Multimodal Dynamics:
Self-Supervised Learning in Perceptual and Motor Systems

by

Michael Harlan Coen

S.M. Computer Science, Massachusetts Institute of Technology (1994)
S.B. Computer Science, Massachusetts Institute of Technology (1991)

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2006

Signature of Author ______________________________

Department of Electrical Engineering and Computer Science
May 25, 2006

Certified by ______________________________

Whitman Richards
Professor of Brain and Cognitive Sciences
Thesis Supervisor

Certified by ______________________________

Howard Shrobe
Principal Research Scientist, Computer Science
Thesis Supervisor

Accepted by ______________________________

Arthur Smith
Chair, Committee on Graduate Students

Multimodal Dynamics: Self-Supervised Learning in Perceptual and Motor Systems

by

Michael Harlan Coen

Submitted to the Department of Electrical Engineering and Computer Science on May 25, 2006 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

ABSTRACT

This thesis presents a self-supervised framework for perceptual and motor learning based upon correlations in different sensory modalities. The brain and cognitive sciences have gathered an enormous body of neurological and phenomenological evidence in the past half century demonstrating the extraordinary degree of interaction between sensory modalities during the course of ordinary perception. We develop a framework for creating artificial perceptual systems that draws on these findings, where the primary architectural motif is the cross-modal transmission of perceptual information to enhance each sensory channel individually. We present self-supervised algorithms for learning perceptual grounding, intersensory influence, and sensory-motor coordination, which derive training signals from internal cross-modal correlations rather than from external supervision. Our goal is to create systems that develop by interacting with the world around them, inspired by development in animals.

We demonstrate this framework with: (1) a system that learns the number and structure of vowels in American English by simultaneously watching and listening to someone speak. The system then cross-modally clusters the correlated auditory and visual data. It has no advance linguistic knowledge and receives no information outside of its sensory channels. This work is the first unsupervised acquisition of phonetic structure of which we are aware, outside of that done by human infants. (2) a system that learns to sing like a zebra finch, following the developmental stages of a juvenile zebra finch. It first learns the song of an adult male and then listens to its own initially nascent attempts at mimicry through an articulatory synthesizer. In acquiring the birdsong to which it was initially exposed, this system demonstrates self-supervised sensorimotor learning. It also demonstrates afferent and efferent equivalence – the system learns motor maps with the same computational framework used for learning sensory maps.

Thesis Supervisor: Whitman Richards
Title: Professor of Brain and Cognitive Sciences

Thesis Supervisor: Howard Shrobe
Title: Principal Research Scientist, EECS

Acknowledgements

It is safe to say that I have spent more time than most as a student at MIT. After finishing my S.B. and S.M. degrees, I spent almost five years running the Artificial Intelligence Laboratory's Intelligent Room project. At the beginning of 2000, I left MIT to join a start-up company on Wall Street in NYC. I returned to CSAIL in January of 2005 and have completed this thesis in the interim.

None of the work in this thesis would exist had I not received the constant encouragement, mentoring, and friendship of Whitman Richards. My debt to him is enormous, and I am at a loss for words to thank him properly; my relationship with Whitman has been the most rewarding of my academic career. Among the great joys of finishing is that we can now collaborate on the many projects dreamed up over our weekly coffees in the past year and a half.

Howard Shrobe provided tremendous encouragement and welcomed me back to MIT with open arms, graciously and generously. Without his support too, this work could not have come to fruition.

In my many years at MIT, a number of faculty members have changed my view of the world. None has had a greater impact than Rodney Brooks, whose approach to AI has thoroughly informed my own in the most basic of ways, the evidence of which can be found throughout this thesis. I have also learned a tremendous amount about science, life, problem solving, and clear thinking from Patrick Winston, Robert Berwick, and Gerald Sussman. They have all provided constant encouragement for which I am deeply and forever grateful. Patrick especially has mentored me through these last months of graduate school, and I continue to learn something new from him at least once a week.

I have also greatly benefited from interactions with many other MIT faculty members. These include Boris Katz, Randall Davis, Hal Abelson, Lynn Stein, Gian-Carlo Rota, and George Boolos. I am indebted to Jim Glass and Eric Grimson, who managed to find time in their schedules to be on my committee and provided invaluable feedback. I am also grateful to the fellow members of the BICA project, particularly Sajit Rao, who have provided a regular source of stimulation, excitement, debate, camaraderie, and encouragement.

In my long tenure here, a number of friends have made a tremendous difference in my life. I particularly want to thank Rajesh and Becky Kasturirangan, Upendra Chaudhari, Stacy Ho, Ye Gu, Jeff Kuo, Yuri Bendana, Andy Cytron, Krzysztof Gajos, Brenton Phillips, Joanna Bryson and Will Lowe, and Kimberley and Damon Asher. I also thank the many students who were involved in the Intelligent Room project, without whom it could not have existed. Marilyn Pierce in the EECS Graduate Office made my transition back to student status far more pleasant and streamlined.

I spent many years living in the community associated with the Chai Odom synagogue in Brookline, MA. They provided a constant source of love and support, and I particularly thank Rabbi and Rebbetzin Dovid Moskowitz.

Above all, I thank my family. My parents and siblings never gave up hope I would finish, even when I was none too sure myself. And foremost among all, I thank my wife Aimee, who patiently tolerated my constant working late into the night. To her, this thesis and my life are dedicated.

In memory of my beloved Grandmother, Mrs. Dora Estrin צ״ל
March 19, 1910 – September 7, 2003

Funding Acknowledgements

The work in this thesis was supported by the United States Government, for which I am extremely grateful, through AFOSR under contract #F49610-03-1-0213 and through AFRL under contract #FA8750-05-2-0274. Thanks to the DARPA/IPTO BICA program and to AFRL.

Contents

> We have sat around for hours and wondered how you look. If you have closed your senses upon silk, light, color, odor, character, temperament, you must be by now completely shriveled up. There are so many minor senses, all running like tributaries into the mainstream of love, nourishing it.
>
> The Diary of Anais Nin (1943)

> He plays by sense of smell.
>
> Tommy, The Who (1969)

Chapter 1

Introduction

This thesis presents a unified framework for perceptual and motor learning based upon correlations in different sensory modalities. The brain and cognitive sciences have gathered a large body of neurological and phenomenological evidence in the past half century demonstrating the extraordinary degree of interaction between sensory modalities during the course of ordinary perception. We present a framework for artificial perceptual systems that draws on these findings, where the primary architectural motif is the cross-modal transmission of perceptual information to structure and enhance sensory channels individually. We present self-supervised algorithms for learning *perceptual grounding*, *intersensory influence*, and *sensorimotor coordination*, which derive training signals from internal cross-modal correlations rather than from external supervision. Our goal is to create perceptual and motor systems that develop by interacting with the world around them, inspired by development in animals.

Our approach is to formalize mathematically an insight in Aristotle's *De Anima* (350 B.C.E.), that *differences in the world are only detectable because different senses perceive the same world events differently*. This implies both that sensory systems need

A glossary of technical terms is contained in Appendix 1. Our usage of the word "sense" is defined in §1.5.

some way to share their different perspectives on the world and that they need some way to incorporate these shared influences into their own internal workings.

We begin with a computational methodology for *perceptual grounding*, which addresses the first question that any natural (or artificial) creature faces: *what different things in the world am I capable of sensing?* This question is deceptively simple because a formal notion of what makes things different (or the same) is non-trivial and often elusive. We will show that animals (and machines) can learn their perceptual repertoires by simultaneously correlating information from their different senses, even when they have no advance knowledge of what events these senses are individually capable of perceiving. In essence, by *cross-modally* sharing information between different senses, we demonstrate that sensory systems can be perceptually grounded by mutually bootstrapping off each other. As a demonstration of this, we present a system that learns the number (and formant structure) of vowels in American English, simply by watching and listening to someone speak and then cross-modally clustering the accumulated auditory and visual data. The system has no advance knowledge of these vowels and receives no information outside of its sensory channels. This work is the first unsupervised acquisition of phonetic structure of which we are aware, at least outside of that done by human infants, who solve this problem easily.

The second component of this thesis naturally follows perceptual grounding. Once an animal (or a machine) has learned the range of events it can detect in the world, *how does it know what it's perceiving at any given moment?* We will refer to this as *perceptual interpretation*. Note that grounding and interpretation are different things. By way of analogy to reading, one might say that *grounding* provides the dictionary and *interpretation* explains how to disambiguate among possible word meanings. More formally, grounding is an ontological process that defines what is perceptually knowable, and interpretation is an algorithmic process that describes how perceptions are categorized within a grounded system. We will present a novel framework for perceptual interpretation called *influence networks* (unrelated to a formalism know as *influence diagrams*) that blurs the distinctions between different sensory channels and allows them to influence one another while they are in the midst of perceiving. Biological perceptual systems share cross-modal information routinely and opportunistically (Stein and Meredith 1993, Lewkowicz and Lickliter 1994, Rock 1997, Shimojo and Shams 2001, Calvert et al. 2004, Spence and Driver 2004); *intersensory influence* is an essential component of perception but one that most artificial perceptual systems lack in any meaningful way. We argue that this is among the most serious shortcomings facing them, and an engineering goal of this thesis is to propose a workable solution to this problem.

The third component of this thesis enables sensorimotor learning using the first two components, namely, perceptual grounding and interpretation. This is surprising because one might suppose that motor activity is fundamentally different than perception. However, we take the perspective that motor control can be seen as perception *backwards*. From this point of view, we imagine that – in a notion reminiscent of a Cartesian theater – an animal can "watch" the activity in its own motor cortex, as if it were a privileged form of *internal* perception. Then for any motor act, there are two associated perceptions – the *internal* one describing the generation of the act and the *external* one describing the self-observation of the act. The perceptual grounding framework described above can then *cross-modally ground* these internal and external perceptions with respect to one another. The power of this mechanism is that it can learn mimicry, an essential form of behavioral learning (see the developmental sections of Meltzoff and Prinz 2002) where one animal acquires the ability to imitate some aspect of another's activity, constrained by the capabilities and dynamics of its own sensory and motor systems. We will demonstrate sensorimotor learning in our framework with an artificial system that learns to sing like a zebra finch by first listening to a real bird sing and then by learning from its own initially uninformed attempts to mimic it.

This thesis has been motivated by surprising results about how animals process sensory information. These findings, gathered by the brain and cognitive sciences communities primarily over the past 50 years, have challenged century long held notions about how the brain works and how we experience the world in which we live. We argue that current approaches to building computers that perceive and interact with the real, human world are largely based upon developmental and structural assumptions, tracing back several hundred years, that are no longer thought to be descriptively or biologically accurate. In particular, the notion that perceptual senses are in functional isolation – that they do not internally structure and influence each other – is no longer tenable, although we still build artificial perceptual systems as if it were.

1.1 Computational Contributions

This thesis introduces three new computational tools. The first is a mathematical model of *slices*, which are a new type of data structure for representing sensory inputs. Slices are topological manifolds that encode dynamic perceptual states and are inspired by surface models of cortical tissue (Dale et al. 1999, Fischl et al. 1999, Citti and Sarti 2003, Ratnanather et al. 2003). They can represent both symbolic and numeric data and provide a natural foundation for aggregating and correlating information. Slices represent the data in a perceptual system, but they are also *amodal*, in that they are not specific to any sensory representation. For example, we may have slices containing visual information and other slices containing auditory information, but it may not be possible to distinguish them further without additional information. In fact, we can equivalently represent either sensory or motor information within a slice. This generality will allow us to easily incorporate the learning of motor control into what is initially a perceptual framework.

The second tool is an algorithm for *cross-modal clustering,* which is an unsupervised technique for organizing slices based on their spatiotemporal correlations with other slices. These correlations exist because an event in the world is simultaneously – but differently – perceived through multiple sensory channels in an observer. The hypothesis underlying this approach is that the world has regularities – natural laws tend to correlate physical properties (Thompson 1917, Richards 1980, Mumford 2004) – and biological perceptory systems have evolved to take advantage of this. One may contrast this with mathematical approaches to clustering where some knowledge of the clusters, e.g., how many there are or their distributions, must be known a priori in order to derive them. Without knowing these parameters in advance, many algorithmic clustering techniques may not be robust (Kleinberg 2002, Still and Bialek 2004). Assuming that in many circumstances animals cannot know the parameters underlying their perceptual inputs, how can they learn to organize their sensory perceptions? Cross-modal clustering answers this question by exploiting naturally occurring intersensory correlations.

The third tool in this thesis is a new family of models called *influence networks* (Figure 1.1). Influence networks use slices to interconnect independent perceptual systems, such as those illustrated in the classical view in Figure 1.1a, so they can influence one another during perception, as proposed in Figure 1.1b. Influence networks dynamically modify percepts within these systems to effect influence among their different components. The influence is designed to increase perceptual accuracy within individual perceptual channels by incorporating information from other co-occurring senses. More formally, influence networks are designed to move ambiguous perceptual inputs into easily recognized subsets of their representational spaces. In contrast with approaches taken in engineering what are typically called *multimodal systems*, influence networks are not intended to create high-level joint perceptions. Instead, they share sensory information across perceptual channels to increase local perceptual accuracy within the individual perceptual channels themselves. As we discuss in Chapter 6, this type of cross-modal perceptual reinforcement is ubiquitous in the animal world.

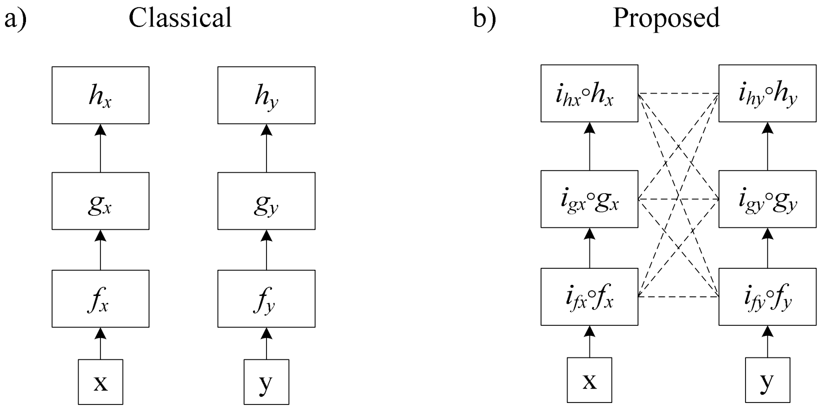


Figure 1.1– Adding an influence network to two preexisting systems. We start in (a) with two pipelined networks that independently compute separate functions. In (b), we compose on each function a corresponding *influence function*, which dynamically modifies its output based on activity at the other influence functions. The interaction among these influence functions is described by an *influence network,* which is defined in Chapter 5. The parameters describing this network can be found via unsupervised learning for a large class of perceptual systems, due to correspondences in the physical events that generate the signals they perceive and to the evolutionary incorporation of these regularities into the biological sensory systems that these computational systems model. Note influence networks are distinct from an unrelated formalism called influence diagrams.

1.2 Theoretic Contributions

The work presented here addresses several important problems. From an engineering perspective, it provides a principled, neurologically informed approach to building complex, interactive systems that can learn through their own experiences. In perceptual domains, it answers a fundamental question in mathematical clustering: *how should an unknown dataset be clustered?* The connection between clustering and perceptual grounding follows from the observation that learning to perceive is learning to organize perceptions into meaningful categories. From this perspective, asking what an animal can perceive is equivalent to asking how it should cluster its sensory inputs. This thesis presents a *self-supervised* approach to this problem, meaning our sub-systems derive feedback from one another cross-modally rather than rely on an external tutor such as a parent (or a programmer). Our approach is also highly nonparametric, in that it presumes neither that the number of clusters nor their distributions are known in advance, conditions which tend to defy other algorithmic techniques. The benefits of self-supervised learning in perceptual and motor domains are enormous because engineered approaches tend to be ad hoc and error prone; additionally, in sensorimotor learning we generally have no adequate models to specify the desired input/output behaviors for our systems. The notion of *programming by example* is nowhere truer than in the developmental mimicry widespread in animal kingdom (Meltzoff and Prinz 2002), and this work is a step in that direction for artificial sensorimotor systems.

Furthermore, this thesis suggests that not only do senses influence each other during perception, which is well established, it also proposes that *perceptual channels cooperatively structure their internal representations*. This mutual structuring is a basic feature in our approach to perceptual grounding. We argue, however, that it is not simply an epiphenomenon; rather, it is a fundamental component of perception itself, because *it provides representational compatibility for sharing information cross-modally* during higher-level perceptual processing. The inability to share perceptual data is one of the major shortcomings in current engineered approaches to building interactive systems.

Finally, within this framework, we will address three questions that are basic to developing a coherent understanding of cross-modal perception. They concern both process and representation and raise the possibility that unifying (i.e. meta-level) principles might govern intersensory function:

1) Can the senses be perceptually grounded by bootstrapping off each other? Is shared experience sufficient for learning how to categorize sensory inputs?
2) How can seemingly different senses share information? What representational and computational restrictions does this place upon them?
3) Could the development of motor control use the same mechanism? In other words, can there be afferent and efferent equivalence in learning?

1.3 A Brief Motivation

The goal of this thesis is to propose a design for artificial systems that more accurately reflects how animal brains appear to process sensory inputs. In particular, we argue against *post-perceptual* integration, where the sensory inputs are separately processed in isolated, increasingly abstracted pipelines and then merged in a final integrative step as in Figure 1.2. Instead, we argue for *cross-modally integrated perception*, where the senses share information during perception that synergistically enhances them individually, as in Figure 1.1b. The main difficulty with the post-perceptual approach is that integration happens after the individual perceptions are generated. Integration occurs *after* each perceptual subsystem has already "decided" what it has perceived, when it is too late for intersensory influence to affect the individual, concurrent perceptions. This is due to information loss from both vector quantization and the explicit abstraction fundamental to the pipeline design. Most importantly, these approaches also preclude cooperative perceptual grounding; the bootstrapping provided by cross-modal clustering cannot occur when sensory systems are independent. These architectures are therefore also at odds with developmental approaches to building interactive systems.

Not only is the post-perceptual approach to integration biologically implausible from a scientific perspective, it is poor engineering as well. The real world is inherently multimodal in a way that most modern artificial perceptual systems do not capture or take advantage of. Isolating sensory inputs while they are being processed prevents the lateral sharing of information across perceptual channels, even though these sensory inputs are inherently linked by the physics of the world that generates them. Furthermore, we will argue that the co-evolution of senses within an individual species provided evolutionary pressure towards representational and algorithmic compatibilities essentially unknown in modern artificial perception. These issues are examined in detail in Chapters 6.

Our work is computationally motivated by Gibson (1950, 1987), who viewed perception as an external as well as an internal event, by Brooks (1986, 1991), who elevated perception onto an equal footing with symbolic reasoning, and by Richards (1988), who described how to exploit regularities in the world to make learning easier. The recursive use of a perceptual mechanism to enable sensorimotor learning in Chapter 4 is a result of our exposure to the ideas of Sussman and Abelson (1983).

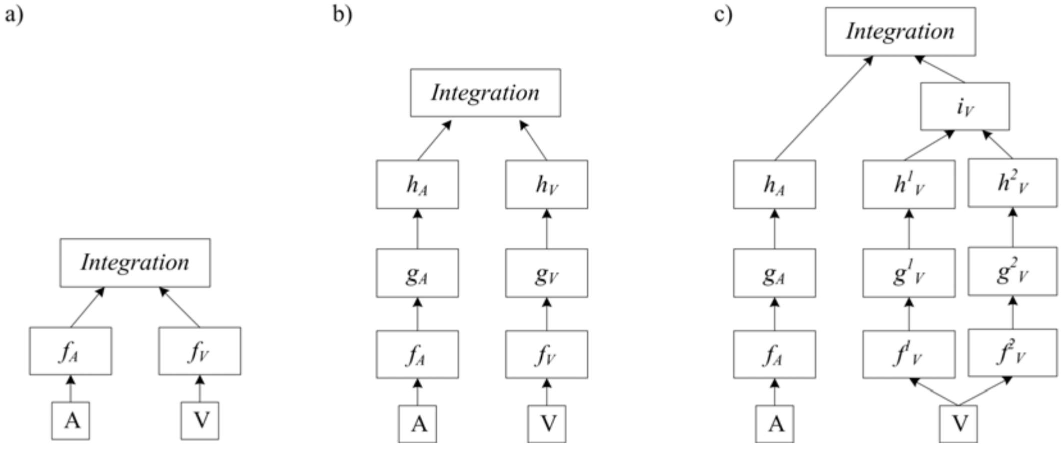


Figure 1.2 – Classical approaches to post-perceptual integration in traditional multimodal systems. Here, auditory (A) and visual (V) inputs pass through specialized unimodal processing pathways and are combined via an integration mechanism, which creates multimodal perceptions by extracting and reconciling data from the individual channels. Integration can happen earlier (a) or later (b). Hybrid architectures are also common. In (c), multiple pathways process the visual input and are pre-integrated before the final integration step; for example, the output of this preintegration step could be spatial localization derived solely through visual input. This diagram is modeled after (Stork and Hennecke 1996).

1.4 Demonstrations

The framework and its instantiation will be evaluated by a set of experiments that explore *perceptual grounding*, *perceptual interpretation*, and *sensorimotor learning*. These will be demonstrated with:

1) **Phonetic learning**: We present a system that learns the number and formant structure of vowels (monophthongs) in American English, simply by watching and listening to someone speak and then cross-modally clustering the accumulated auditory and visual data. The system has no advance knowledge of these vowels and receives no information outside of its sensory channels. This work is the first unsupervised machine acquisition of phonetic structure of which we are aware.

2) **Speechreading**: We incorporate an *influence network* into the cross-modally clustered slices obtained in Experiment 1 to increase the accuracy of perceptual classification within the slices individually. In particular, we demonstrate the ability of influence networks to move ambiguous perceptual inputs to unambiguous regions of their perceptual representational spaces.

3) **Learning birdsong**: We will demonstrate self-supervised sensorimotor learning with a system that learns to mimic a Zebra Finch. The system is directly modeled on the dynamics of how male baby finches learn birdsong from their fathers (Tchernichovski et al. 2004, Fee et al. 2004). Our system first listens to an adult finch and uses cross-modal clustering to learn *songemes*, primitive units of bird song that we propose as an avian equivalent of phonemes. It then uses a vocalization synthesizer to generate its own nascent birdsong, guided by random exploratory motor behavior. By listening to itself sing, the system organizes the motor maps generating its vocalizations by cross-modally clustering them with respect to the previously learned *songeme* maps of its parent. In this way, it learns to generate the same sounds to which it was previously exposed. Finally, we incorporate a standard hidden Markov model into this system, to model the

temporal structure and thereby combine songemes into actual birdsong. The Zebra Finch is a particularly suitable species to use for guiding this demonstration, as each bird essentially sings a single unique song accompanied by minor variations.

We note that the above examples all use real data, gathered from a real person speaking and from a real bird singing. We also present results on a number of synthetic datasets drawn from a variety of mixture distributions to provide basic insights into the algorithms and *slice* data structure work. Finally, we believe it is possible to allow the computational side of this question to inform the biological one, and we will analyze the model, in its own right and in light of these results, to explore its algorithmic and representational implications for biological perceptual systems, particularly from the perspective of how sharing information restricts the modalities individually.

1.5 What Is a "Sense?"

Although Appendix 1 contains a glossary of technical terms, one clarification is so important that it deserves special mention. We have repeatedly used the word *sense*, e.g., sense, sensory, intersensory, etc., without defining what a *sense* is. One generally thinks of a sense as the perceptual capability associated with a distinct, usually external, sensory organ. It seems quite natural to say vision is through the eyes, touch is through the skin, etc. (Notable exceptions include proprioception – the body's sense of internal state – which is somewhat more difficult to localize and vestibular perception, which occurs mainly in the inner ear but is not necessarily experienced there.) However, this coarse definition of *sense* is misleading.

Each sensory organ provides an entire class of sensory capabilities, which we will individually call *modes*. For example, we are familiar with the *bitterness* mode of taste, which is distinct from other taste modes such as *sweetness*. In the visual system, *object segmentation* is a mode that is distinct from *color perception*, which is why we can appreciate black and white photography. Most importantly, individuals may lack particular modes *without other modes in that sense being affected* (e.g., Wolfe 1983), thus demonstrating they are phenomenologically independent. For example, people who like broccoli are insensitive to the taste of the chemical *phenylthiocarbamide* (Drayna et al. 2003); however, we would not say these people are unable to taste – they are simply missing an individual taste mode. There are a wide variety of visual agnosias that selectively affect visual experience, e.g., *simultanagnosia* is the inability to perform visual object segmentation, but we certainly would not consider a patient with this deficit to be blind, as it leaves the other visual processing modes intact.

Considering these fine grained aspects of the senses, we allow intersensory influence to happen between modes even within the same sensory system, e.g., entirely within vision, or alternatively, between modes in different sensory systems, e.g., in vision and audition. Because the framework presented here is *amodal*, i.e., not specific to any sensory system or mode, it treats both of these cases equivalently.

1.6 Roadmap

Chapter 2 sets the stage for the rest of this thesis by visiting an example stemming from the 1939 World's Fair. It intuitively makes clear what we mean by perceptual grounding and interpretation, which until now have remained somewhat abstract.

Chapter 3 presents our approach to perceptual grounding by introducing *slices*, a data structure for representing sensory information. We then define our algorithm for cross-modal clustering, which autonomously learns perceptual categories within slices by considering how the data within them co-occur. We demonstrate this approach in learning the vowel structure of American English by simultaneously watching and listening to a person speak. Finally, we examine and contrast related work in unsupervised clustering with our approach.

Chapter 4 builds upon the results in Chapter 3 to present our approach to perceptual interpretation. We incorporate the temporal dynamics of sensory perception by treating slices as *phase spaces* through which sensory inputs move during the time windows corresponding to percept formation. We define a dynamic activation model on slices and interconnect them through an *influence network*, which allows different modes to influence each other's perceptions dynamically. We then examine using this framework to disambiguate simultaneous audio-visual speech inputs. Note that this mathematical chapter may be skipped on a cursory reading of this thesis.

Chapter 5 builds upon the previous two chapters to define our architecture for sensorimotor learning, based on a Cartesian theater. Our system simultaneously "watches" its internal motor activity while it observes the effects of its own actions externally. Cross-modal clustering then allows it to ground its motor maps using previously clustered perceptual maps. This is possible because slices can equivalently contain perceptual or motor data, and in fact, slices do not "know" what kind of data they contain. The principle example in this chapter is the acquisition of species-specific birdsong.

Chapter 6 connects the work in the computational framework presented in this thesis with a modern understanding of perception in biological systems. Doing so motivates the approach taken here and allows us to suggest how this work may reciprocally contribute towards a better computational understanding biological perception. We also examine related work in multimodal integration and examine the engineered system that motivated much of the work in this thesis. Finally, we speculate on a number of theoretical issues in Intersensory perception and examine how the work in this thesis addresses them.

Chapter 7 contains a brief summary of the contributions of this thesis and outlines future work.

Chapter 2

Setting the Stage

We begin with an example to illustrate the two fundamental problems of perception addressed in this thesis:

1) *Grounding* – how are sensory inputs categorized in a perceptual system?
2) *Interpretation* – how should sensory inputs be classified once their possible categories are known?

The example presented below concerns speechreading, but the techniques presented in later chapters for solving the problems raised here are not specific to any perceptual modality. They can be applied to range of perceptual and motor learning problems, and we will examine some of their nonperceptual applications as well.

2.1 Peterson and Barney at the World's Fair

Our example begins with the 1939 World's Fair in New York, where Gordon Peterson and Harold Barney (1952) collected samples of 76 speakers saying sustained American

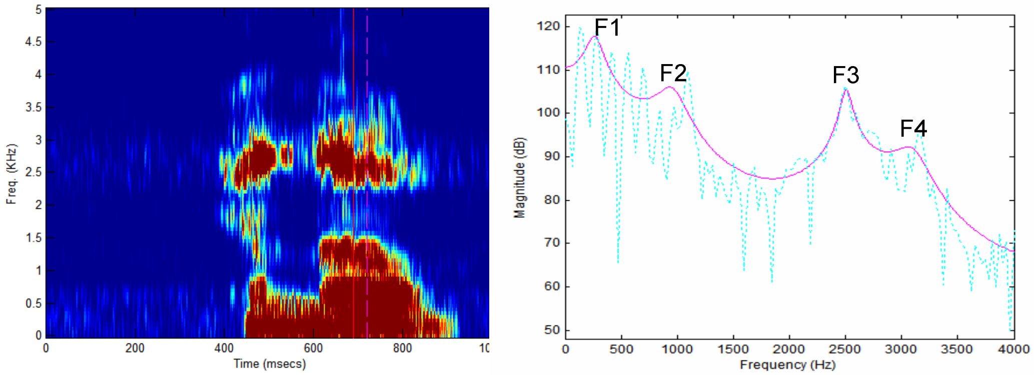


Figure 2.1— On the left is a spectrogram of the author saying, "Hello." The demarcated region (from 690-710ms) marks the middle of phoneme /ow/, corresponding to the middle of the vowel "o" in "hello." The spectrum corresponding to this 20ms window is shown on the right. A 12th order linear predictive coding (LPC) model is shown overlaid, from which the formants, i.e., the spectral peaks, are estimated. In this example: F1 = 266Hz, F2 = 922Hz, and F3 = 2531Hz. Formants above F3 are generally ignored for sound classification because they tend to be speaker dependent. Notice that F2 is slightly underestimated in this example, a reflection of the heuristic nature of computational formant determination.

English vowels. They measured the fundamental frequency and first three formants (see Figure 2.1) for each sample and noticed that when plotted in various ways (Figure 2.2), different vowels fell into different regions of the formant space. This regularity gave hope that spoken language – at least vowels – could be understood through accurate estimation of formant frequencies. This early hope was dashed in part because co-articulation effects lead to considerable movement of the formants during speech (Holbrook and Fairbanks 1962). Although formant-based classifications were largely abandoned in favor of the dynamic pattern matching techniques commonly used today (Jelinek 1997), the belief persists that formants are potentially useful in speech recognition, particularly for dimensional reduction of data.

It has long been known that watching the movement of a speaker's lips helps people understand what is being said (Bender 1981, p41). The sight of someone's moving lips in an environment with significant background noise makes it easier to understand what the speaker is saying; visual cues – e.g., the sight of lips – can alter the signal-to-noise ratio of an auditory stimulus by 15-20 decibels (Sumby and Pollack 1954). The task of lip-reading has by far been the most studied problem in the computational multimodal

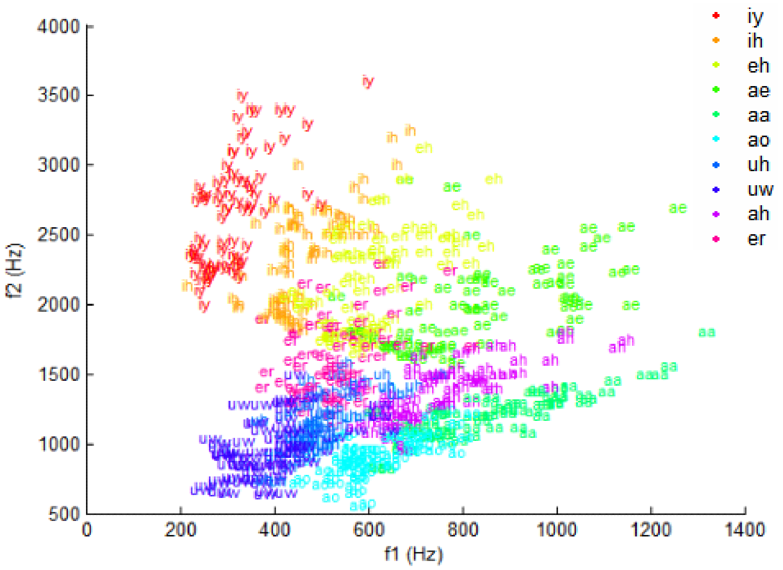


Figure 2.2 – Peterson and Barney Data. On the left is a scatterplot of the first two formants, with different regions labeled by their corresponding vowel categories.

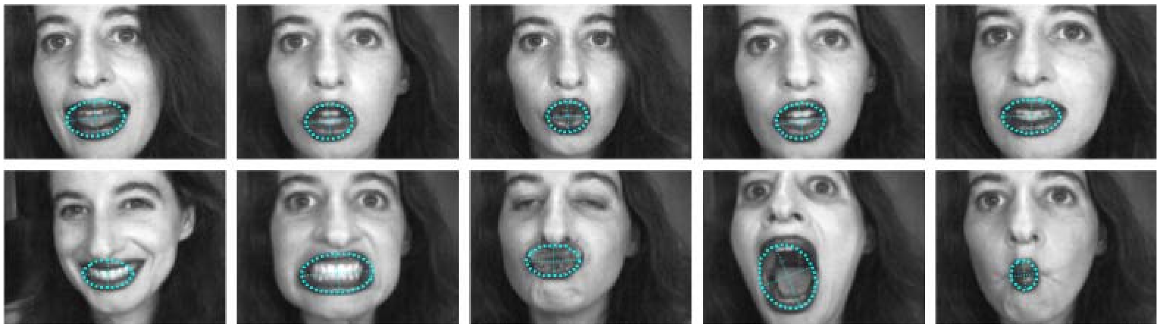


Figure 2.3 – Automatically tracking mouth positions of test subject in a video stream. Lip positions are found via a deformable template and fit to an ellipse using least squares. The upper images contains excerpts from speech segments, corresponding left to right with phonemes: /eh/, /ae/, /uw/, /ah/, and /iy/. The bottom row contains non-speech mouth positions. Notice that fitting the mouth to an ellipse can be non-optimal, as is the case with the two left-most images; independently fitting the upper and lower lip curves to low-order polynomials would yield a better fit. For the purposes of this example, however, ellipses provide an adequate, distance invariant, and low-dimensional model. The author is indebted to his wife for having lips that were computationally easy to detect.

literature (e.g., Mase and Pentland 1990, Huang et al. 2003, Potamianos et al. 2004), due to the historic prominence of automatic speech recognition in computational perception. Although significant progress has been made in automatic speech recognition, state of the art performance has lagged human speech perception by up to an order of magnitude, even in highly controlled environments (Lippmann 1997). In response to this, there has been increasing interest in non-acoustic sources of speech information, of which vision has received the most attention. Information about articulator position is of particular interest, because in human speech, acoustically ambiguous sounds tend to have visually unambiguous features (Massaro and Stork 1998). For example, visual observation of tongue position and lip contours can help disambiguate unvoiced velar consonants /p/ and /k/, voiced consonants /b/ and /d/, and nasals /m/ and /n/, all of which can be difficult to distinguish on the basis of acoustic data alone.

Articulation data can also help to disambiguate vowels. Figure 2.3 contains images of a speaker voicing different sustained vowels, corresponding to those in Figure 2.2. These images are the unmodified output of a mouth tracking system written by the author, where the estimated lip contour is displayed as an ellipse and overlaid on top of the speaker's mouth. The scatterplot in Figure 2.4 shows how a speaker's mouth is represented in this way, with contour data normalized such that a resting mouth

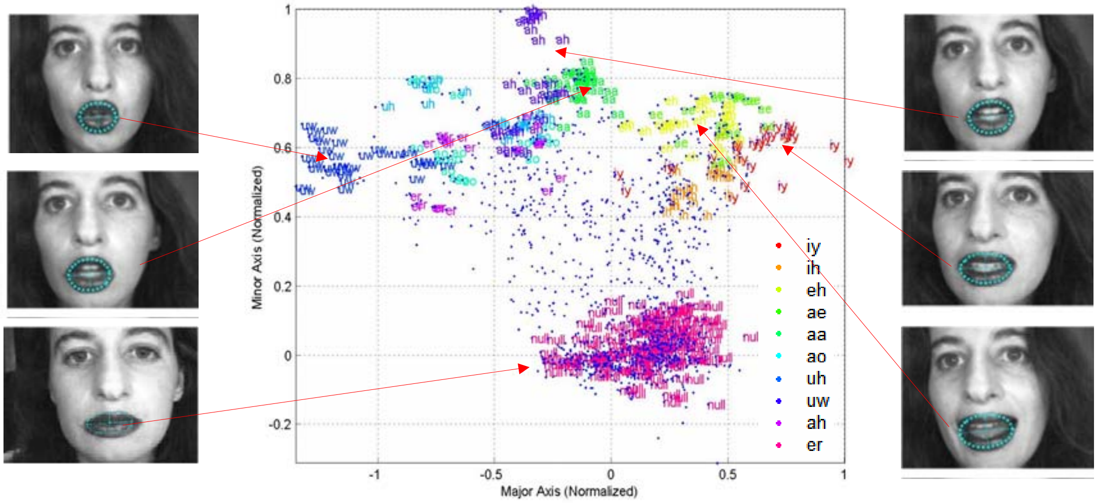


Figure 2.4 -- Modeling lip contours with an ellipse. The scatterplot shows normalized major (x) and minor (y) axes for ellipses corresponding to the same vowels as those in Figure 2.2. In this space, a closed mouth corresponds to a point labeled *null*. Other lip contours can be viewed as offsets from the null configuration and are shown here segmented by color. These data points were collected from video of this woman speaking.

configuration (referred to as *null* in the figure) corresponds with the origin, and other mouth positions are viewed as offsets from this position. For example, when the subject makes an /iy/ sound, the ellipse is elongated along its major axis, as reflected in the scatterplot.

Suppose we now consider the formant and lip contour data simultaneously, as in Figure 2.5. Because the data are conveniently labeled, the clusters within and the correspondences between the two scatterplots are obvious. We notice that the two domains can mutually disambiguate one another. For example, /er/ and /uh/ are difficult to separate acoustically with formants but are easy to distinguish visually. Conversely, /ae/ and /eh/ are visually similar but acoustically distinct. Using these complementary representations, one could imagine combining the auditory and visual information to create a simple speechreading system for vowels.

2.2 Nature Does Not Label Its Data

Given this example, it may be surprising that our interest here is not in building a speechreading system. Rather, we are concerned with a more fundamental problem: how do sensory systems learn to segment their inputs to begin with? In the color-coded plots

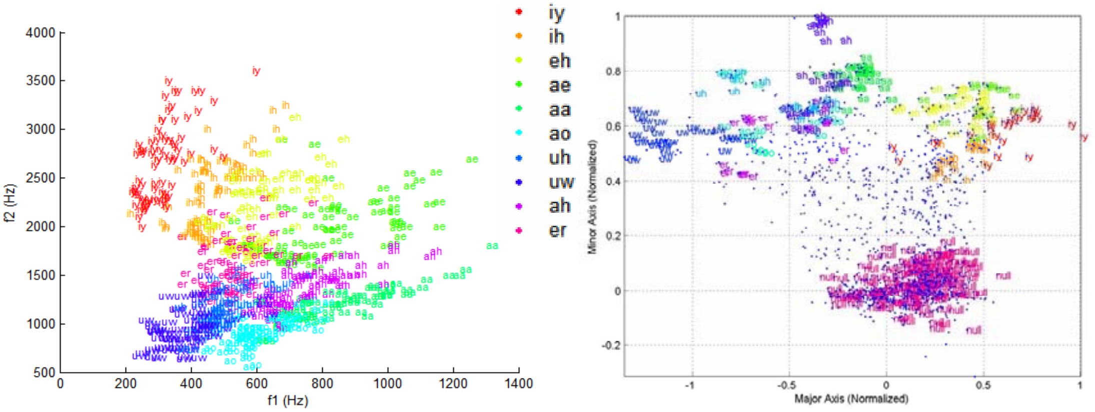


Figure 2.5 – Labeled scatterplots side-by-side. Formant data (from Peterson Barney 1952) is displayed on the left and lip contour data (from the author's wife) is show on the right. Each plot contains data corresponding to the ten listed vowels in American English.

in Figure 2.5, it is easy to see the different represented categories. However, perceptual events in the world are generally not accompanied with explicit category labels. Instead, animals are faced with data like those in Figure 2.6 and must somehow learn to make sense of them. We want to know how the categories are learned in the first place. We note this learning process is not confined to development, as perceptual correspondences are plastic and can change over time.

We would therefore like to have a general purpose way of taking data (such as shown in Figure 2.6) and deriving the kinds of correspondences and segmentations (as shown in Figure 2.5) without external supervision. This is what we mean by *perceptual grounding*

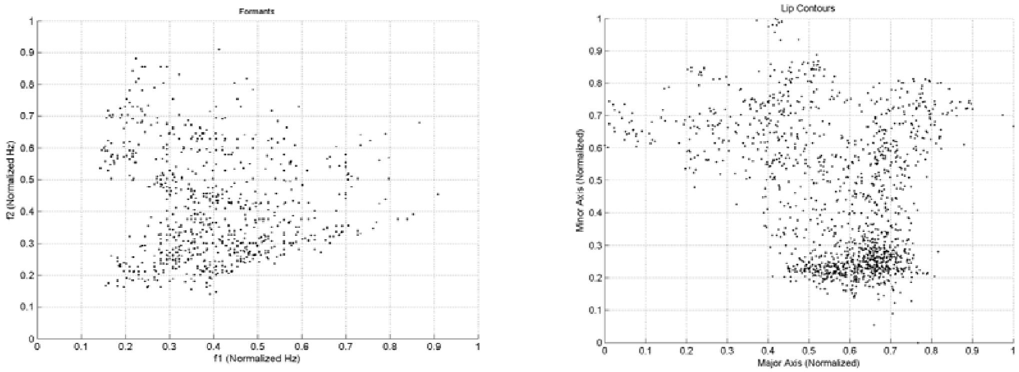


Figure 2.6 – Unlabeled data. These are the same data shown above in Figure 2.5, with the labels removed. This picture is closer to what animals actually encounter in Nature. As above, formants are displayed on the left and lip contours are on the right. Our goal is to learn the categories present in these data without supervision, so that we can automatically derive the categories and clusters such as those show above.

and our perspective here is that it is a clustering problem: animals must learn to organize their perceptions into meaningful categories. We examine below why this is a challenging problem.

2.3 Why Is This Difficult?

As we have noted above, Nature does not label its data. By this, we mean that the perceptual inputs animals receive are not generally accompanied by any meta-level data explaining what they represent. Our framework must therefore assume the learning is unsupervised, in that there are no data outside of the perceptual inputs themselves available to the learner.

From a clustering perspective, perceptual data is highly non-parametric in that both the number of clusters and their underlying distributions may be unknown. Clustering algorithms generally make strong assumptions about one or both of these. For example, the Expectation Maximization algorithm (Dempster et al. 1977) is frequently used a basis for clustering mixtures of distributions whose maximum likelihood estimation is easy to compute. This algorithm is therefore popular for clustering known finite numbers of Gaussian mixture models (e.g., Nabney 2002, Witten and Frank 2005). However, if the number of clusters is unknown, the algorithm tends to converge to a local minimum with the wrong number of clusters. Also, if the data deviate from a mixture of Gaussian (or some expected) distributions, the assignment of clusters degrades accordingly. More generally, when faced with nonparametric, distribution-free data, algorithmic clustering techniques tend not be robust (Fraley and Raftery 2002, Still and Bialek 2004).

Perceptual data are also noisy. This is due both to the enormous amount of variability in the world and to the probabilistic nature of the neuronal firings that are responsible for the perception (and sometimes the generation) of perceivable events. The brain itself introduces a great deal of uncertainty into many perceptual processes. In fact, one may perhaps view the need for high precision as the exception rather than the rule. For example, during auditory localization based on interaural time delays, highly specialized neurons known as the *end-bulbs of Held* – among the largest neuronal structures in the brain – provide the requisite accuracy by making neuronal firings in this section of auditory cortex highly deterministic (Trussell 1999). It appears that the need for mathematical precision during perceptual processing can require extraordinary neuroanatomical specialization.

