Reptilien und Vögeln.
- $P$  Wahrscheinlichkeit
- $P(A, V)$  Abstrakte, unabhängige Wahrscheinlichkeit sensorischer Kodierungen  $\uparrow A$  und  $\uparrow V$ .
- $P(A, V|T)$  Sensormodell: bedingte Wahrscheinlichkeit, dass die sensorischen Kodierungen  $\uparrow A$  und  $\uparrow V$  die An- oder Abwesenheit von Zielen  $\uparrow T$  beschreiben.
- $P(T)$  Priore, unabhängige Wahrscheinlichkeit für die An- oder Abwesenheit von Zielen  $\uparrow T$ .
- $P(T|A, V)$  Posteriore, bedingte Wahrscheinlichkeit für die An- oder Abwesenheit von Zielen  $\uparrow T$  bei gegebenen sensorischen Kodierungen  $\uparrow A$  und  $\uparrow V$  (bisweilen *Belief* genannt).
- $P(T|T_{t-1})$  Bewegungsmodell in rekursiven Algorithmen: bedingte Wahrscheinlichkeit für die An- oder Abwesenheit von Zielen bei gegebener, letzter Belegung der Targetvariable  $\uparrow T_\phi$  und bekannter Bewegungsstatistik der Ziele.
- P-Ganglien Parvozelluläre Sehnervenzellen der Primaten; kodieren vorrangig Farbinformationen im  $\uparrow$ fovealen Bereich (bei Wirbeltieren allg. X-Zellen).
- PCA Hauptkomponentenanalyse (principal component analysis). Basiert auf der Bestimmung der Eigenvektoren der Kovarianzmatrix von Messwerten oder anderen Daten. Wird hier benutzt, um statistische Aussagen über Simulationsergebnisse zu treffen.
- PDF Wahrscheinlichkeitsdichtefunktion (probability density function)
- PET Positronen-Emissions-Tomographie, nichtinvasives, funktionell-bildgebendes Verfahren in der Nuklearmedizin (vergl.  $\uparrow$ fMRI).
- phasisch periodisch oder die Phase betreffend, hier bezogen auf eine reale oder simulierte neuronale Aktivierung, die spezifisch Änderungen eines Stimulus (z.B. Onsets) kodiert.
- PO Parieto-okzipitales Areal.  $\uparrow$ dorsales Kortexareal.
- posterior hinterer bzw. später innervierter Bereich
- PP Posterior parietaler Kortex
- PSP Postsynaptisches Potential
- PSTH Post-stimulus-time Histogram. Darstellung der Spikerate einer Nervenzelle oder neuronalen Region im zeitlichen Verlauf während und nach der Stimulation (meist als Mittelwert der Messungen bei wiederholten Versuchen).
- PVCN Posteroventraler Teil des  $\uparrow$ CN. Von Bedeutung für die  $\uparrow$ tonotope Intensitätskodierung von Geräuschen.
- RF Rezeptives Feld. Bereich des sensorischen Merkmalsraumes (topographische Positionen, akustische Frequenzen etc.), für den ein Neuron sensitiv ist.
- rostal zum Kopfe hin; oben
- SIV (S4) Viertes somatosensorisches Areal, Teil des  $\uparrow$ AES.



Scatterplot	Darstellung einer diskreten Menge von Wertepaaren als Punktwolke. Hier zur Veranschaulichung der Ergebnisse eines $\uparrow$ Benchmarks, dessen Auswertung mittels $\uparrow$ PCA erfolgt.	dellen. 2. Diskrete Menge elementarer akustischer und visueller Ereignisse und Regeln zu deren Kombination in konkreten $\uparrow$ Szenen.
SCs/d	Superior Colliculus (superficial & deep layers), Kern im $\uparrow$ Tectum mit visuellen und multisensorischen Mechanismen zur Steuerung von Aufmerksamkeit und Orientierung sowie prämotorischen Funktionen.	Szene Konkrete, audio-visuelle Sequenz. Instanz eines $\uparrow$ Szenarios und Grundlage von $\uparrow$ Benchmarks.
$S$	Zustandsraum in probabilistischen Modellen; entspricht z.B. der topographischen Abbildung möglicher auditorischer oder visueller Zielpositionen.	$T, T_\phi$ Stochastische Targetvariable zur Beschreibung visueller und auditorischer Ziele (in den Richtungen $\uparrow\phi$ ). Im binären Fall gilt $T=1$ bei Anwesenheit sowie $T=0$ bei der Abwesenheit von Zielen.
$s_i$	Partikelwolke zur Approximation der $\uparrow$ PDF einer stochastischen Variable im Zustandsraum $\uparrow S$ .	TB Trapezoid body. Bereich im $\uparrow$ SOC.
SOC	Superior olivary complex. Erste binaurale Region im Hirnstamm.	Tectum dorsaler Teil des Mittelhirns, bestehend aus $\uparrow$ SC und $\uparrow$ IC.
striär	beschreibt eine gefurchte oder gefaltete Anatomie und die streifenartige Organisation der Merkmalskodierung im primären visuellen Kortex ( $\uparrow$ V1).	temporal zur Schläfe hin; hier der periphere Bereich von Gesichtsfeld oder auditorischem Raum.
STS	Sulcus temporalis superior; kortikales Areal, das in enger Verbindung mit $\uparrow$ IT $\uparrow$ ventrale Klassifikationsleistungen erbringt (Erkennen von Gesichtern), aber auch $\uparrow$ dorsale Verknüpfungen und multisensorische Eigenschaften aufweist.	tonisch andauernd, hier bezogen auf die reale oder simulierte neuronale Aktivierung
Szenario	1. typische Situationen bei Diskussion oder der Anwendung von Wahrnehmungsmo-	tonotop auf akustische Frequenzen bezogene Topologie (Tonotopie).
		$V$ Visuelle, sensorische Kodierung. Entspricht hier der Repräsentation von Bewegungen (beispielsweise Spaltensummen von Differenzbildern).
		V1 Primärer visueller Kortex (auch Areal 17), gekennzeichnet durch streng topographische und $\uparrow$ striäre Kodierung von Intensität, Farbe, Bewegung und Bewegungsrichtung.
		V2 Visuelles Kortexareal; sensitiv für Farbe und orientierte Kanten.

- V3 Visuelles Kortexareal; sensitiv für Form-Merkmale.
- V4 Visuelles Kortexareal im  $\uparrow$ ventralen Pfad. Kodierung von nicht-topographischen Objekteigenschaften.
- V5  $\uparrow$ MT.
- ventral 1. bauchseits, vorne. 2. idealisierte Zusammenfassung kortikaler Klassifikationsleistungen (ventraler „Was?“-Pfad).
- VLSI Very-large-Scale Integration (mikroelektronisches Schaltungsdesign).
- VNLL ventraler Teil der  $\uparrow$ NLL.
- $w_{ij}$  Wichtung der synaptischen Verbindung von Neuron  $j$  zu Neuron  $i$ .
- WTA Winner-Take-All Netzwerkstruktur bzw. Selektionsprozess in einer topologischen Merkmalsrepräsentation.
- $z_j(t)$   $\uparrow$ PSP an der Synapse  $j$ .
- $z(r, t)$  Zustand eines dynamischen Neurons an der Position  $r$  zum Zeitpunkt  $t$ .

# Literaturverzeichnis

- [AC01] P. Azzopardi and A. Cowey. Motion discrimination in cortically blind patients. *Brain*, 124(1):30–46, Jan 2001.
- [Ama77] Shun-ichi Amari. Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields. *Biological Cybernetics*, 27:77–87, 1977.
- [AP03] T.J. Anastasio and Patton; P.E. A two-stage unsupervised learning algorithm reproduces multisensory enhancement in a neural network model of the corticotectal system. *J Neurosci*, 23(17):6713–6727, 2003.
- [APBB00] T. J. Anastasio, P. E. Patton, and K. Belkacem-Boussaid. Using bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Comput*, 12(5):1165–1187, May 2000.
- [AR98] H Aizawa and Wurtz R.H. Reversible inactivation of monkey superior colliculus. i. curvature of saccadic trajectory. *J Neurophysiol.*, 79(4):2082–96, Apr 1998.
- [BC78] J. Blauert and W. Cobben. Some consideration of binaural cross-correlation analysis. *Acustica*, 39:96–103, 1978.
- [BCMM01] A. H. Bell, B. D. Corneil, M. A. Meredith, and D. P. Munoz. The influence of stimulus properties on multisensory processing in the awake primate superior colliculus. *Can J Exp Psychol*, 55(2):123–132, Jun 2001.
- [BCPR05] R.M. Burger, K.S. Cramer, J.D. Pfeiffer, and E.W. Rubel. Avian superior olivary nucleus provides divergent inhibitory input to parallel auditory pathways. *J. Comp. Neurol.*, (481):618, 2005.
- [BECN04] G. Benedek, G. Eordeghe, Z. Chadaide, and A. Nagy. Distributed population coding of multisensory spatial information in the associative cortex. *Eur J Neurosci*, 20(2):525–9, Jul 2004.
- [Ber88a] D.M. Berson. Convergence of retinal w-cell and corticotectal input to cells of the cat superior colliculus. *J Neurophysiol*, 60(6):1861–73, 1988.
- [Ber88b] D.M. Berson. Retinal and cortical inputs to cat superior colliculus: composition, convergence and laminar specificity. *Prog Brain Res*, 75:17–26, 1988.
- [Ber94] P Bertelson. The cognitive architecture behind auditory-visual interaction in scene analysis and speech identification. *Cah Psychol Cogn*, 13:69–75, 1994.
- [BGPV01] A. Blake, M. Gangnet, P. Pérez, and J. Vermaak. Integrated tracking with vision and sound. Technical report, Microsoft Research, [www.research.microsoft.com/vision](http://www.research.microsoft.com/vision), 2001.