Perhaps most importantly, perceptual grounding is difficult because there is no objective mathematical definition of "coherence" or "similarity." In many approaches to clustering, each cluster is represented by a prototype that, according to some well-defined measure, is an exemplar for all other data it represents. However, in the absence of fairly strong assumptions about the data being clustered, there may be no obvious way to select this measure. In other words, it is not clear how to formally define what it means for data to be objectively similar or dissimilar. In perceptual and cognitive domains, it may also depend on why the question of similarity is being asked. Consider a classic AI conundrum, "*what constitutes a chair?*" (Winston 1970, Minsky 1974, Brooks 1987). For many purposes, it may be sufficient to respond, "*anything upon which one can sit*." However, when decorating a home, we may prefer a slightly more sophisticated answer. Although this is a higher level distinction than the ones we examine in this thesis, the principle remains the same and reminds us why similarity can be such a difficult notion to pin down.

Finally, even if we were to formulate a satisfactory measure of similarity for static data, one might then ask how this measure would behave in a dynamic system. Many perceptual (and motor) systems are inherently dynamic – they involve processes with complex, non-linear temporal behavior (Thelen and Smith 1994), as can been seen during perceptual bistability, cross-modal influence, habituation, and priming. Thus, one may ask whether a similarity metric captures a system's temporal dynamics; in a clustering domain, the question might be posed: *do points that start out in the same cluster end up in the same cluster?* We know from Lorentz (1964) that it is possible for arbitrarily small differences to be amplified in a non-linear system. It is quite plausible that a static similarity metric might be oblivious to a system's temporal dynamics, and therefore, sensory inputs that initially seem almost identical could lead to entirely different percepts

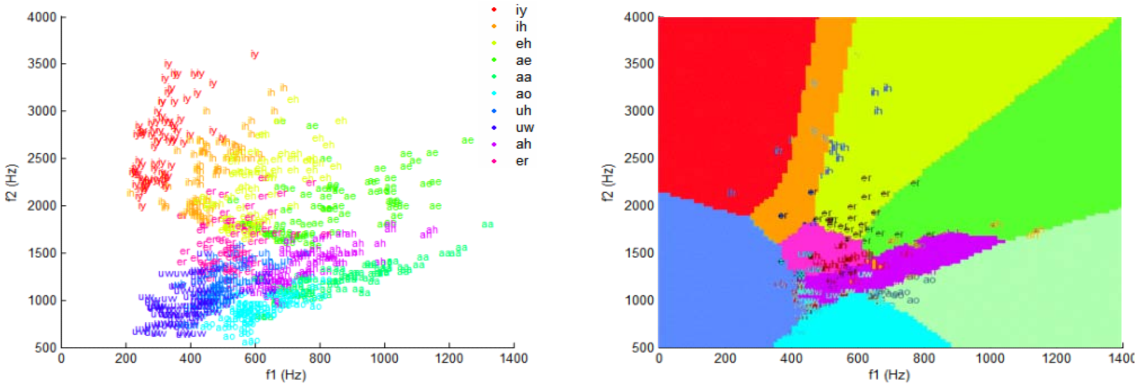


Figure 2.7 – On the left is a scatterplot of the first two formants, with different regions labeled by their corresponding vowel categories. The output of a backpropagation neural network trained on this data is shown on the right and displays decision boundaries and misclassified points. The misclassification error in this case is 19.7%. Other learning algorithms, e.g., AdaBoost using C4.5, Boosted stumps with LogitBoost, and SVM with a 5th order polynomial kernel, have all shown similarly lackluster performance, even when additional dimensions (corresponding to F0 and F3) are included (Klautau 2002). (Figure on right is derived from ibid. and used with permission.)

being generated. This issue will be raised in more detail in Chapter 4, where we will view clusters as fixed points in representational phase spaces in which perceptual inputs follow trajectories between different clusters.

In Chapter 3, we will present a framework for perceptual grounding that addresses many of the issues raised here. We show that animals (and machines) can learn how to cluster their perceptual inputs by simultaneously correlating information from their different senses, even when they have no advance knowledge of what events these senses are individually capable of perceiving. By *cross-modally* sharing information between different senses, we will demonstrate that sensory systems can be perceptually grounded by bootstrapping off each other.

2.4 Perceptual Interpretation

The previous section outlined some of the difficulties in unsupervised clustering of nonparametric sensory data. However, even if the data came already labeled and clustered, it would still be challenging to classify new data points using this information. Determining how to assign a new data point to a preexisting cluster (or category) is what we mean by *perceptual interpretation*. It is the process of deciding what a new input actually represents. In the example here, the difficultly is due to the complexity of partitioning formant space to separate the different vowels. This 50 year old classification problem still receives attention today (e.g., Jacobs et al. 1991, de Sa and Ballard 1998, Clarkson and Moreno 1999) and Klautau (2002) has surveyed modern machine learning algorithms applied to it, an example of which is shown on the right in Figure 2.7.

A common way to distinguish classification algorithms is by visualizing the different spaces of possible decision boundaries they are capable of learning. If one closely examines the Peterson and Barney dataset (Figure 2.8), there are many pairs of points that are nearly identical in any formant space but correspond to different vowels in the actual data, at least according to the speaker's intention. It is difficult to imagine any accurate partitioning that would simultaneously avoid overfitting. There are many factors that contribute to this, including the information loss of formant analysis (i.e., incomplete data is being classified), computational errors in estimating the formants, lack of

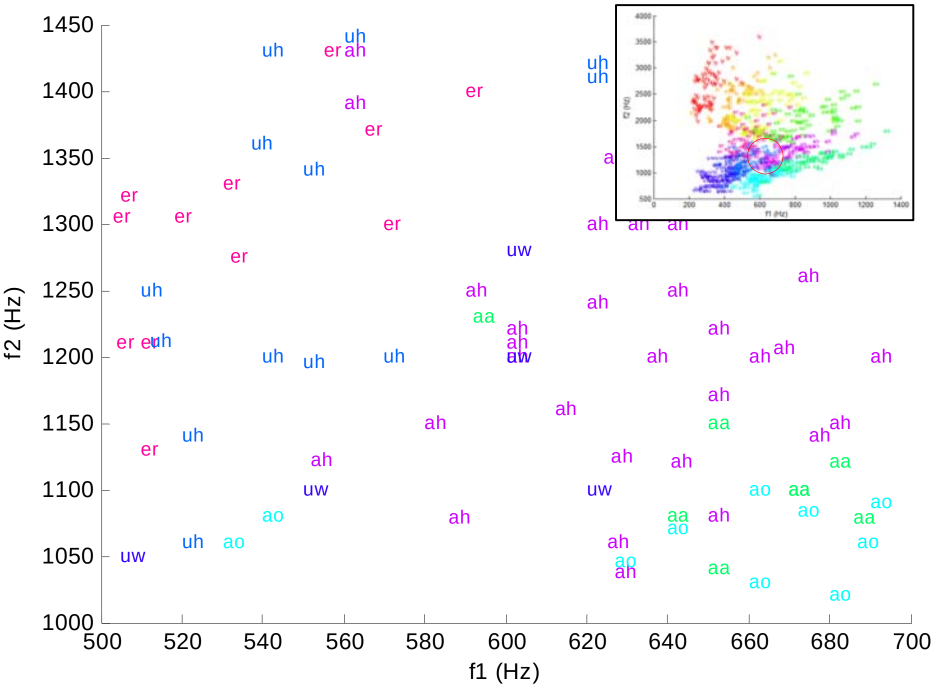


Figure 2.8 – Focusing on one of many ambiguous regions in the Peterson-Barney dataset. Due to a confluence of factors described in the text, the data in these regions are not obviously separable.

differentiation in vowel pronunciation in different dialects of American English, variations in prosody, and individual anatomical differences in the speakers' vocal tracts. It is worth pointing out the latter three of these for the most part exist independently of how data is extracted from the speech signal and may present difficulties regardless of how the signal is processed.

The curse of dimensionality (Bellman 1961) is a statement about exponential growth in hypervolume as a function of a space's dimension. Of its many ramifications, the most important here is that many low dimensional phenomena that we are intuitively familiar with do not exist in higher dimensions. For example, the natural clustering of uniformly distributed random points in a two dimensional space becomes extremely unlikely in higher dimensions; in other words, random points are relatively far apart in high dimensions. In fact, transforming nonseparable samples into higher dimensions is a general heuristic for improving separation with many classification schemes. There is a flip-side to this high dimensional curse for us: ***low dimensional spaces are crowded***. It can be difficult to separate classes in these spaces because of their tendency to overlap. However, blaming low dimensionality for this problem is like the proverbial cursing of darkness. Cortical architectures make extensive use of low dimensional spaces, e.g., throughout visual, auditory, and somatosensory processing (Amari 1980, Swindale 1996, Dale et al. 1999, Fischl et al. 1999, Kaas and Hackett 2000, Kardar and Zee 2002, Bednar et al. 2004), and this was a primary motivating factor in the development of Self Organizing Maps (Kohonen 1984). In these crowded low-dimensional spaces, approaches that try to implicitly or explicitly refine decision boundaries such as those in Figure 2.8 (e.g., de Sa 1994) are likely to meet with limited success because there may be no good decision boundaries to find; perhaps in these domains, decision boundaries are the wrong way to think about the problem.

Rather than trying to improve classification boundaries directly, one could instead look for a way to move ambiguous inputs into easily classified subsets of their representational spaces. This is the essence of the *influence network* approach presented in Chapter 4 and is our proposed solution to the problem of perceptual interpretation. The goal is to use cross-modal information to "move" sensory inputs within their own state spaces to make them easier to classify. Thus, we take the view that perceptual interpretation is inherently a dynamic – rather than static – process that occurs during some window of time. This approach relaxes the requirement that the training data be separable in the traditional machine learning sense; unclassifiable subspaces are not a problem if we can determine how to move out of them by relying on other modalities, which are experiencing the same sensory events from their unique perspectives. We will show that this approach is not only biologically plausible, it is also computationally efficient in that it allows us to use lower dimensional representations for modeling sensory and motor data.

> It might be asked why we have more senses than one. [Had it been otherwise],… everything would have merged for us into an indistinguishable identity.
>
> Aristotle, *De Anima* (350 B.C.E)

Chapter 3

Perceptual Grounding

Most of the enormous variability in the world around us is unimportant. Variations in our sensory perceptions are not only tolerated, they generally pass unnoticed. Of course, some distinctions are of paramount importance and learning which are meaningful as opposed to which can be safely ignored is a fundamental problem of cognitive development. This process is a component of *perceptual grounding*, where a perceiver learns to make sense of its sensory inputs. The perspective taken here is that this is a clustering problem, in that each sense must learn to organize its perceptions into meaningful categories. That animals do this so readily belies its complexity. For example, people learn phonetic structures for languages simply by listening to them; the phonemes are somehow extracted and clustered from auditory inputs even though the listener does not know in advance how many unique phonemes are present in the signal.

Contrast this with a standard mathematical approach to clustering, where some knowledge of the clusters, e.g., how many there are or their distributions, must be known a priori in order to derive them. Without knowing these parameters in advance, algorithmic clustering techniques may not be robust (Fraley and Raftery 2002, Kleinberg 2002, Still and Bialek 2004). Assuming that in many circumstances animals cannot know the parameters underlying their perceptual inputs, how then do they learn to organize their sensory perceptions reliably?

This chapter presents an approach to clustering based on observed correlations between different sensory modalities. These cross-modal correlations exist because perceptions are created through physical processes governed by natural laws (Thompson 1917, Richards 1980, Mumford 2004). An event in the world is simultaneously perceived through multiple sensory pathways in a single observer; while each pathway may have a unique perspective on the event, their perspectives tend to be correlated by regularities in the physical world (Richards and Bobick 1988). We propose here that these correspondences play a primary role in organizing the sensory channels individually. Based on this hypothesis, we develop a new framework for grounding artificial perceptual systems.

Towards this, we will introduce a mathematical model of *slices*, which are topological manifolds that encode dynamic perceptual states and are inspired by surface models of cortical tissue (Dale et al. 1999, Fischl et al. 1999, Citti and Sarti 2003, Ratnanather et al. 2003). Slices partition perceptual spaces into large numbers of small regions (hyperclusters) and then reassemble them to construct clusters corresponding to the actual sensory events being perceived. This reassembly is performed by *cross-modal clustering*, which uses temporal correlations between slices to determine which hyperclusters within a slice correspond to the same sensory events. The cross-modal clustering algorithm does not presume that either the number of clusters in the data or their distributions is known beforehand. We examine the outputs and behavior of this algorithm on simulated datasets, drawn from a variety of mixture distributions, and on real data gathered in computational experiments. Some of the work in this chapter has appeared in (Coen 2005, Coen 2006).

3.1 The Simplest Complex Example

As in Chapter 2, we proceed here by first considering an example. We will return to using real datasets towards the end of this chapter, but for the moment, it is helpful to pare down the subject matter to its bare essentials.

Let us consider two hypothetical sensory modes, each of which is capable of sensing the same two events in the world, which we call the *red* and *blue* events. These two modes are illustrated below in Figure 3.1, where the dots within each mode represent its perceptual inputs and the blue and red ellipses delineate the two events. For example, if a "red" event takes place in the world, each mode would receive sensory input that

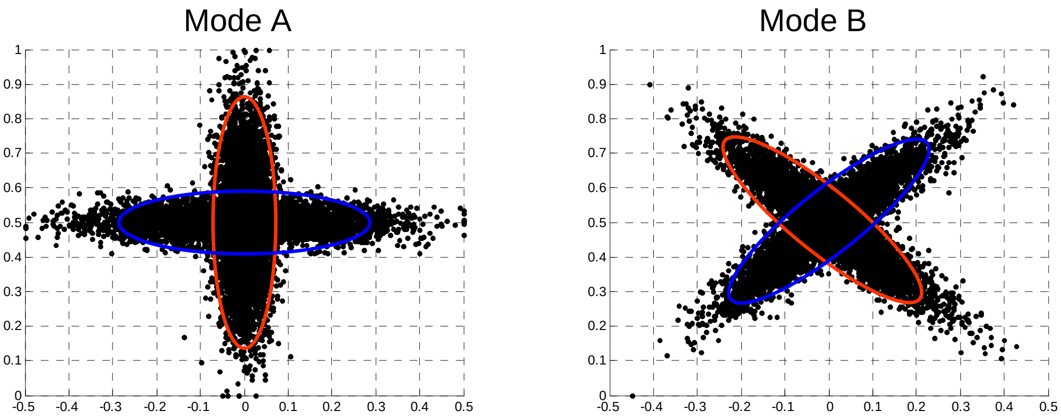


Figure 3.1 – Two hypothetical co-occurring perceptual modes. Each mode, unbeknownst to itself, receives inputs generated by a simple, overlapping Gaussian mixture model. To make matters more concrete, we might imagine Mode A is a simple auditory system that hears two different events in the world and Mode B is a simple visual system sees those same two events, which are indicated by the red and blue ellipses.

(probabilistically) falls within its red ellipse. Notice that events within each mode overlap, and they are in fact represented by a mixture of two overlapping Gaussian distributions. We have chosen this example because it is simple – each mode perceives only two events – but it has the added complexity that the events overlap – meaning there is likely to be some ambiguity in interpreting the perceptual inputs.

Keep in mind that while *we* know there are only two events (red and blue) in this hypothetical world, the *modes* themselves do not "know" anything at all about what they can perceive. The colorful ellipses are solely for the reader's benefit; the only thing the modes receive is their raw input data. Our goal then is to learn the perceptual categories in each mode – e.g., to learn that each mode in this example senses these two overlapping events – by exploiting the temporal correlations between them.

3.2 Generating Codebooks

We are going to proceed by hyperclustering each perceptual space into a codebook. This simply means that we are going to generate far more clusters than are necessary for representing the actual number of perceptual events in the data. In this case, that would

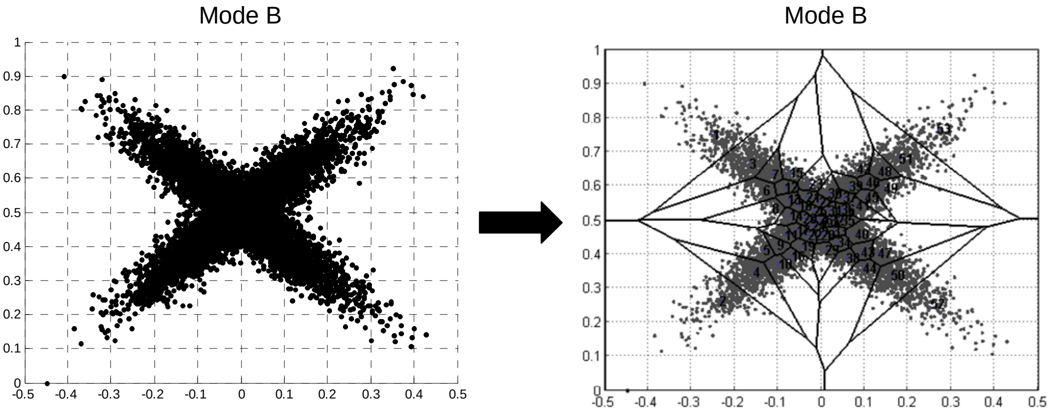


Figure 3.2 – Hyperclustering Mode B with the algorithm given below. Mode B is shown hyperclustered on the right. Here, we specified *k*=30 and the algorithm ended up generating 53 clusters after normalizing their densities.

be two, but instead, we will employ a (much) larger number. For the rest of this discussion, we will refer to two different types of clusters:

1) *codebook clusters* (or *hyperclusters*) are generated by hyperclustering and are illustrated by the Voronoi regions show in Figure 3.2 on the right.

2) *perceptual clusters* refer to actual sensory events and are outlined with the colored ellipses in Figure 3.1.

Our goal will be to combine the *codebook* clusters to "assemble" the *perceptual* clusters. We note that while perceptual clustering is quite difficult, for reasons outlined in the previous chapter, hyperclustering is quite easy because there is no notion of perceptual correctness associated with it. Although we must determine how many codebook clusters to generate, we will show this number influences the amount of training data necessary rather than the correctness of the derived perceptual clusters. In other words, this approach is not overly sensitive to the hyperclustering: generating too many hyperclusters simply means learning takes longer, not that the end results are incorrect. Generating too few hyperclusters tends not to happen because of the density normalization described below. It is also sometimes possible to detect that too few clusters have been generated by using cross-modal information, a technique we examine later in this chapter.

To generate the codebooks, we will use a variant of the Generalized Lloyd Algorithm (GLA) (Lloyd 1982). We modify the algorithm to normalize the point densities within the hyperclusters, which otherwise can vary enormously. Many clustering algorithms, including GLA, optimize initially random codebooks by minimizing a strongly Euclidean distance metric between cluster centroids and their members. A cluster with a large numbers of nearby points may be viewed as equivalent to (from the perspective of the optimization) a cluster with a small number of distant points. It is therefore possible to have substantial variance in the number of points assigned to each codebook cluster. This is problematic because our approach will require that each *perceptual* cluster be represented by multiple *codebook* clusters, from which it is "assembled." The Euclidean bias introduced by the distance metric used for codebook optimization means that "small" perceptual events may be relegated to a single codebook cluster. This would prevent them from ever being detected.

There are many ways one could imagine achieving this density normalization. For example, we could explicitly add inverse cluster size to the minimization calculation performed during codebook refinement. This would leave the number of codebook clusters constant overall but introduce pressure against wide variation in the number of points assigned to each one. Rather than take an approach that preserves the overall number of clusters, we will instead modify the algorithm to recluster codebook regions that have been assigned "too many" points. This benefit of this is that we leave the GLA algorithm intact but now invoke it recursively on subregions where its performance is unsatisfactory. By keeping the basic structure of GLA, many of the mathematic properties of the generated codebooks remain unchanged. The downside of this approach is that the recursive reclustering increases the total number of generated hyperclusters. Thus, the algorithm generates at least as many codebook clusters as we specify and sometimes many more. This increase in codebook size can affect the computational complexity of algorithms operating over these codebooks, which we investigate later in this chapter. We note, however, that adding these additional clusters does not tend to require gathering more training data, an issue raised above. This is because the extra clusters are generated in regions that already have high point densities.

Our hyperclustering algorithm for generating (at least) *k* codebook regions over dataset $D \subseteq \mathbb{R}^N$ is:

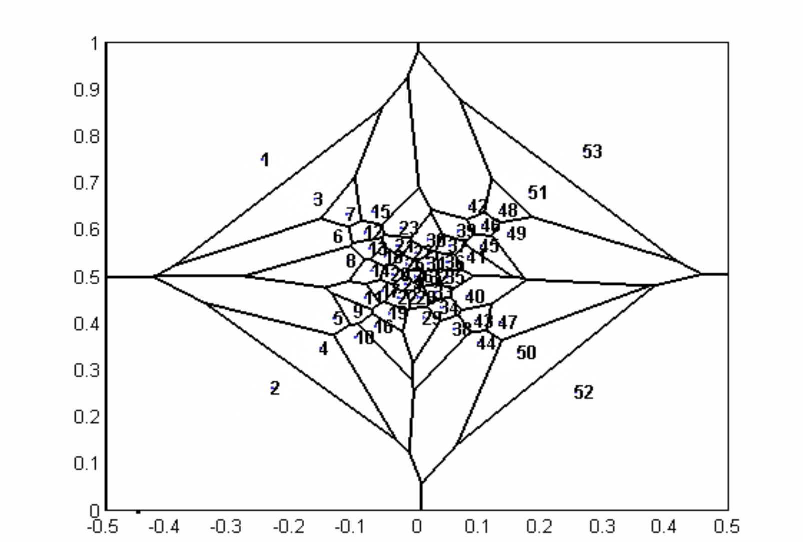


Figure 3.3 – The hyperclusters generated for the data in Mode B, with the data removed. The number identifying each cluster is located at its centroid. Notice how the number of clusters increases in the region corresponding to the overlap of the two Gaussian distributions, where the overall point density is highest.

1) Let $s = |D|/k$. This is our goal size for the number of data points per cluster.

2) Let $P = \{P_1, P_2, ..., P_k\}$, $P_i \subset \mathbb{R}^N$ be a Lloyd partitioning of *D* over *k* clusters. This is the output of the Generalized Lloyd Algorithm.

3) For each cluster $P_i \in P$:

   If $|P_i| > s$ (the cluster has too many points), then recursively partition P_i:

   a. Let $Q = \{P_1, P_2\}$, $P_i \subset \mathbb{R}^N$ be a Lloyd partitioning of P_i over 2 clusters.

   b. Set $P = (P \cup Q)/P_i$. Add the two new partitions and remove the old one.

   End if statement

4) Repeat step 3 until no new partitions are added. Then, return the centroids of the sets in *P* as the final hyperclustering. Empirically, we find that $k < |P| < 2k$.

The output of this algorithm on the data in Mode B is shown above in Figure 3.3. Notice how the number of clusters increases in the region corresponding to the overlap of the two Gaussian distributions, which is due to the density normalization. We note that any number of variations on this algorithm is possible. For example, in the reclustering step

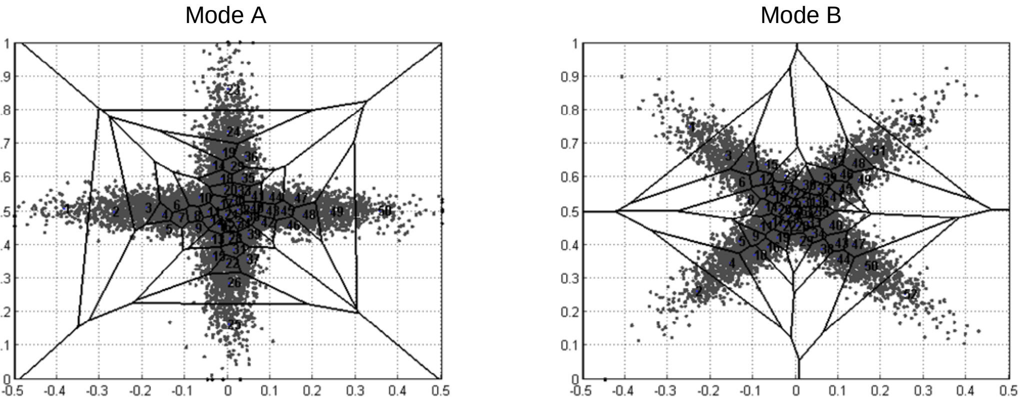


Figure 3.4 – Slices generated for Modes A and B using the hyperclustering algorithm in the previous section. Our goal is to combine the codebook clusters to reconstruct the actual sensory events perceived within the slices.

in (3), we might recursively generate $|P_i|/s$ rather than 2 clusters. We could also modify the goal size s to change the degree of density normalization. In any event, we have found that our approach is not particularly sensitive to the precise details of the codebook's generation; we confirm this statement later in this chapter, when we consider hyperclustering other mixture distributions. At present, the most important consideration is that the cluster densities are normalized, which minimizes the Euclidean bias inherent in the centroid optimization performed by the Lloyd algorithm.

3.3 Generating Slices

We now introduce a new data structure called *slices* that are constructed using the codebooks defined in the previous section. Figure 3.4 illustrates slices constructed for Modes A and B from our example above. Slices are topological manifolds that encode dynamic perceptual states and are inspired by surface models of cortical tissue (Citti and Sarti 2003, Ratnanather et al. 2003). They are able to represent both symbolic and numeric data and provide a natural foundation for aggregating and correlating information. Intuitively, a slice is a codebook with a non-Euclidean distance metric defined between its cluster centroids. In other words, distances within each cluster are Euclidean, whereas distances between clusters are not. A topological manifold is simply a manifold "glued" together from Euclidean spaces, and that is exactly what a slice is.

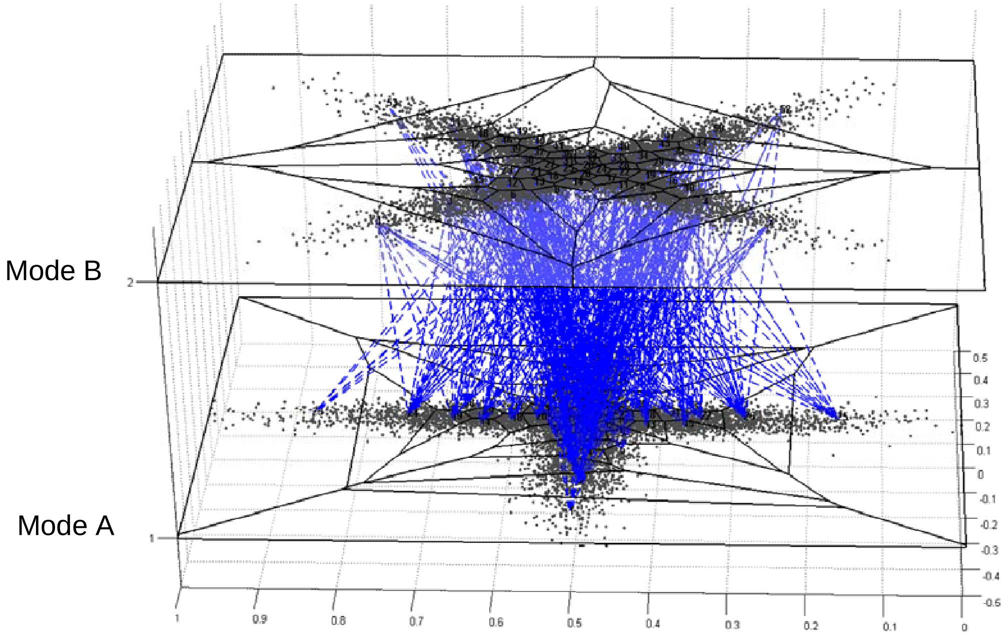


Figure 3.5 – Viewing Hebbian linkages between two different slices. The modes have been vertically stacked here to make the correspondences clearer. The blue lines indicate that two codebook regions temporally co-occur with each other. Note that these connections are weighted based on their strengths, which is not visually represented here, and that these weights are additionally asymmetric between each pair of connected regions.

Our goal is to combine the codebook regions to "reconstruct" the larger perceptual regions within a slice. To do this, we will define a non-Euclidean distance metric between codebook regions that reflects how much we think they are part of the same perceptual event. In this metric, codebook regions corresponding to the same perceptual event will be closer together and those corresponding to different events will be further apart. Towards defining this metric, we first collect co-occurrence data between the codebook regions in different modes. We want to know how each codebook region in a mode temporally co-occurs with the codebook regions in other modes.

This data can be easily gathered with the classical sense of Hebbian learning (Hebb 1949), where connections between regions are strengthened as they are simultaneously active. The result of this process is illustrated in Figure 3.5, where the modes are vertically stacked to make the correspondences clearer. We will exploit the spatial structure of this Hebbian co-occurrence data to define the distance metric within each mode.

3.4 Hebbian Projections

In this section, we define the notion of a *Hebbian projection*. These are spatial probability distributions that provide an intuitive way to view co-occurrence relations between different slices. We first give a formal definition and then illustrate the concept visually.

Consider two slices $M_A, M_B \subseteq \mathbb{R}^n$, with associated codebooks $C_A = \{p_1, p_2, ..., p_a\}$ and $C_B = \{q_1, q_2, ..., q_b\}$, where cluster centroids $p_i, q_j \in \mathbb{R}^N$.

For some event *x*, we define $h(x) = \#$ of times event *x* occurs. Similarly, for events *x* and *y*, we define $h(x, y) =$ # of times events *x* and *y* co-occur. For example, $h(p_1)$ is the number of times inputs that belong to cluster p_1 were seen during some time period of interest. So, we see that $\Pr(x \mid y) = h(p, q) / h(p)$.

We define the *Hebbian projection* of a codebook cluster $p_i \in C_A$ onto mode M_B:

$$\vec{H}_A^B(p_i) = \left[\Pr(q_1 \mid p_i), \Pr(q_2 \mid p_i), ..., \Pr(q_b \mid p_i)\right] \tag{3.1}$$

When the modes are clear from context, we will simply refer to the projection by $\vec{H}(p_i)$.

A Hebbian projection is simply a conditional spatial probability distribution that lets us know what mode M_B probabilistically "looks" like when a region p_i is active in co-occurring mode M_A. This is visualized in Figure 3.6.

We can equivalently define a Hebbian projection for a region $r \subseteq M_A$ constructed out of a subset of its codebook clusters $C_r = \{p_{r1}, p_{r2}, ..., p_{rk}\} \subseteq C_A$:

$$\vec{H}_A^B(r) = \left[\Pr(q_1 \mid r), \Pr(q_2 \mid r), ..., \Pr(q_b \mid r)\right] \tag{3.2}$$

We will also define the notion of a *reverse Hebbian projection*, which projects a Hebbian projection back onto its source mode. It lets us measure – from the perspective of another modality – which other codebook regions in a slice appear similar to a reference region.

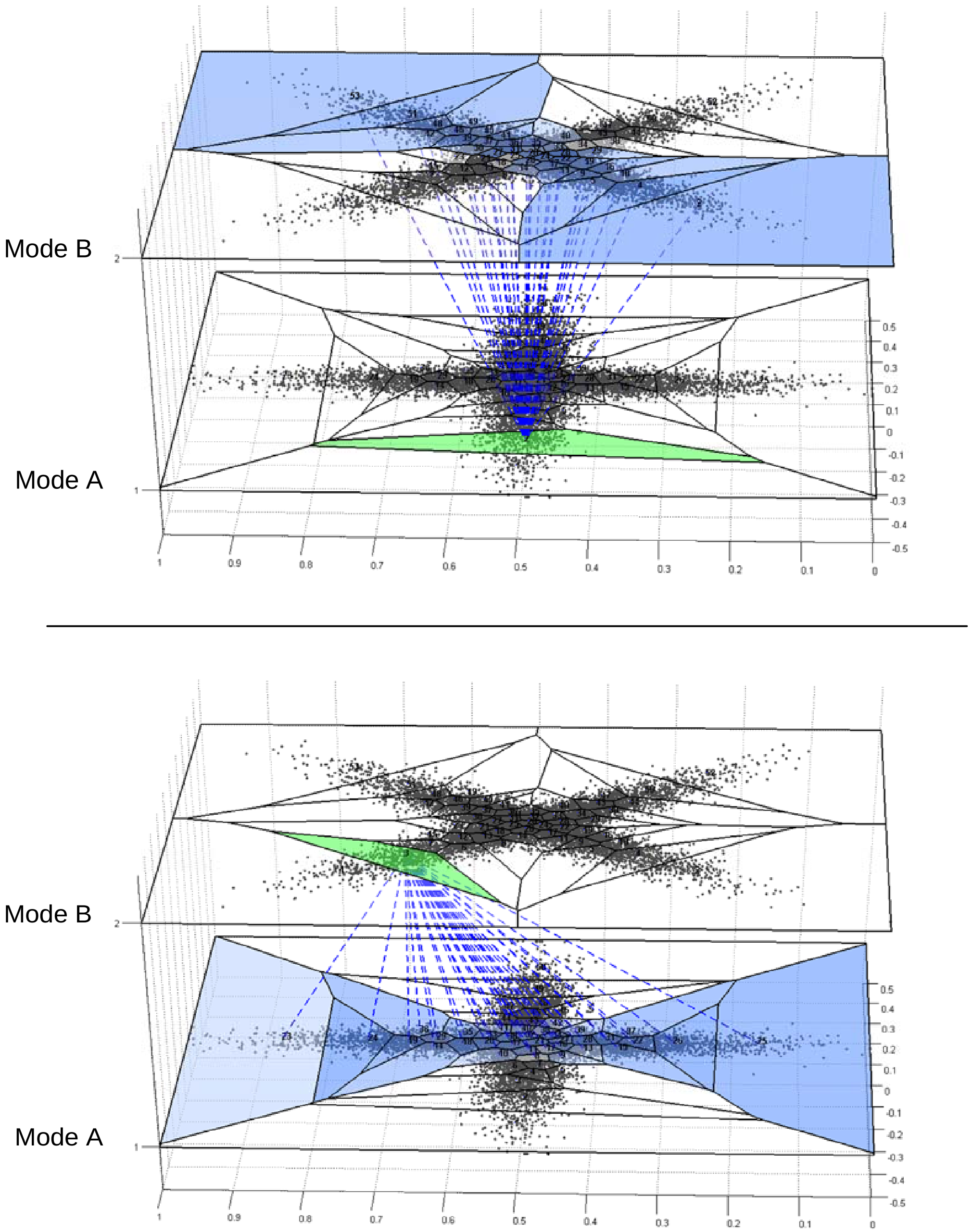


Figure 3.6 – A visualization of two Hebbian projections. On the top, we project from a cluster p_i in Mode A onto Mode B. The dotted lines correspond to Hebbian linkages and the blue shading in each cluster q_j in Mode B is proportional to $\Pr(q_j|p_i)$. A Hebbian projection lets us know what Mode B probabilistically "looks" like when some prototype in Mode A is active. On the bottom, we see a projection from a cluster in Mode B onto Mode A.

To do this, we first define *weighted* versions of the functions defined above for a set of weights ω. Consider a region r, $|r| = k$, where each cluster is assigned some weight ω_i. We assume that $\sum \omega_i = 1$.

$$\begin{aligned}
h_\omega(r) &= \sum_{p \in r} \omega_p h(p), \text{ where } p \text{ is a codebook cluster in region } r \\
\Pr_\omega(q,r) &= h_\omega(r,q)/h_\omega(r) = \sum_{p \in r} \omega_p h(p,q) \Big/ \sum_{p \in r} \omega_p h(p) \\
\vec{H}_\omega(r) &= \left[\Pr_\omega(q_1 \mid r), \Pr_\omega(q_2 \mid r), \ldots, \Pr_\omega(q_n \mid r)\right]
\end{aligned}$$

The reverse Hebbian projection $\overset{\frown}{H}{}_A^B(r)$ of a region $r \subseteq M_A$ onto mode M_B is then defined:

$$\overset{\frown}{H}{}_A^B(r) = \vec{H}_{\vec{H}(r)}(M_B) \tag{3.3}$$

$$= \left[\Pr_{H(r)}(p_1 \mid M_B), \Pr_{H(r)}(p_2 \mid M_B), \ldots, \Pr_{H(r)}(p_m \mid M_B)\right] \tag{3.4}$$

Again, when the modes are clear from context, we will simply refer to this as $\overset{\frown}{H}(r)$.

This distribution has a simple interpretation: the reverse Hebbian projection from mode M_A onto mode M_B for some region $r \subseteq M_A$ is the Hebbian projection of *all of mode* M_B onto mode M_A, weighted by the forward Hebbian projection of region *r*, as shown in equation (3.3). This process is visualized in Figure 3.7. Note that we are projecting an entire mode M_B here. This might seem initially surprising, but it simply corresponds to a projection of a region that contains all the codebook clusters for a given slice.

The reverse Hebbian projection $\overset{\frown}{H}(r)$ answers the question: *what other regions does mode* M_B *think region r is similar to in mode* M_A*?* It can therefore be viewed as a distribution that measures *cross-modal confusion*. For this reason, it provides a useful optimization tool, because we will only need to disambiguate regions that appear in each other's reverse Hebbian projections, i.e., they have a non-zero (or above some threshold) probability of being confused for one another by other modalities.

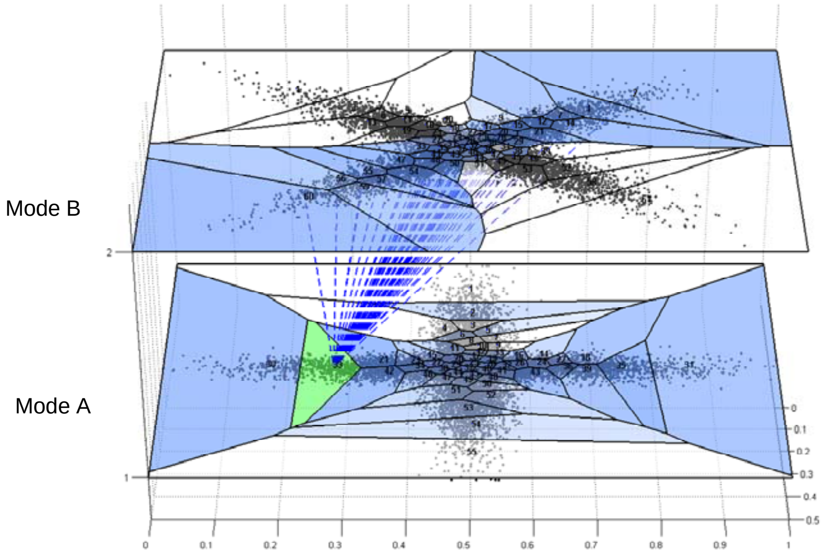


Figure 3.7 – Visualizing a reverse Hebbian projection. We first generate the Hebbian projection of the green cluster p_i in Mode A onto Mode B. This projection is represented by the shading of each region q_j in Mode B, corresponding to $\Pr(q_j|p_i)$. We then project all of Mode B back onto Mode A, weighting the contributions of each cluster q_i by $\Pr(q_j|p_i)$. This generates the reverse Hebbian projection, which is indicated by the shading of regions in Mode A.

3.5 Measuring Distance in a Slice

Let us briefly review where we stand at this point. We have introduced the idea of a *slice*, which breaks up a representational space into many smaller pieces that are generated by hyperclustering it. We would like to assemble these small hyperclusters into larger regions that represent actual perceptual categories present in the input data. In this section, we define the non-Euclidean distance metric between the hyperclusters that helps make this possible.

Consider the colored regions in Figure 3.8. We would like to determine that the blue and red regions are part of their respective *blue* and *red* events, indicated by the colored ellipses. It is important to recall that the colors here are simply for the reader's benefit. There is no labeling of regions or perceptual events within the slice itself. We will proceed by formulating a distance metric that minimizes the distance between codebook regions that are actually within the same perceptual region and maximizes the distance

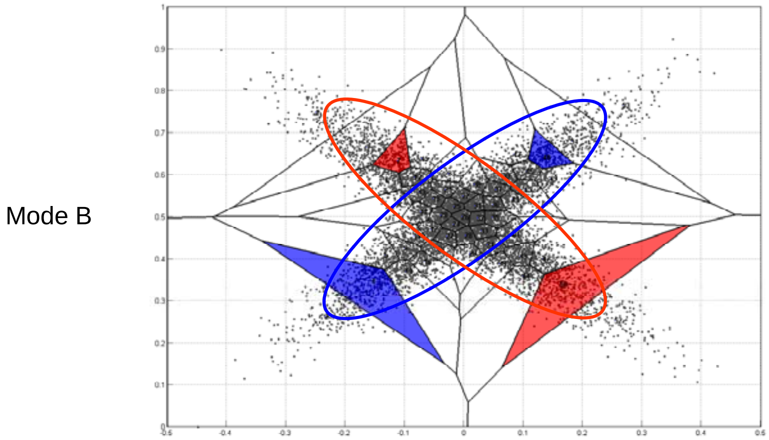


Figure 3.8 – Combining codebook regions to construct perceptual regions. We would like to determine that regions within each ellipse are all part of the same perceptual event. Here, for example, the two blue codebook regions (probabilistically) correspond the *blue* event and the red regions correspond to the *red* event.

between codebook regions that are in different regions. That this metric must be non-Euclidean is clear from looking at the figure. Each highlighted region is closer to one of a different color than it is to its matching partner.

We are going to use the Hebbian projections defined in the previous section to formulate this similarity metric for codebook regions. This will make the metric inherently cross-modal because we will rely on co-occurring modalities to determine how similar two regions within a slice are. Our approach is to compare codebook regions by comparing their Hebbian projections onto co-occurring slices. This process is illustrated in Figure 9.

The problem of measuring distances between prototypes is thereby transformed into a problem of measuring similarity between spatial probability distributions. The distributions are spatial because the codebook regions have definite locations within a slice, which are subspaces of $\mathbb{R}^n$. Hebbian projections are thus spatial distributions on *n*-dimensional data. It is therefore not possible to use one dimensional metrics, e.g., Kolmogorov-Smirnov distance, to compare them because doing so would throw away the essential spatial information within each slice.

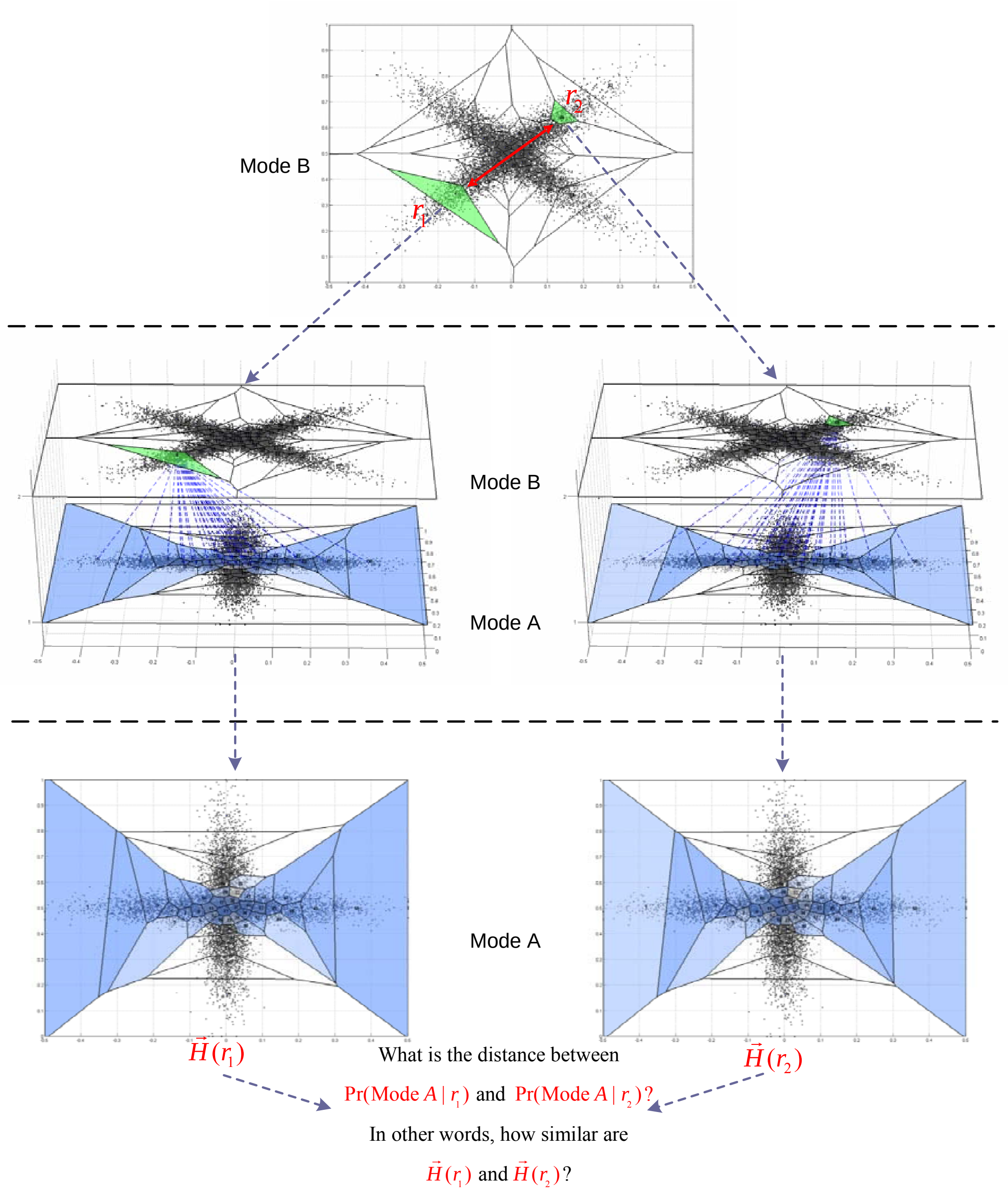

Mode B
r_2
r_1
Mode B
Mode A
Mode A
$\vec{H}(r_1)$
$\vec{H}(r_2)$
What is the distance between
$\Pr(\text{Mode } A \mid r_1)$ and $\Pr(\text{Mode } A \mid r_2)$?
In other words, how similar are
$\vec{H}(r_1)$ and $\vec{H}(r_2)$?

Figure 3.9 – Our approach to computing distances cross-modally. To determine the distance between codebook regions r_1 and r_2 in Mode B on top, we project them onto a co-occurring modality (Mode A) as shown in the middle. We then ask: *how similar are their Hebbian projections onto Mode A?*, as shown on the bottom. We have thereby transformed a question about distance between regions into a question of similarity between the spatial probability distributions provided by their Hebbian projections.

3.6 Defining Similarity

What does it mean for two things to be similar? This deceptively difficult question is at the heart of mathematical clustering and perceptual categorization and is common to a number of fields, including computer vision, statistical physics, and information and probability theory. The goal of measuring similarity between different things is often cast as a problem of measuring distances between multidimensional distributions on descriptive features. For example, in computer vision, finding minimum matchings between image feature distributions is a common approach to object recognition (Belongie et al. 2002).

In this section, we present a new metric for measuring similarity between spatial probability distributions, i.e., distributions on multidimensional metric spaces. We will use this metric to compute distances between codebook regions by comparing their Hebbian projections onto co-occurring modalities, as shown above in Figure 3.9. Our approach is therefore inherently multimodal – although we may be unable to determine a priori how similar two codebook regions are in isolation (i.e., unimodally), we can measure their similarity by examining how they are viewed from the perspectives of other co-occurring sensory channels. We therefore want to formulate a similarity metric on Hebbian projections that tells us not how far apart they are but rather, how similar they are to one another. This will enable perceptual bootstrapping by allowing us to answer a fundamental question:

> *Can any other modality distinguish between two regions in the same codebook? If not, then they represent the same percept.*

3.6.1 Introduction

There are a wide variety of metrics available to quantify distances between probability distributions (see the surveys in Rachev 1991, Gibbs and Su 2002). We may contrast these in many ways, including whether they are actually metrics (i.e., symmetric and satisfy the triangle inequality), the properties of their state spaces, their computational complexity, whether they admit practical bounding techniques, etc. For example, the common χ^2 distance is not a metric because it is asymmetric. In contrast, Kolmogorov-Smirnov distance is a metric but is defined only over $\mathbb{R}^1$. Choosing an appropriate metric for a given problem is a fundamental step towards solving it and can yield important insights into its internal structure.

In discussions of probability metrics, the notion of similarity generally follows directly from the definition of distance. Two distributions are deemed similar if the distance between them is small according to some metric; conversely, they are deemed dissimilar when the metric determines they are far apart. In our approach, we will reverse this dependency. We first intuitively describe our notion of similarity and then formulate a metric that computes it in a well-defined way. We call this metric the *Similarity distance* and it is the primary contribution of this section. Our approach is applicable to comparing distributions over any metric space and has a number of interesting properties, such as scale invariance, that make it additionally useful for work beyond the scope of this thesis.

3.6.2 Intuition

We begin by first examining similarity informally. Consider the two simple examples shown in Figure 3.10. Each shows two overlapping Gaussian distributions, whose similarity we would like to compare. Intuitively, we would say the distributions in Example A (on the left) are more similar to one another than those in Example B (on the right), because we will think of similarity as a measure of the overlap or proximity of spatial density. We are not yet concerned with formally defining similarity, but the intuition in these examples is exactly what we are trying to capture. Notice that the distributions in Example A cover roughly two orders of magnitude more area than those in Example B. Therefore, if we were to derive similarity from distance, the strong Euclidean bias incorporated into a wide variety of probability metrics would lead us to the opposite of our desired result. Namely, because the examples in B are much closer than those in A, we would therefore deem them more similar, thereby contradicting our

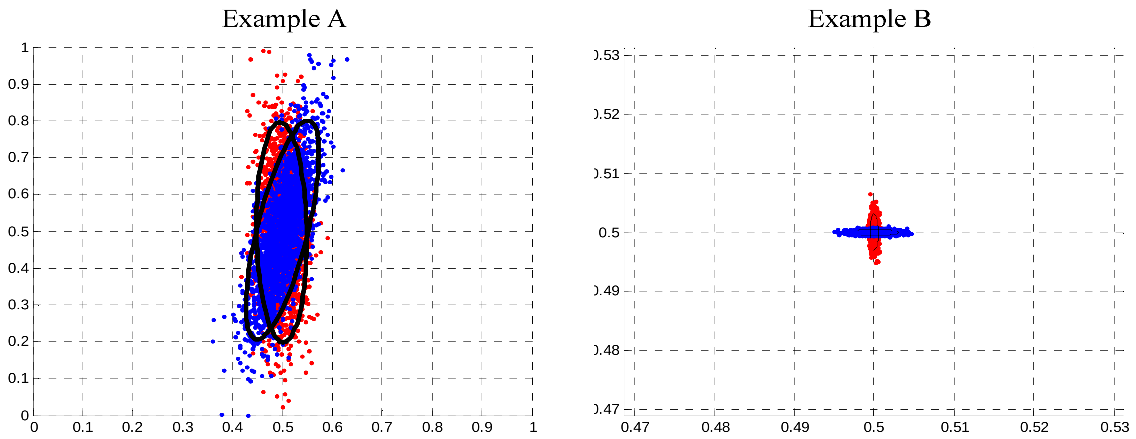


Figure 3.10 – Intuitively defining similarity. We consider the two distributions illustrated in Example A to be far more similar to one another than those in Example B, even though many metrics would deem them further apart due to inherent Euclidean biases. Notice that the distributions in Example A cover roughly two orders of magnitude more area than those in Example B. Note that simply normalizing the distributions before computing some metric on them would be ad hoc, very sensitive to outliers, and make common comparisons difficult.

desired meaning. Because we are looking for a distance measure based on similarity – and not a direct similarity measure – *it should smaller for things that are more similar and larger for things that are less similar*. This is the opposite of what one would expect were we directly formulating a measure of similarity, which presumably would be higher for more similar distributions.

Note that we cannot simply normalize pairs of distributions before computing some metric on them because our results would be highly sensitive to outliers. Doing so would also make common comparisons difficult, which is particularly important when demonstrating convergence in a sequence of probability measures. Finally, we want our similarity metric to be distribution-free and make no assumptions about the underlying data, which would make generalizing a simple normalization schema difficult.

3.6.3 Probabilistic Framework

We begin with some formal definitions. Our approach will be to define *Similarity distance* D_S as the ratio between two other metrics. These are the Kantorovich-Wasserstein distance and a new metric we introduce called the *one-to-many* distance. For each of these, we will provide a definition over continuous distributions and then present equivalent formulations for discrete weighted point sets. These are more computationally efficient for computing Similarity distance on the slice data structures introduced earlier. After this formal exposition, we intuitively explain and motivate these metrics in Section 3.6.4 and then show how Similarity distance is derived from them.

3.6.3.1 Kantorovich-Wasserstein Distance

Let μ and ν and be distributions on state space $\Omega = \mathbb{R}^n$. The Kantorovich-Wasserstein distance D_W (Kantorovich 1942, Gibbs and Su 2002) between μ and ν may be defined:

$$D_W\left(\mu,\nu\right) = \inf_J \left\{D(x,y):\ L(x)=\mu,\ L(y)=\nu\right\} \tag{3.5}$$

where the infimum is taken over all joint distributions *J* on *x* and *y* with marginals $L(x)=\mu$ and $L(y)=\nu$ respectively. For brevity, we will refer to D_W simply as the Wasserstein distance. Notice that in order to compute the Wasserstein distance, we already need to have a distance metric *D* defined to calculate the infimum. Where does *D* come from? In fact, in the approach described above, isn't *D* supposed to be the Similarity distance D_S, because we are proposing to use Similarity distance to measure distances within slices? Thus, we seem to have a “chicken and egg” problem from the start. We will sidestep this by defining *D* recursively through an iterative function system on D_S. This will allow us to compute Similarity distance by incrementally refining our calculation of it.

The definition in (3.5) assumes the distributions are continuous. Hebbian projections, however, are discrete distributions (i.e., weighted point sets) because they are over the codebooks within a slice. We may therefore simplify our computation by carrying it out directly over these codebooks. To do so, we define the Wasserstein distance on weighted point sets corresponding to discrete probability distributions. Consider finite sets $r_1, r_2 \subset \Omega$ with point densities φ_1, φ_2 respectively. Then we have:

$$D_W\left(\left\langle r_1,\varphi_1\right\rangle,\left\langle r_2,\varphi_2\right\rangle\right) = \inf_J \left\{D(x,y):\ L(x)=\left\langle r_1,\varphi_1\right\rangle,\ L(y)=\left\langle r_2,\varphi_2\right\rangle\right\}$$

which by (Levina and Bickel 2001) is equal to:

$$D_W\left(\langle r_1,\varphi_1\rangle,\langle r_2,\varphi_2\rangle\right)=\tfrac{1}{m}\min_{j_1,\ldots,j_m}\sum_{i=1}^{m}\left[D\left(\langle r_1,\varphi_1\rangle_i,\langle r_2,\varphi_2\rangle_{j_i}\right)^2\right]^{1/2} \tag{3.6}$$

where *m* is the maximum of the sizes of r_1 and r_2, the minimum is taken over all *permutations* of $\{1,\ldots,m\}$, and $\langle r_a,\varphi_a\rangle_i$ is the i^{th} element of set $\langle r_a,\varphi_a\rangle$. We note that (ibid.) has shown this is equivalent to the *Earth Mover's distance* (Rubner et al. 1998), a popular empirical measure used primarily in the machine vision community, when they are both computed over probability distributions.

We can now define the Wasserstein distance between Hebbian projections of $r_1,r_2\subseteq M_A$ onto M_B as:

$$D_W\left(\vec{H}(r_1),\vec{H}(r_2)\right)=\inf_J\left\{D(x,y):\ L(x)=\vec{H}(r_1),\ L(y)=\vec{H}(r_2)\right\} \tag{3.7}$$

where the infimum is taken over all joint distributions *J* on *x* and *y* with marginals $\vec{H}(r_1)$ and $\vec{H}(r_2)$. By (3.6), we have this is equal to:

$$D_W\left(\vec{H}(r_1),\vec{H}(r_2)\right)=\tfrac{1}{m}\min_{j_1,\ldots,j_m}\sum_{i=1}^{m}\left[D\left(\vec{H}(r_1)_i,\vec{H}(r_2)_{j_i}\right)^2\right]^{1/2} \tag{3.8}$$

where *m* is the number of codebook regions in M_B, the minimum is taken over all *permutations* of $\{1,\ldots,m\}$, and $\vec{H}(r)_i = i^{th}$ component of $\vec{H}(r)$.

We note that the Wasserstein distance presented above is not a candidate for measuring similarity. In fact, referring back to Figure 3.10, the red and blue distributions in Example A here are further apart as determined by the Wasserstein distance than those in Example B, i.e., $D_W(\text{Example A}) > D_W(\text{Example B})$, which does not capture our intended meaning of similarity. The additional measure needed will be described shortly, following a brief discussion of how D_W can be computed efficiently.

3.6.3.2 Computational Complexity

The optimization problem in (3.6) was first proposed by Monge (1781) and is known as the Transportation Problem. It involves combinatorial optimization because the minimum is taken over $O(2^m)$ different permutations and can be solved by Kuhn's Hungarian method (1955, see also Frank 2004). However, by treating it as a flow problem, we instead use the Transportation Simplex method introduced by Dantzig (1951) and subsequently enhanced upon by Munkres (1957), which has worst case exponential time but in practice is quite efficient (Klee and Minty 1972).

To get some insight into the structure of this problem, we take a moment to examine the complexity of determining exact solutions to it according to (3.8). Although this is not necessary in practice, it is instructive to see how the choice of mixture distributions influences the complexity of the problem and the implications this has for selecting perceptual features. Notice that in the minimization in equation (3.8), the vast majority of permutations can be ignored because we only need examine regions that have non-zero probabilities in the Hebbian projections. In other words, we could choose to ignore any region $q_i \subseteq M_B$ where $\max(\vec{H}(r_1)_i, \vec{H}(r_2)_i) \leq \varepsilon$ for some small ε. A conservative approach would set $\varepsilon = 0$, however, one can certainly imagine using a slightly higher threshold to simultaneously reduce noise and computational complexity.

We may estimate the running time of calculating D_W exactly by asking how many non-zeros values we expect to find in the Hebbian projections onto mode M_B of two regions in mode M_A. Let us suppose that mode M_B actually has *d* events (of equal likelihood) distributed over *m* codebook regions. How many codebook regions are there within each event? If the events do not overlap, then we expect that each perceptual event is covered by *m/d* codebook regions, due to the density normalization performed during codebook generation. In this case, the minimization must be performed over $O(2^{1+m/d})$ permutations. Alternatively, it is possible for all of the sensory events to overlap, giving an upper bound, worst case of *m* regions per event and $O(2^m)$ running time. Thus, the running time is a function of the event mixture distributions as much as it is the number of codebooks. When Hebbian distributions are "localized," in the sense they are confined to subsets of the codebook regions, the worst-case running time is closer to $O(2^{1+m/d})$.

We can optimize the computation by taking advantage of the fact that the projections are over identical codebooks, i.e., their spatial distributions are over the same set of points generated by the hyperclustering of M_B. In the general statement of the Transportation Problem, this need not be the case. We can therefore reduce the number of codebook regions involved by removing the intersection of the projections from the calculation. Where they overlap, namely, the distribution described by a normalized $\min\left(\vec{H}(r_1), \vec{H}(r_2)\right)$, we know the Wasserstein distance between them is 0. Therefore, let:

$$\vec{H}'(r_1) \quad = \vec{H}(r_1) - \min\left(\vec{H}(r_1), \vec{H}(r_2)\right) \tag{3.9}$$

$$\vec{H}'(r_2) \quad = \vec{H}(r_2) - \min\left(\vec{H}(r_1), \vec{H}(r_2)\right) \tag{3.10}$$

$$\Delta \qquad = 1 - \sum \min\left(\vec{H}(r_1), \vec{H}(r_2)\right) \tag{3.11}$$

We then have:

$$D_W\left(\vec{H}(r_1), \vec{H}(r_2)\right) = \Delta \, D_W\left(\vec{H}'(r_1), \vec{H}'(r_2)\right) \tag{3.12}$$

In other words, the Wasserstein distance computed over a common codebook (3.12) is equal to the distance computed on the distributions ((3.9) and (3.10)) over their non-intersecting mass (3.11). (Note that we must normalize (3.9) and (3.10) to insure they remain probability distributions, but the reader may assume this normalization step is always implied when necessary.) When the distributions overlap strongly, which we previously identified as the worst case scenario, we can typically use this optimization to cut the number of involved codebooks regions in half. When the distributions do not overlap, this optimization provides no benefit, but as we have already noted, this is a best case scenario and optimization is less necessary. As a further enhancement, we could also establish thresholds for Δ to avoid calculating D_W altogether. For example, in the case where their non-intersecting mass is extremely small, we might chose to define $D_W = 0$ or some other approximation .

In summary then, the computational complexity of exactly computing the Wasserstein distance very much rests on the selection of mixture distributions over which it is computed. These in turn depend upon the *feature selection* used in our perceptual algorithms, which directly determine the distributions of sensory data within a slice. We say that "good" features are ones that tend to restrict Hebbian projections to smaller subsets of slices and to reduce the amount of overlap among detectable perceptual events. (We suspect one can directly formulate a measure of the entropy in features based on these criteria but have not done so here.) Empirically, "good" features for computing the Wasserstein distance tend to be similar to the ones we naturally select when creating artificial perceptual systems. "Bad" features provide little information because their values are difficult to separate, i.e., they have high entropy. Later in the thesis, we will draw biological evidence for these theses idea from (Ernst and Banks 2002) .

3.6.3.3 The One-to-Many Distance

We now introduce a new distance metric called the *one-to-many* distance. Afterwards, we examine this metric intuitively and show how it naturally complements the Wasserstein distance. We will use these metrics together to formalize our intuitive notion of similarity.

Let *f* and *g* be the respective density functions of distributions μ and ν on state space $\Omega = \mathbb{R}^n$. Then the one-to-many distance (D_{OTM}) between μ and ν is:

$$\begin{aligned} D_{OTM}(\mu,\nu) &= \int_\mu f(x)\cdot D_W(x,\nu)\; dx \\ &= \int_\mu \int_\nu f(x)\cdot g(y)\cdot d(x,y)\; dxdy \\ &= \int_\nu g(y)\cdot D_W(\mu,y)\; dy = D_{OTM}(\nu,\mu) \end{aligned}$$

We define this over weighted pointed sets as:

$$\begin{aligned} D_{OTM}\left(\langle r_1,\varphi_1\rangle,\langle r_2,\varphi_2\rangle\right) &= \sum_{p_i\in r_1}\varphi_i\ D_W\left(p_i,\langle r_2,\varphi_2\rangle\right) \\ &= \sum_{p_i\in r_1}\varphi_i\sum_{p_j\in r_2}\varphi_j\ d(p_i,p_j) \\ &= \sum_{p_j\in r_2}\varphi_j\sum_{p_i\in r_1}\varphi_i\ d(p_i,p_j) \\ &= D_{OTM}\left(\langle r_2,\varphi_2\rangle,\langle r_1,\varphi_1\rangle\right) \end{aligned}$$

The one-to-many distance is the weighted sum of the Wasserstein distances between each individual point within a distribution and the entirety of another distribution. It is straightforward to directly calculate D_W between a point and a distribution and computing D_{OTM} requires $O(m^2)$ time, where *m* is the maximum number of weighted points in the distributions. Also, we note from these definitions that D_{OTM} is symmetric.

From this definition, we can now formulate D_{OTM} between Hebbian projections of $r_1, r_2 \subseteq M_A$ onto M_B:

$$D_{OTM}\left(\vec{H}(r_1),\vec{H}(r_2)\right) = \sum_{i\in\vec{H}(r_1)}\omega_i\ D_W\left(i,\vec{H}(r_2)\right) \quad (3.13)$$

$$= \sum_{i\in\vec{H}(r_1)}\omega_i\sum_{j\in\vec{H}(r_2)}\omega_j D(i,j) \quad (3.14)$$

where $\omega_i = \vec{H}(r_1)_i$ and $\omega_j = \vec{H}(r_2)_j$. Recall that $\vec{H}(r)_i = i^{th}$ component of $\vec{H}(r)$.

We see this is symmetric:

$$D_{OTM}\left(\vec{H}(r_1),\vec{H}(r_2)\right) = D_{OTM}\left(\vec{H}(r_2),\vec{H}(r_1)\right)$$

The one-to-many distance is a weighted sum of the Wasserstein distances between each individual region in a projection and the other projection in its entirety. The weights are taken directly from the original Hebbian projections. This is represented by the term $\omega_i\ D_W\left(i,\vec{H}(r_2)\right)$ in (3.13). We note that the Wasserstein distance between a single region and a distribution is trivial to compute directly, as shown in (3.14), and it can be calculated in $O(m)$ time, where *m* is the number of codebooks in Mode B. Therefore, D_{OTM} can be computed in $O(m^2)$ time.

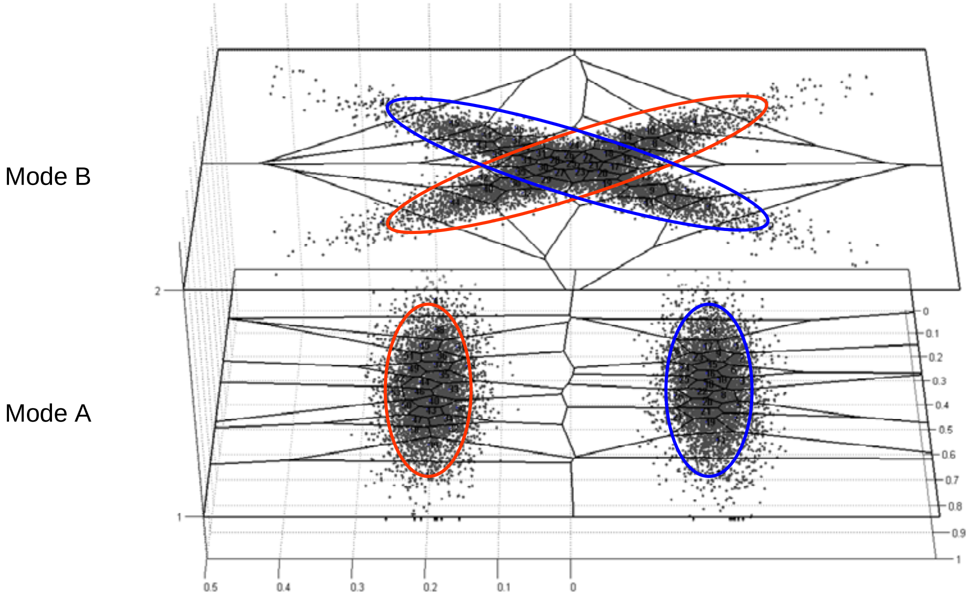


Figure 3.11 – A simpler example. Mode B is the same as in previous examples but the events in Mode A have been separated so they do not overlap. The colored ellipses show how the external *red* and *blue* events probabilistically appear within each mode.

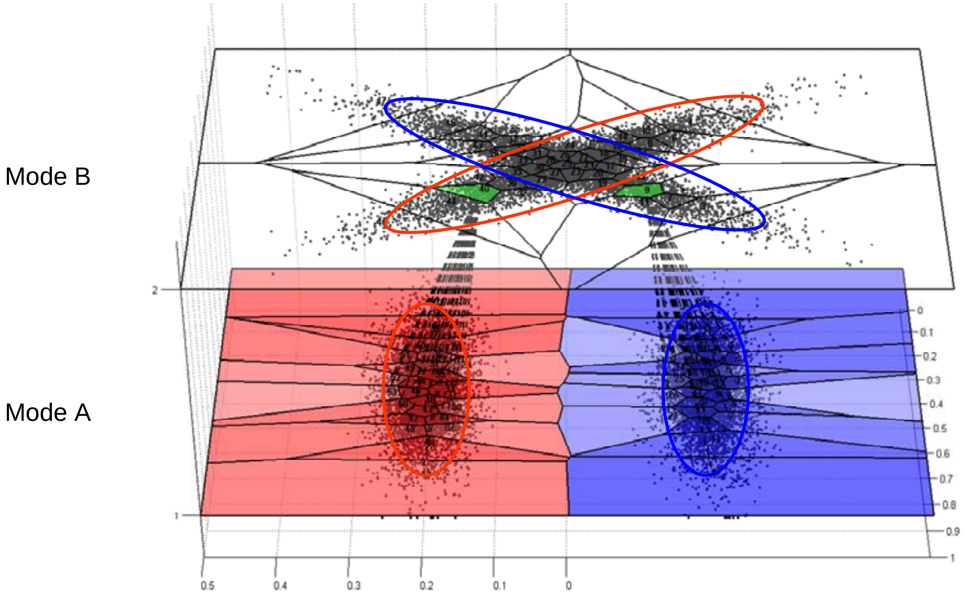


Figure 3.12 – Hebbian projections from regions in Mode B onto Mode A. Notice that the Hebbian projections have no overlap, which simplifies the visualization and discussion in the text. However, this does not affect the generality of the results.

3.6.4 Visualizing the Metrics: A Simpler Example

Consider the example in Figure 3.11. We have modified Mode A, on the bottom, so that its events no longer overlap. (Mode B on top remains unchanged.) This will simplify the presentation but does not affect the generality of the results presented here. As before, the two world events perceived by each mode are delineated with colored ellipses for the benefit of the reader but the modes themselves have no knowledge of them. The Hebbian projections from two codebook regions in Mode B are shown in Figure 3.12. We see this example was designed so that the projections have no overlap, making it easy to view them independently.

We now give an intuitive interpretation of the two distance metrics, D_W and D_{OTM}, based on the classic statement of the Transportation Problem (Monge 1781). This problem is more naturally viewed with discrete distributions, but the presentation generalizes readily to continuous distributions. Consider the Hebbian projections from our example in isolation, as show in Figure 3.13. On the left, $\vec{H}(r_1)$ is shown in red, and $\vec{H}(r_2)$ is shown in blue on the right. The shading within each Voronoi region is proportional to its weight (i.e., point density) within its respective distribution.

In the Transportation Problem, we imagine the red regions (on the left) need to deliver supplies to the blue regions (on the right). Each red region contains a mass of supplies proportional to its shading and each blue region is expecting a mass of supplies proportional to its shading. (We know that mass being "shipped" is equal to the mass being "received" because they are described by probability distributions.) The one-to-many distance is how much work would be necessary to deliver all the material from the red to blue regions, if each region had to independently deliver its mass proportionally to all regions in the other distribution. Work here is defined as $mass \times distance$.

The Wasserstein distance computes the minimum amount of work that would be necessary if the regions cooperate with one another. Namely, red regions could deliver material to nearby blue regions on behalf of other red regions, and blue regions could receive material from nearly red regions on behalf of other blue regions. Nonetheless, we maintain the restriction that each region has a maximum amount it can send or receive, corresponding to its point density. This is why the Wasserstein distance computes the solution to the Transportation Problem, which is directly concerned with this type of delivery optimization.

Thus, we may summarize that D_{OTM} computes an unoptimized Transportation Problem, where cooperation is forbidden, and that D_W computes the optimized Transportation Problem, where cooperation is required.

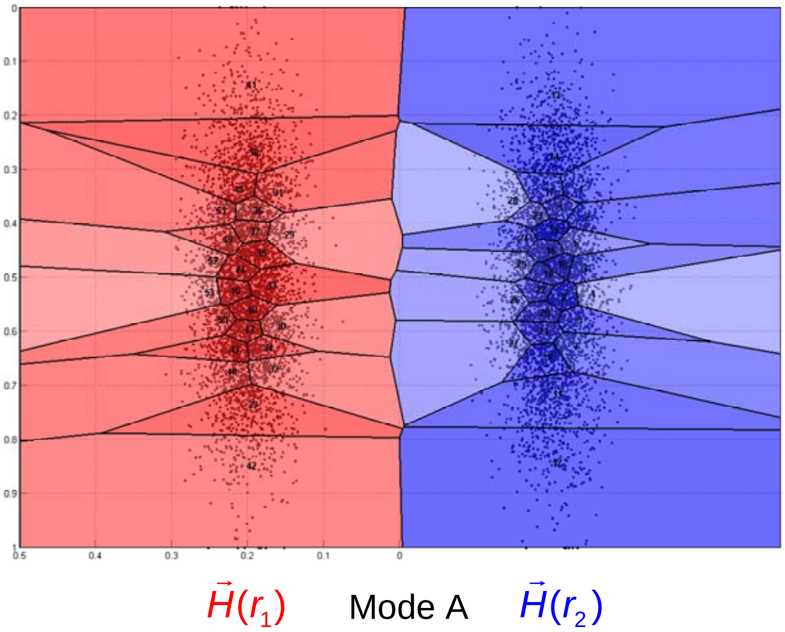


Figure 3.13 – Visualizing D_W and D_{OTM} through the Transportation Problem. We examine the Hebbian projections onto Mode A shown in Figure 3.12. $\vec{H}(r_1)$ is shown in red on the left and $\vec{H}(r_2)$ is shown in blue on the right. Each region is shaded according to its point density. In the Transportation Problem, we want to move the "mass" from one distribution onto the other. If we define $work = mass \times distance$, then D_{OTM} computes the work required if each codebook region must distribute its mass proportionally to all regions in the other distribution. D_W computes the work required if the regions cooperatively distribute their masses, to minimize the total amount of work required.

3.6.5 Defining Similarity

We are now in a position to formalize the intuitive notion of similarity presented above. We define a new metric called the *Similarity distance* (D_S) between continuous distributions μ and ν:

$$D_S(\mu,\nu) = \frac{D_W(\mu,\nu)}{D_{OTM}(\mu,\nu)}$$

and over weighted point sets:

$$D_S(\langle r_1,\varphi_1\rangle,\langle r_2,\varphi_2\rangle) = \frac{D_W\left(\langle r_1,\varphi_1\rangle,\langle r_2,\varphi_2\rangle\right)}{D_{OTM}\left(\langle r_1,\varphi_1\rangle,\langle r_2,\varphi_2\rangle\right)}$$

We thereby define the Similarity distance between Hebbian projections of $r_1, r_2 \subseteq M_A$ onto M_B:

$$D_S\left(\vec{H}(r_1),\vec{H}(r_2)\right) = \frac{D_W\left(\vec{H}(r_1),\vec{H}(r_2)\right)}{D_{OTM}\left(\vec{H}(r_1),\vec{H}(r_2)\right)} \tag{3.15}$$

The Similarity distance is the ratio of the Wasserstein to the one-to-many distance. It measures the optimization gained when transferring the mass between two spatial probability distributions if cooperation is allowed. Intuitively, it normalizes the Wasserstein distance. It is scale invariant (see Figure 3.14) and captures our desired notion of similarity.

An important note to avoid confusion: Because D_S is a distance measure based on similarity – and not a similarity measure – *it is smaller for things that are more similar and larger for things that are less similar*. So, for any distribution ν, $D_S(\nu,\nu)=0$, expressing the notion that anything is (extremely) similar to itself.

We briefly examine the behavior of D_S at and in between its limits. Let $\vec{H}(r_1)$ and $\vec{H}(r_2)$ be identical Hebbian projections separated by some distance Δ. Then we have the following properties:

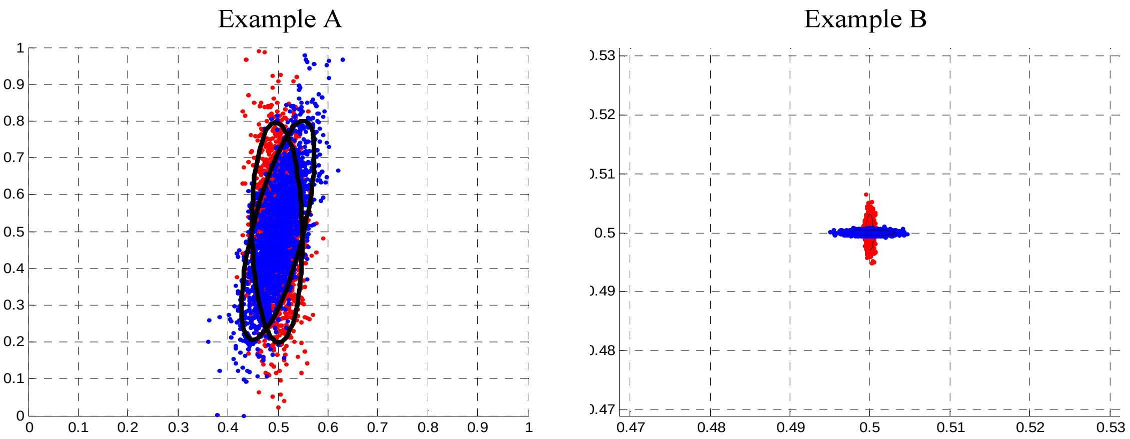


Figure 3.14– Examining Similarity distance. Comparing the distributions in the two examples, we have $D_S\text{(Example A)} \ll D_S\text{(Example B)}$, which captures our intuitive notion of similarity.

1) We see that as the distributions are increasingly separated, the optimization provided in the Wasserstein calculation disappears:

$$\lim_{\Delta\to\infty} D_W\left(\vec{H}(r_1),\vec{H}(r_2)\right) = D_{OTM}\left(\vec{H}(r_1),\vec{H}(r_2)\right)$$

and therefore:

$$\lim_{\Delta\to\infty} D_S\left(\vec{H}(r_1),\vec{H}(r_2)\right) = 1$$

2) As the distributions are brought closer together, the Wasserstein distance decreases much faster than the One-To-Many distance:

$$\frac{\partial}{\partial\Delta} D_W\left(\vec{H}(r_1),\vec{H}(r_2)\right) > \frac{\partial}{\partial\Delta} D_{OTM}\left(\vec{H}(r_1),\vec{H}(r_2)\right)$$

3) As they approach, it eventually dominates the calculation:

$$\lim_{\Delta\to 0} D_W(\vec{H}(r_1),\vec{H}(r_2)) = 0 \text{ and therefore, } \lim_{\Delta\to 0} D_S(\vec{H}(r_1),\vec{H}(r_2)) = 0$$

4) So, we see that $0 \le D_S(\vec{H}(r_1),\vec{H}(r_2)) \le 1$ and D_S varies non-linearly between these limits.

On the next two pages, we visualize the dependence of D_S on the distance Δ and the angle θ between pairs of samples drawn from different distributions. We examine samples drawn from Gaussian and Beta distributions in Figure 3.15 and Figure 3.16 respectively.

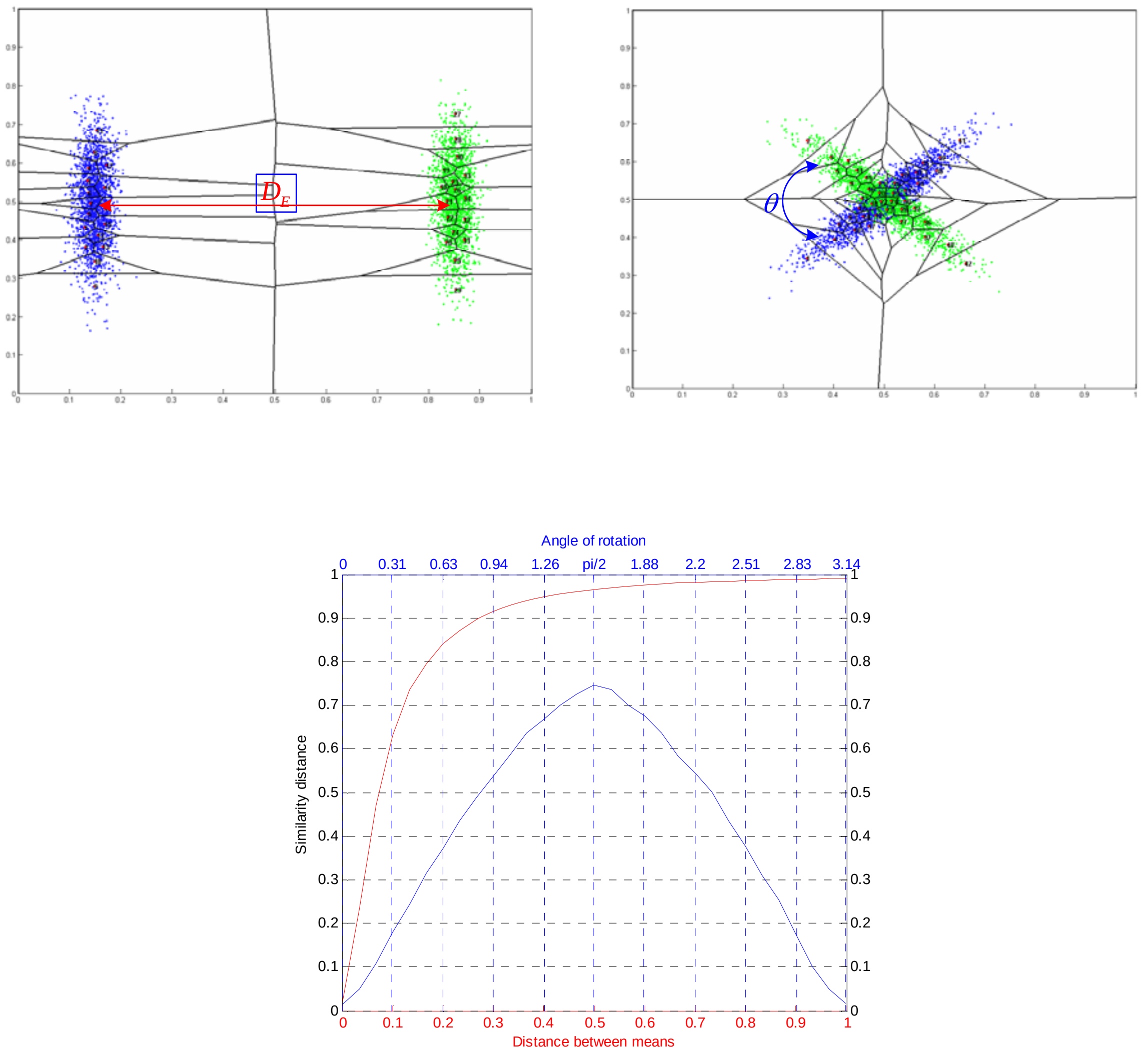


Figure 3.15 – Effects on D_S as functions of the distance D_E and angle θ between Gaussian distributions. As the distance D_E or angle between θ two Gaussian distributions decreases, we see how their corresponding similarity distance D_S decreases non-linearly in the graph on the bottom. D_E is shown in red and θ is shown in blue.

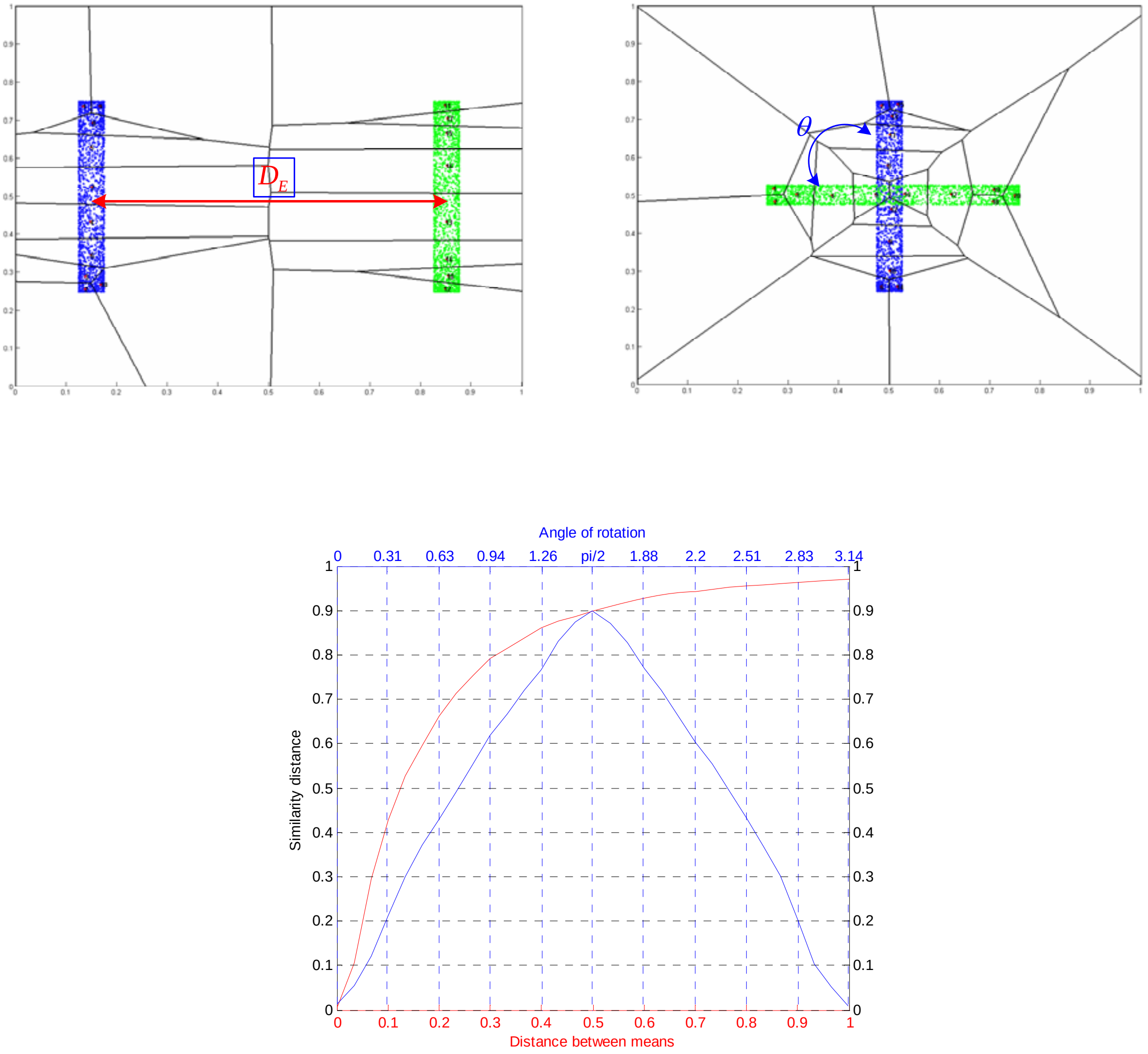


Figure 3.16 –Effects on D_S as functions of the distance D_E and angle θ between Beta distributions. As the distance D_E or angle between θ the two Beta distributions decreases, we see how their corresponding similarity distance D_S decreases non-linearly in the graph on the bottom. D_E is shown in red and θ is shown in blue.