- [BJ98] Jochen Braun and Bela Julesz. Withdrawing attention at little or no cost: detection and discrimination tasks. *Perception & Psychophysics*, 60(1):1–23, January 1998.
- [BJA02] M.J. Beal, N. Jojic, and H. Attias. A self-calibrating algorithm for speaker tracking based on audio-visual statistical models. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing - ICASSP 2002.*, 2002. IEEE.
- [BJA03] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 25(7):828–836, 2003.
- [BK96] M. Behan and N. M. Kime. Intrinsic circuitry in the deep layers of the cat superior colliculus. *Vis Neurosci*, 13(6):1031–1042, Nov 1996.
- [Bla80] J. Blauert. In G. v.d. Brink and F.A. Bilsen, editors, *Psychophysical, Physiological, and Behavioural Studies in Hearing*, chapter Modelling of interaural time and intensity difference discrimination., pages 421–424. Delft University Press, 1980.
- [Bla96] J. Blauert. Spatial hearing : The psychophysics of human sound localization. chapter 2. Spatial Hearing with One Sound Source. MIT Press, 1996.
- [BMC02] A. Battaglia-Mayer and R. Caminiti. Optic ataxia as a result of the breakdown of the global tuning fields of parietal neurones. *Brain.*, 125(2):225–37, Feb 2002.
- [BMFM<sup>+</sup>00] A. Battaglia-Mayer, S. Ferraina, T. Mitsuda, B. Marconi, A. Genovesio, P. Onorati, F. Lacquaniti, , and R. Caminiti. Early coding of reaching in the parietooccipital cortex. *J Neurophysiol*, 83(4):2374–2391, April 2000.
- [Bod93] Markus Bodden. Modeling human sound-source localization and the cocktail-party-effect. *acta acustica*, 1:43–55, Februar/April 1993.
- [Bre90] A.S. Bregman, editor. *Auditory scene analysis: the perceptual organization of sound*. MIT Press, Cambridge, MA, 1990.
- [BRS02] E. Ben-Reuven and Y. Singer. Discriminative binaural sound localization. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems, vol. 15*, pages 1229–1236. MIT Press, Cambridge, 2002.
- [BS80] V. Bruns and E. Schmieszek. Cochlear innervation in the greater horseshoe bat: demonstration of an acoustic fovea. *Hearing Research*, 3(1):27–43, Jul 1980.
- [BVB<sup>+</sup>98] B. A. Bacon, J. Villemagne, A. Bergeron, F. Lepore, and J. P. Guillemot. Spatial disparity coding in the superior colliculus of the cat. *Exp Brain Res*, 119(3):333–344, Apr 1998.
- [BXB<sup>+</sup>05] N.E. Barraclough, D. Xiao, C.I. Baker, M.W. Oram, and Perrett D.I. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J Cogn Neurosci.*, 17(3):377–91, Mar 2005.

- [CC01] G.A. Calvert and R. Campbell. Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex*, 11(12):1110–1123, 2001.
- [CD04] H. Colonius and A. Diederich. Why aren't all deep superior colliculus neurons multisensory? a bayes' ratio analysis. *Cogn Affect Behav Neurosci*, 4(3):344–53, Sep 2004.
- [CF99] C. E. Carr and M. A. Friedman. Evolution of time coding systems. *Neural Comput*, 11(1):1–20, Jan 1999.
- [CK90] C. E. Carr and M. Konishi. A circuit for detection of interaural time differences in the brain stem of the barn owl. *J Neurosci*, 10(10):3227–3246, Oct 1990.
- [Col73] H. S. Colburn. Theory of binaural interaction based on auditory-nerve data. i. general strategy and preliminary results on interaural discrimination. *J. Acoust. Soc. Am.*, 54:1458–1470, 1973.
- [Col77] H. S. Colburn. Theory of binaural interaction based on auditory-nerve data. ii. detection of tones in noise. *J. Acoust. Soc. Am.*, 61:525–533, 1977.
- [Cow99] Robert Cowell. An introduction to inference for Bayesian networks. In Jordan [Jor99], pages 9–26.
- [CRS85] C. Cavada and F. Reinoso-Suarez. Topographical organization of the cortical afferent connections of the prefrontal cortex in the cat. *J Comp Neurol*, 242(3):293–324, Dec 1985.
- [CS91] A Cowey and P. Stoerig. The neurobiology of blindsight. *Trends Neurosci.*, 14(5):140–5, Apr 1991.
- [CS94] P. Churchland and T.S. Sejnowski. *The Computational Brain*. MIT Press, Cambridge, 1994.
- [CVWMVO02] B.D. Corneil, M. Van Wanrooij, D.P. Munoz, and A.J. Van Opstal. Auditory-visual interactions subserving goal-directed saccades in a complex scene. *J Neurophysiol.*, 88(1):438–54, Jul 2002.
- [CW99a] C.I. Cheng and G.H. Wakefield. Introduction to head-related transfer functions (hrtf's): Representations of hrtf's in time, frequency, and space (invited paper). In *Proceedings of the Audio Engineering Society 107th Convention, New York: 1999.*, 1999.
- [CW99b] C.I. Cheng and G.H. Wakefield. Spatial frequency response surfaces: An alternative visualization tool for head-related transfer functions (hrtfs). In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP99), Phoenix, Arizona: 1999*, 1999.
- [Dan76] G. Dannenbring. Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology*, 30:99–114, 1976.

- [DBDU93] C. Distler, D. Boussaoud, R. Desimone, and L.G. Ungerleider. Cortical connections of inferior temporal area teo in macaque monkeys. *J Comp Neurol.*, 334(1):125–50, Aug 1993.
- [DBSK00] T. P. Doubell, J. Baron, I. Skaliora, and A. J. King. Topographical projection from the superior colliculus to the nucleus of the brachium of the inferior colliculus in the ferret: convergence of visual and auditory information. *Eur J Neurosci*, 12(12):4290–4308, Dec 2000.
- [DD95] Robert Desimone and John Duncan. Neural Mechanisms of selective visual attention. *Annual Review Neuroscience*, 18:193–222, 1995.
- [DH89] John Duncan and Glyn W. Humphreys. Visual Search and Stimulus Similarity. *Psychological Review*, 96(3):433–458, 1989.
- [DH92] John Duncan and Glyn Humphreys. Beyond the Search Surface: Visual Search and Attentional Engagement. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2):578–588, 1992.
- [DHW97] J. Duncan, G.W. Humphreys, and R. Ward. Competitive brain activity in visual attention. *Current Opinion in Neurobiology*, 7:255–261, 1997.
- [DKEM02] M.C. Dorris, R.M. Klein, S. Everling, and D.P. Munoz. Contribution of the primate superior colliculus to inhibition of return. *J Cogn Neurosci.*, 15(14):1256–63, Nov 2002.
- [DMS96] J. Dudel, R. Menzel, and R.F. Schmidt. *Neurowissenschaft: vom Molekül zur Kognition*. Springer-Verlag Berlin, 1996.
- [DP85] C. Downing and S. Pinker. The spatial structure of visual attention. In *M. Posner, & O.S.M. Marin (Eds.), Attention and Performance*. London: Erlbaum., pages 171–187, 1985.
- [DP04] S. Deneve and A Pouget. Bayesian multisensory integration and cross-modal spatial links. *J Physiol Paris. 2004 Jan-Jun;98(1-3)*, 98(1-3):249–58, Jan–Jun 2004.
- [DPM97] M.C. Dorris, M. Pare, and D.P. Munoz. Neural activity in monkey superior colliculus related to the initiation of saccadic eye movements. *J Neurosci*, 17:8566–8579, 1997.
- [DR93] N. Dieringer and Meier R.K. Evidence for separate eye and head position command signals in unrestrained rats. *Neurosci Lett.*, 162, 1993.
- [EG01] J. A. Edelman and M. E. Goldberg. Dependence of saccade-related activity in the primate superior colliculus on visual target presence. *J Neurophysiol*, 86(2):676,691, Aug 2001.
- [Ehr97] G. Ehret. The auditory midbrain, a "shunting yard" of acoustical information processing. In *The Central Auditory System* [ER97], pages 287–295.
- [ER97] G. Ehret and R. Romand. *The Central Auditory System*. Oxford University Press, 1997.

- [EW90] G. Ettliger and W.A. Wilson. Cross-modal performance: behavioural processes, phylogenetic considerations and neural mechanisms. *Behav Brain Res*, 40(3):169–92, Nov 1990.
- [FCBK02] A. Falchier, S. Clavagnier, P. Barone, and H. Kennedy. Anatomical evidence of multimodal integration in primate striate cortex. *J Neurosci*, 22(13):5749–59, Jul 2002.
- [FCK96] Y. Fang, M. Cohen, and T.G Kincaid. Dynamics of a winner-take-all dynamical neural network. *Neural Networks*, 9(7):1141–1154, 1996.
- [FK91] I. Fujita and M. Konishi. The role of gabaergic inhibition in processing of interaural time difference in the owl’s auditory system. *J Neurosci.*, 11(3):1722–39, 1991.
- [FK01] D. C. Fitzpatrick and S. Kuwada. Tuning to interaural time differences across frequency. *J Neurosci*, 21(13):4844–4851, Jul 2001.
- [FK03] C.K. Friesen and A. Kingstone. Covert and overt orienting to gaze direction cues and the effects of fixation offset. *Neuroreport*, 14(3):489–493, 2003.
- [FKB00] D. C. Fitzpatrick, S. Kuwada, and R. Batra. Neural sensitivity to interaural time differences: beyond the jeffress model. *Journal of Neuroscience*, 20(4):1605–1615, Feb 2000.
- [Fla60] J. L. Flanagan. Models for approximating basilar membrane displacement. *Bell Syst. Tech. Journal*, 39:1163–1192, Sept 1960.
- [FS84] J.M. Ferro and M.E. Santos. Associative visual agnosia: a case study. *Cortex*, 20(1):121–34, Mar 1984.
- [FVO98] M. A. Frens and A. J. Van Opstal. Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus. *Brain Res Bull*, 46(3):211–224, Jun 1998.
- [FVOvdW95] M. A. Frens, A. J. Van Opstal, and R.F. Van der Willigen. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Percept. Psychophys.*, 57(6):802–16, Aug 1995.
- [Gai93] W. Gaik. Combined evaluation of interaural time and intensity differences: psychoacoustic results and computer modeling. *J. Acoust. Soc. Am.*, 94:98–110, 1993.
- [GG89] R. Groner and M.T. Groner. Attention and eye movement control: an overview. *Eur. Arch. Psychiatry Neurol.*, 239(1):9–16, 1989.
- [GJP68] L.J. Garey, E.G. Jones, and T.P. Powell. Two visual corticotectal systems in cat. *J Neurol Neurosurg Psychiatry*, 31(2, month=), 1968.
- [GK02] W. Gerstner and W.M. Kistler. Part i: Single neuron models. In *Spiking Neuron Models. Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

- [GPFM02] D.R. Gitelman, T.B. Parrish, K.J. Friston, and M.M. Mesulam. Functional anatomy of visual search: regional segregations within the frontal eye fields and effective connectivity of the superior colliculus. *Neuroimage*, 15(4):970–82., Apr 2002.
- [GRvH93] W. Gerstner, R. Ritz, and J. L. van Hemmen. Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns. *Biological Cybernetics*, 69:503–515, 1993.
- [GS92] P. Glimcher and D.L. Sparks. Movement selection in advance of action: saccade-related bursters of the superior colliculus. *Nature*, 355:542–545, 1992.
- [GSD00] D. Gonzalo, T. Shallice, and R. Dolan. Time-dependent changes in learning audiovisual associations: a single-trial fMRI study. *NeuroImage*, (11):243–255, 2000.
- [GTU<sup>+</sup>01] J. M. Groh, A. S. Trause, A. M. Underhill, K. R. Clark, and S. Inati. Eye position influences auditory responses in primate inferior colliculus. *Neuron*, 29(2):509–518, Feb 2001.
- [GUGMM04] G. Garcia, I. Uria, I. Gerrikagoitia, and L. Martinez-Millan. Connection from the dorsal column nuclei to the superior colliculus in the rat: topographical organization and somatotopic specific plasticity in response to neonatal enucleation. *J Comp Neurol.*, 468(3):410–24, Jan 2004.
- [GvO99] H.H. Goossens and A.J. van Opstal. Influence of head position on the spatial representation of acoustic targets. *J Neurophysiol*, 81(6):2720–36, Jun 1999.
- [GW72] M.E. Goldberg and R.H. Wurtz. Activity of superior colliculus in behaving monkey. ii. effect of attention on neural responses. *J Neurophysiol*, 35:560–574, 1972.
- [GZ96] A.W. Gummer and H.P. Zenner. Central processing of auditory information. In R. Greger and U. Windhorst, editors, *Comprehensive Human Physiology. From Cellular Mechanisms to Integration.*, pages 729–756. Springer Verlag Berlin Heidelberg, 1996.
- [HA97] Robert H. Helfert and Andreas Aschoff. Superior olivary complex and nuclei of the lateral lemniscus. In *The Central Auditory System [ER97]*, pages 193–258.
- [Ham03] F.H. Hamker. The reentry hypothesis: linking eye movements to visual perception. *Journal of Vision*, 3(11):808–16, Dec 2003.
- [Ham05] F.H. Hamker. The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cereb. Cortex.*, 15(4):431–47, Apr 2005.
- [Har38] H. K. Hartline. The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology*, 121:400–415, 1938.
- [HBM00] J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun. The neural mechanisms of top-down attentional control. *Nat.Neurosci.*, 3(3):284–291, Mar 2000.



- [HE01] P.S. Hyde and Knudsen E.I. A topographic instructive signal guides the adjustment of the auditory space map in the optic tectum. *J Neurosci.*, 21(21), 2001.
- [HE02] P.S. Hyde and Knudsen E.I. The optic tectum controls visually guided adaptive plasticity in the owl's auditory space map. *Nature.*, 415:73–6, Jan 2002.
- [HM01] M. G. and Fletcher E. M. Hopfinger, J. B. and Woldorff and G.R. Mangun. Dissociating top-down attentional control from selective perception and action. *Neuropsychologia*, 39(12):1277–1291, 2001.
- [HN99] G. D. Horwitz and W. T. Newsome. Separate signals for target selection and movement specification in the superior colliculus. *Science*, 284(5417):1158–1161, May 1999.
- [HN01] G. D. Horwitz and W. T. Newsome. Target selection for saccadic eye movements: direction-selective visual responses in the superior colliculus. *J Neurophysiol.*, 86(5):2527–42, 2001.
- [HRLNF94] H.C. Hughes, P.A. Reuter-Lorenz, G. Nozawa, and R. Fendrich. Visual-auditory interactions in sensorimotor processing: saccades versus manual responses. *J Exp Psychol Hum Percept Perform*, 20(1):131–53, Feb 1994.
- [HRST98] L. Herrero, F. Rodriguez, C. Salas, and B. Torres. Tail and eye movements evoked by electrical microstimulation of the optic tectum in goldfish. *Exp Brain Res.*, 120(3):291–305, Jun 1998.
- [Hub90] D. H. Hubel. *Auge und Gehirn*. Spektrum-der-Wissenschaft-Verlagsgesellschaft, Heidelberg, 2 edition, 1990.
- [HvO03] M. Hofman and A.J. van Opstal. Binaural weighting of pinna cues in human sound localization. *Exp Brain Res.*, 148(4):458–70, Feb 2003.
- [IHM01] N. J. Ingham, H. C. Hart, and D. McAlpine. Spatial receptive fields of inferior colliculus neurons to auditory apparent motion in free field. *J Neurophysiol*, 85(1):23–33, 2001.
- [Irv92] D.R. Irvine. Physiology of the auditory brainstem. In Popper and Fay [PF92], pages 153–231.
- [ISP99] R. Ižák, G. Scarbata, and P. Paschke. Sound source localization with an integrate-and-fire neural system. In *Proc. of 7th International Conference on Microelectronics for Neural, Fuzzy, and Bio-Inspired Systems MicroNeuro'99*, pages 103–109, Granada, Spain, April 1999. IEEE Computer Society.
- [Jef48] L. A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol.*, 41:35–39, 1948.
- [JGJS99] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In Jordan [Jor99], pages 105–162.
- [Joh95] Mark H. Johnson. The Development of Visual Attention: A Cognitive Neuroscience Perspective. In Michael S. Gazzaniga, editor, *The Cognitive Neurosciences*, A Bradford Book, pages 735–747. The MIT Press, 1995.