3.6.5.1 A Word about Generality

We can use Similarity distance to compare arbitrary discrete spatial probability distributions. Several such comparisons are illustrated in Figure 3.17. These examples are important, because our being able to compute Similarity distances between these distributions means that we will be able to perceptually ground events drawn from their mixtures. That all our examples have so far involved mixtures of Gaussians has simply been for convenience. We will demonstrate later in this chapter that we can separate events corresponding to a wide assortment of mixture distributions such as the ones shown here.

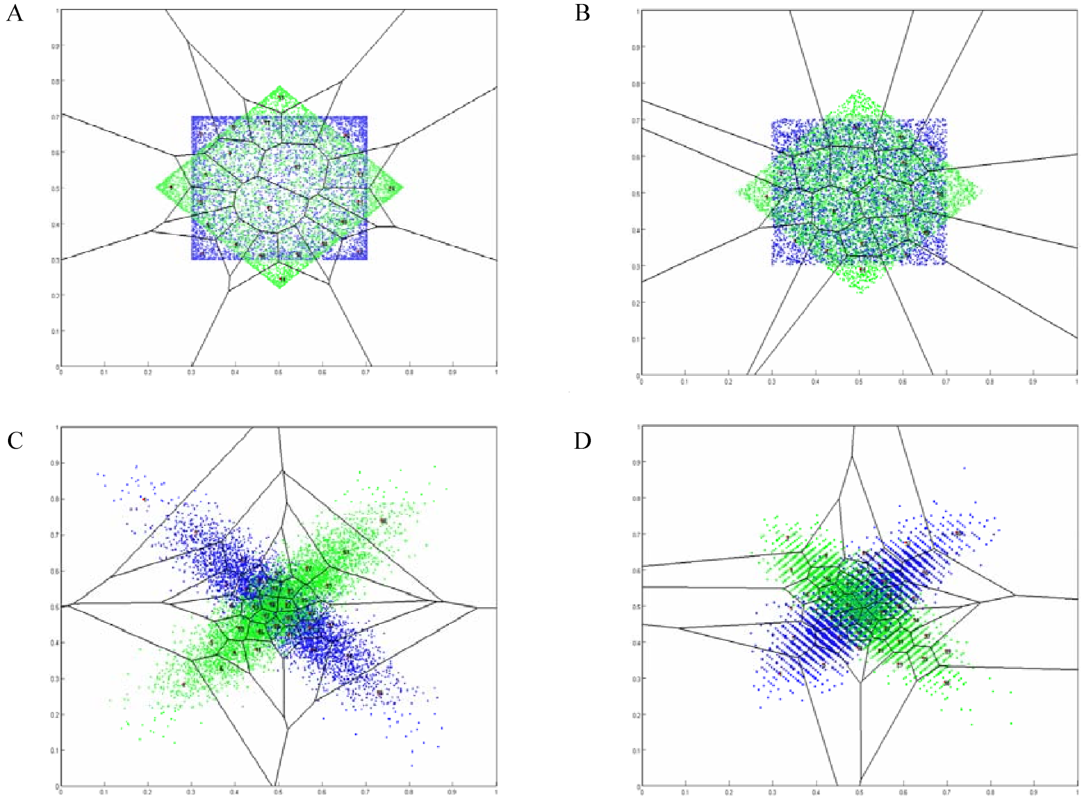


Figure 3.17 – Comparing spatial probability distributions. In each slice, the green and blue points represent samples drawn from equivalent but rotated 2-dimensional distributions. For each example, we identify the source distribution and the Similarity distance between the green and blue points. (A) A 2-D beta distributions with a, b = 4. Note the low density of points in the center of the distributions and the corresponding sizes of the codebook regions. $D_S = 0.22$. (B) A 2-D uniform distribution. $D_S = 0.11$. (C) A 2-D Gaussian distribution with $\sigma = (.15, .04)$. $D_S = 0.67$. (D) A 2-D Poisson distribution, with $\lambda = (50,30)$ and scaled by (.8, .3). $D_S = .55$.

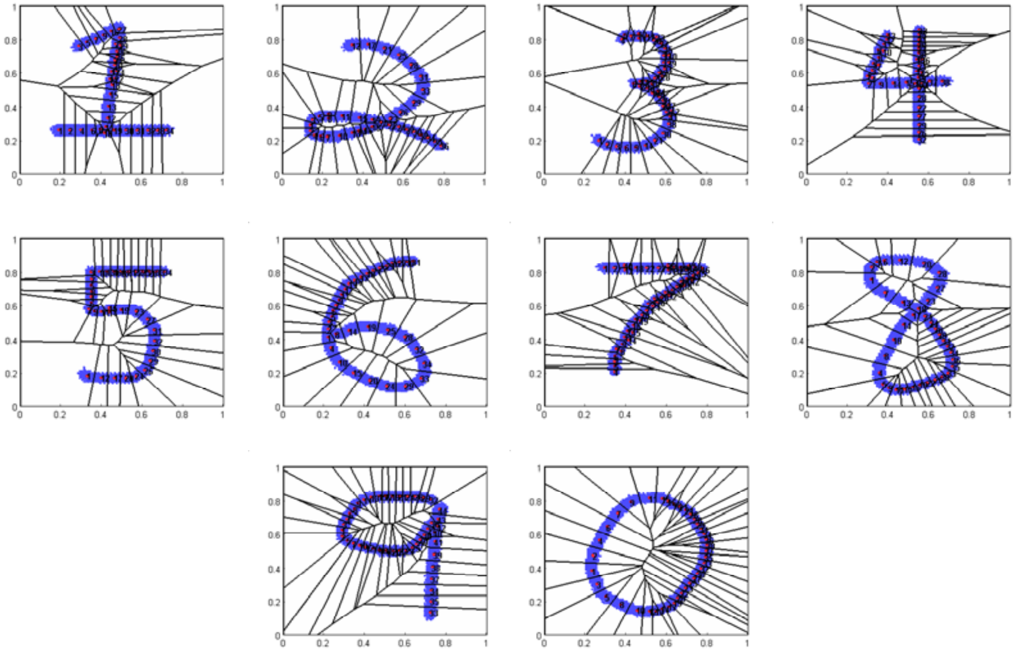


Figure 3.18 – Some familiar non-parametric probability distributions within codebooks.

We can also apply the Similarity distance to non-parametric distributions. For example, consider the familiar distributions shown in Figure 3.18. We have examined using D_S for handwriting recognition, and one can notice some of the well known properties of codebooks constructed on contours in the above images, for example, how they capture the medial axes.

3.7 Defining the Distance between Regions

We now use Similarity distance to define the *Cross-Modal distance* (D_{CM}) between two regions $r_1, r_2 \in M_A$ with respect to mode M_B:

$$D_{CM}(r_1, r_2) \quad = \left[(1-\lambda)\left[D_E(r_1,r_2)\right]^2 \; + \; \lambda \left[\sqrt{2}\; D_S(\vec{H}(r_1), \vec{H}(r_2)) \right]^2 \right]^{1/2} \tag{3.16}$$

$$= \left[(1-\lambda) \; \left[D_E(r_1,r_2)\right]^2 + 2\lambda \left(\frac{D_W\left(\vec{H}(r_1), \vec{H}(r_2)\right)}{D_{OTM}\left(\vec{H}(r_1), \vec{H}(r_2)\right)} \right)^2 \right]^{1/2} \tag{3.17}$$

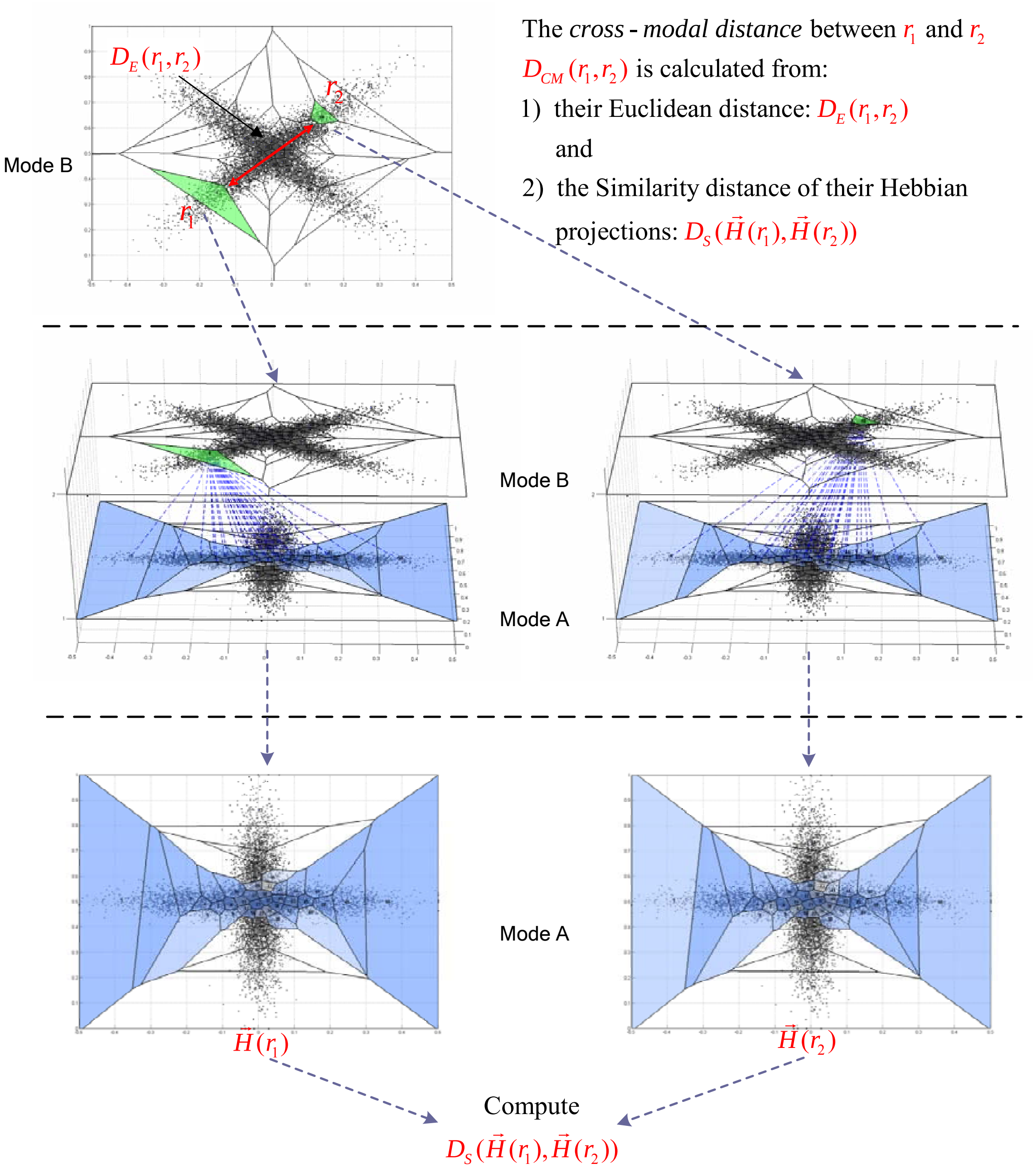
where D_E is Euclidean distance and λ is the relative importance of cross-modal to Euclidean distance. Thus the distance between two regions within a slice is defined to have some component $(1-\lambda)$ of their Euclidean distance and some component (λ) of the Similarity distance between their Hebbian projections. This is illustrated in Figure 3.19.

In almost all uses of cross-modal distance in this thesis, we set $\lambda = 1$ and ignore Euclidean distance entirely. However, in some applications, e.g., hand-writing or drawing recognition, spatial locality within a slice is important because it is a fundamental component of the phenomenon being recognized. If so, we can use a lower of λ. Determining the proper balance between Euclidean and Similarity distances is an empirical process for such applications.

So far, we have only considered two co-occurring modes simultaneously to keep the examples simple. However, it is straightforward to generalize the definition to incorporate additional modalities, and the calculation scales linearly with the number of modalities involved. To define the cross-modal distance (D_{CM}) between two regions $r_1, r_2 \in M_A$ with respect to a set of co-occurring modes $M_I \in \mathrm{M}$, we define:

$$D_{CM}(r_1, r_2) = \left[(1-\lambda_E)\left[D_E(r_1, r_2)\right]^2 \; + \; 2\sum_{M_I \in \mathrm{M}} \lambda_I \; \left[\; D_S(\vec{H}_A^I(r_1), \vec{H}_A^I(r_2))\right]^2 \right]^{1/2} \qquad (3.18)$$

where the contributions of each mode M_I is weighted by λ_I and we set $\lambda_E + \sum \lambda_I = 1$. For guidance in setting the values of the λ_I, we can turn to (Ernst and Banks 2002), who found that in intersensory influence, people give preference to senses which minimize the variance in joint perceptual interpretations, confirming an earlier prediction by (Welch and Warren 1986) about sensory dominance during multimodal interactions. This lends credence to our hypothesis in section 3.6.3.2 regarding the computational value of entropy minimization in the selection of perceptual features. We reexamine these issues in the dynamic model presented in Chapter 4.



The *cross-modal distance* between r_1 and r_2 $D_{CM}(r_1,r_2)$ is calculated from:

1) their Euclidean distance: $D_E(r_1,r_2)$ and
2) the Similarity distance of their Hebbian projections: $D_S(\vec{H}(r_1),\vec{H}(r_2))$

Figure 3.19 – Calculating the *cross-modal distance* between codebook regions in a slice. The distance is a function of their local Euclidean distance and the how similar they appear from the perspective of a co-occurring modality. To determine this for regions r_1 and r_2 in Mode B on top, we project them onto Mode A, as shown in the middle. We then compute the Similarity distance of their Hebbian projections, as shown on the bottom.

3.7.1 Defining a Mutually Iterative System

In this section, we show how to use the cross-modal distance function defined above to calculate the distances between regions within a slice. This statement may seem surprising. Why is any elaboration required to use D_{CM}, which we just defined? There are two remaining issues we must address:

1) We have yet to specify the distance function *D* used to define the Wasserstein distance in equations (3.7) and (3.8), which was also "inherited" in our definition of the one-to-many distance in equation (3.14).

2) By defining distances cross-modally, we have created a mutually recursive system of functions. Consider any two regions r_1, r_2 in mode M_A. When we calculate $D_{CM}(r_1, r_2)$, we are relying on knowing the distances between regions within another mode M_B, which are used to calculate $D_S\left(\vec{H}(r_1), \vec{H}(r_2)\right)$. However, the distances between regions in M_B are calculated exactly the same way but with respect to M_A. So, every time we calculate distances in a mode, we are implicitly changing the distances within every other mode that relies upon it. And of course, this means its own inter-region distances may change as a result! How do we account for this and how do we know such a system is stable?

We will approach both of these issues simultaneously. Suppose we parameterize the distance function *D* in all of our definitions:

$$D_W\left(\vec{H}(r_1), \vec{H}(r_2), D\right) \quad = \tfrac{1}{m} \min_{j_1,..,j_m} \sum_{i=1}^{m} \left[D\left(\vec{H}(r_1)_i, \vec{H}(r_2)_{j_i}\right)^2 \right]^{1/2} \qquad (3.19)$$

$$D_{OTM}\left(\vec{H}(r_1), \vec{H}(r_2), D\right) \quad = \sum_{i \in \vec{H}(r_1)} \omega_i \, D_W\left(i, \vec{H}(r_2), D\right) \qquad (3.20)$$

$$D_S\left(\vec{H}(r_1),\vec{H}(r_2),D\right) \quad = \frac{D_W\left(\vec{H}(r_1),\vec{H}(r_2),D\right)}{D_{OTM}\left(\vec{H}(r_1),\vec{H}(r_2),D\right)} \tag{3.21}$$

$$D_{CM}\left(r_1,r_2,D\right) = \left[(1-\lambda)\left[D_E(r_1,r_2)\right]^2 + 2\lambda \left[D_S(\vec{H}(r_1),\vec{H}(r_2),D)\right]^2\right]^{1/2} \tag{3.22}$$

We now define an iterative function system on modes M_A and M_B that *mutually* calculates D_{CM} over their regions:

Let $\Delta_t^X = D_{CM}$ in mode M_X at time *t*. Recall that D_E is Euclidean distance.

For all pairs of regions $r_i, r_j \in M_A$ and $q_i, q_j \in M_B$, we define:

$$\Delta_0^A(r_i,r_j) \quad = D_E(r_i,r_j) \tag{3.23}$$

$$\Delta_0^B(q_i,q_j) \quad = D_E(q_i,q_j) \tag{3.24}$$

$$\Delta_t^A(r_i,r_j) \quad = D_{CM}\left(r_i,r_j,\Delta_{t-1}^B\right) \tag{3.25}$$

$$\Delta_t^B(q_i,q_j) \quad = D_{CM}\left(r_i,r_j,\Delta_{t-1}^A\right) \tag{3.26}$$

Thus, we are start by assuming in (3.23) and (3.24) that the distances between regions in a slice are Euclidean, in the absence of any other information. (We later eliminate this assumption in the intermediate steps of cross-modal clustering, where we have good estimates on which to base the iteration.) The iterative steps are shown in (3.25) and (3.26) , where at time *t*, we recalculate the distances within each slice based upon the distances in the other slice at time $t-1$. For example, notice how the definition of $\Delta_t^A(r_i,r_j)$ calculates D_{CM} using Δ_{t-1}^B in (3.25). After all pairs of distances have been computed at time *t*, we can then proceed to compute them for time $t+1$. As we did in equation (3.18), we can easily generalize this system to include any number of mutually recursive modalities. The complexity again scales linearly with the number of modalities involved.

We stop the iteration when Δ_t^A and Δ_t^B begin to converge, which empirically tends to happen very quickly. Thus, we stop iterating on mode M_X at time *t* when:

$$\max_{r_i, r_j \in M_X} \frac{\left| \Delta_t^X(r_i, r_j) - \Delta_{t-1}^X(r_i, r_j) \right|}{\Delta_t^X(r_i, r_j)} < \kappa, \text{ for } \kappa = .9, \text{ we typically have } t \leq 4.$$

We will refer to this final value of Δ_t^X for any regions $r_i, r_j \in M_X$ as $\tilde{D}_{CM}\left(r_i, r_j\right)$.

With this, we complete our formal definition of the *slice* data structure. The final component necessary for specifying the topological manifold defined by a slice was the non-Euclidean distance metric between the hyperclustered regions. We now define this distance to be $\tilde{D}_{CM}$.

3.8 Cross-Modal Clustering

Recall that our goal has been to combine codebook regions to "reconstruct" the larger perceptual regions within a slice. The definition of the iterated cross-modal distance $\tilde{D}_{CM}$ in the previous section allows us to proceed, because it suggests how to answer the following fundamental question:

> *Can any other modality distinguish between two regions in the same codebook? If not, then they represent the same percept.*

Because $\tilde{D}_{CM}$ represents the distance between two regions *from the perspective of other modalities*, we will use it to define a metric that determines whether to combine them or not. If $\tilde{D}_{CM}\left(r_i, r_j\right)$ is sufficiently small for two regions $r_1, r_2 \subseteq M_A$, then we will say they are *indistinguishable* and therefore part of the same perceptual event. If $\tilde{D}_{CM}\left(r_i, r_j\right)$ between two regions is large, we will say they are *distinguishable* and therefore, cannot be part of the same perceptual event. These criteria suggest the general structure of our cross-modal clustering algorithm. One important detail remains: how small must $\tilde{D}_{CM}\left(r_i, r_j\right)$ be for us to say it is "sufficiently" small? How do we define the threshold for merging two regions? An earlier version of this work appeared in (Coen 2005).

3.8.1 Defining Self-Distance

We define the notion of *self-distance*, which measures the internal value of $\tilde{D}_{CM}$ within an individual region. Thus, rather than measure the distance between two different regions, which has been our focus so far, *self-distance* measures the internal cross-modal distance between points within a single region. It is a measure of internal coherence and will allow us to determine whether two different regions are sufficiently similar to merge.

Suppose we are considering merging two regions r_i and r_j within some slice *M*, where we have already determined their cross-modal distance $\tilde{D}_{CM}\left(r_i, r_j\right)$. For example, on the left in Figure 3.20 we are considering merging the green and blue regions. Let us hypothetically assume that we did merge them to create a new region r'. We will now

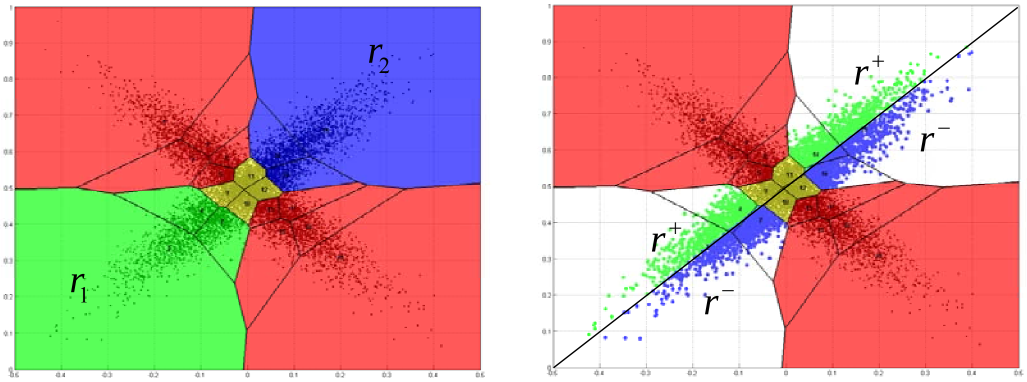


Figure 3.20 -- Determining when to merge two regions. On the left, we are considering merging the green and blue regions r_1 and r_2. To determine whether or not to proceed, we divide the candidate resulting region using a hyperplane generated by its principal components, as shown on the right. This generates two new regions r^+ and r^-. If $\tilde{D}_{CM}\left(r^+,r^-\right) << \tilde{D}_{CM}\left(r_i,r_j\right)$, then we determine the regions should not be merged. See the accompanying text for details.

immediately split this new region into two pieces, r^+ and r^-, as show on the right in Figure 3.20. We now ask: what is the cross-modal distance $\tilde{D}_{CM}\left(r^+,r^-\right)$ between these new regions? The intuition here is that if r_1 and r_2 really *are* part of the same region, then if we split r' differently, we should find the two new regions have approximately the same cross-modal distance $\tilde{D}_{CM}\left(r_i,r_j\right)$. In other words, the Hebbian projections of the new regions r^+ and r^- should be roughly as similar as the Hebbian projections of the original regions r_1 and r_2 we are considering merging, because they should co-occur in the same way with other modalities. If the distance $\tilde{D}_{CM}\left(r^+,r^-\right)$ is much less than $\tilde{D}_{CM}\left(r_i,r_j\right)$, then we have merged two regions that are actually different from one another. Why? *Because* we *would now be averaging the Hebbian projections of two genuinely different regions, which would drastically increase their similarity and therefore make* $\tilde{D}_{CM}\left(r^+,r^-\right)$ *much smaller than* $\tilde{D}_{CM}\left(r_i,r_j\right)$. The question remains then, how do we divide r'?

We will partition r' by fitting a linear orthogonal regression onto it. For a slice $M \subseteq \mathbb{R}^N$, this will generate an $(N-1)$-dimensional hyperplane that divides r' into two sets, r^+ and r^-, minimizing the perpendicular distances from them to the hyperplane.

Note that because the data are drawn from independent distributions, there is no error-free predictor dimension that generates the other dimensions according to some function. This is equivalent to the case where all variables are measured with error, and standard least squares techniques do not work in this circumstance. We therefore perform principal components analysis on the points in region r' and generate the hyperplane by retaining its $N-1$ principal components. This computes what is known as the orthogonal regression and works even in cases where all the data in r' are independent.

We use this hyperplane to partition r' into r^+ and r^-, as shown in Figure 3.20. We define the *self-distance* (D_{self}) of region r':

$$D_{self}(r') = \frac{\tilde{D}_{CM}\left(r^+, r^-\right)}{\tilde{D}_{CM}\left(r_1, r_2\right)} \tag{3.27}$$

If $D_{self}(r')$ < ½, then we do not combine r_1 and r_2, because this indicates we would be averaging two dissimilar Hebbian projections were the merger to occur. At the moment, this remains an empirical statement, but we note that this threshold is not a parameter of the cross-modal clustering algorithm and it is fixed throughout the results in this thesis. In practice, the self-distance value (D_{self}) tends towards either zero or one, which motivated our selection of ½ as the merger threshold. A more theoretical investigation of this empirical criterion is among our future work.

3.8.2 A Cross-Modal Clustering Algorithm

We now present an algorithm for combining codebook clusters into regions that represent the sensory events within a slice. This is done in a greedy fashion, by combining the closest regions according to $\tilde{D}_{CM}$ within each slice. We use the definition of *self-distance* to derive a threshold for insuring regions are sufficiently close to merge. Afterwards, we examine the algorithm and some examples of its output.

Cross-Modal Clustering:

Given: A set of slices M and λ, the parameter for weighting Euclidean to Similarity distances. For each slice $M_i \in M$, we will call its codebook $C_i = \left\{ p_1, ..., p_{k_i} \right\}$.

Initialization: For each slice $M_i \in M$, initialize a set of *regions* $R_i = C_i$. Each slice will begin with a set of regions based on its codebook. We will merge these regions together in the algorithm below.

Algorithm:

```
Calculate $\tilde{D}_{CM}$ over the slices in set M.

While (true) do:
    Calculate $\tilde{D}_{CM}$ over the slices in set M.  Use current $\tilde{D}_{CM}$ as t=0 value
    For each slice $M_i \in M$ :
        Sort the pairs of regions in $M_i$, $r_a, r_b \in R_i$, by $\tilde{D}_{CM}(r_a, r_b)$
            For each pair $r_a, r_b \in R_i$, in sorted order:
                If  $D_{self}(r') \geq .5$, where $r' = r_a \cup r_b$ :
                    Merge($r_a, r_b$)
                    Exit inner for loop.

                For each codebook cluster $p_i$ in $C_i$ :
                    Let $r = \min_{r \in R_i} \arg\left[ \tilde{D}_{CM}(p_i, r) \right]$
                    Move $p_i$ into region r

    If no regions were merged in any slice
        Either wait for new data or stop

Procedure Merge($r_a, r_b$):
    $r_a = r_a \cup r_b$.
    $R_i = R_i / r_b$.
    For all regions $r_c \in R_i$, set $\tilde{D}_{CM}(r_a, r_c) = \min\left( \tilde{D}_{CM}(r_a, r_c), \tilde{D}_{CM}(r_b, r_c) \right)$.
```

The cross-modal clustering algorithm initially creates a set of regions in each slice corresponding to its codebook. The goal is to merge these regions based on their cross-modal distances. The algorithm proceeds in a two-step greedy fashion:

1) For each slice, consider its regions in pairs, sorted by $\tilde{D}_{CM}$. If we find two regions r_a and r_b satisfying $D_{self}\left(r'\right) \geq .5$, where $r' = r_a \cup r_b$, we merge them and move onto the next step.

2) If as a result of this merger, some codebook cluster p_i is now closer to another region, we simply move it there.

When we merge two regions, we set the pairwise distances to other regions to be the minimum of the distances to the original regions, because we now view them as all part of the same underlying perceptual event and therefore equivalent to one another. At the end of each loop, we recompute $\tilde{D}_{CM}$ using the current value as the starting point in the iteration, which propagates the effects of mergers to the other slices in M. In the event no mergers are made in any slices, we can choose to either wait for new data, which will update the Hebbian linkages, or we can terminate the algorithm, if we assume sufficient training data has already been collected.

Most clustering techniques work by iteratively refining a model subject to an optimization constraint. The iterative refinement in our algorithm occurs in the recalculation of $\tilde{D}_{CM}$, which is updated after each round of mergers within the slices. This spreads the effect of a merger within a slice by changing the Similarity distances between Hebbian projections onto it. This in turn changes the distances between regions in other slices and so forth, as discussed in section 3.7.1. The optimization constraint is that we do not create regions whose internal self-distances violate the above constraint, where a drastic decrease in self-distance would indicate the regions under consideration are viewed differently by other co-occurring modalities.

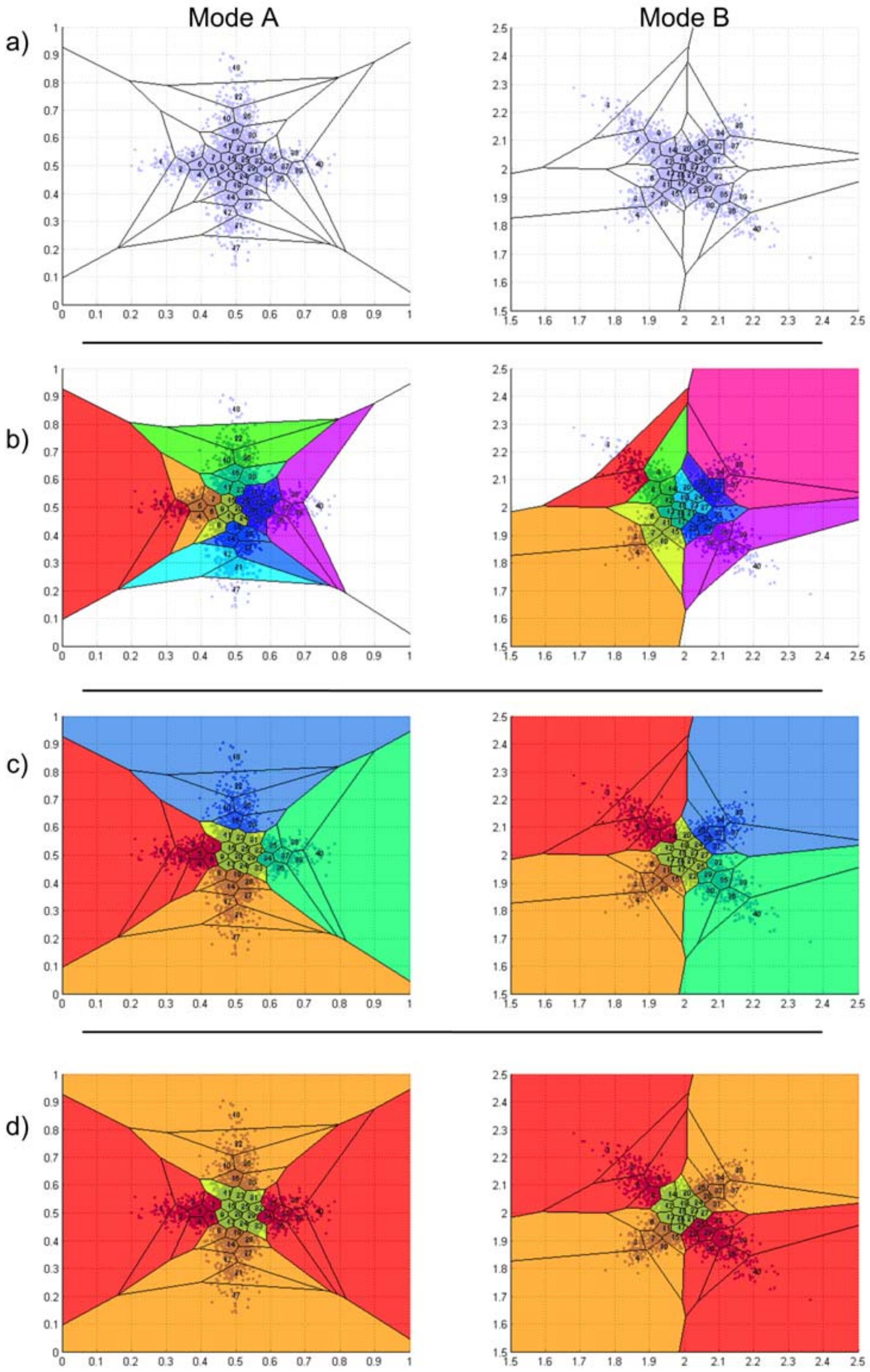
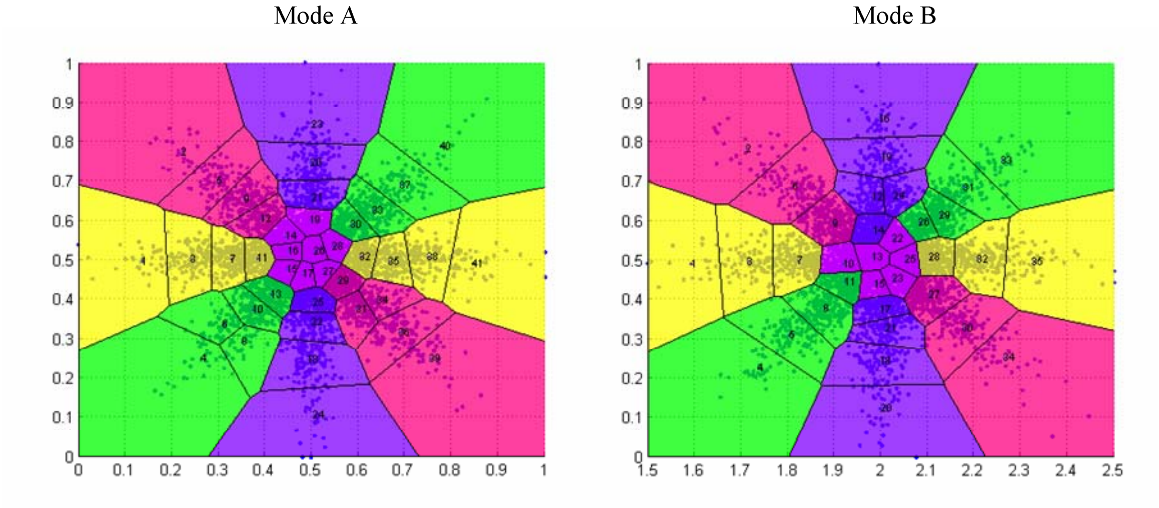
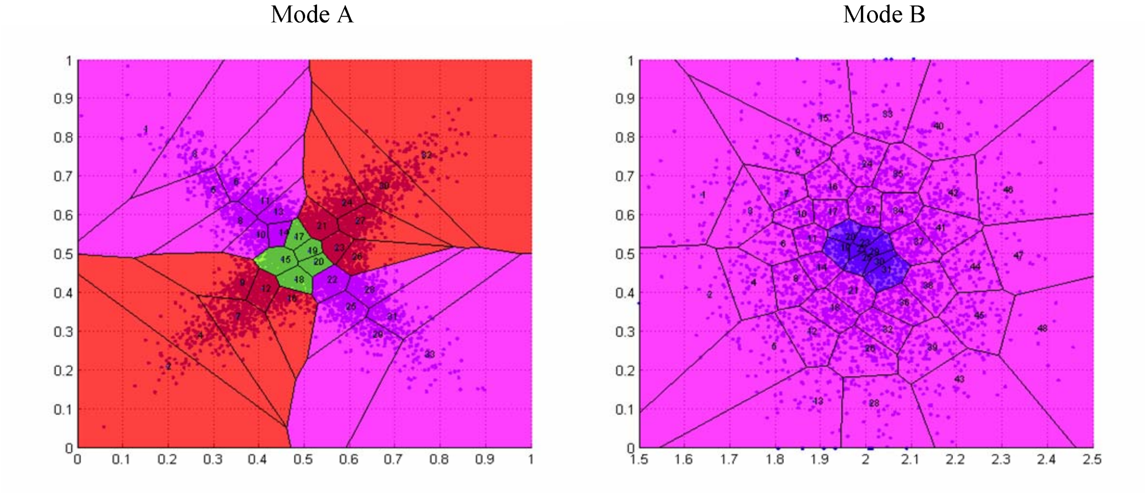


Figure 3.21 – The progression of the cross-modal clustering algorithm. (A) shows the initial codebook creation in each slice. (B) and (C) show intermediate region formation. (D) shows the correctly clustered outputs, with the confusion region between the categories indicated by the yellow region in the center. Note in this example, we set $\lambda = .7$ to make region formation easier to see by favoring spatial locality. The final clustering was obtained by setting $\lambda = 1$.


Mode A

Mode B

Figure 3.22 – The output of cross-modally clustering four overlapping Gaussian distributions in each slice. The confusion region between them is indicated in the center of the clusters.


Mode A

Mode B

Figure 3.23 – Finding one cluster embedded in another. In mode B, cross-modal clustering is able both to detect the small cluster embedded in the larger one and to use this separation of clusters to detect those in mode A. This is due to the non-Euclidean scale invariance of Similarity distance, which is used for determining the cross-modal distance between regions. Thus, region size is unimportant in this framework, and "small" regions are as effective in disambiguating other modes as are "large" regions.

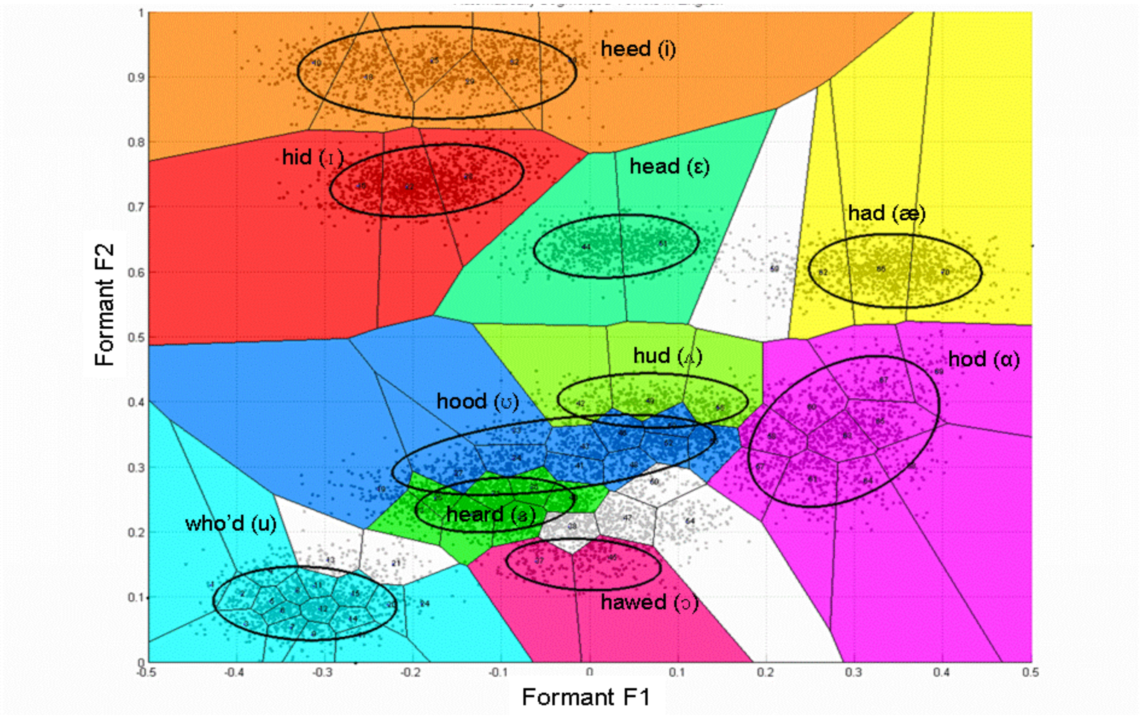


Figure 3.24 – Self-supervised acquisition of vowels (monophthongs) in American English. This work is the first unsupervised acquisition of human phonetic data of which we are aware. The identifying labels were manually added for reference and ellipses were fit onto the regions to aid visualization. All data have been normalized. Note the correspondence between this and the Peterson-Barney data show below.

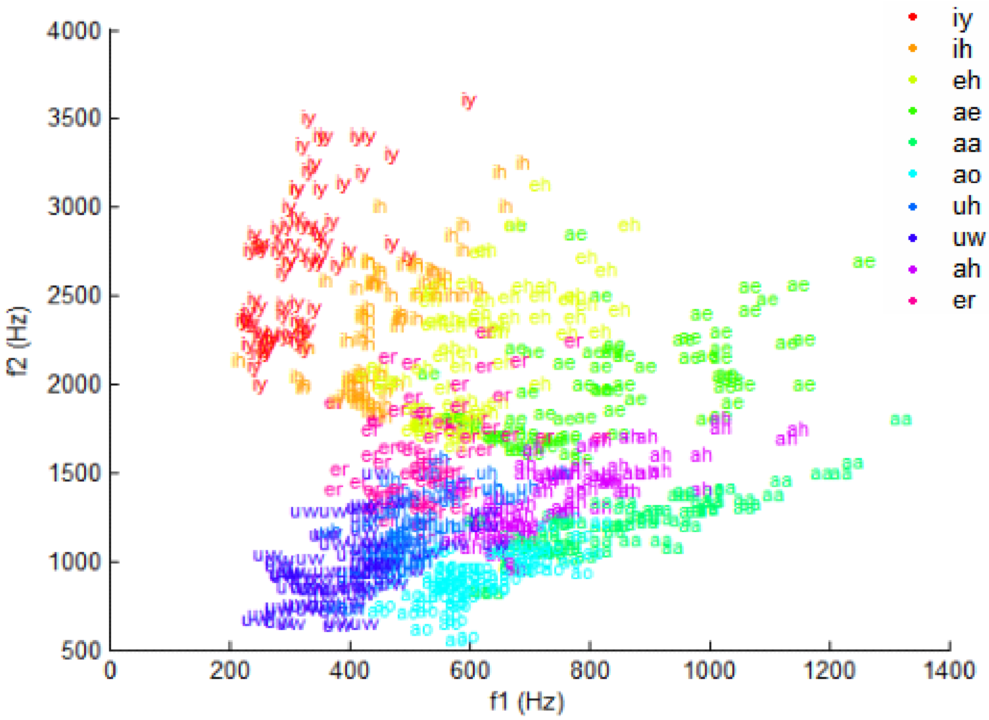


Figure 3.25—The Peterson-Barney dataset. Note the correspondence between this and Figure 3.24.

The progression of the algorithm starting with the initial codebook is shown in Figure 3.21. Setting $\lambda < 1$ includes Euclidean distances in the calculation of $\tilde{D}_{CM}$. This favors mergers between adjacent regions, which makes the algorithm easier to visualize. At the final step, setting $\lambda = 1$ and thereby ignoring Euclidean distance allows the remaining spatially disjoint regions to merge. Had we been uninterested in visualizing the intermediate clusterings, we would have set $\lambda = 1$ at the beginning. Doing so yields an identical result with this dataset, but the regions merge in a different, disjoint order. Note in general, however, it is not the case that different values of λ yield identical clusterings. All other examples in this thesis use $\lambda = 1$ exclusively.

Figure 3.22 demonstrates that the algorithm is able to resolve multiple overlapping clusters, in this case, two mixtures of four Gaussian distributions. Figure 3.23 show an important property of the Similarity distance, namely, it is scale invariant. The smaller cluster in Mode B is just as "distinct" as the larger one in which it is embedded. It is both detected and used to help cluster the regions in Mode A.

3.9 Clustering Phonetic Data

In Chapter 2, we asked the basic question of how categories are learned from unlabelled perceptual data. In this section, we provide an answer to this question using cross-modal clustering. We present a system that learns the number (and formant structure) of vowels (monophthongs) in American English, simply by watching and listening to someone speak and then cross-modally clustering the accumulated auditory and visual data. The system has no advance knowledge of these vowels and receives no information outside of its sensory channels. This work is the first unsupervised machine acquisition of phonetic structure of which we are aware.

For this experiment, data was gathered using the same pronunciation protocol employed by (Peterson and Barney 1952). Each vowel was spoken within the context of an English word beginning with [h] and ending with [d]; for example, /ae/ was pronounced in the context of "had." Each vowel was spoken by an adult female approximately 90-140 times. The speaker was videotaped and we note that during the recording session, a small number of extraneous comments were included and analyzed with the data. The auditory and video streams were then extracted and processed.