- [Jor99] M.I. Jordan, editor. *Learning in graphical Models*. MIT Press, 1999.
- [JSY98] P. X. Joris, P. H. Smith, and T. C. Yin. Coincidence detection in the auditory system: 50 years after jeffress. *Neuron*, 21(6):1235–8, Dec 1998.
- [JvSBC03] C. Jin, A. van Schaik, V. Best, and S. Carlile. Perceptual spatial-audio coding. In *Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, July 6-9, 2003*, pages 255–258, 2003.
- [JWJ<sup>+</sup>01] W. Jiang, M. T. Wallace, H. Jiang, J. W. Vaughan, and B. E. Stein. Two cortical areas mediate multisensory integration in superior colliculus neurons. *J Neurophysiol*, 85(2):506–522, Feb 2001.
- [Kat96] P.S. Katz. Neurons, networks, and motor behavior. *Neuron.*, 16(2):245–53, Feb 1996.
- [KC76] C.H. Knapp and G.C Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on ASSP*, 4(24):320–327, 1976.
- [KC01] A. J. King and G. A. Calvert. Multisensory integration: perceptual grouping by eye and ear. *Current Biology*, 11(8):R322–5, Apr 2001.
- [Ken86] C. Kennard. Higher control mechanisms of saccadic eye movements. *Trans. Ophthalmol. Soc. UK*, 105(6):705–8, 1986.
- [KGWvH99] R. Kempster, W. Gerstner, H. Wagner, and J. L. van Hemmen. Quality of coincidence detection and itd tuning: A theoretical framework. In T. Dau, V. Hohmann, and B. Kollmeier, editors, *Psychophysics, Physiology and Models of Hearing*, pages 185–194. World Scientific, Singapore, 1999.
- [KH00] J. H. Kaas and T. A. Hackett. How the visual projection map instructs the auditory computational map. *J Comp Neurol*, 421(2):143–145, May 2000.
- [Kin97] A. J. King. Signal selection by cortical feedback. *Current Biololgy*, 7(2), 1997.
- [Kin99] A. J. King. Sensory experience and the formation of a computational map of auditory space in the brain. *Bioessays*, 21(11):900–911, Nov 1999.
- [KJM98] A. J. King, Z. D. Jiang, and D. R. Moore. Auditory brainstem projections to the ferret superior colliculus: anatomical contribution to the neural coding of sound azimuth. *J Comp Neurol*, 390(3):342–365, 1998.
- [KK79] K Kawamura and T. Konno. Various types of corticotectal neurons of cats as demonstrated by means of retrograde axonal transport of horseradish peroxidase. *Exp Brain Res.*, 35(1):161–75, Mar 1979.
- [KK94] Samuel Kaski and Teuvo Kohonen. Winner-Take-All Networks for Physiological Models of Competitive Learning. *Neural Network*, 7(6/7):973–984, 1994.
- [KMMS94] C.Q. Kao, J.G. McHaffie, M.A. Meredith, and B.E. Stein. Functional development of a central visual map in cat. *J Neurophysiol.*, 72(1):266–72, Jul 1994.

- [Knu99] E.I. Knudsen. Mechanisms of experience-dependent plasticity in the auditory localization pathway of the barn owl. *J Comp Physiol.*, 185(4), 1999.
- [Koh93] T. Kohonen. Physiological interpretation of the self-organizing map algorithm. *Neural Networks*, 6:895–905, 1993.
- [Kon93] M. Konishi. Die Schallortung der Schleiereule. *Spektrum der Wissenschaft*, pages 58–71, Juni 1993.
- [Kon00] M. Konishi. Study of sound localization by owls and its relevance to humans. *Comp Biochem Physiol A Mol Integr Physiol*, 126(4):459–469, Aug 2000.
- [KS95] Klaus Kopecz and Gregor Schöner. Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biological Cybernetics*, 73(1):49–60, Jun 1995.
- [KSC<sup>+</sup>96] A. J. King, J. W. Schnupp, S. Carlile, A. L. Smith, and I. D. Thompson. The development of topographically-aligned maps of visual and auditory space in the superior colliculus. *Prog Brain Res*, 112:335–350, 1996.
- [KST98] A. J. King, J. W. Schnupp, and I. D. Thompson. Signals from the superficial layers of the superior colliculus enable the development of the auditory space map in the deeper layers. *J Neurosci*, 18(22):9394–9408, 1998.
- [KT96] C. H. Keller and T. T. Takahashi. Binaural cross-correlation predicts the responses of neurons in the owl's auditory space map under conditions simulating summing localization. *J Neurosci*, 16(13):4300–4309, Jul 1996.
- [KT04] S. Kasderidis and J. G. Taylor. Attention-based learning. In *Proceedings Int. Joint Conf. on Neural Networks - IJCNN - 2004*. IEEE Computer Society, 2004.
- [KU85] Koch, Christof and Ullman, Shimon. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, 4:219–227, 1985.
- [KVV02] M. Kubovy and D. Van Valkenburg. Auditory and visual objects. In Scholl [Sch02c], pages 97–126.
- [KVV<sup>+</sup>97] D. C. Kadunce, J. W. Vaughan, M. T. Wallace, G. Benedek, and B. E. Stein. Mechanisms of within-modality and cross-modality suppression in the superior colliculus. *J Neurophysiol*, 78(6):2834–2847, Dec 1997.
- [KVWS01] D. C. Kadunce, J. W. Vaughan, M. T. Wallace, and B. E. Stein. The influence of visual and auditory receptive field organization on multisensory integration in the superior colliculus. *Exp Brain Res 2001*, 139(3):303–310, Aug 2001.
- [Lab95] David Laberge. Computational and Anatomical Models of Selective Attention in Object Identification. In Michael S. Gazzaniga, editor, *The Cognitive Neurosciences*, pages 649–663. 1995.
- [LBB03] T.M. Lock, J.S. Baizer, and D.B. Bender. Distribution of corticotectal cells in macaque. *Exp Brain Res*, 151(4):455–70, Aug 2003.

- [LD04] L.J. Lanyon and S.L. Denham. A model of active visual search with object-based attention guiding scan paths. *Neural Netw. - Special issue Vision and brain*, 17(5-6):873–97, Jun-Jul 2004.
- [LGW00] H. Luksch, B. Gauger, and H. Wagner. A candidate pathway for a visual instructional signal to the barn owl's auditory system. *J Neurosci. (Rapid Communication)*, 20(8-RC70):1–4, Apr 2000.
- [LH87] M.S. Livingstone and D.H. Hubel. Psychophysical evidence for separate channels for the perception of form, color, movement and depth. *J Neurosci*, (7):3416–3468, 1987.
- [LHP94] J.C Lynch, J.E Hoover, and Strick P.L. Input to the primate frontal eye field from the substantia nigra, superior colliculus, and dentate nucleus demonstrated by transneuronal transport. *Exp Brain Res.*, 100(1):181–6, 1994.
- [Lin86a] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. i. simulation of lateralization for stationary signals. *J. Acoust. Soc. Am.*, 80(6):1608–1622, Dec 1986.
- [Lin86b] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition.ii. the law of the first wave front. *J. Acoust. Soc. Am.*, 80(6):1623–1630, December 1986.
- [LKB97] Dale K. Lee, Christof Koch, and Jochen Braun. Spatial vision thresholds in the near absence of attention. *Visual Research*, 37(17):2409–2418, 1997.
- [LM88] R. F. Lyon and C. A. Mead. An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, 36(7):1119–1134, July 1988.
- [LM95] J. Lazzaro and C. Mead. A silicon model of auditory localization. In Steven F. Zornetzer, editor, *An introduction to neural and electronic networks*, pages 185–203. San Diego: Academic Press, 1995.
- [LRMM89] J. Lazzaro, S. Ryckebusch, M. A. Mahowald, and C.A. Mead. Winner-take-all networks of  $o(n)$  complexity. In D. Touretzky, editor, *Advances in Neural information processing systems, Vol. 1*, pages 703–711. San Mateo,CA: Morgan Kaufmann, 1989.
- [LV93] E.A. Lachica and Casagrande V.A. The morphology of collicular and retinal axons ending on small relay (w-like) cells of the primate lateral geniculate nucleus. *Vis Neurosci.*, 10(3):403–18, May-Jun 1993.
- [LY98a] R. Y. Litovsky and T. C. Yin. Physiological studies of the precedence effect in the inferior colliculus of the cat. i. correlates of psychophysics. *J Neurophysiol*, 80(3):1285–1301, Sep 1998.
- [LY98b] R. Y. Litovsky and T. C. Yin. Physiological studies of the precedence effect in the inferior colliculus of the cat. ii. neural mechanisms. *J Neurophysiol*, 80(3):1302–1316, Sep 1998.
- [Lyo97] Richard F. Lyon. All-pole models of auditory filtering. In E. Lewis et al., editors, *Diversity in Auditory Mechanics*, pages 205–211. World Scientific Publishing, Singapore, 1997.