Formant analysis was done with the Praat system (Goedemans 2001, Boersma and Weenink 2005), using a 30ms FFT window and a 14th order LPC model. Lip contours were extracted using a system written by the author described in Chapter 2. Time-stamped formant and lip contour data were fed into *slices* in an implementation of the work in this thesis written by the author in Matlab and C. This implementation is able to visually animate many of the computational processes described here. This capability was used to generate most of the figures in this thesis, which represent actual system outputs.

Figure 3.24 shows the result of cross-modally clustering formant data with respect to lip contour data. Notice the close correspondence between the formant clusterings in Figures 22 and 23, which displays the Peterson-Barney dataset introduced earlier. We see the cross-modal clustering algorithm was able to derive the same clusters with the same spatial topology, without knowing either the number of clusters or their distributions.

The formant and lip slices are shown together in Figure 3.26, where the colors show region correspondences between the slices. This picture exactly captures what we mean by mutual bootstrapping. Initially, the slices "knew" nothing about the events they perceive. Cross-modal clustering lets them mutually structure their perceptual representations and thereby learn the event categories that generated their sensory inputs. The black lines in the figure connect neighboring regions within each slice and the red lines connect corresponding regions in different slices. They show a graph view of the clustering within each slice and illustrate how a higher-dimensional manifold may be constructed out of lower-dimensional slices. This proposes an alternative view of the structures created by cross-modal clustering, which we hope to explore in future work.

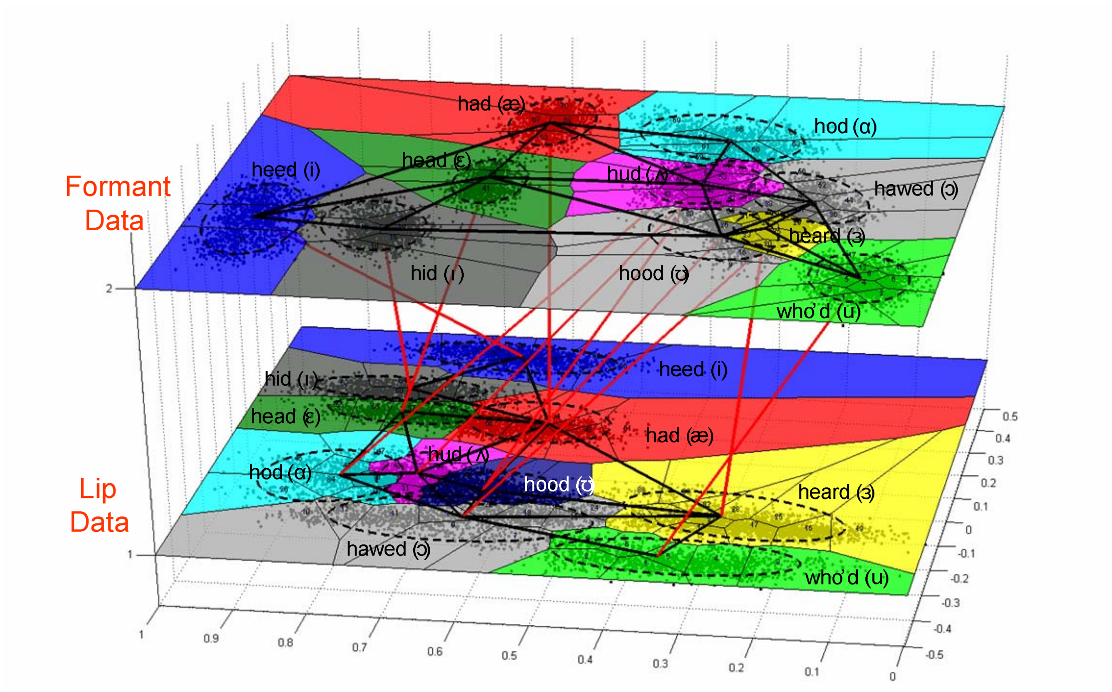


Figure 3.26 – Mutual bootstrapping through cross-modal clustering. This displays the formant and lip slices together, where the colors show the region correspondences that are obtained from cross-modal clustering. Initially, the slices "knew" nothing about the events they perceive. Cross-modal clustering lets them mutually structure their perceptual representations and thereby learn the event categories that generated their sensory inputs. The black lines in the figure connect neighboring regions within each slice and the red lines connect corresponding regions in different slices. The identifying labels were manually added for reference and ellipses were fit onto the regions to aid visualization. All data have been normalized.

3.10 Related Work

There is a vast literature on unsupervised clustering techniques. However, these generally make strong assumptions about the data being clustered, have no corresponding notion of correctness associated with their results, or employ arbitrary threshold values. The intersensory approach taken here is entirely non-parametric and makes no *a priori* assumptions about the underlying distributions or the number of clusters being represented. We examine several of the main avenues of related work below.

3.10.1 Language Acquisition

De Sa (1994) and de Sa and Ballard (1997) have taken a similar approach to the work presented here. Namely, they are interested in unsupervised learning of linguistic category boundaries based on cross-modal co-occurrences. However, their approach requires that the number of categories be known beforehand. For example, to separate the vowels in the English, they would need to know in advance that ten such vowels exist. Thus, their approach cannot solve the perceptual grounding problem presented in this chapter. However, their work is capable of refining boundaries once the categories have been learned, which would be a useful addition to our framework. Thus, one can easily imagine combining the two approaches to yield something more powerful than either of them alone.

De Marcken (1996) has studied the unsupervised acquisition of English from audio streams. However, he uses the phonetic model of Young and Woodland (1993) implemented in the HTK HMM toolkit to perform phonetic segmentation, which has already been trained to segregate phonemes. Thus, this aspect of his work is heavily supervised.

Other unsupervised linguistic learning systems (e.g., McCallum and Nigam 1998) are built around the Expectation Maximization algorithm (Dempster et al. 1977), which we discuss below. These approaches make very strong assumptions about the parametric nature of the data being clustering.

3.10.2 Machine Vision

The vast majority of research in the machine vision has employed supervised learning techniques. Some notable unsupervised approaches include Bartlett's (2001) work on using independent component analysis for face recognition, based upon assumed statistical dependencies among image features. Although there are no obvious metrics for correctness associated with the clustering itself, her selection of features has yielded performance comparable to that of humans in face recognition. One may therefore view her approach and ours as complementary. Her work would be an ideal basis for guiding the feature selection that generates inputs into slices in our approach.

Stauffer (2002) has studied using multiple sensors to learn object and event segmentation. The unsupervised learning component of his system assumes that the data are represented by mixtures of Gaussians, which is quite reasonable given the enormous amount of data collected by his sensor networks and their intended applications. In this, his framework makes strong parametric assumptions. It is also difficult to gauge the absolute correctness of his results in that it does not seem to have been applied to problems that have well-defined metrics. As this is not the goal of his framework, this comment should be viewed more as a differing characteristic than a criticism of his work.

3.10.3 Statistical Clustering

There is an enormous body of literature on statistical clustering techniques, which used assumed properties of the data being clustering to guide the segmentation process. For example, the Expectation Maximization (EM) algorithm (Dempster et al. 1977) is widely used a basis for clustering mixtures of distributions whose maximum likelihood estimation is easy to compute. This algorithm is therefore popular for clustering known finite numbers of Gaussian mixture models (e.g., Nabney 2002, Witten and Frank 2005). However, if the number of clusters is unknown, the algorithm tends to converge to a local minimum with the wrong number of clusters. Also, if the data deviate from a mixture of Gaussian (or some expected) distributions, the assignment of clusters degrades accordingly.

Wide ranges of clustering techniques (e.g., Cadez and Smyth 1999, Taskar et al. 2001) base their category assignment upon the EM algorithm. As such, they tend to require large amounts of data corresponding to presumed distributions, and they generally require the number of expected clusters be known in advance. While these assumptions may be reasonable for a wide range of applications, they violate the clustering requirements stated here.

3.10.4 Blind Signal Separation

Separating a set of independent signals from a mixture is a classic problem in clustering. For example, from a recording of an orchestra, we might like to isolate the sounds contributed by each individual instrument. This is an extraordinarily difficult problem in its general form. However, the assumption that the signals are mutually independent, which is not always realistic, supports a number of approaches. Principal components analysis (PCA) is a frequent tool used for separating signals, and while it is useful for reducing the dimensionality of a dataset, it does not necessarily provide a notion of correct clustering.

A number of interesting results have been found using PCA for data mining, for example, Honkela and Hyvärinen's (2004) work on linguistic data extraction, but our approach is founded on the notion that our cross-modal signals are mutually *dependent*, not independent. In that sense, our work makes a very different set of assumptions than most efforts in signal separation. Instead of looking for principal components in a high dimensional space, which for example might correspond to the vowels in the example in this chapter, we use a larger number of low dimensional spaces (i.e., slices) in parallel. Thus, we approach the problem from an entirely different angle, and as such, our framework appears to be more biologically realistic in that it is dimensionally compact.

3.10.5 Neuroscientific Models

The neuroscience community has proposed a number of approaches to unsupervised clustering (e.g., Becker and Hinton 1995, Rodriguez et al. 2004, Becker 2005). These are frequently based upon standard statistical approaches and often involve threshold values that are difficult to justify independently. Most problematically, they generally have no measure of absolute correctness. So, for example, the work of (Aleksandrovsky et al. 1996) learns multiple phonetic representations of audio signals without actually knowing how many unique phonemes are represented. Although one may argue that each phoneme is represented by multiple *phones* in practice, there is a clear consensus that it is meaningful to speak about the number of actual number of phonemes within in a language, which these techniques do not capture.

3.10.6 Minimally Supervised Methods

Blum and Mitchell (1988) have proposed using small amount of supervision to help bootstrap unsupervised learning systems. Although their approach is clearly supervised, one may argue that this supervision could come from phylogenetic development (Tinbergen 1951) and thereby provide support for innate models. We return to this discussion in Chapter 5, and note that it is entirely complementary to our approach. By pre-partitioning slices, we can easily incorporate varying amounts of supervision into our framework.

3.11 Summary

This chapter introduced *slices*, a neurologically inspired data structure for representing sensory information. Slices partition perceptual spaces into codebooks and then reassemble them to construct clusters corresponding to the actual sensory events being perceived. To enable this, we defined a new metric for comparing spatial probability distributions called *Similarity distance*; this allows us to measure distances within slices through cross-modal Hebbian projections onto other slices.

We then presented an algorithm for *cross-modal clustering*, which uses temporal correlations between slices to determine which hyperclusters within a slice correspond to the same sensory events. The cross-modal clustering algorithm does not presume that either the number of clusters in the data or their distributions is known beforehand and has no arbitrary thresholds.

We also examined the outputs and behavior of this algorithm on simulated datasets and on real data gathered in computational experiments. Finally, using cross-modal clustering, we learned the ten vowels in American English without supervision by watching and listening to someone speak. In this, we have shown that sensory systems can be perceptually grounded by bootstrapping off each other.

Chapter 4

Perceptual Interpretation

We saw in Chapter 3 that sensory systems can mutually structure one another by exploiting their temporal co-occurrences. We called this process of discovering shared sensory categories *perceptual grounding* and suggested that it is a fundamental component of cognitive development in animals; it answers the first question that any natural (or artificial) creature faces: *what different events in the world can I detect?*

The subject of this chapter follows naturally from this question. Once an animal (or a machine) has learned the set of events it can detect in the world, *how does it know what it is perceiving at any given moment?* We refer to this as *perceptual interpretation*. We will take the view that perceptual interpretation is inherently a dynamic – rather than static – process that occurs during some window of time. This approach relaxes the requirement that our perceptual categories be separable in the traditional machine learning sense; unclassifiable subspaces are not a problem if we can determine how to move out of them by relying on other modalities. We will argue that this approach is not only biologically plausible, it is also computationally efficient in that it allows us to use lower dimensional representations for modeling sensory and motor data.

4.1 Introduction

In this chapter, we introduce a new family of models called *influence networks*, which incorporate temporal dynamics into our framework. Influence networks connect cross-modally clustered slices and modify their sensory inputs to reflect the perceptual states within other slices. This *cross-modal influence* is designed to increase perceptual accuracy by fusing together information from co-occurring senses, which are all experiencing the same sensory events from their unique perspectives. This type of cross-modal perceptual reinforcement is commonplace in the animal world, as we discuss in Chapter 6.

Our approach will be to formulate a dynamic system in which slices are viewed as coupled perceptual state spaces. In these state spaces, sensory inputs move along trajectories determined both locally – due a slice's internal structure – and cross-modally – due to the influence of other co-occurring slices. This addition of temporal dynamics will allow us to define our notion of *perceptual interpretation*, which is the goal of this chapter.

The model presented here is inspired by sensory dynamics in animals, but it does not approach the richness or complexity inherent in biological perception. Our intent, however, is to work toward creating better artificial perceptual systems by providing them with a more realistic sensorial framework.

4.2 The Simplest Complex Example

We will proceed by considering an example. We turn to the hypothetical perceptual modes introduced in Chapter 3, as shown in Figure 4.1. Recall that each mode here is capable of sensing the same two events in the world, which we have called the *red* and *blue* events. The cross-modal clustering of these modes in shown in Figure 4.2. We see that the larger perceptual regions have been assembled out of the codebook clusters, shaded red or blue to indicate their corresponding sensory category. Our interest here, however, is not in these large perceptual regions but rather in the small yellow regions at their intersections. Sensory inputs within the yellow regions are ambiguous – these are the inputs that we cannot classify, at least not without further information.

Although this example was selected for its simplicity, it worthwhile pointing out some of the complexity it presents, and thereby examine some of the assumptions in our framework before we proceed. Notice that because the mixtures of Gaussians here intersect near their means, the "small" yellow regions will receive almost as many sensory inputs as either of the "large" blue and red ones. Estimating from the limits of the density normalization performed during codebook generation, we expect at least 1/4 of the inputs in this example to be unclassifiable because they will fall into these yellow confusion zones. If our goal is to categorize sensory inputs, these yellow regions will prove quite troublesome; we need to find some way of avoiding them.

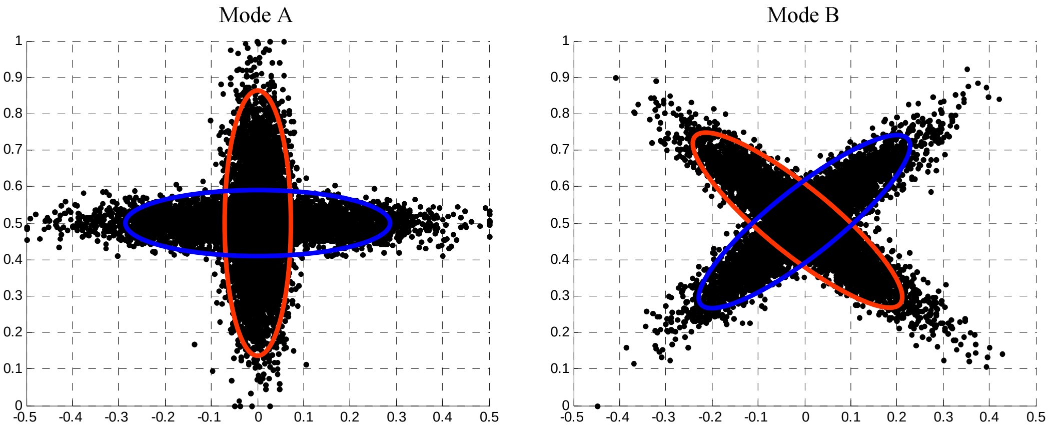


Figure 4.1 -- Two hypothetical co-occurring perceptual modes. Each mode, unbeknownst to itself, receives inputs generated by a simple, overlapping Gaussian mixture model. For example, if a "red" event takes place in the world, each mode would receive sensory input that probabilistically falls within its red ellipse. To make matters more concrete, we might imagine Mode A is a simple auditory system that hears two different events in the world and Mode B is a simple visual system sees those same two events, which are indicated by the red and blue ellipses.

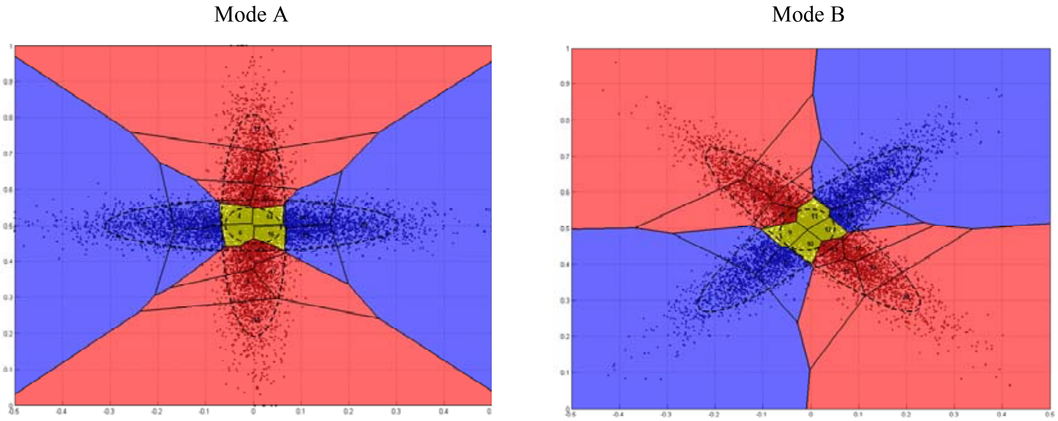


Figure 4.2 – The output of cross-modally clustering the modes in Figure 4.1. The perceptual regions have been constructed out of the codebook clusters and are indicated respectively by the blue and red shading. The confusion region between them is indicated at their intersection in yellow. To assist with visualization, ellipses have been fit using least squares onto the data points within each sensory region, as determined by cross-modal clustering. Note that approximately ¼ of the inputs within each mode fall into the confusion region and are therefore ambiguous.

In fact, as things stand at the moment, we expect the overall classification error rate would be somewhat higher than this. Sensory events are only probabilistically described by the derived perceptual regions. Thus, inputs will not always fall into the "right" area, at least according to the output of cross-modal clustering. We also assume that the feature extraction (e.g., formant estimation) generating the sensory inputs to slices introduces some degree of error; this may be due to noise, heuristic estimation, instability, incorrect or incomplete modeling, limitations from perceptual thresholds, etc. We elaborate on this point in Chapter 6, but it is reasonable to expect these types of processing errors occur in both biological and artificial systems. Although, not a source of error, an additional complexity is that we generally expect many more than two modes will be active simultaneously, corresponding to the fine-grained model of perception we adopted in Chapter 1.

4.2.1 Visualizing an Influence Network

Our goal is to "move" sensory inputs within slices to make them easier to classify. In doing so, we seek to avoid perceptual ambiguity when possible and to recover from errors introduced during perceptual processing.

We have so far considered slices through their codebooks or through the entire perceptual (e.g., red and blue) regions found within them by cross-modal clustering. We now instead look at the local neighboring (connected) components within each of these larger regions; we are going to call these local regions *nodes* (Figure 4.3); to be clear, the *nodes* partition the *modes* (i.e., slices) into locally connected regions. Nodes correspond to the representational areas that are easy to classify, and therefore, they will help us disambiguate perceptual inputs. We note that the lines between nodes represent *perceptual equivalence* as determined by cross-modal clustering. Although they are derived from Hebbian data, the lines do not represent the Hebbian linkages described earlier because they are restricted to regions corresponding to the same perceptual categories.

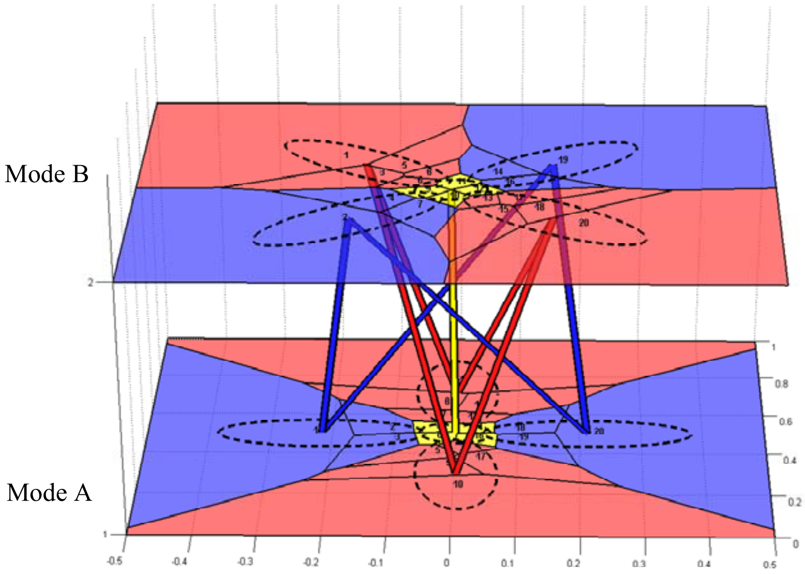


Figure 4.3 – Viewing the nodes within the modes. Nodes are the locally connected components in each perceptual region; they are indicated here by color and with ellipses, fit on their inputs after cross-modal clustering. In this example, each mode has 5 nodes: 2 blue, 2 red, and 1 yellow. The colored lines indicate perceptual equivalence between nodes in different slices.

We would like to use cross-modal correspondences, as indicated by these lines of perceptual equivalence, to define a framework where slices can mutually disambiguate one another. For example, suppose Mode B on top sees an input in a blue node. One could certainly imagine that knowing this might help resolve a simultaneous ambiguous input in Mode A. In a static framework, we could for example implement a disambiguation strategy using posterior probabilities (e.g., Wu et al. 1999). This would be relatively straightforward, at least for a small number of co-occurring modes.

The problem with this solution, however, is that perception in animals has complex temporal dynamics – *generating percepts does not correspond to an instantaneous decision process*. This is evidenced during cross-modal influence (Calvert et al. 2004), perceptual warping (Beale and Keil 1995), interpretative bistability (Blake et al. 2003), habituation (Grunfeld et al. 2000), and priming (Wiggs and Martin 1998), and these effects are particularly prominent during development (Thelen and Smith 1994). We will examine these phenomena in detail in Chapter 6. For the moment, we further note that even unimodal percept generation has complex sensory and temporal threshold dynamics (Nakayama et al. 1986); in fact, different features may have different thresholds, which themselves change dynamically (e.g.. Hamill et al. 1989, Wang et al. 2002).

Most importantly, these phenomena are not simply biological "implementation" details; they are not epiphenomena. Rather, they are fundamental components of perceptual activity and are a large part of why animal perception is so robust. Because the perceptual phenomena we are interested in modeling correspond to temporal processes, we will argue that our models should be similarly dynamic. In particular, static techniques seem poor approaches for understanding the intersensory and temporal complexities of biological interpretative mechanisms.

4.2.2 Towards a Dynamic Model

Instead, our approach will be to view slices as perceptual state spaces, where the nodes correspond to fixed points. Ambiguous nodes, like the yellow ones in our example, will be treated as repellers in this space, and unambiguous nodes, such as the red and blue ones, will be treated as attractors. Perceptual inputs will travel along trajectories defined by these fixed points.

We visualize this state space view of slices on page 101. The repellers (ambiguous nodes) correspond to maxima; the attractors (unambiguous nodes) are the minima and surrounded by their basins of attraction. In this framework, perceptual interpretation will loosely correspond to energy minimization; that is, we would like move sensory inputs into the basins that represent their correct classifications. In this way, the dynamics of the system will perform the sensory classification for us.

Intersensory influence – the ability of one slice to modify perceptions in another – will be realized by having slices induce vector fields upon one another, thereby cross-modally modifying their temporal dynamics. This in turn can modify perceptual classifications by leading sensory inputs into different basins of attraction (or perhaps by even introducing bistability).

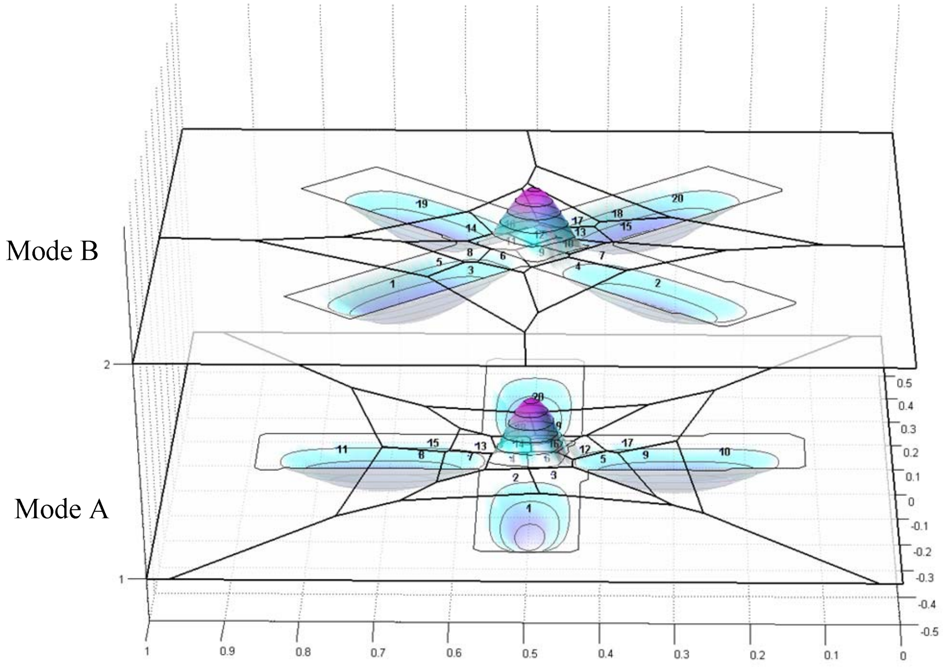


Figure 4.4 -- Visualizing slices as state spaces where the nodes are the fixed points. Repellers, corresponding to ambiguous nodes, are maxima and shown in the center; attractors, corresponding to unambiguous nodes, are minima and are outlined by their basins of attraction. It may be helpful to compare this view to the one in Figure 4.3.

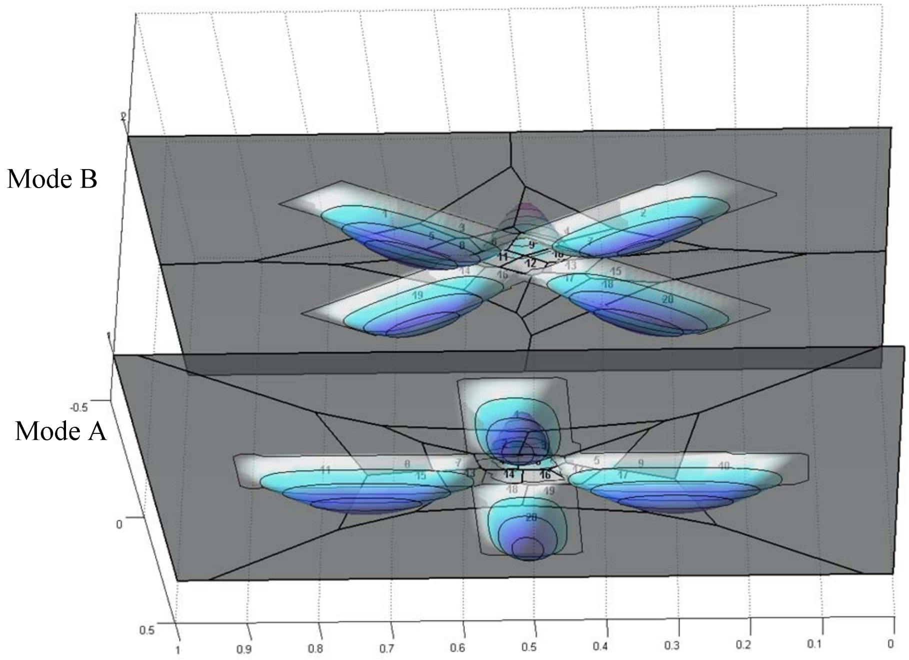


Figure 4.5 -- A view of the basins of attraction from below. This is a rotation of Figure 4.4 into the page to help visualize the basins, which are partially occluded above. The basins here correspond to the blue and red nodes in Figure 4.3.

To formulate a dynamic model, the first step is to incorporate some model of time into our framework. We start by assuming that any perceivable event in the world persists for some (perhaps variable) temporal duration, e.g., 10 – 1000 milliseconds. During this interval, the event will be perceptually sampled, i.e., some finite number of sensory inputs will be generated describing it. This assumption is quite reasonable from a biological perspective. For example, this could correspond to a series of neuronal firings in the striate cortex, during the period of time an object is visually observed (Hubel and Wiesel 1962). Alternatively, in an artificial system, a stream of sensory inputs might be generated by sliding a fast Fourier transform window over an auditory signal. Regardless of how the features are selected and extracted, the important consideration here is that an event in the world generates a stream of perceptual inputs, rather than a unitary data point. We will use the time this provides to effect intersensory influence.

We also need to define the state within a slice, namely, the quantity that will be changing during the lifetime of an individual perception. In our model, each slice will receive a stream of sensory inputs when it is stimulated, as described above. After receiving some number of inputs, a slice may eventually "decide" that a recognizable perceptual event has occurred. In the interim, however, a slice maintains an *estimate* of what it might be perceiving. The estimate corresponds to a point in $\mathbb{R}^N$ that moves through the slice's state space. At any given moment, three factors influence the trajectory of this point through state space.

1) The current perceptual input. An estimate tends to move towards the current input.

2) A gradient defined by the fixed points within a slice. An estimate moves towards the attractors, which are the unambiguous nodes in a slice. The gradients for the two slices we have been examining are displayed in Figure 4.6 .

3) Any induced vector fields from co-occurring slices. This is visualized in Figure 4.7 for the ambiguous scenario described above. We discuss this in more detail below.

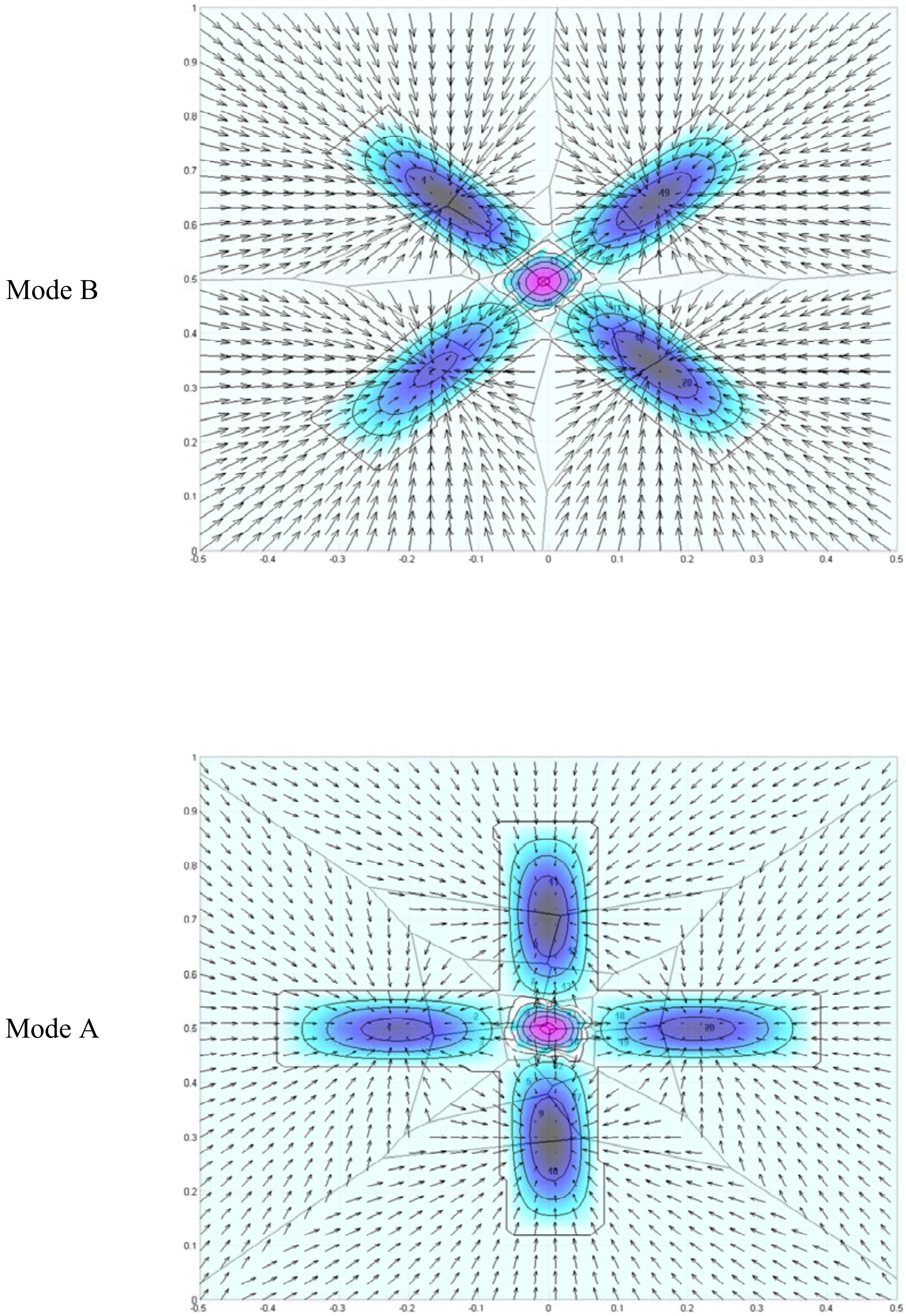


Figure 4.6 – Viewing innate state space dynamics. The nodes define a gradient in the state space. In the absence of other influences, points move towards attractors (unambiguous nodes) and away from repellers (ambiguous nodes). The generation of these gradients is discussed in § 4.3.1.1.

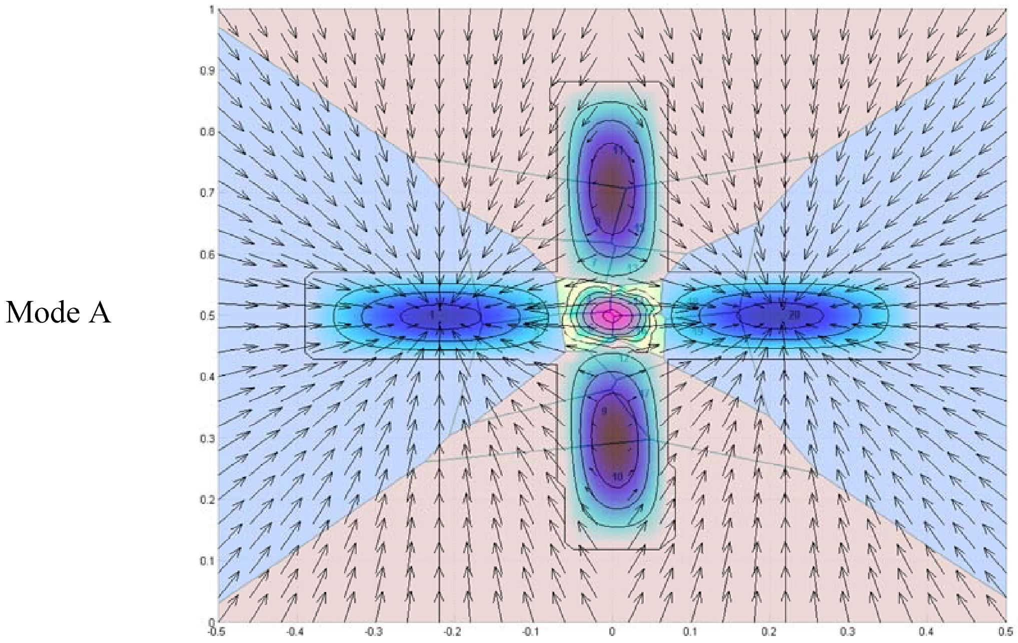


Figure 4.7 – Visualizing cross-modal influence through a vector field induced on Mode A by Mode B. This field modifies the dynamic behavior in Mode A to favor Mode B's perceptual interpretation. This "pulls" the estimate in Mode A towards the blue attractors, corresponding in this case to Mode B's assessment of the world. We have added light red and blue shading in the background to recall the perceptual categories contained in this slice.

4.2.3 Intersensory Influence

Biological perceptual systems share cross-modal information routinely and opportunistically (Stein and Meredith 1993, Lewkowicz and Lickliter 1994, Rock 1997, Shimojo and Shams 2001, Calvert et al. 2004, Spence and Driver 2004); we have suggested that *intersensory influence* is an essential component of perception but one that most artificial perceptual systems lack in any meaningful way. In our framework, cross-modal influence is realized by having modes bias the temporal dynamics of other modes, thereby sharing their views of the world.

Recall the ambiguous scenario described on page 99. We imagined that in Figure 4.3, Mode B saw an input corresponding to blue node. We asked how this might help disambiguate a simultaneous event in Mode A. Our solution to this problem is illustrated in Figure 4.7. We allow Mode B to induce a vector field on Mode A, thereby modifying its dynamic behavior. This vector field incorporates Mode B's perceptual perspective into Mode A. This "pulls" the estimate in Mode A towards the blue attractors, corresponding to Mode B's assessment of what is being perceived. The "interlingua" that makes this influence possible is their shared perceptual categories, which are obtained through cross-modal clustering and provide a common frame of sensorial reference.

Therefore, we see that the movement of the estimate within a slice is governed: (1) externally, by events happening in the world; (2) innately, by the previously learned structure of events with the slice; and (3) cross-modally, from the perspectives of other perceptual channels viewing the same external events. We can in fact adjust the influences of these individual contributions dynamically. For example, in noisy conditions, an auditory slice may prefer to discount perceptual inputs and rely more heavily on cross-modal inputs. These kinds of tradeoffs are common in biological perception (Cherry 1953, Sumby and Pollack 1954). We discuss this further below.

4.3 Influence Networks

In our model of perception, an event in the world corresponds to a path taken through a slice. In this section, we define both the rules that govern this path and how a slice decides something has been perceived, a concept we have not encountered so far in this chapter.

By formulating this problem dynamically, we lose one of the primary attractions of a static framework – that we know with certainty something has "actually" occurred. In static frameworks, this decision is usually made independently by some precursor to the interpretative mechanism; in other words, the decision must be made but it is done somewhere else. This usually corresponds to fixed sets of determination criteria within the independent perceptual components.

In our framework, how far down its path must an input travel for a slice to know that an event is actually occurring? Presumably, a single sampled input is insufficient. Generally, both natural and artificial perceptual systems have thresholds – they require some minimum amount of stimulation to register an event in the world (Hughes 1946, Fitzpatrick and McCloskey 1994). Rather than assume this happens externally, a slice must make this determination itself, but it is not necessarily doing so alone. This decision itself is now a dynamic process open to intersensory influence – consistent interpretations among slices will lower their sensory and temporal thresholds, whereas inconsistent interpretations will raise them. Cross-modal perception is known to significantly improve upon unimodal response times in humans (Hershenson 1962, Frens 1995, Calvert et al. 2000), a phenomenon that is captured by our model.

We now proceed by defining the state-space view of slices that has been motivated informally so far. State within each slice has two distinct components:

1) The position of its estimate. This is what the slice "thinks" it might be perceiving.

2) The *activation potential* of its nodes. Nodes have internal potentials, which are increased whenever the perceptual estimate falls within them.

Below, we define the state space equations that govern how these quantities change during the lifetime of a perception. Towards this, we begin with some preliminary definitions covering the concepts raised earlier in this chapter.

Note that we will limit some our of presentation to 2-dimensional slices, because $\mathbb{R}^2$ is easy to describe and a reasonable model for cortical representations. $\mathbb{R}^3$ is the largest space in which we have constructed slices. Higher dimensional spaces may generalize from these aspects of the presentation but that is an unexamined hypothesis. Also, because cross-modal clustering connects regions in different slices, it lets us approximate higher-dimensional manifolds. This may reduce the need for directly implementing higher dimensional perceptual representations.

We note that the dynamic framework below is implemented in the examples in this thesis by a discrete simulation that uses a 10 millisecond time step.

4.3.1 Preliminary Concepts

Consider a slice $M_A \subseteq \mathbb{R}^n$ with associated codebook $C_A = \left\{ p_1, p_2, ..., p_a \right\}$. Suppose M_A is cross-modally clustered with respect to slice M_B into *k* perceptual categories. We call the set of perceptual categories $C = \{c_1, ..., c_k\}$. To represent the assignment of codebooks to categories, we define a *multi-partitioning MP* of M_A as a mapping $MP : C_A \to 2^{\{1,2,...,k\}}$. This assigns each codebook cluster to a subset of integers from 1 to *k*, representing the *k* different perceptual categories discovered during cross-modal clustering. If a codebook cluster is in more than one category, then it is deemed ambiguous, as shown in (4.1).

We define a *node* $u \subseteq C_A$ as a locally connected set of codebook clusters that are members of the same perceptual categories; for any two clusters p_b and p_c in *u*, $MP(p_b) = MP(p_c)$. We say $u \in c_i$ if node *u* is in category c_i. We define the *nodebook* N_A as the set of all nodes in slice M_A. Nodebooks are used to represent the connected components in the output of cross-modal clustering. The *density* of node *u* is the percentage of sensory inputs falling within it over some assumed time window.

For node u_i, we define

$$\text{sign}(u_i) = \begin{cases} 1, & \text{if } \left|MP(u_i)\right| = 1 \quad \text{(attractor)} \\ \text{-}1, & \text{if } \left|MP(u_i)\right| > 1 \quad \text{(repeller)} \end{cases} \tag{4.1}$$

which indicates whether u_i will be an attractor or a repeller in the state space, determined by whether it is or is not ambiguous. Let $S_A = [\text{sign}(u_1),\ \text{sign}(\text{u}_2), ..., \text{sign}(u_a)]$ be a vector containing all signs within a slice, which captures the ambiguity within it.

4.3.1.1 Defining a Surface

Towards defining a surface on slice $M_A \subseteq \mathbb{R}^2$, we first define a reference surface from which it will be piecewise constructed. Let $Z(x,y) = \sin(\pi x) \cdot \sin(\pi y)$, corresponding to the shape of our fixed points.

We will fit copies of Z onto M_A's nodebook, towards building a surface over the entire slice. For each node $u_i \in N_A$, let $\langle a,b,x_0,y_0,\theta \rangle_i$ be an ellipse fit onto the node's sensory inputs via (Fitzgibbon et al. 1999). We then define a surface on node u_i, $Z_i = \text{sign}(u_i) \cdot \text{density}(u_i) \cdot Z\left(\langle a,b,x_0,y_0,\theta \rangle_i\right)$, where we stretch, translate, and rotate the reference shape Z onto the node's descriptive ellipse. The term $\text{sign}(u_i)$ determines whether a node is a local maxima or minima.

We define the *innate surface* Z_A over the entire slice by piecewise summing the contributions of the individual nodes, $Z_A = \sum Z_i$. Although this surface is constructed piecewise, we know it is smooth due to the reference surface Z being sine-based, and therefore, it is continuously differentiable within the boundaries of the slice. This surface is illustrated in Figure 4.4 and Figure 4.5, where the repellers correspond to local maxima and the attractors correspond to local minima.

The gradient of this surface

$$\nabla Z_A = \frac{\partial Z_A}{\partial x} + \frac{\partial Z_A}{\partial y} \tag{4.2}$$

captures the innate dynamic behavior of the slice due to cross-modal clustering. Two examples of this are illustrated in Figure 4.6. We can view the effect of this gradient as the slice trying to classify inputs to match its learned events. In other words, perceptual grounding biases slices in favor of particular interpretations, namely, the ones derived from cross-modal clustering. This may correspond with the "perceptual magnet effect," which reduces detectable differences between sensory inputs near previously acquired categories (Kuhl 1991).