- [Maa00a] W. Maass. Neural computation with winner-take-all as the only nonlinear operation. In S. A. Solla, T. K. Leen, and K. R. Müller, editors, *Advances in Neural Information Processing Systems, Vol. 12*. 2000.
- [Maa00b] W. Maass. On the computational power of winner-take-all. *Neural Computation*, 12(11):2519–35, Nov 2000.
- [Maz98] J. A. Mazer. How the owl resolves auditory coding ambiguity. *Proc Natl Acad Sci U S A*, 95(18):10932–10937, 1998.
- [MC89] M. A. Meredith and H. R. Clemo. Auditory cortical projection from the anterior ectosylvian sulcus (field aes) to the superior colliculus in the cat: an anatomical and electrophysiological study. *J Comp Neurol*, 289(4):687–707, Nov 1989.
- [McI90] J.T. McIlwain. Topography of eye-position sensitivity of saccades evoked electrically from the cat’s superior colliculus. *Vis Neurosci*, 4(3):289–98, Mar 1990.
- [McI91] J.T. McIlwain. Distributed spatial coding in the superior colliculus: a review. *Vis Neurosci*, 6(1):3–13, Jan 1991.
- [MCL98] A.R. McIntosh, R.E. Cabeza, and N.J. Lobaugh. Analysis of neural interactions explains the activation of occipital cortex by an auditory stimulus. *J Neurophysiol*, 80:2790–2796, 1998.
- [MCMH00] S Martinez-Conde, S.L. Macknik, and D.H. Hubel. Microsaccadic eye movements and firing of single cells in the striate cortex of macaque monkeys. *Nature Neuroscience*, 3(3):251–8, Mar 2000.
- [Mea89] Carver Mead. *Analog VLSI and neural systems*. Addison-Wesley, 1989.
- [Mer99] M. A. Meredith. The frontal eye fields target multisensory neurons in cat superior colliculus. *Exp Brain Res*, 128(4):460–470, Oct 1999.
- [MF02] D.P. Munoz and J.H. Fecteau. Vying for dominance: dynamic interactions control visual fixation and saccadic initiation in the superior colliculus. *Prog Brain Res.*, 140:3–19, 2002.
- [MG95] A.D. Milner and M.A. Goodale, editors. *The visual Brain in Action*. Oxford Psychology Series. 27, 1995.
- [MH94] J. L. Meador and P. D. Hylander. Pulse coded winner-take-all networks. In M. E. Zghloul, J. Meador, and R. W. Newcomb, editors, *Silicon Implementation of Pulse Coded Neural Networks*, pages 79–99. Kluwer Academic Publishers, Boston, 1994.
- [MI98] D.P. Munoz and P.J. Istvan. Lateral inhibitory interactions in the intermediate layers of the monkey superior colliculus. *J Neurophysiol.*, 79(3):1193–209, Mar 1998.
- [MJCW03] N.G. Muggleton, C.H. Juan, A. Cowey, and V. Walsh. Human frontal eye fields and visual search. *J Neurophysiol.*, 89(6):3340–3, Jun 2003.

- [MK81] A. Moiseff and M. Konishi. Neural and behavioral sensitivity to binaural time differences in the owl. *J Neurosci*, (1):40–48, 1981.
- [MK99] D. R. Moore and A. J. King. Auditory perception: The near and far of sound localization. *Current Biology*, 9(10):R361–R363, May 1999.
- [MK02] R.M. McPeck and E.L. Keller. Saccade target selection in the superior colliculus during a visual search task. *J Neurophysiol*, 88(4):2019–34, Oct 2002.
- [MK04] M.A. Meredith and A.J. King. Spatial distribution of functional superficial-deep connections in the adult ferret superior colliculus. *Neuroscience*, 128(4):861–70, 2004.
- [MNBC82] L. Mucke, M. Norita, G. Benedek, and O. Creutzfeldt. Physiologic and anatomic investigation of a visual cortical area situated in the ventral bank of the anterior ectosylvian sulcus of the cat. *Exp Brain Res*, 46(1):1–11, 1982.
- [MNS87] M. A. Meredith, J.W. Nemitz, and B. E. Stein. Determinants of multisensory integration in superior colliculus neurons. i. temporal factors. *J Neurosci*, 7(10):3215–29, Oct 1987.
- [Mod88] R. Moddemeijer. An information theoretical delay estimator. In K. A. Schouhamer Immink, editor, *Ninth Symposium on Information Theory in the Benelux*, pages 121–128, Enschede (NL), 1988. Werkgemeenschap Informatie-en Communicatietheorie.
- [Moo97] Brian C.J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 4 edition, 1997.
- [MR98] M. A. Meredith and A. S. Ramoa. Intrinsic circuitry of the superior colliculus: Pharmacophysiological identification of horizontally oriented inhibitory interneurons. *J Neurophysiol*, 79(3):1597–1602, Mar 1998.
- [MS80] L.E. Mays and D.L. Sparks. Dissociation of visual and saccade-related responses in superior colliculus neurons. *J Neurophysiol*, 43(1):207–32, Jan 1980.
- [MS86] M. A. Meredith and B. E. Stein. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J Neurophysiol*, 56(3):640–662, Sep 1986.
- [MS96] M. A. Meredith and B. E. Stein. Spatial determinants of multisensory integration in cat superior colliculus neurons. *J Neurophysiol*, 75(5):1843–1857, May 1996.
- [MSND91] J.H. Maunsell, G. Sclar, T.A. Nealey, and D.D. DePriest. Extraretinal representations in area v4 in the macaque monkey. *Vis Neurosci*, 7(6):561–73, Dec 1991.
- [MTA87] D. Minciacchi, G. Tassinari, and A. Antonini. Visual and somatosensory integration in the anterior ectosylvian cortex of the cat. *Brain Res*, 410(1):21–31, Apr 1987.

- [MTS01] J. G. McHaffie, C. M. Thomson, and B. E. Stein. Corticotectal and corticostriatal projections from the frontal eye fields of the cat: an anatomical examination using wga-hrp. *Somatosens Mot Res*, 18(2):117–130, 2001.
- [MUM83] M. Mishkin, L.G. Ungerleider, and K.A. Macko. Objects vision and spatial vision: two cortical pathways. *Trends Neurosci*, (6):414–417, 1983.
- [MW47] O. Marburg and J. F. Warner. The pathway of the tectum of the midbrain in cat. *J. Nerv. Ment. Dis.*, 106:415–446, 1947.
- [MWS92] M. A. Meredith, M. T. Wallace, and B. E. Stein. Visual, auditory and somatosensory convergence in output neurons of the cat superior colliculus: multisensory properties of the tecto-reticulo-spinal projection. *Exp Brain Res*, 88(1):181–186, 1992.
- [NB75] U. Neisser and R Becklen. Selective looking: attending to visually specified events. *Cognitive Psychology*, (7):480–494, 1975.
- [Nei67] U. Neisser, editor. *Cognitive psychology*. Appleton-Century-Crofts, New York, 1967.
- [NLS88] D.P. Northmore, E.S. Levine, and G.E. Schneider. Behavior evoked by electrical stimulation of the hamster superior colliculus. *Exp Brain Res.*, 73(3):595–605, 1988.
- [Nor80] M. Norita. Neurons and synaptic patterns in the deep layers of the superior colliculus of the cat. a golgi and electron microscopic study. *J Comp Neurol.*, 190(1):29–48, Mar 1980.
- [NSM97] T. Niida, B. E. Stein, and J. G. McHaffie. Response properties of corticotectal and corticostriatal neurons in the posterior lateral suprasylvian cortex of the cat. *J Neurosci*, 17(21):8550–8565, Nov 1997.
- [OD99] K. O’Craven and P. Downing. fMRI evidence for objects as the units of attentional selection. *Nature*, 401:584–587, 1999.
- [OMCW04] J. O’Shea, N.G. Muggleton, A. Cowey, and V. Walsh. Timing of target discrimination in human frontal eye fields. *J Cogn Neurosci.*, 16(6):1060–7, Jul-Aug 2004.
- [OMS84] K. Ogasawara, J.G. McHaffie, and B.E. Stein. Two visual corticotectal systems in cat. *J Neurophysiol*, 52(6):1226–45, Dec 1984.
- [PA03] P.E. Patton and T.J. Anastasio. Modeling cross-modal enhancement and modality-specific suppression in multisensory neurons. *Neural Comput.*, 15(4):783–810, 2003.
- [PB97] C. K Peck and J. A. Baro. Discharge patterns of neurons in the rostral superior colliculus of cat: activity related to fixation of visual and auditory targets. *Exp Brain Res*, 113(2):291–302, 1997.
- [PB03] L. Petit and M.S. Beauchamp. Neural basis of visually guided head movements studied with fmri. *J Neurophysiol*, 89(5):2516–27, May 2003.