An inefficiency introduced by representing data in Euclidean Voronoi regions is that sometimes space is "wasted," particularly around the boundaries. We can see this in Figure 4.2, where the data points are fairly sparse. We note this is a synthetic data set with only two events, so the sparsity is exaggerated here. Nonetheless, we would prefer a "warped" representation, where unused areas of the map are reduced in favor of expanding or stretching the areas that are more perceptually prominent. It is clear that cortical maps are continuously modified in animals, where regions are allocated proportionally with their use (Buonomano and Merzenich 1998, Kaas 2000). For example, it has been found that the fingers of the left hand in right-handed violinists have increased cortical representation (Elbert et al 1995), corresponding to their development of fine motor control during fingering.

Because our implementation does not support spatial plasticity, there may be sections of Voronoi regions which are generally devoid of inputs. However, over time, we expect that inputs will fall within them, albeit somewhat infrequently. To account for these stray inputs, we induce a weak, constant magnitude vector field over the surface of each slice, specifically for the benefit of these "empty" regions. The field within each node u_i points towards or away from its fixed point, depending upon $\text{sign}(u_i)$. This reflects each node's innate perceptual bias throughout its entire Voronoi region.

4.3.1.2 Hebbian Gradients

Towards defining cross-modal influence, we parameterize the approach above. Consider a slice $M_A \subseteq \mathbb{R}^n$ with associated codebook $N_A = \{u_1, u_2, ..., u_a\}$. Let $S = \{s_1, s_2, ..., s_a\}$, where each $s_i \in [-1, 1]$. We call *S* a *sign set* and use it to provide signs for each node in N_A, i.e., whether it attracts or repels. Although signs were previously restricted to be either -1 or 1, a continuous range here reflects confidence in the assignments.

We define:

$$Z_i(S) \quad = S(u_i) \cdot \text{density}(u_i) \cdot Z\left(\left\langle a, b, x_0, y_0, \theta \right\rangle_i\right) \tag{4.3}$$

$$Z_A(S) \quad = \sum Z_i(S) \tag{4.4}$$

$$\nabla H(S) \quad = \frac{\partial Z_A(S)}{\partial x} + \frac{\partial Z_A(S)}{\partial y} \tag{4.5}$$

We call $\nabla H(S)$ the *Hebbian gradient* induced on M_A by sign set *S*. Slices will induce vector fields derived from Hebbian gradients on one another to create cross-modal influence. The influence is effected by assigning values in sign set *S* that correspond to their perceptual outlooks. This is possible because cross-modal clustering provides slices with a common frame of reference for making these assignments. Figure 4.7 contains an example of such a field.

4.3.1.3 Activation Sets

Finally, we define the notion of an *activation set* $\{e_1, e_2, ..., e_j\}$, where $0 \le e_i \le 1$. Activation sets will be used to represent *activation potentials* in our state space models, when we define leaky integrate and fire networks. The activation potential for each node corresponds to a temporal-integration of its inputs over some time period, as detailed in section 4.3.3. In anticipation of applying Hebbian gradients in the next section, we show here how to project the activation potentials in one mode onto another, by using their shared categories as a common frame of reference.

For mode M_A, let $N_A = \{v_1, v_2, ..., v_a\}$ be its nodebook and let N_A contain *k* distinct categories, as determined by cross-modal clustering with some mode M_B. We call this set of perceptual categories $C = \{c_1, ..., c_k\}$.

Let $E_A = \{e_1, e_2, ..., e_a\}$ be an activation set on M_A. We define the activation set on *C* derived from E_A as $\Phi = \{\phi_1, \phi_2, ..., \phi_k\}$, where ϕ_i is simply the sum of the activation potentials of the individual nodes within category *i* :

$$\phi_i = \sum_{v \in c_i} e_v \tag{4.6}$$

We will refer to the mapping in (4.6) as $\overline{\Phi}(E_A)$, which projects nodebook activations onto category activations.

Let $\Phi = \{\phi_1, \phi_2, ..., \phi_k\}$ be an activation set on the categories in *C*. We define an activation set $E_A = \{e_1, e_2, ..., e_a\}$ on a node M_A derived from Φ by reversing the above process:

$$e_i = \sum_{c_j \in MP(a)} \phi_j \tag{4.7}$$

We will refer to the mapping in (4.7) as $\overline{E}_A(\Phi)$, which projects category activations onto nodebook activations.

Being able to move back and forth between activations in nodebooks and categories is quite useful. It allows us to share state between two different slices, using their categories as an interlingua, and thereby define the Hebbian gradients used in intersensory influence.

4.3.2 Perceptual Trajectories

During the time window corresponding to a sensory event, a slice $M_A \subseteq \mathbb{R}^N$ integrates its sensory inputs into a single estimate $h \in R^N$ that models what it is in the midst of perceiving. This estimate changes over time, and its movement is governed: (1) externally, by events happening in the world; (2) innately, by the previously learned structure of events with the slice; and (3) cross-modally, from the perspectives of other perceptual channels viewing the same external events.

We now define the state space dynamics due to each of these components.

(1) Events in the world reach the slice through a stream of sensory inputs. (We discuss integration of slices into perceptual pipelines later in this chapter.) We call this input stream $I = <d_1, t_1>, <d_2, t_2>, ..., <d_k, t_k>$, where input $d_i \in R^N$ arrives at time t_i.

Let $\Delta I_i = (d_i - d_{i-1})$, which is a vector measuring the difference between successive inputs. Let $\delta_W^A(t_i)$ be a unit impulse function that is nonzero if an input arrives at M_A at time t_i.

We define the change in the estimate due to events in the world as:

$$W(t) = \delta_W^A(t_i) \cdot \Delta I_i \tag{4.8}$$

$W(t)$ provides the gradient of the input change at time *t*, assuming an input occurred. Otherwise, it is zero. The "W" in this function reminds us it due to changes in the world. In the absence of other factors, $\int W(t)\, dt$ tracks the inputs exactly.

(2) The learned category structure within each slice also influences its perceptual inputs. In equation (4.4), we defined a parameterized surface $Z_A(S)$ constructed over slice M_A. The sign values in set S determine the heights of the fixed points corresponding to the nodes; these range from -1 to 1, corresponding to repellers or attractors respectively. Values within this range reduce the strength of the corresponding gradient.

Recall vector $S_A = \{\text{sign}(u_1),\ \text{sign}(\text{u}_2),...,\text{sign}(u_a)\}$, which captures the ambiguity of M_A and is used to define the innate surface Z_A.

Let $A_t(M_A)$ be a vector of activation potentials for M_A at time *t*, $A_t(M_A) = [e_1, e_2, ..., e_a]$, where $0 \le e_i \le 1$. (This was defined above in § 4.3.1.3 and receives a more detailed treatment in § 4.3.4.)

We are going to scale S_A by $A_t(M_A)$ to generate a gradient corresponding to the innate dynamics of slice M_A. Let $S_A(t) = \{e_1 \cdot \text{sign}(u_1), e_2 \cdot \text{sign}(u_2), ..., e_a \cdot \text{sign}(u_a)\}$.

We define:

$$I(t) = \left(\nabla Z_A\left(S_A(t)\right)\right)(h) \tag{4.9}$$

This is the component of the estimate's change that is due to a slice's preference to favor previously learned categories. This influence increases as the activation potentials in the nodes increase; the intuition here is that the nodes become increasingly "sticky" as they grow closer to generating a perception. We call this refer to this influence as the slice's innate dynamics, and the "I" in this function is a mnemonic for "innate."

(3) To define the component of state space dynamics due to intersensory influence, let M_B be a slice that has been cross-modally clustered with M_A.

We are going to use M_B to construct a Hebbian gradient $\nabla H(S)$ on M_A. To do this, we need to determine a sign set *S* that reflects M_B's perceptual preferences. As we touched upon at the beginning of this section, we will be defining an activation potential model on the nodebooks in each slice. We are going to use this model to determine the sign set for calculating $\nabla H(S)$.

Let $A_t(M_B)$ be the set of activation potentials for M_B. Defining the sign set *S* is a two step process. First, we use $A_t(M_B)$ to determine the potentials Φ of the mutual categories *C* between M_A and M_B. Then, we use Φ to determine signs for the nodes in M_A.

The simplest definition of $S = \{e_1, e_2, ..., e_a\}$ is:

$$S = \overline{E}_A\left(\overline{\Phi}\left(A_t(M_B)\right)\right) \tag{4.10}$$

This projects the activation potentials in M_B directly onto M_A, by using *C* as a common reference. Because $e_i \geq 0$, this means M_B can only induce a vector field corresponding to attractors on M_A, per equation (4.1). Thus, it can encourage the recognition of certain categories but it cannot discourage the recognition of others.

We can imagine a less permissive strategy, where some nodes are encouraged and some discouraged, as in:

$$S = \bar{E}_A\left(2\tan^{-1}\left(\alpha\cdot\left(\bar{\Phi}(A_t(M_B)) - \text{mean}(\bar{\Phi}(A_t(M_B)))\right)\right)/\pi\right),\ \alpha \ge 2 \qquad (4.11)$$

Here, nodes corresponding to categories with above mean potentials are treated as attractors and those with below mean potentials are treated as repellers.

Finally, let us consider the scenario where M_B determines what it is perceiving before M_A does. In the model defined below, this will correspond to $\max(\Phi) \ge 1-\varepsilon$. Namely, the activation potential of some category reaches a threshold value, at least from M_B's perspective, which "resets" the potentials in all other categories. Perhaps in this case we would like to discourage all of the non-recognized categories in M_A. We can do that by substituting:

$$S = \bar{E}_A\left(2\tan^{-1}\left(\alpha\cdot\bar{\Phi}\left(A_t(M_B)\right)\right)/\pi\right),\ \text{for some large } \alpha \qquad (4.12)$$

In practice, we use (4.11) and dynamically substitute in (4.12) when a co-occurring node fires.

We can now define the change in the estimate due to intersensory influence as:

$$C(t) = \left(\nabla H(S)\right)(h) \qquad (4.13)$$

The "C" here is a reminder this represents cross-modal influence.

4.3.3 Perceptual Dynamics

We can now define the state space equation for estimate *h*:

$$\begin{aligned}\frac{dh}{dt} &= \alpha\cdot W(t) + \beta\cdot I(t) + \gamma\cdot C(t)\\ \frac{dh}{dt} &= \alpha\cdot\delta_W^A(t_i)\cdot\Delta I_i(h) + \beta\cdot\left(\nabla Z_A\left(S_A(t)\right)\right)(h) + \gamma\cdot\left(\nabla H(S)\right)(h)\end{aligned} \qquad (4.14)$$

which combines the dynamic effects from the world (W), its innate structure (I), and cross-modal influence (C), where *S* is defined as above. We define $h = d_1$ at time t_1.

How do we set the constants α, β, and γ? By default, we set $\alpha = 1$ so that each slice "receives" its full inputs stream. We also set $\beta + \gamma = 1$ so that by the time each slice is ready to generate its input, its innate structure and cross-modal influences are as important as its inputs in determining which category is recognized. In particular, we set $\gamma > \beta$, so that intersensory influences dominate innate ones. The reason for this is that we suppose *the world is already enforcing the innate structure of events in generating them*. From the outlooks of both Gibson (1950) and Brooks (1987), dynamically reinforcing the world's structure within a slice is redundant. Nonetheless, we view it is as an important component of interpretative stability. While it may not be necessary from a theoretical Gibsonian perspective, incorporating innate perceptual dynamics seems to have clear computational advantages in real world conditions. It stabilizes sensory inputs by biasing perceptual interpretation to favor events the slice has previously learned.

Ideally, it would be optimal to adjust these parameters dynamically. For example, in noisy conditions, an auditory slice may prefer to discount its perceptual inputs in favor of visual cross-modal one, realizing what is known as the *cocktail party effect* (Cherry 1953, Sumby and Pollack 1954); raising γ and lowering α in (4.14) would have this effect. This condition is likely detectable, particularly where we suppose one slice is not registering events and another co-occurring one is. Although we have not implemented an automatic mechanism for dynamically balancing these parameters, doing so does not seem implausible (e.g., according to maximum-likelihood estimation approach of Ernst and Banks, 2002).

4.3.4 An Activation Potential Model

Let us examine where we stand at this point. We have defined how to calculate the path of a slice's perceptual estimate within its state space. However, we have not yet specified how this path dynamic generates a percept. We now return to the question posed earlier: how far down its path must an input travel for a slice to know that an event has occurred?

There is a vast body of literature on perceptual thresholds tracing back to the seminal work of Weber and Fechner in the 19th century. Of particular interest here are the notion of temporal thresholds – that sensory stimuli are temporally integrated to generate perceptions (Hughes 1946, Green 1960, Watson 1986, Sussman et al. 1999). We will use this to motivate a leaky integrate and fire network (Gerstner and Kistler 2002) over the nodes within a slice. Each node is modeled as a leaky integrator; when it fires, a perception corresponding to its category is generated.

For each node u_i in the nodebook of slice M_A, we refer to its activation potential at time t as $A_t(u_i)$, which ranges between 0 and 1. If $A_t(u_i) \geq 1-\varepsilon$, we say the node *fires*, and the quantity $1-\varepsilon$ is called its *firing threshold*.

We define a mapping function $V : R^N \to N_A$, which assigns a point to its nearest node in nodebook N_A. Namely, *V* describes the Voronoi regions over the nodebook. For each node u_i, we associate a function $u_i(p)$ where $u_i(p)=1$ if and only if $V(p)=u_i$. Otherwise, it is equal to zero. When $u_i(p)=1$, we say $u_i(p)$ is *active*.

A node's activation increases when the perceptual estimate falls within it. We define the change in potential due to direct activation:

$$D(t) = \frac{1}{\eta}\delta_W^A(t) \cdot u_i\left(h(t)\right) \tag{4.15}$$

The variable η determines the temporal threshold for the category represented by u_i. It answers the question: how many times must the nodes in a single category be active for that category to be perceived? Although much has been written documenting temporal thresholds, it is not always clear how (or if) they are acquired. In our framework, we make the following fairly weak assumption: during development, a sufficient number of unambiguous sensory inputs are observed such that we can derive (e.g., 90%) confidence intervals on their lengths. Then for some confidence interval, $[t_{min}, t_{max}]$, we take the lower bound t_{min} as the temporal threshold. If the inputs are sampled at λ Hz, then we define $\eta = \lambda \cdot t_{min}$. We note there is some evidence for simplified (unambiguous) inputs during development in people, where parents exaggerate inflection and pitch in speech (in what is traditionally known as "motherese"), presumably to help infants analyze sounds (Garnicia 1977).

To this, we also add a leakage term. Leakage provides that in the absence of stimulation, each node drifts to its rest value, which we uniformly define as zero in this model. This is necessary to counter the effect of noise in the perceptual inputs and of spurious intersensory activations. We define the leakage at time *t*:

$$\mathrm{L}(t) = \exp\left(-(t-t_0)/\tau\right) \tag{4.16}$$

where time constant τ is equal to t_{max} in the confidence interval above and t_0 is the most recent time the node was active.

Therefore, for each node u_i, we have:

$$\begin{aligned} \frac{dA_t(u_i)}{dt} &= D(t) + \dot{L}(t) \\ \frac{dA_t(u_i)}{dt} &= \frac{1}{\eta}\delta_W^A(t)\cdot u_i\left(h(t)\right) - \frac{1}{\tau}\exp\left(-(t-t_0)/\tau\right) \end{aligned} \tag{4.17}$$

We thereby define a system of $|N_A|$ simultaneous state equations for mode M_A.

Based on these node activation potentials, we derive activation potentials for perceptual categories:

$$A_t(c_i) = \sum_{u_j \in c_i} A_t(u_j) \tag{4.18}$$

That is, the activation potential of a category is the sum of the activation potentials of its nodes. If $A_t(c_i) \geq 1-\varepsilon$, then we say that category c_i has *fired* and the slice outputs this category as its perception. At this point, the activation potentials for all nodes are reset to zero. Assuming we do not implement a refractory period, the node is free to await its next input. In practice, we set $\varepsilon = 10^{-4}$ to account for numerical rounding errors during our simulation.

4.3.4.1 Hebbian Activations

Finally, we note an important variation on this model by allowing activation potentials to spread among nodes in different slices. In this way, we can incorporate Hebbian influence directly into activation potentials and thereby reduce the temporal thresholds for perceptual events.

For $A_t(M_B)$, M_B's activation potentials at time *t*, let:

$$S = \bar{E}_A\left(f\left(\bar{\Phi}\left(A_t(M_B)\right)\right)\right) = \{e_1, e_2, ..., e_a\} \tag{4.19}$$

be the activation set projected onto mode M_A from M_B via (4.11). We define the *Hebbian activation potential* $H_t(u_i)$ induced on node u_i at time *t*:

$$H_t(u_i) = \frac{1}{\eta_b}\,\delta_W^B(t)\cdot e_i \tag{4.20}$$

where η_b is defined with respect to M_B. This allows M_A to directly incorporate M_B's activation potentials into its own, yielding new state space equations for each u_i:

$$\begin{aligned}\frac{dA_t(u_i)}{dt} &= D(t) + \kappa H_t(u_i) + \dot{L}(t)\\ \frac{dA_t(u_i)}{dt} &= D(t) + \frac{\kappa}{\eta_b}\,\delta_W^B(t)\cdot e_i + \dot{L}(t)\end{aligned} \tag{4.21}$$

We fully expand (4.21) to show the contribution of $A_t(M_B)$, the activation potentials in M_B:

$$\frac{dA_t(u_i)}{dt} = D(t) + \left[\frac{\kappa}{\eta_b}\,\delta_W^B(t)\cdot \bar{E}_A\left(f\left(\bar{\Phi}\left(\boxed{A_t(M_B)}\right)\right)\right)\right]_i + \dot{L}(t) \tag{4.22}$$

The system defined by (4.22) reduces the effective temporal thresholds in perceptual systems by spreading activation potentials between their nodes. We vary κ between 0 and 1 to adjust this cross-modal *sensory acceleration,* in proportion to the observed agreement between slices. Cross-modal perception is known to significantly improve upon unimodal response times in humans (Hershenson 1962, Frens 1995, Calvert et al. 2000); this is captured by our model here.

4.4 Summary

In this chapter, we have introduced a new family of models called *influence networks*, which incorporate temporal dynamics into a perceptual framework. These networks fuse together information from the outside world, innate perceptual structure, and intersensory influence to increase perceptual accuracy within slices. The cross-modal aspect of this is illustrated in Figure 4.8, where auditory formant data influences the visual perception of lip contours by inducing a Hebbian gradient on it. This has the effect of "moving" the visual sensation to be in closer accord with the auditory perception.

Influence networks incorporate several basic biological phenomena into a computational model, including: cross-modal influence; dynamic adjustment of sensory and temporal thresholds; cross-modal substitution; and bistability. We will show in Chapter 6 that while these are fundamental perceptual features in biological systems, they largely tend to be absent in artificial ones. Influence networks are a step toward creating more capable artificial perceptual systems by providing them with a more realistic sensorial framework. Figure 4.9 illustrates how an influence network can be incorporated into a

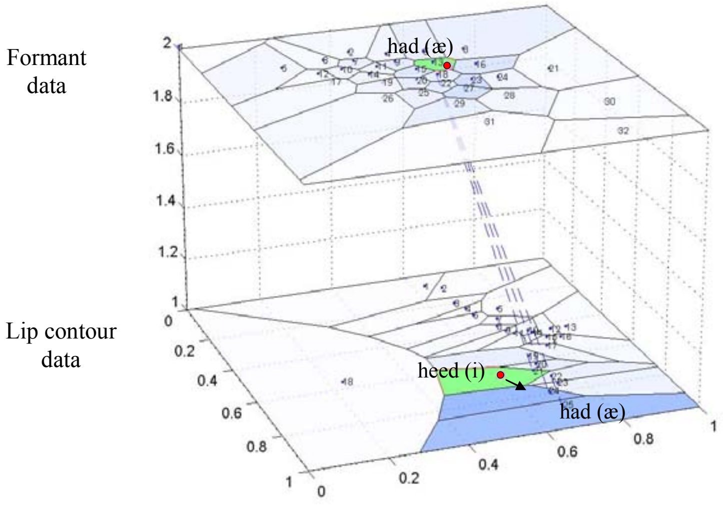


Figure 4.8 -- Cross-modal activations in an influence network. The auditory perception in formant space on top increases activations in the lip contour space on the bottom. This example is unusual because it shows an auditory modality influencing a visual one, which is biologically realistic but unusual in an artificial perceptual system. The effect of this influence is that the visual perception is modified due to an induced Hebbian gradient. Shading here corresponds to activation potential levels.

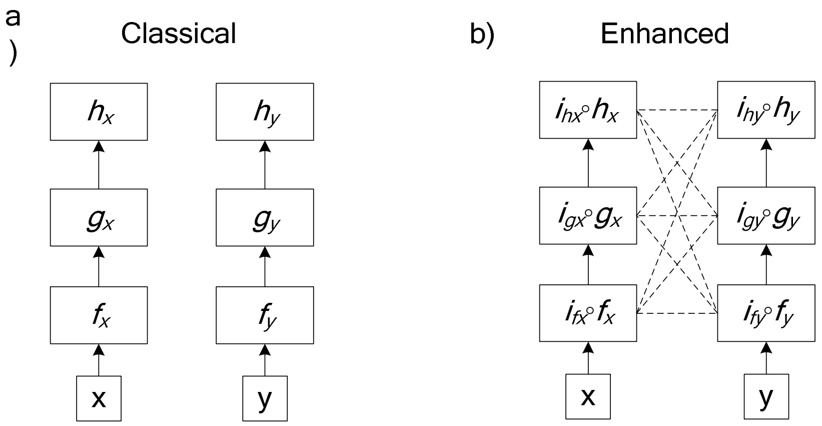


Figure 4.9– Adding an influence network to two preexisting systems. We start in (a) with two pipelined networks that independently compute separate functions. In (b), we incorporate an influence network into this architecture that interconnects the functional components of the pipelines and enables them to dynamically modify their percepts.

preexisting artificial perceptual system to enable a new type of cross-modal influence, designed to increase overall perceptual consistency.

Chapter 5

Sensorimotor Learning

Up to this point, we have been concerned with learning to *recognize* events in the world. We now turn to the complementary problem of learning to *generate* events in the world. That these two problems are interrelated is well established – animal behaviors are frequently learned through observation, particularly in vertebrates (Bloedel et al. 1996). In this chapter, we will propose a computational architecture for acquiring intentional motor control guided by sensory perception. Our approach addresses two basic questions:

1) **What is the role of *perceptual* grounding in learning *motor* activity?** In other words, how does the categorization of sensory events assist in the acquisition of voluntary motor behaviors?

2) **Can motor systems internally reuse perceptual mechanisms?** Specifically, we examine the possibility that the perceptual framework presented in Chapter 3 can be applied to learning motor coordination. This is the most important issue addressed in this chapter, because it generalizes to suggest how higher level cognitive structures may be iteratively bootstrapped off lower level perceptual inputs. In doing so, it suggests a framework for realizing the embodied cognition approaches of Brooks (1991a), Lakoff (1987), and Mataric (1997).

Towards answering these questions, we present a computational architecture for *sensorimotor* learning, where an animal (or a machine) acquires control over its motor systems by observing the effects of its own actions. Sensory feedback can both initially guide juvenile development and then subsequently refine adult motor activity. This type of self-supervised learning is thought to be among the most powerful developmental mechanisms available both to natural creatures (Thorndike 1898, Piaget 1971, Hall and Moschovakis 2004) and to artificial ones (Weiner 1948, Maes and Brooks 1990).

Our approach is to reapply the framework in Chapter 3. We will treat the motor component of sensorimotor learning as if it were a perceptual problem. This is surprising because one might suppose that motor activity is fundamentally different than perception. However, we take the perspective that motor control can be seen as perception *backwards*. We imagine that – in a notion reminiscent of a Cartesian theater – an animal can "watch" the activity in its own motor cortex, as if it were a privileged form of *internal* perception. Then for any motor act, there are two associated perceptions – the *internal* one describing the generation of the act and the *external* one describing the self-observation of the act. The perceptual grounding framework described in Chapter 3 can then *cross-modally ground* these internal and external perceptions with respect to one another. The insight behind this approach is that *a system can develop motor control by learning to generate the events it has previously acquired through perceptual grounding.*

A benefit of this framework is that it can learn *imitation*, a fundamental form of biological behavioral learning (Byrne and Russon 1998, Meltzoff and Prinz 2002). In imitative behaviors – sometimes known as *mimicry* – an animal acquires the ability to reproduce some aspect of another's activity, constrained by the capabilities and dynamics of its own sensory and motor systems. This is widespread in the animal kingdom (Galef 1988) and is thought to be among the primary enablers for creating self-supervised intelligent machines (Schaal 1999, Dautenhahn and Nehaniv 2002).

We will demonstrate sensorimotor learning in this framework with an artificial system that learns to sing like a zebra finch. Our system first listens to the song of an adult finch; it cross-modal clusters this input to learn *songemes*, primitive units of bird song that we propose as an avian equivalent of phonemes. It then uses a vocalization synthesizer to generate its own nascent birdsong, guided by random exploratory motor behavior. The motor parameters describing this exploratory vocal behavior are fed into *motor slices* – as if they corresponded to external perceptual inputs. By simultaneously listening to itself sing, the system organizes these *motor* slices by cross-modally clustering them with respect to the previously learned *songeme* slices. During this process, the fact that the motor data were derived internally from innate exploratory behaviors, rather than from external perceptual events, is irrelevant. By treating the motor data as if they were derived perceptually, the system thereby learns to reproduce the same sounds to which it was previously exposed. This approach is modeled on the dynamics of how male juvenile finches learn birdsong from their fathers (Tchernichovski et al. 2004, Fee et al. 2004).

The model presented here is inspired by sensorimotor learning in animals, and it shares a number of features with several prominent approaches to modeling biological sensorimotor integration (e.g., Massone and Bizzi 1990, Stein and Meredith 1994, Wolpert et al. 1995). However, it is intended as an abstract computational model; as such, it does not approach the richness or complexity inherent in biological sensorimotor systems. Our goal in this chapter is to demonstrate that motor learning can be accomplished through iterative perceptual grounding. In other words, we show that perceptual and motor learning are unexpectedly similar processes and can be achieved within a common mathematical framework. This surprising result also suggests an approach to grounding higher level cognitive development, by iteratively reapplying this technique of *internal perception*. We discuss these issues below and will examine them again in Chapter 6.

5.1 A Sensorimotor Architecture

We begin by examining abstract models of innate sensory and motor processing in isolation. Initially, the isolated sensory system simply categorizes the different events to which it is exposed, using cross-modal clustering. One may view this as an unsupervised learning phase, in which perceptual categories are acquired through passive observation. Subsequently, the isolated motor system generates innately specified behaviors using an open-loop control system (Prochazka 1993). In other words, it receives no initial feedback. To enable sensorimotor learning, we must close this loop by interconnecting these isolated systems.

For this purpose, we will reuse the perceptual machinery of Chapter 3. Specifically, we introduce the notion of *internal perception*, which allows a system to watch the generation of its internal motor parameters as if they were coming from the outside world. By treating the motor parameters as perceptual inputs, they can be cross-modally clustered, regardless of their internal origins. The system thereby learns to generate events it has previously learned how to recognize, by associating motor parameters with their observed effects.

5.1.1 A Model of Sensory Perception

Our framework begins with the model of afferent sensory perception outlined in Figure 5.1, which schematically diagrams an abstract computational sensory cortex. In this model, external events in the world impinge upon sensory organs. These receptors in turn generate perceptual inputs, which feed into specialized perceptual processing channels. A primary outcome of this processing is the extraction of descriptive features (e.g., Muller and Leppelsack 1985, Hubel 1995), which capture abstracted sensory detail. This process occurs in parallel within multiple sensory pathways, as illustrated in Figure

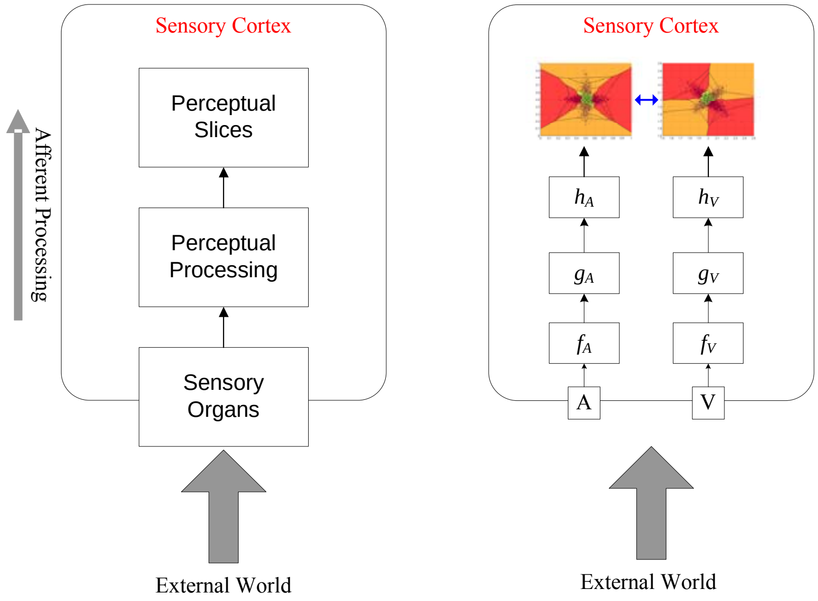


Figure 5.1 – An abstract model of sensory processing in our framework. A schematic view is shown on the left, which is expanded upon in the example of the right. Events in the world are detected by sensory organs, here labeled A and V, representing auditory and visual receptors. These are fed into processing pipelines shown here by the composition of functional units. The features extracted from these pipelines are fed into slices, which are then cross-modally clustered with respect to one another. We discuss this model further in the text.

5.1 on the right. This hypothetical example shows auditory and visual receptors that provide inputs to their respective perceptual pathways. These channels extract features from their perceptual input streams, which are fed into the slices displayed on top. These slices are then cross-modally clustered with respect to one another, as described in Chapter 3.

Later in the thesis, we reexamine the structure of these sensory pipelines. In biological systems, sensory channels are highly interconnected and display complex temporal dynamics; we will modify our perceptual model to reflect that. In Chapter 6, we examine the biological implausibility of assuming independence between perceptual channels. However, the sensorimotor learning in this chapter assumes only that slices can be cross-modally clustered, an assumption which remains valid in the subsequent elaborations later in this thesis.

5.1.2 A Simple Model of Innate Motor Activity

We now present an abstract model of innate efferent motor activity, which is sometimes called *reflexive behavior*. It is well established that young animals engage in a range of involuntary motor activities; much of this appears to facilitate the acquisition of cognitive and motor functions, leading to the development of voluntary, intentional behaviors (Pierce and Cheney 2003, Chapter 3). Behavioral learning is therefore not a passive phenomenon; instead, it is often guided by phylogenetically "programmed" activities that have been specifically selected to satisfy the idiosyncratic developmental requirements of an individual species (Tinbergen 1951).

An abstract model describing the generation of innate efferent motor activity is shown in Figure 5.2. In a sense, this model is the reverse of the one displayed in Figure 5.1. Instead of the outside world generating events, we assume an innate generative mechanism stimulates a motor control center. This in turn evokes coordinated activity in a muscle or effector system, leading to the generation of an external event in the world.

In our model, the innate specification of developmental behaviors is represented by a joint probability distribution over a set of parameters governing motor activity. This is motivated by Keele and Summers (1976), where a motor program is described by a descriptive parameterization. Alternatively, one could assume the existence of a set of deterministic motor schemas, corresponding to predetermined patterns of activity (Arbib 1985). From the perspective of our model, this distinction makes little difference; we simply assume some mechanism (or set thereof) is responsible for producing the innate behaviors that will eventually generate feedback for sensorimotor learning.

To give a clearer sense of this process, we examine the diagram on the right in Figure 5.2. This presents an example of human vocal articulation, motivated by (Rubin et al. 1981). Although we will focus primarily on avian vocalization later in this chapter, first examining human articulation has strong intuitive and didactic appeal, and it sets the

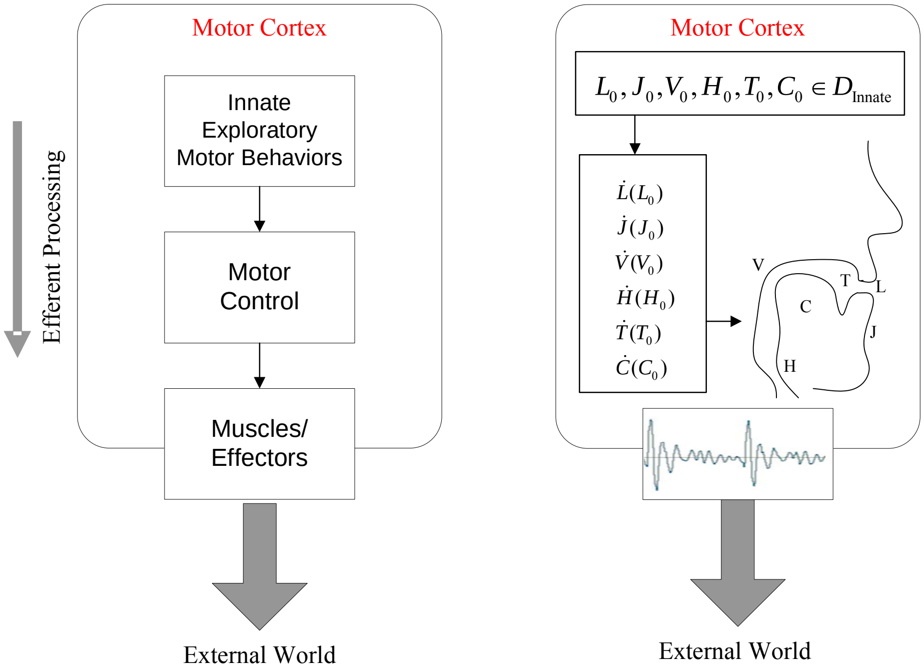


Figure 5.2 – An abstract model of innate motor activity. A schematic view is shown on the left, which is expanded upon in the example on the right. On the bottom right is a model of human vocal articulation. This is parameterized by articulator positions at the lips (L), tongue tip (T), jaw (J), tongue center (C), velum (V), and hyoid (H). Motor control corresponds to a set of state equations on the left governing the constrained movement of these articulators over some time period. Parameters describing this movement are selected from some assumed innate distribution on the top right. The selection of parameters in this model is based on (Rubin et al. 1981).

stage for discussing birdsong generation below. In this example, a set of parameters are selected from some innate distribution D_{Innate}; these parameters describe the trajectories of six primary human vocal articulators, as illustrated in the figure. The selection of parameters thereby corresponds to a primitive speech act. The actual movements of the articulators during a speech act are modeled by a set of partial differential equations, determined by the biomechanical constraints on the musculoskeletal apparatus of the human vocal tract. In other words, these equations model the physical characteristics and limitations of human vocal production independently of what is being said. Together, the parameters and the state equations lead to the generation of a speech act. In this domain, learning how to produce meaningful sounds corresponds to selecting parameters that appropriately generate them, given the physical and dynamic constraints of human vocalization.

5.1.3 An Integrated Architecture

Towards sensorimotor learning, we now interconnect these isolated sensory and motor systems. To do this, we introduce the notion of *internal perception*, which allows a system to "watch" the generation of its internal motor parameters as if they were coming from the outside world. Thus, we will create *motor slices* that are populated with *behavioral* data, in exactly the same way we created *perceptual slices*, which were populated with *sensory* data. The resulting slices do not "know" if their data were generated internally or externally, and for the purposes of cross-modal clustering, it makes no difference. We can thereby learn *motor categories* that correspond to previously acquired *perceptual categories*.

In our model, internal perception occurs through the addition of a *Cartesian theater* (Dennett 1991), so named because it provides a platform for internal observation. Pursuing this philosophical metaphor a bit further, the *homunculus* in our theater will be replaced by cross-modal clustering. As we saw in Chapter 3, this is an unsupervised learning technique. We may therefore employ the notion of a Cartesian theater without engendering the associated dualistic criticisms of Ryle (1941) or Dennett (1991). In fact,

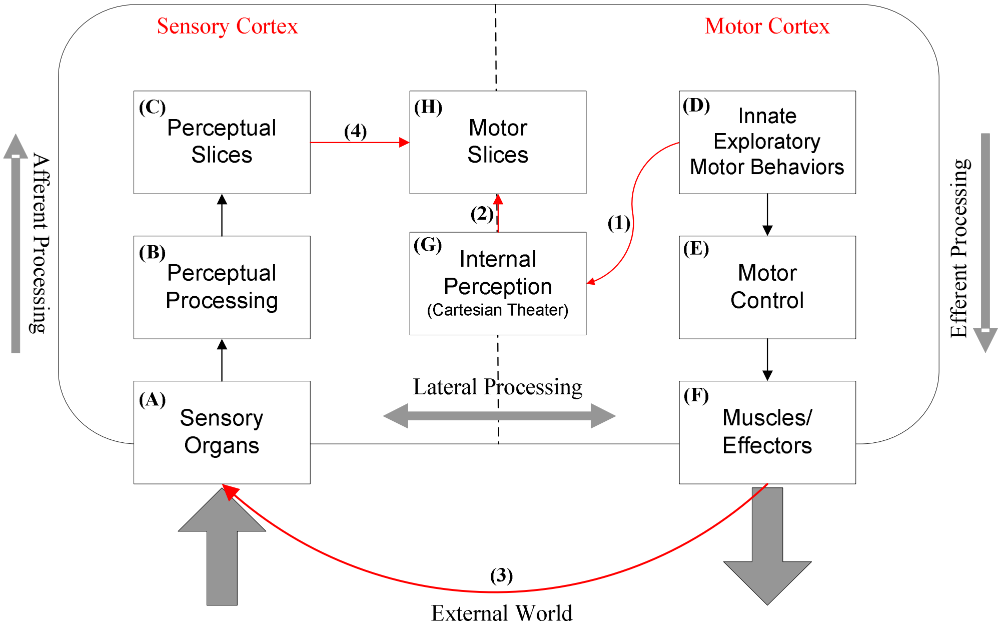


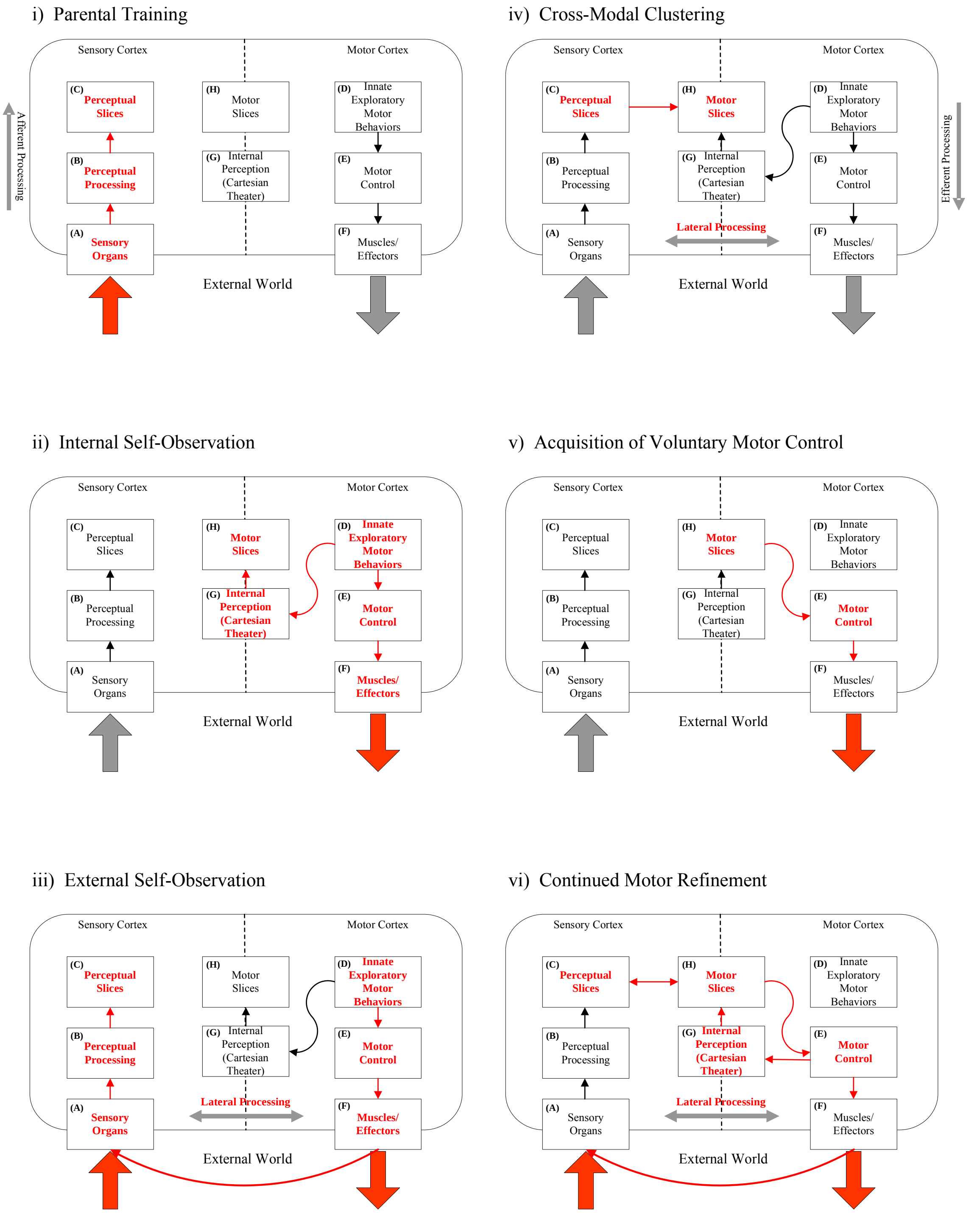
Figure 5.3 – An integrated sensory motor framework. We connect the isolated sensory and motor systems with the addition of a Cartesian theater (G), which receives data via (1), corresponding to innate exploratory behaviors generated in (D). These data are fed into motor slices (H) via (2). These exploratory behaviors also trigger motor activity via the efferent pathways in (E) and (F). Most importantly, the system is able to perceive its own actions, as shown by (3). These inputs feed into the afferent sensory system, where features are extracted and fed into perceptual slices (C). We thereby learn the Hebbian linkages between the codebook clusters in perceptual slices (C) and motor slices (H), which describe the generation of these perceptions. In the final step, we cross-modally cluster the motor slices (H) with respect to the perceptual slices (C); we thereby learn the motor categories that generate previously acquired sensory categories learned when the system was perceptually grounded.

we will argue that internal perception is a useful framework for higher level cognitive bootstrapping, where cross-modal clustering replaces a homunculus and any notions of "intentionality" (Dennett and Haugeland 1991) are attributed to innate phylogenetic structures and tendencies. In other words, we will keep the theater but eliminate the metaphysical audience.

Our integrated sensorimotor framework is shown in Figure 3. We briefly outline this architecture and then examine the stages of sensorimotor learning in detail below. In the above diagram, we see the independent sensory and motor components described above on the left and right respectively. In addition, there is now a Cartesian theater (G), which receives inputs corresponding to innate exploratory behaviors generated by (D). These internally produced data are fed to *motor slices* (H), which are thereby populated with behavioral rather than perceptual data. These data are also simultaneously fed into a motor control system (E), which lead to the generation of perceivable events in the external world.