- [PBBA02] P.E. Patton, K. Belkacem-Boussaid, and T.J. Anastasio. Multimodality in the superior colliculus: an information theoretic analysis. *Cognitive Brain Research*, 14(6):10–19, 2002.
- [PBW93] C. K. Peck, J. A. Baro, and S. M. Warder. Sensory integration in the deep layers of superior colliculus. *Prog Brain Res*, pages 91–102, 1993.
- [PBW95] C. K. Peck, J. A. Baro, and S. M. Warder. Effects of eye position on saccadic eye movements and on the neuronal responses to auditory and visual stimuli in cat superior colliculus. *Exp Brain Res*, 103(2):227–242, 1995.
- [PDMNM05] Ch. Pierrot-Deseilligny, R.M. Muri, T. Nyffeler, and D. Milea. The role of the human dorsolateral prefrontal cortex in ocular motor behavior. *Annals of the New York Acad. Sci.*, (1039):239–51, Apr 2005.
- [Pec96] C. K. Peck. Visual-auditory integration in cat superior colliculus: implications for neuronal control of the orienting response. *Prog Brain Res*, 112:167–177, 1996.
- [PF92] A. Popper and R. Fay, editors. *The Mammalian Auditory Pathway: Neurophysiology*. Springer, New York, 1992.
- [PFR<sup>+</sup>04] P. Pietrini, M.L. Furey, E. Ricciardi, M.I. Gobbini, W.H. Wu, L. Cohen, M. Guazzelli, and J.V. Haxby. Beyond sensory images: Object-based representation in the human ventral pathway. *Proc Natl Acad Sci U S A.*, 101(15):5658–63, Apr 2004.
- [PK00] J. L. Pena and M. Konishi. Cellular mechanisms for resolving phase ambiguity in the owl’s inferior colliculus. *Proc Natl Acad Sci USA*, 97(22):11787–11792, 2000.
- [PKHG04] T.J. Park, A. Klug, M. Holinstat, and B. Grothe. Interaural level difference processing in the lateral superior olive and the inferior colliculus. *J Neurophysiol*, 92(1):289–301, Jul 2004.
- [PMA88] S.E. Petersen, F.M. Miezin, and J.M. Allman. Transient and sustained responses in four extrastriate visual areas of the owl monkey. *Exp Brain Res.*, 70(1):55–60, 1988.
- [Poh73] W. Pohl. Dissociation of spatial discrimination deficits following frontal and parietal lesions in monkeys. *J. Comp. Physiol. Psychol.*, (82):227–239, 1973.
- [Pri88] D. Price. *Psychological and Neural Mechanisms of Pain*. Raven, New York, 1988.
- [PSD80] M.I. Posner, C.R.R. Snyder, and B.J. Davidson. Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109:160–174, 1980.
- [PTY04] L. C. Populin, D.J. Tollin, and T. C. Yin. Effect of eye position on saccades and neuronal responses to acoustic stimuli in the superior colliculus of the behaving cat. *J Neurophysiol.*, 92(4):2151–67, Oct 2004.
- [PVAK96] J. L. Pena, S. Viète, Y. Albeck, and M. Konishi. Tolerance to sound intensity of binaural coincidence detection in the nucleus laminaris of the owl. *J Neurosci*, 16(21):7046–7054, Nov 1996.



- [PY98] L. C. Populin and T. C. Yin. Behavioral studies of sound localization in the cat. *J Neurosci*, 18(6):2147–60, 1998.
- [PY02] L. C. Populin and T. C. Yin. Bimodal interactions in the superior colliculus of the behaving cat. *J Neurosci*, 22(7):2826–34, 2002.
- [QAOR98] C. Quaia, H Aizawa, L.M. Optican, and Wurtz R.H. Reversible inactivation of monkey superior colliculus. ii. maps of saccadic deficits. *J Neurophysiol.*, 79(4):2097–110, Apr 1998.
- [RA97] R. Romand and P. Avan. Anatomical and functional aspects of the cochlear nucleus. In *The Central Auditory System* [ER97], pages 97–192.
- [Rad94] M. Radeau. Auditory-visual spatial interaction and modularity. *Curr. Psychol. Cogn.*, 13(1):117–23, Feb 1994.
- [RC93] M.H. Rowe and J.F. Cox. Spatial receptive-field structure of cat retinal w cells. *Vis Neurosci*, 10(4):765–79, Jul-Aug 1993.
- [REW99] M. Rucci, G. M. Edelman, and J. Wray. Adaption of orienting behavior: From the barn owl to a robotic system. *IEEE Transactions on Robotics and Automation*, 15(1):96–110, 1999.
- [RFC03] T. Ro, A. Farne, and E. Chang. Inhibition of return and the human frontal eye fields. *Exp Brain Res.*, 150(3):290–6, Jun 2003.
- [RG92] W.S. Rhode and S. Greenberg. Physiology of the cochlear nucleus. In Popper and Fay [PF92], pages 94–152.
- [RGA89] H.R. Rodman, C.G. Gross, and T.D. Albright. Afferent basis of visual response properties in area mt of the macaque. i. effects of striate cortex removal. *J Neurosci.*, 9(6):2033–50, Jun 1989.
- [RGA90] H.R. Rodman, C.G. Gross, and T.D. Albright. Afferent basis of visual response properties in area mt of the macaque. i. effects of superior colliculus removal. *J Neurosci.*, 10(4):1154–64, Apr 1990.
- [RH99] M.N. Radford and G.E. Hinton. A view to the EM algorithm that justifies incremental, sparse and other variants. In Jordan [Jor99], pages 355–368.
- [RM89] D.L. Robinson and J.W. McClurkin. The visual superior colliculus and pulvinar. *Rev. Oculomot. Res.*, 3:337–60, 1989.
- [RMF85] R.W. Rhoades, R.D. Mooney, and S.E. Fish. Subcortical projections of area 17 in the anophthalmic mouse. *Brain Res.*, 349(1-2):171–81, Jan 1985.
- [RMKP91] D.L. Robinson, J.W. McClurkin, C. Kertzman, and S.E. Petersen. Visual responses of pulvinar and collicular neurons during eye movements of awake, trained macaques. *J Neurophysiol.*, 66(2):485–96, Aug 1991.
- [RMR<sup>+</sup>89] R.W. Rhoades, R.D. Mooney, W.H. Rohrer, M.M. Nikolettseas, and S.E. Fish. Organization of the projection from the superficial to the deep layers of the hamster’s superior colliculus as demonstrated by the anterograde transport of phaseolus vulgaris leucoagglutinin. *J Comp Neurol.*, 283(1):54–70, May 1989.

- [Rou97] E.M. Rouiller. Functional organization of the auditory pathways. In *The Central Auditory System* [ER97], pages 3–96.
- [RP92] D.L. Robinson and S.E. Petersen. The pulvinar and visual salience. *Trends Neurosci.*, 15(4), Apr 1992.
- [RP95] M.H. Rowe and L.A. Palmer. Spatio-temporal receptive-field structure of phasic w cells in the cat retina. *Vis Neurosci*, 12(1):117–39, Jan-Feb 1995.
- [RT00] J. P. Rauschecker and B. Tian. Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc Natl Acad Sci USA*, 97(22):11800–6, Oct 2000.
- [RW76] D.L. Robinson and R.H. Wurtz. Use of an extraretinal signal by monkey superior colliculus neurons to distinguish real from self-induced stimulus movement. *J Neurophysiol.*, 39(4):852–70, Jul 1976.
- [RWE00] M. Rucci, J. Wray, and G. M. Edelman. Robust localization of auditory and visual targets in a robotic barn owl. *Journal of Robotics and Autonomous Systems*, 30(1-2):181–193, 2000.
- [RXMO99] S.E. Raiguel, K Xiao, D, V.L. Marcar, and G.A. Orban. Response latency of macaque area MT/V5 neurons and its relationship to stimulus parameters. *J Neurophysiol*, 82(4):1944–56, Oct 1999.
- [SCF04] S. Sdoia, A. Couyoumdjian, and F. Ferlazzo. Opposite visual field asymmetries for egocentric and allocentric spatial judgments. *Neuroreport*, 17(8):1303–5, Jun 2004.
- [Sch85] H.P. Schiller. A model for the generation of visually guided saccadic eye movements. In D. Rose and V.G. Dobson, editors, *Models of the visual cortex*, pages 62–70. Wiley, 1985.
- [Sch98] C. Schauer. Akustische Quellenlokalisierung auf Basis biologienaher Modelle von Hirnstrukturen mit spikenden Neuronen. Diplomarbeit, Technische Universität Ilmenau, 1998.
- [Sch02a] J.D. Schall. The neural selection and control of saccades by the frontal eye field. *Philos. Trans. R. Soc. Lond. - Biol. Sci.*, 357:1073–82, 2002.
- [Sch02b] C. Schauer. Verfahren und Anordnung zur passiven Schallortung. Offenlegungsschrift zum Patent: DE 10043055A1. Deutsches Patent- und Markenamt, 2002.
- [Sch02c] Brian J. Scholl, editor. *Objects and Attention*. Elsevier Science Publishers, Amsterdam, 2002.
- [Sch02d] Brian J. Scholl. Objects and attention: the state of the art. In *Objects and Attention* [Sch02c], pages 1–45.
- [SDR86] N. Sahibzada, P. Dean, and P. Redgrave. Movements resembling orientation or avoidance elicited by electrical stimulation of the superior colliculus in rats.. *J Neurosci.*, 6(3):723–33, Mar 1986.