Most importantly, the system observes its own actions. Innately generated events impinge upon the sensory organs and are fed into the sensory apparatus on the left. Features extracted from these data are fed into sensory slices (C). This process thereby creates Hebbian linkages between the sensory slices (C) and the motor slices (D).

We point out that slices are what may be deemed *agnostic* data structures – they neither "know" nor "care" what type of data they contain. We can therefore cross-modally cluster the motor slices (D), based on the categories acquired during the perceptual grounding of the sensory slices (C). Note, that this is a one-way process. In other words, we fix the sensory categories and only cluster the motor data. We thereby learn *motor categories* that correspond to previously acquired *perceptual categories*. We discuss relaxing this one-way restriction below, to allow increasing motor sophistication to assist in restructuring perceptual categories. This type of perceptual refinement as a consequence of fine motor development has been observed in humans, particularly in musicians (e.g., Ohnishi et al. 2001).



i) Parental Training

iv) Cross-Modal Clustering

ii) Internal Self-Observation

v) Acquisition of Voluntary Motor Control

iii) External Self-Observation

vi) Continued Motor Refinement

Figure 5.4 – Developmental stages in our model. i) The juvenile acquires perceptual structures from its parent. ii) Motor acts are observed *internally* through a Cartesian Theater. iii) The effects of motor acts are observed *externally* through perceptual channels. iv) Motor slices are cross-modally clustered with respect to perceptual slices. The juvenile thereby learns how to generate the events it learned in stage (i). v) Random exploratory behaviors are disconnected and motor slices take over the generation of motor activity. The juvenile is now able to intentionally generate the sensory events acquired from its parent. vi) Internal perception can be used subsequently in non-juveniles to refine motor control.

5.2 Stages of Sensorimotor Development

We now examine the developmental stages of our framework, as illustrated in Figure 5.4. Note that one may introduce any number of complexities into these stages, a few of which we examine below. However, we sidestep a number of issues relevant to any developmental model that do not uniquely distinguish our approach. For example, we make no commitment to the role of innate vs. learned knowledge (e.g., Chomsky 1957, Meltzoff and Moore 1977, Fodor 1983) and believe we can incorporate either perspective. The primary goal here is to outline how our model computationally learns imitative behaviors, towards examining the acquisition of birdsong in the next section. We therefore avoid a number of important albeit orthogonal topics, some of which we will examine in Chapter 6.

5.2.1 Parental Training

In the first stage of our model, we assume the system corresponds to a neonate. Its sensory processing channels are sufficiently developed to extract features from perceptual inputs, provided by its parents, other animals (particularly conspecifics), or directly from the environment. The goal of the system at this point is to collect sufficient data to begin cross-modal clustering and thereby become perceptually grounded. Let us consider the outcome of this grounding process, as illustrated in Figure 5.5. Here, a set of four different events in the world has been acquired using the framework of Chapter 3. The goal then is for the system to learn how to generate these events by itself.

We note that perceptually grounding may involve a gradual progression rather than a sudden transition. Within a single modality, some categories may be easier to learn than others, due to their inherent structures, availability of training data, degree of perceptual redundancy, individual variations, etc. In contrast, some perceptual features may be higher-level aggregates of simpler ones. As such, their acquisition depends upon that of their components, which themselves may be differentially acquired. For example. phonological development in children proceeds in a number of well-defined, interconnected stages that constructively build upon each other (Vihman 1996). During

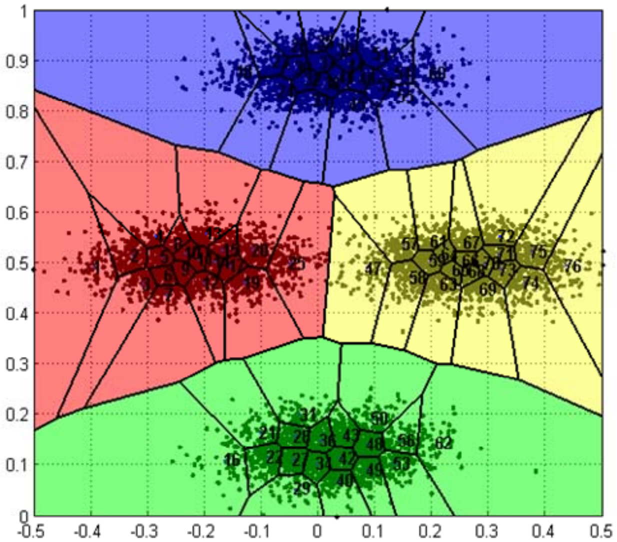


Figure 5.5 – A hypothetical sensory system that has learned four events in the world. These are acquired through cross-modal clustering, using the framework in the previous chapter. For simplicity, only a single sensory mode is illustrated here.

this lengthy process, children need constant exposure to language, during which they extract features corresponding to their current state of development. More generally, a juvenile may undergo multiple stages of perceptual grounding, during which it acquires increasingly abstract knowledge from its tutors or environment.

Although we assume that slices are initially unstructured, i.e., *tabula rasa*, it is certainly possible to incorporate *innate pre-partitioning* of slices into our framework. Also, our model does not specify how perceptual features are selected in the first place. We assume this is specified genetically in the biological world and programmatically in the artificial world.

5.2.2 Internal and External Self-Observation

The second and third stages of our model correspond to a system's observation of its own innate, exploratory motor activity. Although reflexive behaviors are phylogenetically selected in animals to satisfy their individual motor requirements (Tinbergen 1951), in artificial systems, we must specify how these innate behaviors are generated. While it

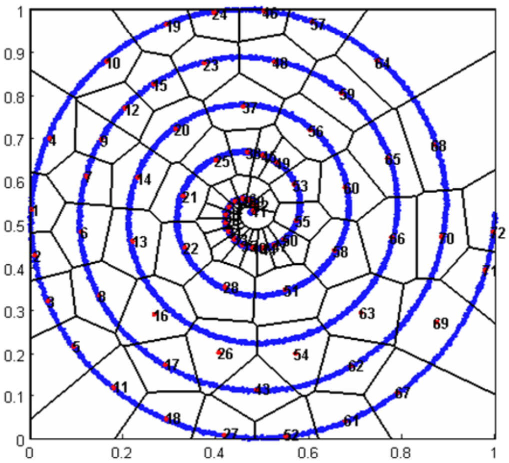


Figure 5.6 – Internal perception of exploratory motor behavior corresponding to an Archimedean spiral. These data correspond to the parameters used to generate motor activity.

may often be reasonable to design exploratory behaviors that are predetermined to satisfy a set of motor goals, we examine a generic strategy here. Our goal is simply to explore a motor space and in doing so, simultaneously observe the effects *internally* through the Cartesian theater and *externally* through normal perceptual channels.

Consider, for example, the problem of generating pairs of exploratory parameters (*x*,*y*) in a hypothetical motor system. We have found it useful to select these parameters and thereby explore motor spaces according to an Archimedean spiral. We could therefore

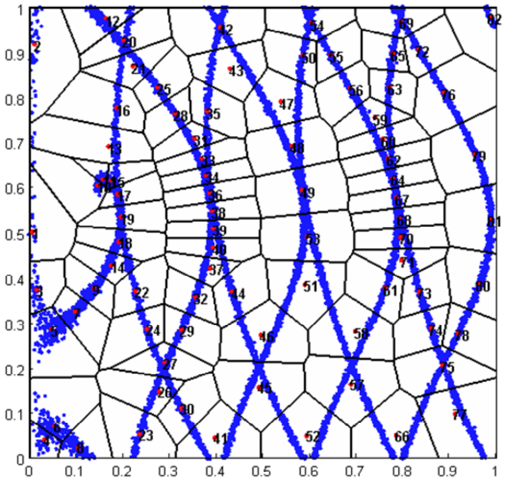


Figure 5.7 – External perception of exploratory motor behavior. This slice perceives the events generated by the motor activity described by Figure 5.6. These data correspond to perceptual features describing sensory observations.

specify that the parameter values are drawn from a distribution described by $x = \alpha_1 \theta \cdot \cos(\alpha_2 \theta)$ and $y = \alpha_1 \theta \cdot \sin(\alpha_2 \theta)$. In this case, the *internal perception* of this motor activity might be represented by the slice in Figure 5.6.

Note that the slice representing the *external* perception of this motor activity may "look" entirely different than the motor slice representing its generation. In other words, there is no reason to expect any direct correspondence or isomorphism between motor and perceptual slices. The motor parameters indirectly generate perceptual events through an effector system, which may be non-linear, have discontinuities, or display complex dynamics or artifacts.

We see this phenomenon in the slice in Figure 5.7, which visualizes perception of the motor activity described by Figure 5.6. For the purpose of this example, we generated a non-bijective mapping between the motor parameters and the perceptual features of events they generate in order to illustrate the degree to which corresponding motor and perceptual slices may appear incongruous. Thus, even though these two slices may represent the same set of percepts abstractly, they should not be expected to bear any superficial resemblance to each other. This was also the case with the solely perceptual slices in Chapter 3; however, one might have assumed that a stronger correspondence would exist here due to the generative coupling between the slices, which is not the case.

5.2.3 Cross-Modal Clustering

The fourth stage of our model allows us to learn the correspondences between motor and perceptual slices. Figure 5.8 shows the interconnection of the three slices introduced above. On the bottom left in (A), we see the categories acquired during perceptual grounding, as shown in Figure 5.5. Using these, the system can categorize external observations of its own activities as shown in (B); this corresponds to the cross-modal clustering of the slice in Figure 5.7. It can then subsequently use these classifications to categorize internally observed motor behaviors, which were responsible for generating

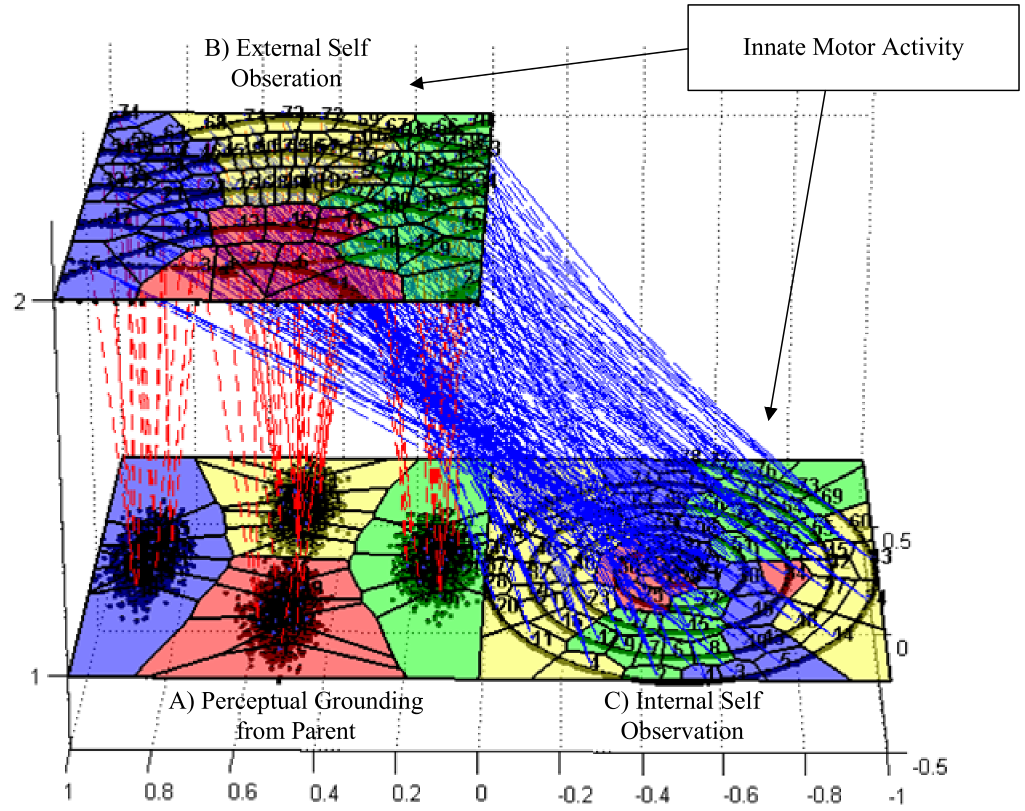


Figure 5.8 – Stages of cross-modal clustering. Starting from the acquisition of perceivable events in (A), we learn to classify the effects of our own behaviors in terms of these events in (B). Finally, we can then relate this back to the innate motor activity generating our actions, as in (C). There are thereby three stages of cross-modal clustering in this model.

this activity to begin with. This is show in (C) and corresponds to the cross-modal clustering of the slice in Figure 5.6.

Although the discussion here presumes these steps happen sequentially, i.e., the are three "rounds" of cross-modally clustering represented in Figure 5.8, taking us from stages (A) through (C), it is possible to imagine them overlapping. In other words, motor learning may co-occur with perceptual grounding and one may propose a continuum between these two temporal alternatives. Particularly in systems that have limited access to training inputs, self-supplementing these data by generating them independently, as perceptual capabilities are differentially acquired, may help to overcome a paucity in external stimulation. This may be additionally helpful in determining ambiguous subsets of perceptual and motor space, which were discussed in the previous chapter.

5.2.4 Voluntary Motor Control

The fifth stage of our model replaces innate, exploratory behavior with voluntary, intentional motor control. Figure 5.9 displays the bottom two slices of Figure 5.8, where the motor map on the right is now clustered according to the perceptions it externally generates – by way of some effector system – in the sensory map on the left. Significantly, the system is now capable of generating the events to which it was exposed during parental training.

As we saw in Chapter 4, motor and sensory events do not happen instantaneously in time. They are not discrete, discontinuous phenomena. Thus, rather than select individual points in a motor map to trigger behaviors, it is far more plausible to imagine a system "moving" through a motor map during a time period corresponding to sustained activity. We previously examined how to move through sensory maps to avoid perceptual ambiguity. We note, however, that one may also wish to incorporate other types of constraints into motor systems, for example, to minimize energy or maximize stability.

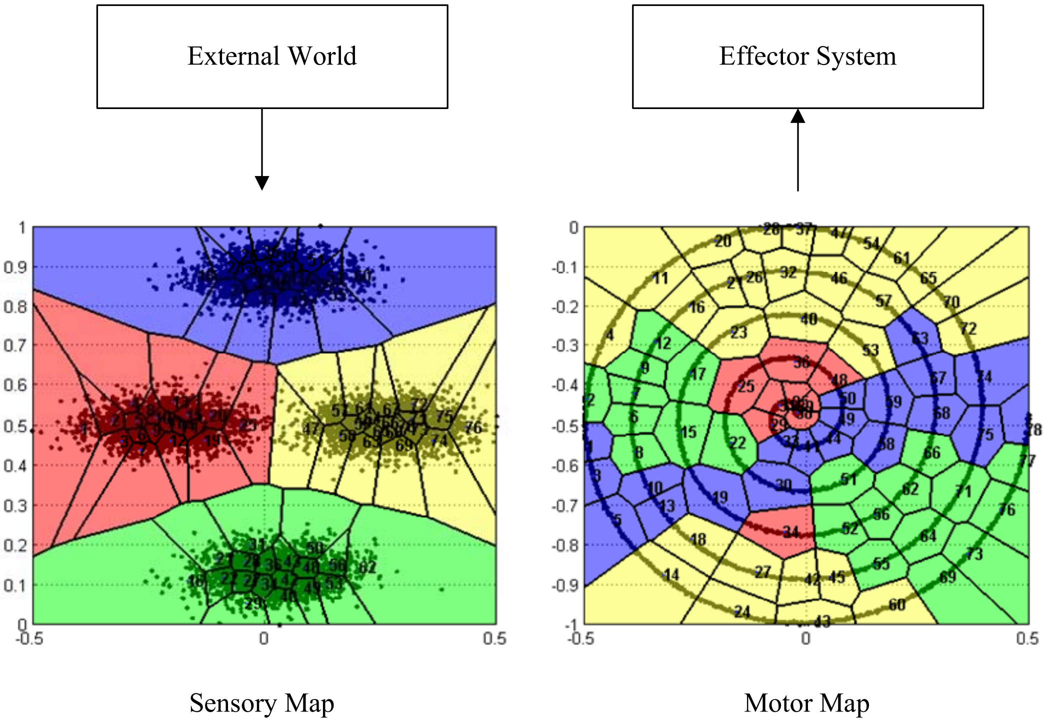


Figure 5.9 – Acquisition of voluntary motor control. Regions in the motor map on the right are now labeled with the perceptual events they generate in the sensory map on the right.

5.3 Learning Birdsong

We now propose to use the framework in the previous section for self-supervised learning of birdsong. Our presentation will focus on song learning in the zebra finch, a popular species for studying oscine songbird vocal production. We begin with a brief introduction to this species, focusing on the structure of its song and the developmental stages of its acquisition. Towards building a computational model, we next introduce the notion of *songemes*, primitive units of birdsong that we propose as avian equivalents of phonemes. Finally, we present a system that learns to imitate an adult zebra finch in a developmentally realistic way, modeled on the dynamics of how male juvenile finches learn birdsong from their fathers (Tchernichovski et al. 2004, Fee et al. 2004).

Our system first listens to an adult male finch and uses cross-modal clustering to learn the *songemes* comprising the song of its "father." It then uses an articulatory synthesizer to generate its own nascent birdsong, guided by random exploratory motor behavior. By listening to itself sing, the system organizes the motor maps generating its vocalizations by cross-modally clustering them with respect to the previously learned *songeme* maps of its parent. In this way, it learns to generate the same sounds to which it was previously exposed.

We are indebted to Ofer Tchernichovski (2005, 2006) for providing the adult zebra finch song recordings used to train our system. We are also grateful to Heather Williams for making available a generationally-indexed birdsong library, which provided an additional source of inputs using during initial testing (Williams 1997, 2006). Other sources of birdsong files are cited individually below.

5.3.1 Introduction to the Zebra Finch

The zebra finch is an extremely popular species for researching birdsong acquisition. (For surveys on birdsong learning, see Brenowitz et al. 1997 and Ziegler and Marler 2004, and Nottebohm 2005). In part, this is due to the ease of maintaining and breeding

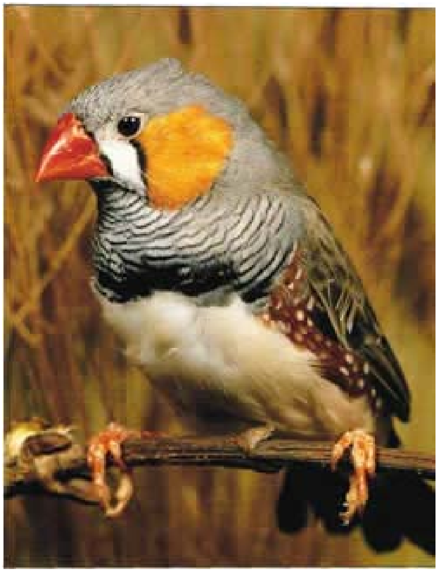


Figure 5.10 – An adult male zebra finch (*Taeniopygia guttata*). Zebra finches are small, unusually social songbirds that grow to approximately 10cm (4 inches) tall. They are extremely popular both as pets and as research subjects for studying neural, physiological, evolutionary, social, and developmental aspects of birdsong acquisition.

---

the species in captivity and its rapid sexual maturity. However, there are two additional characteristics of zebra finches that make them particularly attractive for study.

The first of these is the noisy spectral quality of their songs, which are distinct from the whistled, tonal characteristics of most other songbirds, such as sparrows or canaries. We can see this harmonic complexity in Figure 5.11, which displays spectrograms of songs from several oscine species, along with a human vocalization for reference. It is hypothesized that spectral complexity of a zebra finch's song reflects its physical prowess and aids in sexual selection (Kroodsma and Byers 1991). The complexity of their vocalizations, along with a range of behavioral and neurological similarities, has prompted many researchers to propose studying song learning in zebra finches as a model for understanding speech development in humans (Marler 1970, Nottebohm 1972, Doupe and Kuhl 1999, Brainard and Doupe 2002, Goldstein et al. 2003). Perhaps supporting this notion, it has been determined that human FOXP2, the first gene linked to speech and vocal production, has a protein sequence that is 98% identical to the same gene in the zebra finch (Haesler et al. 2004, see also Webb and Zhang 2005).

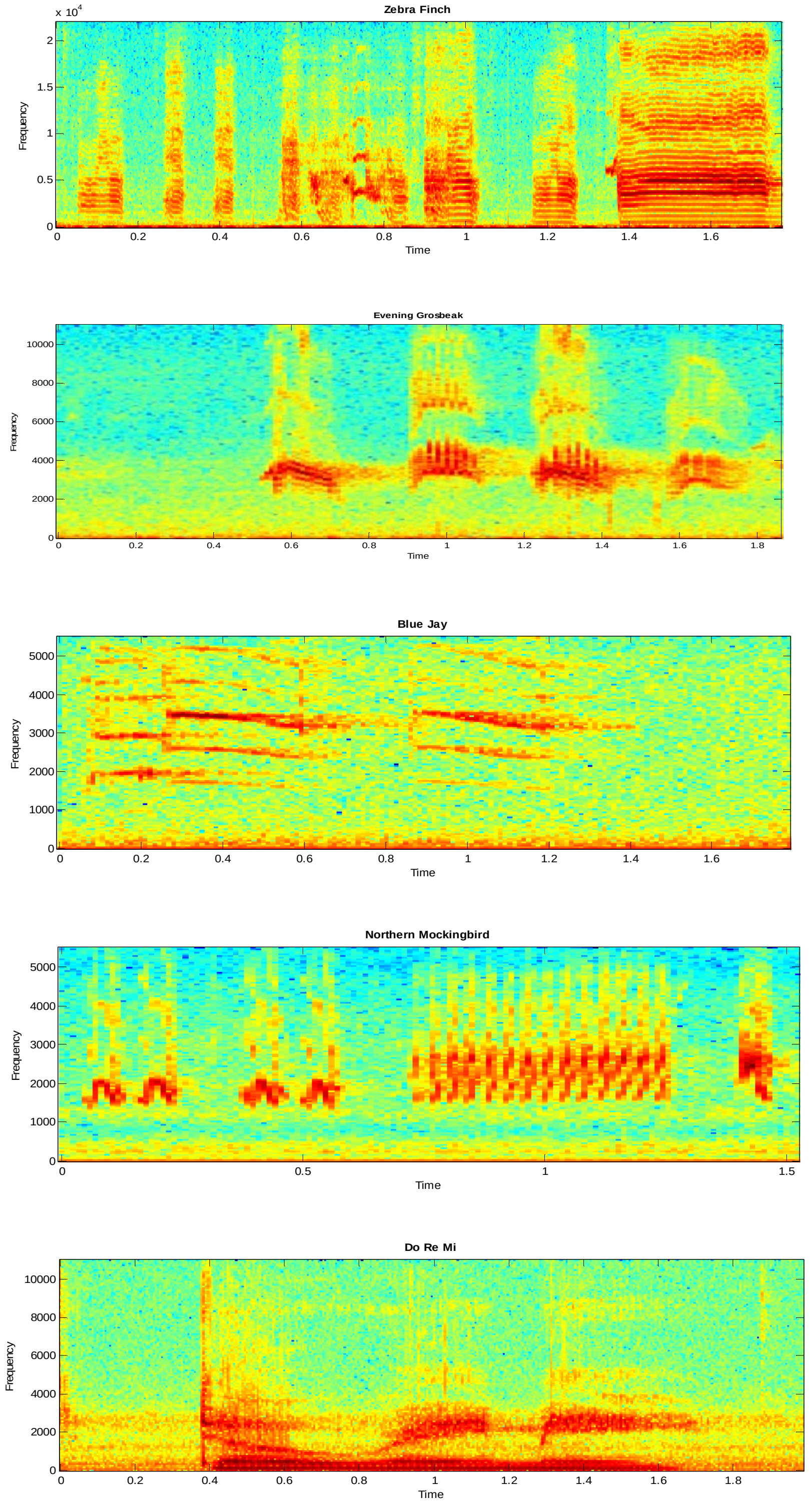


Figure 5.11 – Spectrograms of songs from five different species. These include: i) a zebra finch; ii) an evening grosbeak; iii) a blue jay; iv) a northern mockingbird; and v) the author singing "Do Re Mi" to provide a reference with human vocalization. Notice the complex harmonic structure of the zebra finch's song compared to those of the other birds. Song in (i) provided by (Tchernichovski 2006**)**. Songs in (ii)-(iv) were obtained from the U.S. National Park Service (2005). Note that the frequency ranges are different in each of these spectrograms.

The second characteristic that makes zebra finches popular models for song learning is the well-defined developmental process through which their song is acquired (Slater et al. 1988). Fledglings of both sexes begin learning their father's song early in life. Approximately one month after hatching, male juveniles begin producing nascent, squeaky sounds and then proceed through a series of stages of vocal refinement (Immelmann 1969, Tchernichovski and Mitra 2002). The goal of this process for a zebra finch is to learn to approximately reproduce its father's song, accompanied by idiosyncratic, individual variations unique to each bird. Therefore, a son sounds similar but not identical to his father. At 90 days of age, a bird's song crystallizes as it simultaneously reaches sexual maturity, and its song template remains unchanged for the remainder of its life (Doupe and Kuhl 1999). Thus, an adult male zebra finch adheres to a single song motif, where each vocalization is constructed from a fixed set of acquired components. We note that as with many other oscine songbird species, females zebra finches do not sing. Instead, they make simple vocalizations known as distance calls, through which birds of both sexes are able to individually recognize one another (Miller 1979, Vignal et al. 2004).

The juvenile development of song generation is heavily guided by auditory feedback (Konishi 1965). In fact, as a bird begins to vocalize, it no longer requires exposure to its tutor's song but instead, it must be able to hear itself sing. Adult birds also need feedback to maintain their singing ability (Nordeen and Nordeen 1992). Thus, even though adult males cannot learn new songs, they require auditory feedback to maintain the neural song patterns acquired as juveniles (Brainard and Doupe 2000).

How juveniles learn from auditory feedback is unknown. Marler (1997) outlines three models of sensorimotor-based song development. These range from fully open, instructive tutoring to the assumption of innate neural templates for highly constrained, conspecific song patterns. He argues for an intermediate approach, incorporating song memorization into a phylogenetically constrained framework that has been selected to facilitate rapid learning within an individual species.

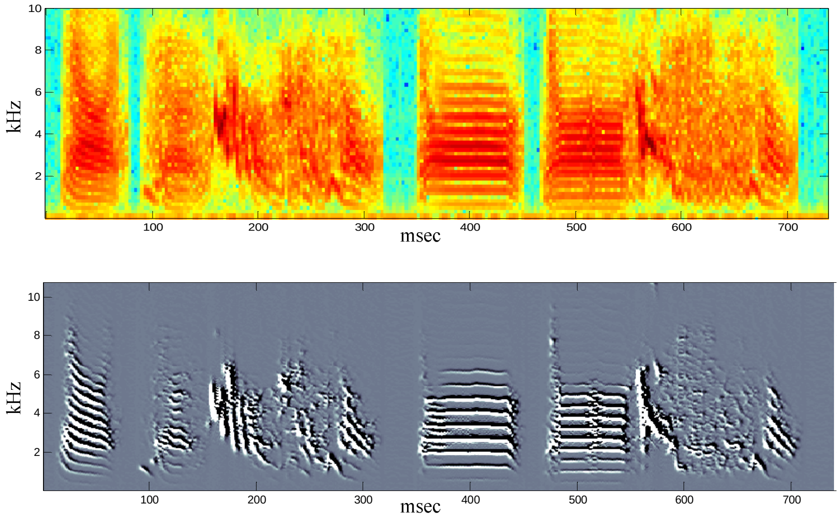


Figure 5.12 – Comparing a spectrogram to a spectral derivative display for the song of a zebra finch. A spectrogram for a zebra finch song is displayed on top. The spectral derivative of that song is shown on the bottom and provides much clearer visual detail. It also provides a framework for subsequent harmonic analysis.

5.3.2 Songemes

We now introduce the notion of *songemes*, primitive units of birdsong that we propose as avian equivalents of phonemes. Marler (1997, p508) notes that although "*it is the subject of score of studies of song learning and its neural basis, … little effort has been directed to descriptive studies of zebra finch song structure.*" In the research literature, birdsong is generally divided into components known as *syllables*, which are continuous sounds bounded by silent intervals (e.g., Williams and Staples 1992, Ölveczky et al. 2005). It is not, however, generally broken down into smaller units. The goal of this section is to define primitive units we call *songemes*, which we argue correspond more closely to basic elements of physiological song production.

Before proceeding, we briefly discuss the multitaper spectral analysis methods (Thomson 1982) that were introduced into the analysis of birdsong by (Tchernichovski et al. 2000). Traditional spectrograms show the power at different frequency components of a signal over time, as in the top of Figure 5.12. Multitaper methods compute spectrograms while simultaneously providing estimates of spectral derivatives. That is, rather than only measuring power, they also measure instantaneous changes in power, and as such,

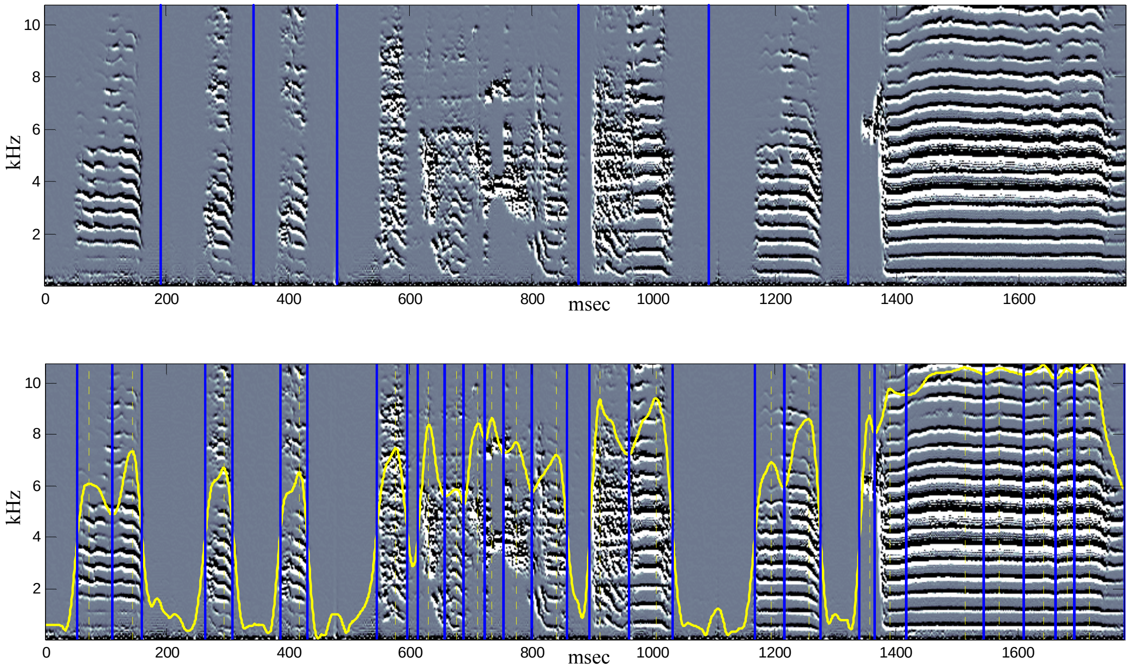


Figure 5.13 – Breaking a birdsong down into constituent songemes. On the top, the song is displayed divided into seven syllables. The 22 derived songemes, defined via peaks of the song's smoothed log(power), are shown on the bottom. The peaks are indicated by the dotted vertical yellow lines. The blue lines indicate songeme boundaries, determined by locally adjacent minima. We note the long vocalization at the end of the song corresponds to a distance call.

perform a sort of edge detection on the spectrogram, making contours easier to detect (see the bottom of Figure 5.12). They also provide a framework for harmonic analysis (ibid, Tchernichovski and Mitra 2004). In the rest of this chapter, we use spectral derivatives in place of spectrograms for examining acoustic analyses of birdsongs.

Returning to our discussion of syllables in bird song, let us consider the song displayed in Figure 5.13. On the top of the figure, we have partitioned the song according to the traditional syllabic breakdown, as determined by the intervals of silence. On the bottom, we have partitioned the song into *songemes*, which captures the fine structure in the song. The songeme partitioning is computed by finding the peaks in the smoothed log(power) of the signal between 860 and 8600Hz, corresponding to the expected vocalization range of a zebra finch. The smoothing is done with a low-pass Savitzky-Golay filter, using a 2nd order polynomial over a window corresponding to approximately 40msec. The songeme boundaries are determined by finding the local minima adjacent to

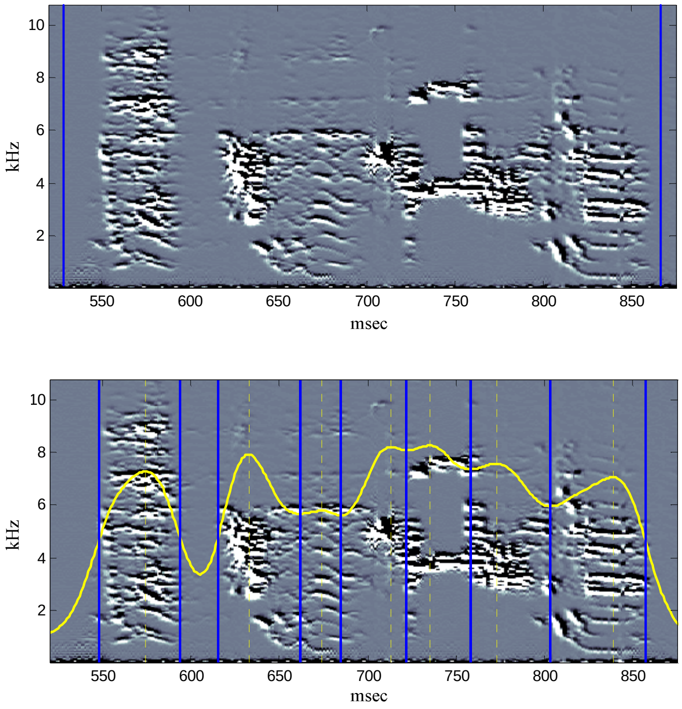


Figure 5.14 – Partitioning a single syllable into songemes. This figure displays the segment of birdsong in Figure 5.13 between 520 and 875 msecs. The single syllable shown on top has been automatically partitioned into seven songemes on the bottom, which correspond more closely to the changes in vocalization during this interval. This example supports our belief that the widespread syllabic approach to studying birdsong is a poor model for capturing its internal complexity.

these peaks, and the boundaries are shared between songemes when they are temporally adjacent. We examine the partitioning of an individual complex syllable into seven songemes in Figure 5.14. This example supports our belief that the widespread syllabic approach to studying birdsong is a poor model for capturing its internal complexity.

We have used this technique to automatically extract approximately 10,000 songemes from wav files recorded from two different zebra finches provided by (Tchernichovski 2005, 2006). Of these, we heuristically rejected any songeme of duration less than 10 msecs, which eliminated approximately 1000 of them. Many of these are due to non-verbal sounds, e.g., from a bird moving in its cage or other background noises, that are audible on the recordings.

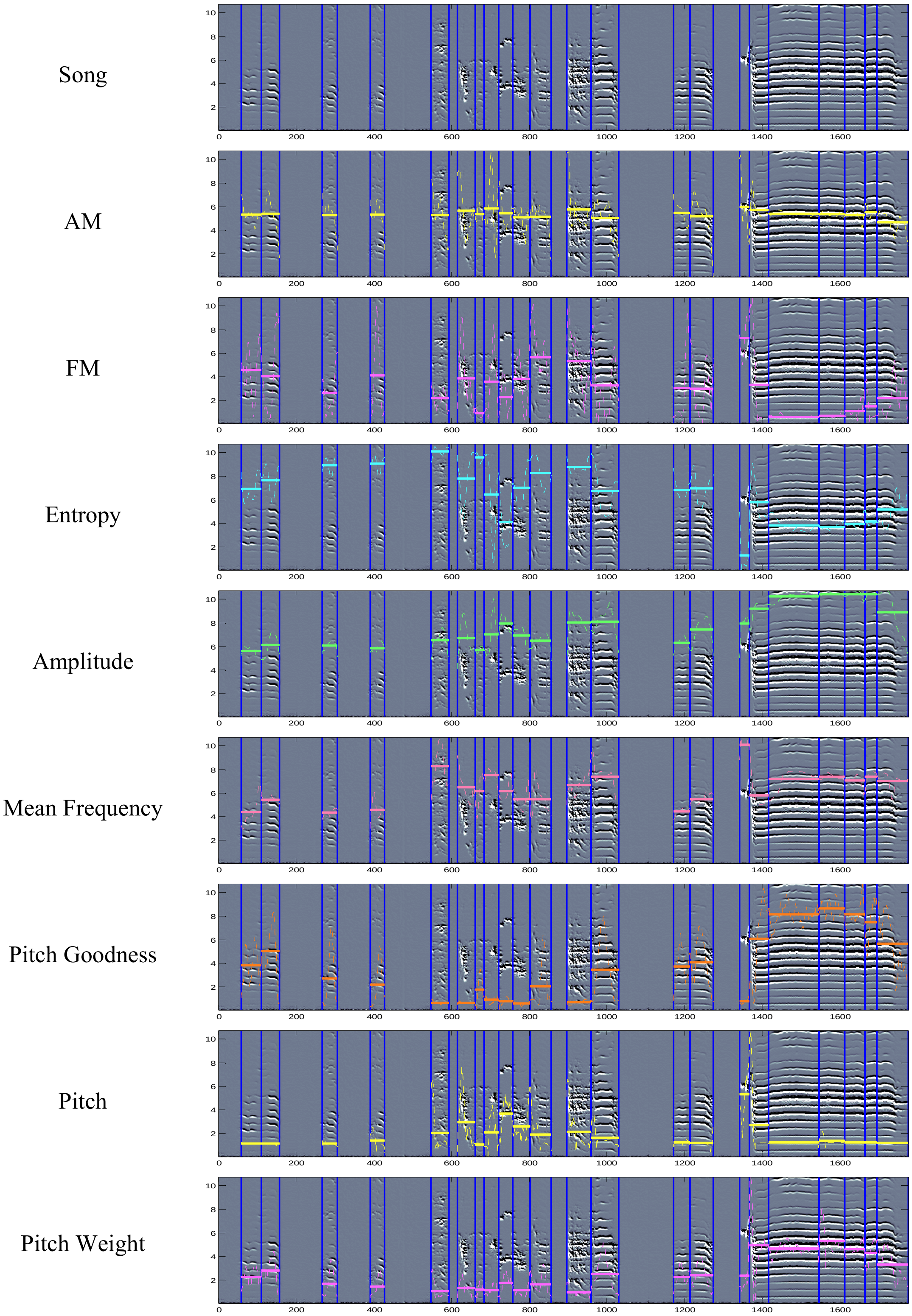


Figure 5.15 – Feature extraction for a zebra finch song partitioned into songemes. The solid line within each songeme shows the mean value for the corresponding feature within it. These values are cross-modally clustered to learn the structure of the birdsong. The dotted line within each songeme shows the actual feature data, which is smoothed with a low-pass Savitzky-Golay filter. The feature values have been normalized to fit within each plot. We note this is the same song as shown in Figure 5.13

We note several other efforts at separating complex animal sounds into primitive components. Kogan and Margoliash (1997) used dynamic time warping and hidden Markov models (HMMs) to derive simple units of birdsong for testing automated recognition of birds. Mellinger and Clark (1993) used HMMS for detecting and identifying bowhead whales, and Clemins and Johnson (2003) used HMMs for recognizing vocalizations in African elephants.

5.3.3 Birdsong Analysis

We extracted streams of acoustic features from our birdsong sound files using a customized version of the Sound Analysis For Matlab software (Saar 2005). The extracted features include: 1) amplitude modulation; 2) frequency modulation; 3) entropy; 4) amplitude; 5) mean frequency; 6) pitch goodness; 7) pitch; and 8) pitch weight as shown in Figure 5.15. We note the song in this figure is the one shown Figure 5.13.

For each songeme, we average the values of the features within it to obtain a compact acoustic description. These average values are shown by the solid horizontal lines within each songeme in Figure 5.15a. The dotted lines within each songeme display the actual feature values, which have been smoothed as described above. The average feature values for approximately 9,000 songemes, derived from our training sound files, were fed into as assembly of interconnected slices, to be discussed below.

We see two of the outputs of this clustering in Figure 5.16. Among the most interesting of our technical results, we can interpret the upper slice as demonstrating the system has learned there are three different types of vocalizations: 1) the blue region corresponds to noisy sounds perhaps generated by chaotic activity in the avian syringeal sound generator. This is similar to chaotic (e.g., fricative) speech in humans; 2) the green region corresponds to pure tones, such as whistles; and 3) the orange region corresponds to harmonic sounds, such as the in the distance call. We note the relative scarcity of pure tones in zebra finch song, as reflected by the sparsity of data in the green region.

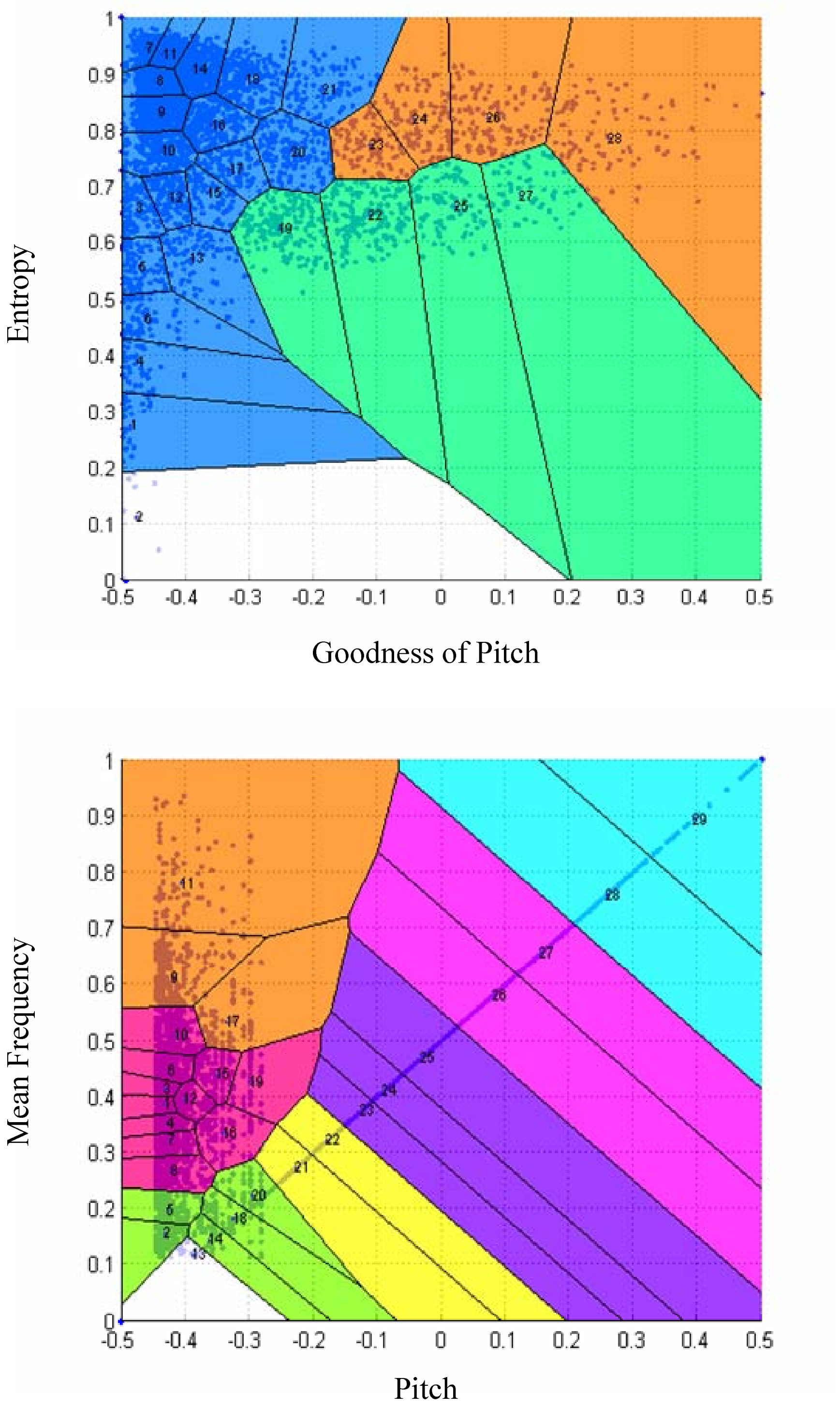


Figure 5.16 -- Cross-modally clustered zebra finch slices. We can interpret the upper slice as demonstrating the system has learned there are three different types of vocalizations: 1) the blue region corresponds to noisy sounds, perhaps generated by chaotic activity in the avian syringeal sound generator. This is similar to aspirated speech in humans; 2) the green region corresponds to pure tones, such as whistles; and 3) the orange region corresponds to harmonic sounds, such as in the distance call. The lower slice shows the system has learned the pitch structure for seven different component vocalizations.