- [SFCG01] D. L. Sparks, E.G. Freedman, L.L. Chen, and N.J. Gandhi. Cortical and subcortical contributions to coordinated eye and head movements. *Vision Res*, 41(25), 2001.
- [SG96] S. M. Sherman and R. W. Guillery. Functional organization of thalamocortical relays. *J Neurophysiol*, 76(3):1367–1395, Sep 1996.
- [SG03] C. Schauer and H.-M. Gross. A computational model of early auditory-visual integration. *Proc. of the 25th Pattern Recognition Symposium (DAGM 2003), Magdeburg. Lecture Notes in Computer Science*, 2781:362–369, 2003.
- [SG04] C. Schauer and H.-M. Gross. Design and optimization of Amari neural fields for early auditory-visual integration. In *Int. Joint Conf. on Neural Networks (IJCNN 2004), Budapest, Conference Proceedings 04CH37541C*, pages 2523–2528. IEEE, 2004.
- [SH98] J.D. Schall and D.P. Hanes. Neural mechanisms of selection and control of visually guided eye movements. *Neural Netw.*, 11(7), 1998.
- [Sha48] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423 & 623–656, Jul, Oct 1948.
- [Sha85] S. A. Shamma. Speech processing in the auditory system ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *J. Acoust. Soc.Am.*, 78(5):1622–1632, Nov 1985.
- [SHB00] A. J. Sach, N. I. Hill, and P. J. Bailey. Auditory spatial attention using interaural time differences. *J Exp Psychol Hum Percept Perform*, 26(2):717–729, 2000.
- [Shi03] S. Shipp. The functional logic of cortico-pulvinar connections. *Philosophical Transactions: Biological Sciences*, 358:1605–1624, Oct 2003.
- [SI04] Y. Saito and T. Isa. Laminar specific distribution of lateral excitatory connections in the rat superior colliculus. *J Neurophysiol.*, 92(6):3500, Dec 2004.
- [Sie70] W. M. Siebert. Frequency discrimination in the auditory system: place or periodicity mechanisms. *Proceedings of the IEEE*, 58:723–730, 1970.
- [SK90] S.M. Sherman and S. Koch. Thalamus. In G.M. Shepherd, editor, *The synaptic organisation of the brain*. Oxford University Press, Inc., 3. edition edition, 1990.
- [SL93] M. Slaney and F. R. Lyon. On the importance of time - a temporal representation of sound. In *Visual Representations of Speech Signals*, pages 95–116. 1993.
- [Sla98] M. Slaney. Auditory toolbox. Technical Report 1998-010, Interval Research Corporation, 1998.
- [SM93] B. E. Stein and M. A. Meredith. *The Merging Of The Senses*. The MIT Press, Cambridge, Massachusetts, 1993.
- [SMKB95] J.D. Schall, A. Morel, D.J. King, and J. Bullier. Topography of visual cortex connections with frontal eye field in macaque: convergence and segregation of processing streams. *J Neurosci.*, 15(6):4464–87, Jun 1995.

- [SP99] C. Schauer and P. Paschke. A spike-based model of binaural sound localization. *Int J Neural Syst*, 9(5):447–52, 1999.
- [Spa99] D. L. Sparks. Conceptual issues related to the role of the superior colliculus in the control of gaze. *Current Opinion in Neurobiology*, 9(6):698–707, Dec 1999.
- [SPG01a] C. Schauer, P. Paschke, and H.-M. Gross. Model and application of a binaural 360 degree sound localization system. In *IEEE-INNS Intern. Joint Conference on Neural Networks (IJCNN 2001)*, Washington, Conference Proceedings, pages 1132–1137. IEEE Omnipress, 2001.
- [SPG01b] C. Schauer, P. Paschke, and H.-M. Gross. A model of horizontal 360 degree object localization based on biaural hearing and monocular vision. *Intern. Conference on Artificial Neural Networks (ICANN 2001)*, Vienna. *Lecture Notes in Computer Science*, 2130:1141–1146, 2001.
- [SRZ00] D. L. Sparks, W. H. Rohrer, and Y. Zhang. The role of the superior colliculus in saccade initiation: a study of express saccades and the gap effect. *Vision Res*, 40(20):2763–2777, 2000.
- [SS00] S. Sudkamp and M. Schmidt. Response characteristics of neurons in the pulvinar of awake cats to saccades and to visual stimulation. *Exp. Brain Res.*, 133(2):209–18, Jul 2000.
- [SSG89] S. A. Shamma, N. M. Shen, and P. Gopalaswamy. Stereausis: binaural processing without neural delays. *J Acoust Soc Am*, 86(3):989–1006, 1989.
- [SST+96] J.W. Scannell, F. Sengpiel, M.J. Tovee, P.J. Benson, C. Blakemore, and M.P. Young. Visual motion processing in the anterior ectosylvian sulcus of the cat. *J Neurophysiol*, 76(2):895–907, Aug 1996.
- [ST95] M.S. Stern and C. Trahiotis. Models of binaural interaction. In Brian C.J. Moore, editor, *Handbook of Perception and Cognition, Volume 6: Hearing*, pages 347–386. Academic Press, San Diego, London, 1995.
- [ST03] P.H. Schiller and E.J. Tehovnik. Cortical inhibitory circuits in eye-movement generation. *Eur J Neurosci.*, 18(11):3127–33, Dec 2003.
- [Ste75] S. Stevens. *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects*. Wiley, New York, 1975.
- [Ste78] B. E. Stein. Nonequivalent visual, auditory, and somatic corticotectal influences in cat. *J Neurophysiol.*, 41(1):55–64, Jan 1978.
- [Ste98] B. E. Stein. Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. *Exp Brain Res*, 123(1-2):124–135, Nov 1998.
- [SW96] B. E. Stein and M. T. Wallace. Comparisons of cross-modality integration in midbrain and cortex. *Prog Brain Res*, 112:289–299, 1996.
- [SW00] M.A. Sommer and R. H. Wurtz. Composition and topographic organization of signals sent from the frontal eye field to the superior colliculus. *J Neurophysiol*, 83(4):1979–2001, Apr 2000.

- [SW04a] M.A. Sommer and R. H. Wurtz. What the brain stem tells the frontal cortex. ii. role of the sc-md-fef pathway in corollary discharge. *J Neurophysiol*, 91(3):1403–23, Mar 2004.
- [SW04b] M.A. Sommer and R. H. Wurtz. What the brain stem tells the frontal cortex. i. oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. *J Neurophysiol*, 91(3):1381–402, Mar 2004.
- [SWD<sup>+</sup>03] M.A. Schoenfeld, M. Woldorff, E. Duzel, H. Scheich, H.J. Heinze, and G.R. Mangun. Form-from-motion: Meg evidence for time course and processing sequence. *J Cogn Neurosci*, 15(2):157–72, Feb 2003.
- [SZPG00a] C. Schauer, T.P. Zahn, P. Paschke, and H.-M. Gross. Binaural sound localization in an artificial neural network. In *IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP 2000), Istanbul, vol. 2*, pages 865–868. IEEE Press, 2000.
- [SZPG00b] C. Schauer, T.P. Zahn, P. Paschke, and H.-M. Gross. A model of binaural sound localization in real environs. In *Proc. 15th European Meeting on Cybernetics and Systems Research (EMCSR 2000), Vienna, 2000, vol. 2*, pages 559–563. Austrian Society for Cybernetic Studies, 2000.
- [TB05] K.G. Thompson and N.P. Bichot. A visual salience map in the primate frontal eye field. *Prog Brain Res.*, 147:251–62, 2005.
- [TDMK01] T. P. Trappenberg, M. C. Dorris, D. P. Munoz, and R. M. Klein. A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *J Cogn Neurosci*, 13(2):256–271, Feb 2001.
- [TDPC05] E. Tardif, B. Delacuisine, A. Probst, and S. Clarke. Intrinsic connectivity of human superior colliculus. *Exp. Brain Res.*, (Epub ahead of print), Jul 20 2005.
- [TF04] J. G. Taylor and N. Fragopanagos. Modelling human attention and emotions. In *Proceedings Int. Joint Conf. on Neural Networks - IJCNN - 2004*. IEEE Computer Society, 2004.
- [TG80] Anne M. Treisman and Garry Gelade. A feature-integration Theory of Attention. *Cognitive Psychology*, 12:97–136, 1980.
- [TMS<sup>+</sup>01] A.S. Tolia, T. Moore, E.J. Smirnakis, S.M. ans Tehovnik, A.G. Siapas, and P.H. Schiller. Eye movements modulate visual receptive fields of v4 neurons. *Neuron*, 29(3):757–67, Mar 2001.
- [TRSL80] A. Tortelly, F. Reinoso-Suarez, and A. Llamas. Projections from non-visual cortical areas to the superior colliculus demonstrated by retrograde transport of hrp in the cat. *Brain Res.*, 188(2):543–9, Apr 1980.
- [TSE92] G Tononi, O. Sporns, and G.M. Edelman. Reentry and the problem of integrating multiple cortical areas: simulation of dynamic integration in the visual system. *Cereb Cortex.*, 2(4):310–35, Jul-Aug 1992.

- [UN96] M. Usher and E. Niebur. Modelling the temporal dynamics of it neurons in visual search: A mechanism for top-down selective attention. *Journal of Cognitive Neuroscience*, 8:311–327, 1996.
- [VEAO94] D.C. Van Essen, C.H. Anderson, and B.A. Olshausen. Dynamic routing strategies in sensory, motor, and cognitive processing. In *Large Scale Neuronal Theories of the Brain*. 1994.
- [VGBP01] J. Vermaak, M. Gangnet, A. Blake, and P. Pérez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Proceedings of the 11th International Conference on Image Analysis and Processing - ICIAP 2001*. IEEE Computer Society, 2001.
- [VK89] S.F. Volman and M. Konishi. Spatial selectivity and binaural responses in the inferior colliculus of the great horned owl. *J Neurosci*, 9(9):3083–96., Sep 1989.
- [VOJ89] A.J. Van Opstal and Gisbergen J.A. A nonlinear model for collicular spatial interactions underlying the metrical properties of electrically elicited saccades. *Biol. Cybern.*, 60(3):171–83, 1989.
- [vSFVM97] A. van Schaik, E. Fragüere, E. Vittoz, and R. Meddis. Analogue vlsi building blocks for an electronic auditory pathway. In *11th International Symposium on Hearing, Grantham, UK, 1st - 6th August 1997*, pages 157–163, 1997.
- [vSJC99] A. van Schaik, C. Jin, and S. Carlile. Human localisation of band-pass filtered noise. *Int J Neural Syst*, 9(5):441–6, Oct 1999.
- [War82] R. Warren. *Auditory perception: a new synthesis*. Pergamon, New York, 1982.
- [WBS96] T. Wesarg, B. Brückner, and C. Schauer. A three-stage model of speech processing and recognition in the auditory system. In Amari et.al., editor, *Progress in Neural Information Processing, Vol. 1*, pages 306–310. Springer, Singapore, 1996.
- [WCF89] Jeremy M. Wolfe, Kyle R. Cave, and Susan L. Franzel. Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology; Human Perception and Performance*, 15(3):419–433, 1989.
- [WFG95] M.F. Walker, E.J. Fitzgibbon, and M.E. Goldberg. Neurons in the monkey superior colliculus predict the visual result of impending saccadic eye movements. *J Neurophysiol.*, 73(5):1988–2003, May 1995.
- [Wie59] T.N. Wiesel. Recording inhibition and excitation in the cat’s retinal ganglion cells with intracellular electrodes. *Nature.*, 183(4):264–5, Jan 1959.
- [WKR<sup>+</sup>90] G.W. Westby, K.A. Keay, P. Redgrave, P. Dean, and M. Bannister. Output pathways from the rat superior colliculus mediating approach and avoidance have different sensory properties. *Exp Brain Res.*, 81(3):626–38, 1990.
- [WMS93] M. T. Wallace, M. A. Meredith, and B. E. Stein. Converging influences from visual, auditory, and somatosensory cortices onto output neurons of the superior colliculus. *J Neurophysiol*, 69(6):1797–1809, Jun 1993.

- [WMS96] L. K. Wilkinson, M. A. Meredith, and B. E. Stein. The role of anterior ectosylvian cortex in cross-modality orientation and approach behavior. *Exp Brain Res*, 112(1):1–10, Nov 1996.
- [WMS98] M. T. Wallace, M. A. Meredith, and B. E. Stein. Multisensory integration in the superior colliculus of the alert cat. *J Neurophysiol*, 80(2):1006–1010, Aug 1998.
- [WO98] W. W. Wilson and W. E. O’Neill. Auditory motion induces directionally dependent receptive field shifts in inferior colliculus neurons. *J Neurophysiol*, 79(4):2040–2062, Apr 1998.
- [Wol91] Siegbert Wolf. *Untersuchungen zur Lokalisation von Schallquellen in geschlossenen Räumen*. PhD dissertation, Ruhr-Universität Bochum, Fakultät für Elektrotechnik, 1991.
- [Wol94] Jeremy M. Wolfe. Guided Search 2.0. *Psychonomic Bulletin & Review*, 1(2):203–238, 1994.
- [WPHS04] M. T. Wallace, T.J. Jr. Perrault, W.D. Hairston, and B. E. Stein. Visual experience is necessary for the development of multisensory integration. *J Neurosci.*, 24(43):9580–4, Oct 2004.
- [WRH<sup>+</sup>04] M.T. Wallace, G.E. Roberson, W.D. Hairston, B.E. Stein, J.W. Vaughan, and J.A. Schirillo. Unifying multisensory signals across time and space. *Exp Brain Res.*, 158(2):252–8, Sep 2004.
- [WS94] M. T. Wallace and B. E. Stein. Cross-modal synthesis in the midbrain depends on input from cortex. *J Neurophysiol*, 71(1):429–432, Jan 1994.
- [WSG05] M.M Walton, D.L. Sparks, and N.J. Gandhi. Simulations of saccade curvature by models that place superior colliculus upstream from the local feedback loop. *J Neurophysiol*, 93(4):2354–8., Apr 2005.
- [WTK87] H. Wagner, T. Takahashi, and M. Konishi. Representation of interaural time difference in the central nucleus of the barn owl’s inferior colliculus. *J Neurosci.*, 7(10):3105–16, Oct 1987.
- [WW93] F.R. Waugh and R. M. Westervelt. Analog neural networks with local competition. i. dynamics and stability. *Phys. Rev.*, E 47(6):45244536, Jun 1993.
- [WWB<sup>+</sup>04] W.J. Waleszczyk, C. Wang, G. Benedek, W. Burke, and B Dreher. Motion sensitivity in cat’s superior colliculus: contribution of different visual processing channels to response properties of collicular neurons. *Acta Neurobiol Exp*, 64(2):209–28, 2004.
- [WWS96] M. T. Wallace, L. K. Wilkinson, and B. E. Stein. Representation and integration of multiple sensory inputs in primate superior colliculus. *J Neurophysiol*, 76(2):1246–1266, Aug 1996.
- [YC90] T.C. Yin and J.C. Chan. Interaural time sensitivity in medial superior olive of cat. *J Neurophysiol.*, 64(2):465–88, Aug 1990.

- [YG95] A. Yuille and D. Geiger. Winner-take-all mechanisms. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1056–1060. The MIT Press, Cambridge, 1995.
- [Yin94] T.C. Yin. Physiological correlates of the precedence effect and summing localization in the inferior colliculus of the cat. *J Neurosci.*, 14(9):5170–86, Sep 1994.
- [YP97] L. Yao and C. K. Peck. Saccadic eye movements to visual and auditory targets. *Exp Brain Res*, 115(1):25–34, Jun 1997.
- [Zah03] T.P. Zahn. *Neural Architecture for Echo Suppression during Sound Localization based on Spiking Neural Cell Models*. PhD thesis, Technische Universität Ilmenau, Oct 2003.
- [ZDD02] D.N. Zotkin, R. Duraiswami, and L.S. Davis. Joint audio-visual tracking using particle filters. *Eurasip JASP*, (11):1154–1164, 2002.
- [Zek93] S. Zeki, editor. *A vision of the brain*. Blackwell Scientific Publications, Oxford, 1993.
- [Zen94] Hans-Peter Zenner. *Hören: Physiologie, Biochemie, Zell- und Neurobiologie*. Thieme, Stuttgart, New York, 1994.
- [ZK98] J. Zihl and H-O. Karnath. Zerebrale Sehstörungen. In T. Brandt, J. Dichgans, and H.C. Diener, editors, *Therapie und Verlauf neurologischer Erkrankungen.*, pages 226–236. Verlag Kohlhammer, Stuttgart, 1998.
- [Zor95] Steven F. Zornetzer. *An introduction to neural and electronic networks*. San Diego: Acad. Press, 1995.
- [ZvC79] J. Zihl and D. von Cramon. The contribution of the 'second' visual system to directed visual attention in man. *Brain.*, 102(4):835–56, Dec 1979.
- [ZVvO04] M.P. Zwiers, H. Versnel, and A.J. van Opstal. Involvement of monkey inferior colliculus in spatial hearing. *J. Neurosci.*, 24(17):4145–56, Apr 2004.