The slice at the bottom of Figure 5.16 demonstrates that the system has acquired the pitch structure of its parent's birdsong, corresponding with seven different base "notes," out of which songemes are constructed by varying other acoustic features such as harmonic complexity and frequency modulation.

A view of all the slices used for birdsong learning is illustrated in Figure 5.17. The low-level features, such as those in Figure 5.15, supplied data to these slices, which manually selected and interconnected. Although these types of interconnections are likely phylogenetically determined in nature, a more sophisticated artificial system might

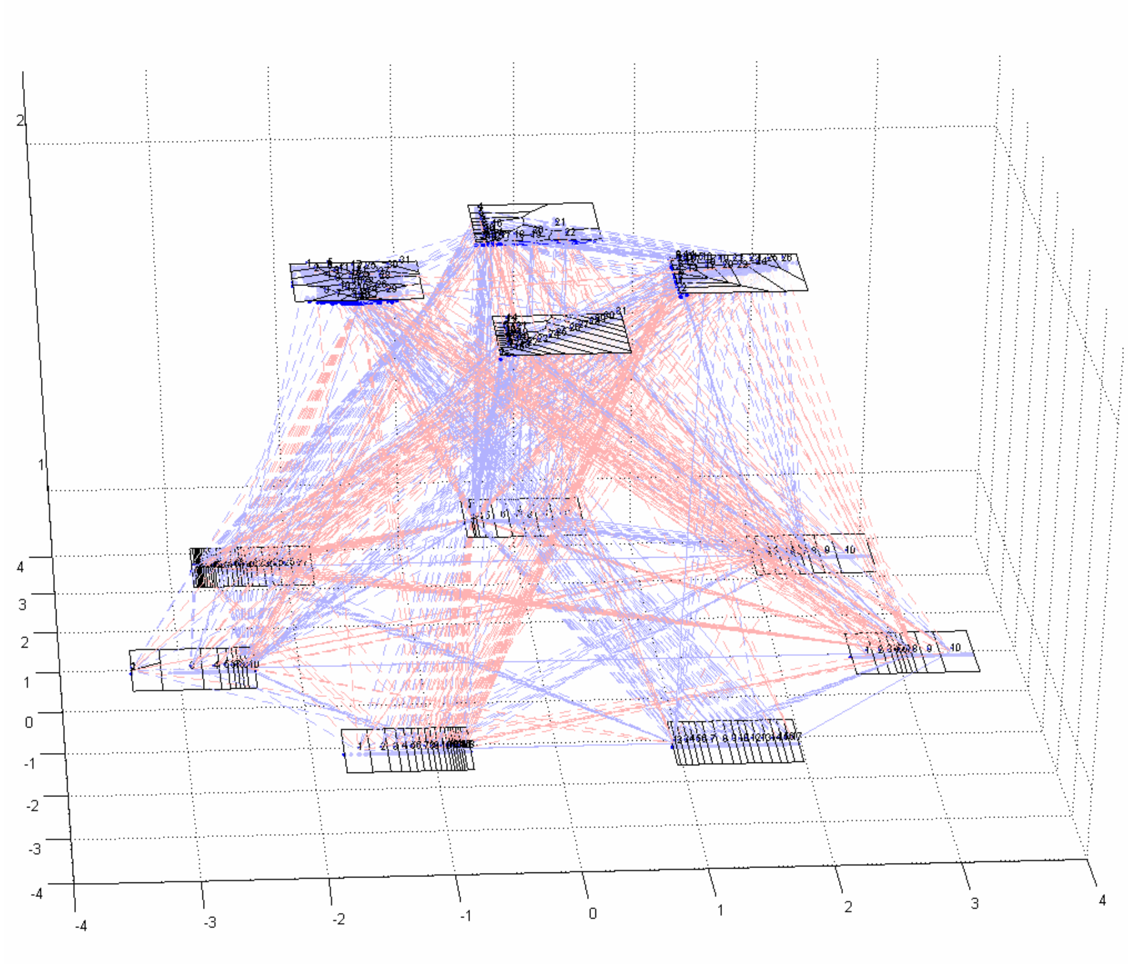


Figure 5.17 – Slices for birdsong learning. On the bottom, one dimensional slices feed songeme feature values into the two dimensional slices on the top. The colored lines represent learned Hebbian linkages. The slices are then cross-modally clustered to learn songeme categories. This perceptually grounds the system with respect to its "parent's" song. Detailed views of two of these slices are contained in Figure 5.16. See the text for additional details of this architecture.

employ techniques such as (Bartlett 2001) to automatically select interconnections between acoustic features based on independent component analysis. However, given our task here is to demonstrate acquisition of sensorimotor control using perceptual mechanisms, the precise details of feature selection and interconnection were not of great concern. Nonetheless, it should be acknowledged that a fair amount of manual effort went into architecting the birdsong learner illustrated in Figure 5.17, which then was cross-modally clustered based on the input song of its parent, as described in §5.2.1.

5.3.4 Articulatory Synthesis

To implement the motor component of this system, we created a naive articulatory synthesizer for generating birdsong, based on the additive synthesizer in the Common Lisp Music System (Taube 1989) and translated into Matlab by (Strum 2001). The motor parameters in our model correspond to: (1) syringeal excitation; (2) pitch; (3) power; and (4) temporal, frequency and amplitude envelopes corresponding to simple models of avian vocalization. In our implementation, chaotic syringeal excitation is realized by phase and amplitude perturbations of a vocalization's harmonic components. We note that this does not correspond with a biologically realistic syringeal mechanism, which would be complicated to model accurately. However, our goal here is not to model birdsong with perfect accuracy but rather to demonstrate self-supervised sensorimotor learning within our framework. Making the synthesizer sounds generally realistic was sufficient for our purposes, as we discuss below.

We refer to the nascent activity of the system as babbling. Some examples of increasingly complex babbling are shown in Figure 5.18. These demonstrate the system's acquisition of harmonic complexity in response to auditory feedback and comparison with its parent’s song. The initial babbling corresponds to uninformed, innate motor behavior as described above. As the system simultaneously listened to its own outputs while “watching” the internal generation of motor activity, it was able to determine which regions of its motor maps were responsible for providing harmonic complexity matching its parents through cross-modal clustering, as described in §5.2.3.

The increase in complexity is due to continued refinement of the system's motor maps. In other words, as it explored its motor capabilities more completely, it was able to classify codebook regions in its motor maps to learn which generated harmonically complex outputs. We manually implemented the strategy that our system select different aspects of its parent song to independently master, which reflect the actual strategies taken by zebra finches in the wild (Liu et al. 2004), who sequentially focus on different features of their song, rather than try to master the parent's song in its entirely. A

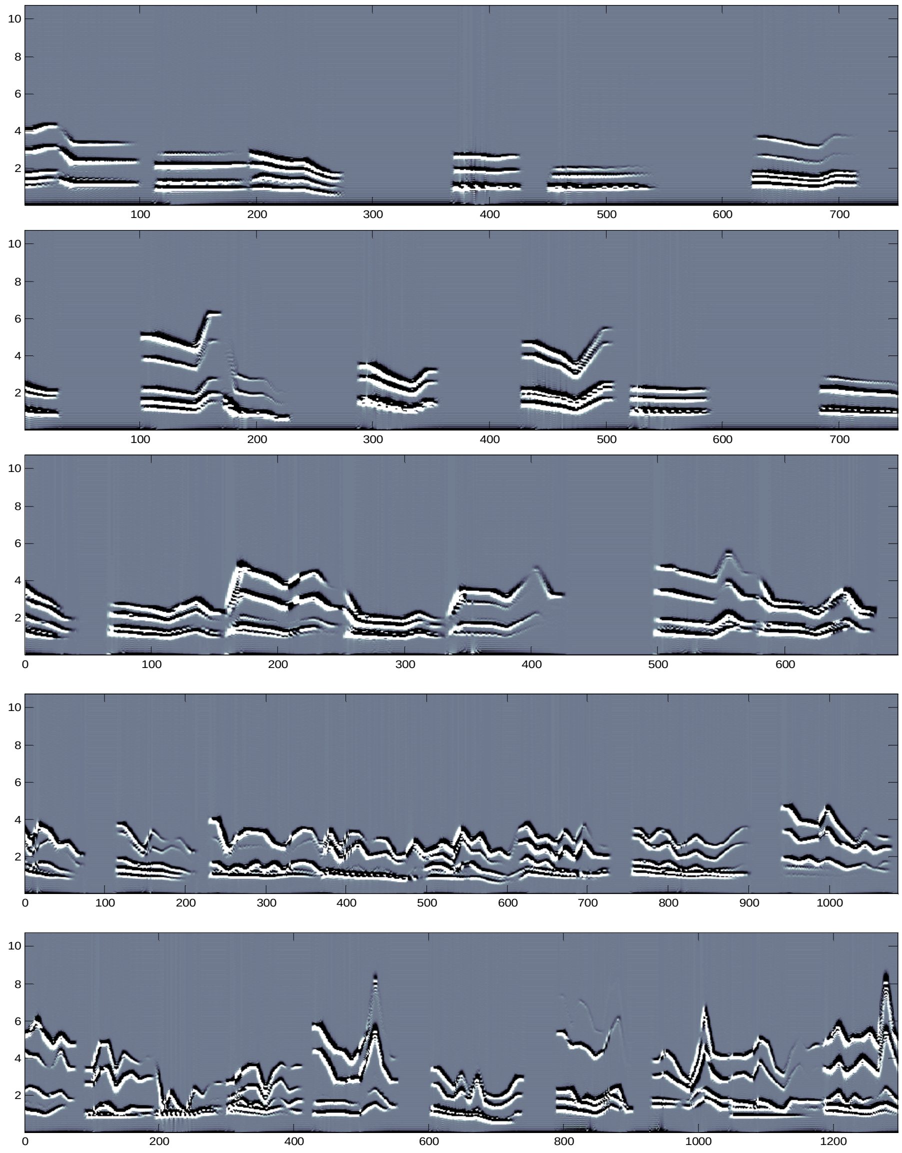


Figure 5.18 – The temporal evolution of bird babbling in our system. This figure illustrates the acquisition of harmonic complexity due to auditory feedback. See the text for explanatory commentary.

presumable benefit of this approach is that it reduces search space complexity enormously by isolating and focusing on individual acoustic features, thereby permits gradual song acquisition and differential acquisition in case other male siblings are also in the midst of their own learning. Thus, this strategy also prevents co-located siblings from confusing one another. For example, two brothers raised together will select (presumably) non-conflicting developmental paths, where one may focus on the temporal aspects of its song while the other focuses on pitch, even though they are both essentially learning the same song. Our system was designed to focus on individual songemes, in essence acquiring the "notes" of its parent's song. As this is a much lower level treatment than is typically given in the oscine developmental literature, which focuses on song syllables, it is difficult to evaluate its developmental realism. Nonetheless, for an artificial system, it seems quite adequate.

Our strategy was to learn mimicry of each acoustic feature in terms of the articulator's model, which formed the structure of our motor slices. For each acquired songeme, we selected the combination of articulations which maximized our approximation of it. We note that given the constraints of our articulator, certain acoustic features of the parent

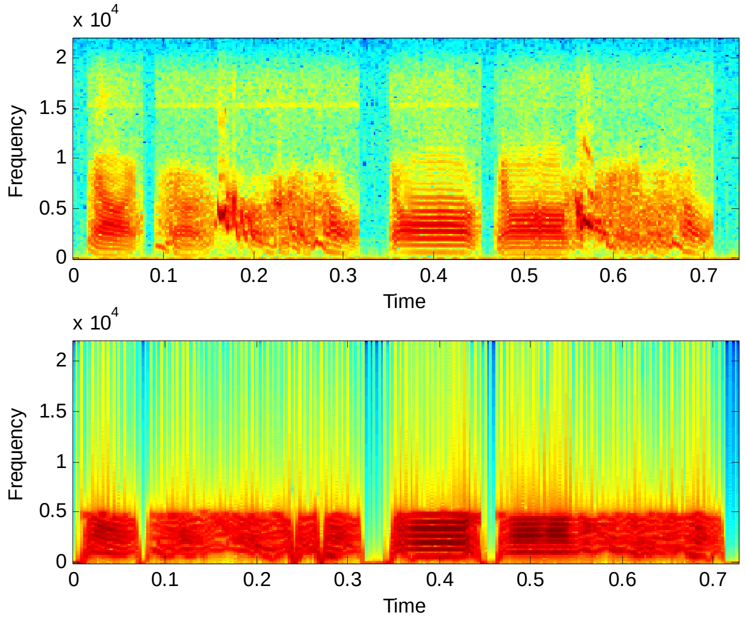


Figure 5.19 – Birdsong mimicry. On the top is a sample of the zebra finch song used as the "parent" for our system. On the bottom is the system's learned imitation, where the acquired songemes have been fit to the template of the parent's song and smoothed.

birdsong were not possible to reproduce well. In particular, the harmonic complexity associated with chaotic syringeal excitation, a common phenomenon in zebra finch song, could only roughly be approximated by our articulatory model. However, the human ear seems far more sensitive to pitch accuracy, which we were able to capture quite well, and tends to be less sensitive to reduced acoustic entropy.

5.3.5 Results

This experiment was in essence empirical. Our goal was to demonstrate that our framework could perform self-supervised sensorimotor learning, using the zebra finch as a testbed. Although the acquired song sounds recognizably like the parent bird's to human ears, does it sounds like a bird to another bird? In other words, it is unclear how to evaluate this mimicry. We note that human auditory range overlaps well with that of the zebra finch, and as mentioned above, the FOXP2 gene thought to be responsible for vocal production is remarkably similar between the two species. Nonetheless, one should assume that auditory feature extraction and subsequent processing in the zebra finch is uniquely tuned to its auditory requirements, which are surely quite different than our own.

Perhaps the best way to evaluate this work would be to use it to train a fledgling and see if it acquires song. In other words, use our system as a parent.[1] This is a fascinating possibility which we are currently investigating. However, it is important to keep in mind that the goal of this chapter is to present an architecture for sensorimotor learning in *artificial* systems; songbirds are a well-studied model for this type of acquisition, and as such, they are helpful guideposts for examining this problem. Nonetheless, our goal is proposing architecture for more sophisticated computational systems rather than precisely imitating the song of a particular bird.

[1] It has been suggested by Chris Atkeson that we attempt mating our system with a female zebra finch, but that would presumably void our computer's warrantee.

5.4 Summary

This chapter has demonstrated that the cross-modal clustering framework presented in Chapter 3 can be recursively reapplied to acquiring self-supervised sensorimotor control. We have developed architecture for this type of learning using an internal *Cartesian theater* to correlate the generation of motor activity with its perceived effects through cross-modal clustering. This is possible because slices neither know nor care whether they represent sensory or motor data. It is thus straightforward to seamlessly move between the two during clustering.

We demonstrated this approach with a system that learns birdsong, following the developmental stages of a fledgling zebra finch. This works suggests a number of other possible applications for learning motor control through observation, among the most common forms of learning in the animal world.

The benefits of self-supervised learning for artificial sensorimotor systems are enormous because engineered approaches tend to be ad hoc and error prone; additionally, in sensorimotor learning we generally have no adequate models to specify the desired input/output behaviors for our systems. As we mentioned in Chapter 1, the notion of *programming by example* is nowhere truer than in the developmental mimicry widespread in animal kingdom, and this chapter is a significant step in that direction for providing that capability to artificial sensorimotor systems.

Chapter 6

Biological and Perceptual Connections

The chapter connects the computational framework introduced in this thesis to a modern understanding of perception in biological systems. In doing so, we motivate the approach taken here and simultaneously suggest how this work may reciprocally contribute towards a better computational understanding of biological perception. We begin by examining the interaction between sensory systems during the course of ordinary perception.

6.1 Sensory Background

Who would question that our senses are distinct? We see, we feel, we hear, we smell, and we taste, and these are qualitatively such different experiences that there is no room for confusion among them. Even those affected with the peculiar syndrome *synesthesia*, in which real perceptions in one sense are accompanied by illusory ones in another, never lose awareness of the distinctiveness of the senses involved. Consider the woman described in (Cytowic 2002), for whom a particular taste always induced the sensation of

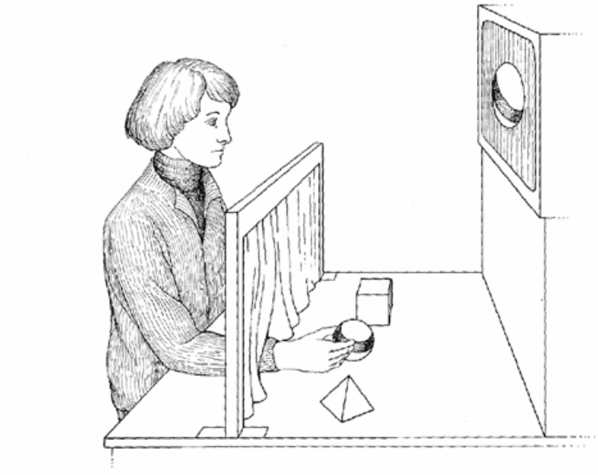


Figure 6.1 - Cross-modal matching: a subject is asked to use haptic (e.g., tactile) cues to select an object matching a visual stimulus. From (Stein and Meredith 1993).

a specific, corresponding geometric object in her left hand. A strange occurrence indeed, but nonetheless, the tasting and touching – however illusory – were never confused; they were never merged into a sensation the rest of us could not comprehend, as would be the case, for example, had the subject said something *tasted* octagonal. Even among those affected by synesthesia, sensory perceptions remain extremely well defined.

Given that our senses appear so unitary, how then does the brain coordinate and combine information from different sensory modalities? This has become known as the *binding problem* (see Wolfe and Cave 2000 for a review), and the classical assumption has been to assume that the sensory streams are abstracted, merged, and integrated in the cortex, at the highest levels of brain functioning. This hypothesis assumed a cognitive developmental process in which children slowly developed high-level mappings between innately distinct modalities through their interactions with the world.

We may examine this assumption in context of the most frequently studied of intersensory phenomena, *cross-modal matching*, which is determining that perceptions in two different sensory modalities could have the same source. Figure 6.1 shows a standard experimental matching task, in which a subject is asked to use tactile cues to select an object matching a visual stimulus. The actual mechanisms that make cross-modal matching possible are unknown, but clearly, sufficient correspondences between the modalities must exist to enable making this type of equivalency judgment. Whatever form these correspondences take – whether through topographic maps, amodal perceptual interlinguas, or other representations – is according to Piaget (1954), among many others, developed slowly through experience. Only when the involved senses have developed to the point of descriptional (i.e., representational) parity, can these interrelations develop and cross-modal matching thereby take place.

This position directly traces back to Helmholtz (1884) and even earlier, to Berkeley (1709) and Locke (1690), who believed that neonatal senses are congenitally separate and interrelated only through experience. According to this viewpoint, the interrelation *does not diminish the distinctiveness of the senses themselves*, it merely accounts for correspondences among them based on perceived co-occurrences. This was seemingly a very intuitive foundational assumption for studying perception. Unlike so much of the rest of human cognition, we *can* be introspective about how we perceive the world. Are not our sensory systems innately designed to bring themselves to our attention?

6.2 Intersensory Perception

Unfortunately for this introspective approach to perception, an overwhelming body of evidence has been gathered in the past half century demonstrating that perceptions themselves emerge as the *integrated* products of a surprising diversity of components. (For surveys, see Stein and Meredith 1993, Lewkowicz and Lickliter 1994, Rock 1997, Shimojo and Shams 2001, Calvert et al. 2004, and Spence and Driver 2004.) Our auditory, olfactory, proprioceptive, somatosensory, vestibular, and visual systems influence one another in complex interactive processes *rarely subject to conscious introspection.* In fact, the results can often be completely counterintuitive. Perception is such a fundamental component of our experience that we seldom give its mechanisms any direct attention. It is the perceptions produced by these mechanisms that draw our attention, not the mechanisms themselves, and we are ill equipped to examine the mechanisms directly without the aid of clever, sometimes even serendipitous, experimentation. In Gibson's (1950) view, perceivers are aware of the world, not their own perceptions.[2]

For example, let us consider the well-known work of McGurk and MacDonald (1976), who studied how infants perceive speech during different periods of development. In preparing an experiment to determine how infants reconcile conflicting information in different sensory modalities, they had a lab technician dub the audio syllable /ba/ onto a video of someone saying the syllable /ga/. Much to their surprise, upon viewing the dubbed video, they repeatedly and distinctly heard the syllable /da/ (alternatively, some hear /tha/), corresponding neither to the actual audio nor video sensory input. Initial

[2] One may contrast this with Russell's (1913) criticism of "materialistic monism," in which he argues that abstract, self-aware mental models are an essential component of perception. This distinction, in the guise of the Physical Symbol System Hypothesis (Miller, Galanter, and Pribram 1960, Newell and Simon 1976), has been the subject of intense scrutiny within the artificial intelligence community (e.g., Johnson-Laird 1983, Brooks 1990a).

assumptions that this was due to an error on the part of the technician were easily discounted simply by shutting their eyes while watching the video; immediately, the sound changed to a /ba/. This surprising fused perception, subsequently verified in numerous redesigned experiments and now known as the *McGurk effect*, is robust and persists even when subjects are aware of it.

The McGurk effect is among the most convincing demonstration of the intersensory nature of face-to-face spoken language and the undeniable ability of one modality to radically change perception in another. It has been one of many components leading to the reexamination of the introspective approach to perception. Although it may appear reasonable to relegate intersensory processing to the cortex for the *reasoning* (as opposed to *perceptual*) processes involved in the cross modal matching experiments described above, it becomes far more implausible in cases where different senses impinge upon each other in ways that locally change the perceptions in the sensory apparatus themselves.

One might object that the McGurk effect is pathological – it describes a perceptual phenomenon outside of ordinary experience. Only within controlled, laboratory conditions do we expect to have such grossly conflicting sensory inputs; obviously, were these signals to co-occur naturally in the real world, we would not call them conflicting. We can refute this objection both because the real world *is* filled with ambiguous, sometimes directly conflicting, perceptual events, and because the McGurk effect is by no means the only example of its kind. There is a large and growing body of evidence that the type of direct perceptual influence illustrated by the McGurk effect is commonplace in much of ordinary human and more generally animal perception, and it

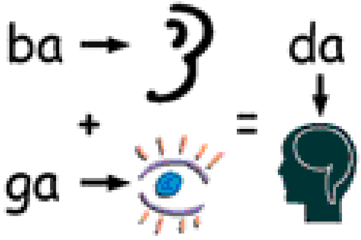


Figure 6.2 - The McGurk Effect**.** Disparate auditory and visual inputs can create perceptions corresponding to neither. Picture is from (Haskins 2005).

strongly makes the case that our perceptual streams are far more interwoven than conscious experience tends to make us aware. In this, the McGurk effect is unusual in that the conflicting audio and visual inputs create perceptions corresponding to neither actual input. More typically, sensory systems influence (e.g., enhance, degrade, change) perceptions in one another and at times, substitute for each other as well, by transferring perceptual information across modalities (Calvert et al. 2004).

An example of one sense enhancing another occurs all the time in noisy environments. The sight of someone's moving lips in an environment with significant background noise makes it easier to understand what the speaker is saying; *visual* cues – e.g., the sight of lips – can alter the signal-to-noise ratio of an *auditory* stimulus by 15-20 decibels (Sumby and Pollack 1954). Tied in with audio source separation, this phenomenon is commonly known as the *cocktail party effect* (Cherry 1953). We see, therefore, that a decrease in auditory acuity can be offset by increased reliance on visual input. In fact, it has long been known that watching the movement of a speaker's lips helps people understand what is being said: Juan Pablo Bonet wrote in his 1620 classic "Simplification of Sounds and the Act of Teaching the Deaf to Speak," (cited in Bender 1981, p41):

> "For the deaf to understand what is said to them by the motions of the lips there is no teaching necessary; indeed to attempt to teach them it would be a very imperfect thing ...to enable the deaf-mute to understand by the lips alone, as it is well known many of them have done, cannot be performed by teaching, but only by great attention on their part…"

Although the neural substrate behind this interaction is unknown, it has been determined that just the *sight* of moving lips – without any audio component – modifies activity in the *auditory* cortex (Sams et al. 1991, Calvert et al. 1997). Furthermore, psycholinguistic evidence has long supported the belief that lip-read and heard speech share a degree of common processing, notwithstanding the obvious differences in their sensory channels (Dodd et al. 1984).

Examples of senses substituting for one another are commonplace, as when auditory and tactile cues replace visual ones in the dark; this is familiar to anyone who has walked through a dark room with outstretched, waving arms and hyper-attentive ears. A far more interesting example is seen in a phenomenon known as the "facial vision" of the blind. In locating objects, blind people often have the impression of a slight touch on their forehead, cheeks, and sometimes chest, as though being touched by a fine veil or cobweb (James 1890, Kohler 1967). Consider this quote contained in James (1890, p.204), from a blind author of the time:

> "Whether within a house or in the open air, whether walking or standing still, I can tell, although quite blind, when I am opposite an object, and can perceive whether it be tall or short, slender or bulky. I can also detect whether it be a solitary object or a continuous fence; whether it be a close fence or composed of open rails, and often whether it be a wooden fence, a brick or stone wall, or a quick-set hedge. …The sense of hearing has nothing to do with it, as when snow lies thickly on the ground objects are more distinct, although the footfall cannot be heard. I seem to perceive objects through the skin on my face, and to have the impressions immediately transmitted to the brain."

The explanation for this extraordinary perceptory capability had long been a subject of fanciful debate. James demonstrated, by stopping up the ears of blind subjects with putty, that audition was behind this sense, which is now known to be caused by intensity, direction, and frequency shifts of reflected sounds (Cotzin and Dallenbach 1950, Hoshino and Kuroiwa 2001). The auditory input is so successfully represented haptically in the case of facial vision that *the perceiver himself cannot identify the source of his perceptions.*

There is no doubt that we constantly make use of intersensory cues during perception and in directing our attention, and when deprived of these cues, through artificially contrived or naturally occurring circumstances, can display marked degradation in what seem to us conceptually and functionally isolated sensory systems. The breadth of these interactions and the range of influences they demonstrate have been so surprising that they have called for radically new approaches to understanding how our individual perceptual systems work, how the brain merges their perceptions, and why these two questions are inseparable.

6.3 The Relevance

Why should this reexamination of sensory independence concern us? We believe the answer stems from the fact that the early assumption of perceptual isolation underlies nearly all modern interactive systems – it remains the primary architectural metaphor for multimodal integration. The unfortunate consequence of this has been making integration a *post-perceptual* process, which assembles and integrates sensory input after the fact, in a mechanism separate from perception itself. Modern psychophysical, neuroanatomical, evolutionary, and phenomenological evidence paints a very different picture of how animals, including humans, merge their senses; the notion that the senses are processed in isolation is highly implausible. Much of this evidence also undermines classical symbol-processing models of perception (e.g., Fodor and Pylyshyn 1998), where cognitivist assumptions can obscure rather than illuminate the subtle intersensory perceptual mechanisms we are interested in here.

6.3.1 Perception in Pipelines

Computational approaches to perception have traditionally been bottom-up, feeding raw perceptual inputs into abstraction pipelines, as show in Figure 6.3. In these frameworks, a pipeline is constructed through functional composition of increasingly high-level feature detectors. This approach is apparent in some of the earliest work in computational object recognition (Horn 1970, Binford 1971, Agin 1972). It is also explicitly reflected in the perceptual theories described by Marr (1976, 1982), Minsky (1987), and Ullman (1998), as well as in the subsumption architecture of Brooks (1986) for sensorimotor control.[3]

Early approaches to artificial perception focused exclusively on modeling aspects of individual modalities in isolation, although the potential for more complex multimodal interactions drew the imaginations of early researchers (Turing 1950, Selfridge 1955). Among the first efforts to create a user interface that combined two independent

[3] Note that both Ullman and Brooks make extensive use of top-down feedback to govern their bottom-up processing.

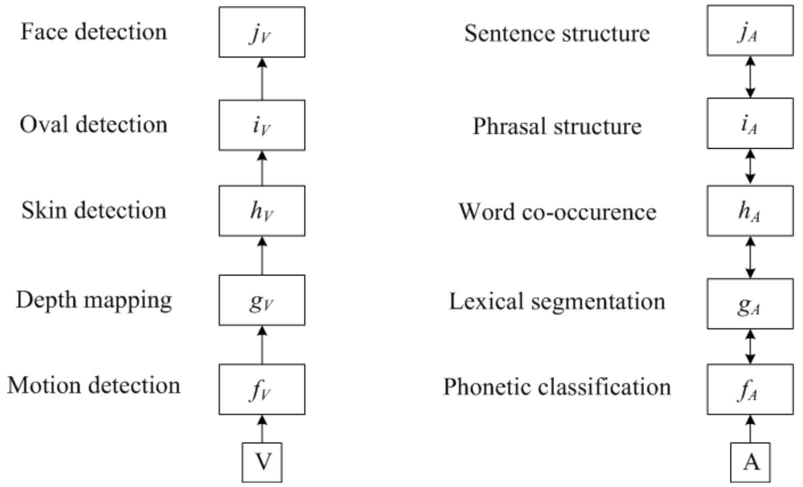


Figure 6.3 – Unimodal processing pathways. Individual modalities are processed in specialized pipelines. The visual pathway on the left uses 3-dimensional depth and color maps to find candidate regions for locating faces. (Modeled after Darrell et al. 1999.) The auditory pathway on the right performs speech recognition. (Modeled after Jelinek 1997.) Notice in this example that higher-level syntactic constraints can feed back into the lower-level morphological and phonetic analyses.

perceptual channels was that of Bolt (1980). His "Put-that-there" system (Figure 6.4) enabled users to interact with projected maps by speaking and pointing simultaneously. Bolt's system resolved spoken *deictic* references involving identity ("*that*") and location ("*there*") by visually observing the pointing gestures that accompanied them. Deictic gesture resolution has subsequently become a very popular application for multimodal user-interface development (e.g., Oviatt et al. 2003, Kumar et al. 2004).

Another significant application combining separate modalities is lip-reading, which has become perhaps the most studied problem in the computational multimodal literature (e.g., Mase and Pentland 1990, Massaro et al. 1993, Brooke and Scott 1994, Bregler and Konig 1994, Benoît 1995, Hennecke et al. 1996, Bangalore and Johnston 2000, Huang et al. 2003, Potamianos et al. 2004). This is due both to the historic prominence of automatic speech recognition in computational perception and more importantly, to the inherent difficulty of recognizing unconstrained speech. We note that the robotics community was perhaps the first to realize that combining multiple sensory channels could increase overall perceptual reliability (Nilson 1984, Flynn 1985, Brooks 1986). However, in these systems, sensors would substitute for one another depending upon context; in other words, one sensor would be used at a time and a robot would dynamically switch between them. Speechreading systems were the first multimodal

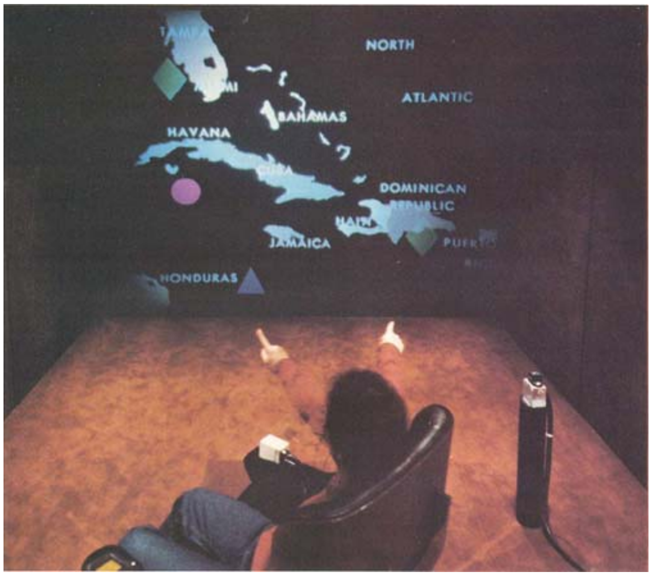


Figure 6.4 – An early multimodal user-interface. The "Put-that-there" system combined speech processing and visual gesture recognition to resolve spoken deictic references to a projected map. From (Bolt 1980).

approaches that sought to increase perceptual accuracy by simultaneously combining information from complementary sensory channels. This is in contrast with the more typical goal of *multimodal user-interface design*, which is to make human-computer interaction more natural for people by providing additional input modalities such as pointing or context-awareness (Dey et al. 2001).

In recent years, a wide range of multimodal research areas has emerged, many of which were inspired by Weiser's (1991, 1994) notion of *ubiquitous computing*. These include intelligent environments (Coen 1999, Vanderheiden et al. 2005), wearable computing (Starner 1999), physiological monitoring (Intille et al. 2003, Sung and Pentland 2005), ambient intelligence (Remagnino et al. 2005), multimodal design (Adler and Davis 2004), and affective computing (Picard 1997). Additionally, we note that roboticists have a long tradition of combining multimodal perception with sensorimotor control (see the references above, along with Brooks et al. 1998, Thrun et al. 2005).

6.3.2 Classical Architectures

Surprisingly, even though all of the above applications address a diverse assortment of computational problems, their implementations have similar – sometimes almost identical – architectures. Namely, they share a common framework where sensory inputs are individually processed in isolated, specialized pathways (Figure 1.2). Each perceptual pipeline then outputs an abstract description of what it sees, hears, senses, etc. This description captures detail sufficient for higher-level manipulation of perceptions, while omitting the actual signal data and intermediate analytic representations. Typically, the perceptual subsystems are independently developed and trained on unimodal data; in other words, each system is designed to work in isolation. They are then interconnected through some fusive mechanism that combines temporally proximal, abstract unimodal inputs into some integrated event model.

The integration itself may be effected in many different ways. These include: multilayered neural networks (Waibel et al. 1995); hidden Markov models (Stork and Hennecke 1996); coupled hidden Markov models (Nefian et al. 2002); dynamic Bayesian

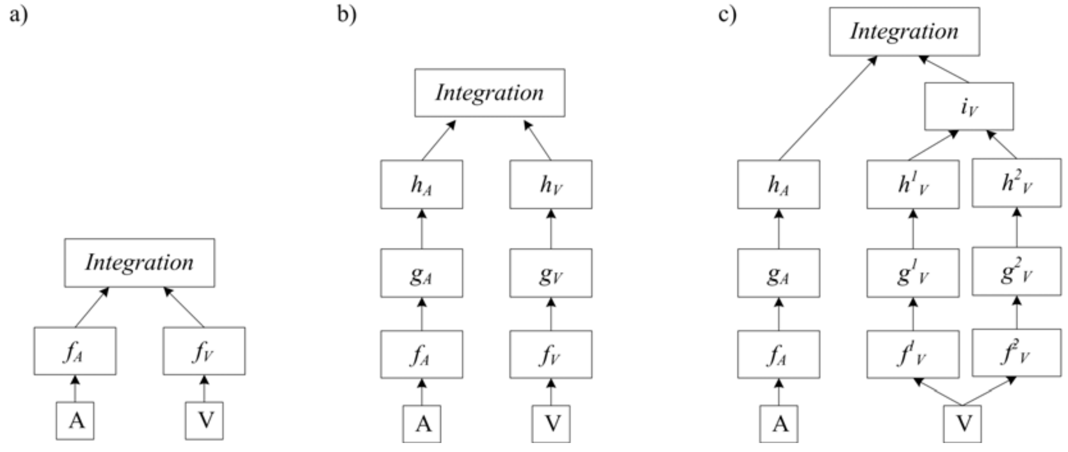


Figure 6.5 – Classical post-perceptual integration in multimodal systems. Here, auditory (A) and visual (V) inputs pass through specialized unimodal processing pathways and are combined via an integration mechanism, which creates multimodal perceptions by extracting and reconciling data from the individual channels. Integration can happen earlier (a) or later (b). Hybrid architectures are also common. In (c), multiple pathways process the visual input and are pre-integrated before the final integration step; for example, the output of this preintegration step could be spatial localization derived solely through visual input. This diagram is modeled after (Stork and Hennecke 1996).

networks (Wang et al. 2005); unification logics (Cohen et al. 1997); Harmony theory (Smolensky 1986); posterior probabilities (Wu et al. 1999); fuzzy logic (Massaro 1998); maximally informative joint subspaces (Fisher et al. 2001); recurrent mixture models (Hsu and Ray 1999); subsumption architectures (Brooks 1986); agent hierarchies (Minsky 1986); partially observable Markov decision processes (Lopez et al 2003); layered topographic maps (Coen 2001); and various ad hoc techniques, which tend to be the most popular (e.g., Bobick et al. 1998). The integrated events themselves have specialized representations, which may be multimodal, in that they explicitly represent information gathered from separate modalities, or they may be amodal, in that they represent features not tied to any specific perceptual channel. For example, *intensity* is an amodal feature, because it can be measured independently of any sensory system, whereas *deictic references* are inherently multimodal, because they consist of co-occurrences of two distinct modal cues, namely, of speech and gesture. The output of the integration process, whether amodal or multimodal, is then fed into some higher-level interpretative mechanism – the architectural equivalent of a cortex.

6.3.3 What's Missing Here?

This post-perceptual approach to integration is commonplace in engineered systems. It suffers, however, from a number of serious flaws:

1) As we saw earlier in this chapter, the evidence is strongly against animals perceiving this way. Perceptual modalities interact constantly during ordinary perception, and even unimodal perception has strong multimodal components. That is not to say that all perception *must* be multimodal. Nonetheless, symbiotic interactions between sensory systems go a long way toward explaining why perception is so robust in biological systems, in marked contrast with their engineered counterparts. Because the individual components of multimodal systems in these architectures tend to be independently designed, *they are both representationally and algorithmically incompatible* with one another. Therefore,

it is often extraordinarily difficult to enable information sharing among them after the fact. We return to this point below.

2) Integration in these models happens too late. Integration occurs *after* each system has already "decided" what it has perceived, when it is too late for intersensory influence to affect the individual concurrent perceptions. This is due to the information loss from both vector quantization and the explicit abstraction fundamental to the pipeline design.

3) Biological sensory systems are *perceptually impoverished* – they are incapable of simultaneously detecting all sensible events to which they are exposed. Instead, they exhibit attentive focusing, which narrows their windows of sensory awareness and thereby increases sensory discrimination. This prevents seemingly chaotic composite signals from impinging upon the sensory cortex, and this mechanism is thought to be responsible for the coherence in our generation of perceptual *scenes.* However, it simultaneously requires that something guides this focus. For example, saccadic behavior in the eye enables perception of a large image, overcoming the retina's narrow foveal limits. However, without highly informed sampling, key details may be missed simply because they are never observed. Many of our perceptual channels have similar attentive mechanisms (Naeaetaenen et al. 2001), of which we are generally not consciously aware. Importantly for us, the cues which guide attentive focusing are frequently generated cross-modally (Driver et al. 2005). The role of attentive focusing in eliminating extraneous sensory detail appears to be a basic component of robust animal perception. Representational and algorithmic incompatibilities make this type of cross-modal influence implausible in many engineered systems.

4) The independence between sensory components in classical architectures precludes mutual bootstrapping, such as with the cross-modal clustering of Chapter 3. Since these systems tend to be developed separately and connected only at their outputs, there is no possibility of perceptually grounding them based on naturally occurring temporal correspondences, to which they may remain

totally oblivious. Furthermore, the derivation of common perceptual categories provided by cross-modal clustering would help alleviate the representational incompatibility raised above in (1). Preventing this common grounding simply exacerbates the problem.

We note that early integration (Figure 6.5a) would seem to bypass perceptual isolation by combining sensory streams at a low-level feature (or sometimes, even signal) level. It does so, however, at the cost of losing the domain structuring, high-level feature extraction, and dimensional reduction explicitly provided by the abstraction-based, pipeline architectures that are ubiquitous in encephalized species. Early integration has shown much promise with problems amenable to information theoretic and statistical approaches where there is little available domain knowledge. However, the majority of multisensory problems of interest in this thesis do not appear to fit into this category. When combining sensory information, *structured domain knowledge*, as described by Wertheimer (1923), Chomsky and Halle (1968), and Marr (1982) along with many others – and as implied by the Ugly Duckling Theorem (Watanabe 1985) – is essential for reducing the size of perceptual search spaces. It is doubtful that a wide range of interesting sensory phenomena can be detected without it, particularly given the roles of context and expectation in perceptual interpretation, even in relatively simple tasks (Bruner and Postman 1949), and the computational intractability of directly describing complex perceptual phenomena in terms of raw sensory input or low level features. Early integration generally also precludes the possibility of top-down processing models, such as with the visual *sequence seeking* approach in (Ullman 1996) or in various approaches to auditory scene analysis (e.g., Slaney 1995).

The approach taken in this thesis is not motivated by the idea that computational systems should use biologically-inspired mechanisms simply for the sake of doing so. Rather, there are two justifications for the path taken here: (1) the perceptual phenomena we are interested in computationally understanding are complex amalgams of mutually interacting sensory input streams – they are not end-state combinations of unimodal abstractions or features. Therefore, they cannot be accurately represented or described by mechanisms that make this assumption; and (2) biological systems use low-level sensory integration to handle a vast array of perceptual ambiguity and "errors." In designing robust artificial perceptual systems, it would be a poor design to ignore Nature's schema for interpretative stabilization. This is all the more so given the relatively fragility and high degree of error in our best computational methods for processing unimodal data streams. We believe that the current approach to building multimodal interfaces is an artifact of how people like to build symbol processing systems and not at all well-suited to dealing with the cross-modal interdependencies of perceptual understanding. Perception does not seem to be entirely amenable to the clear-cut abstraction barriers that computer scientists find so valuable for solving other problems, and approaching it as if it were has lead to the fragility of current multimodal systems.

Modern multimodal systems tend towards being inflexible and unpredictable and require structured, typically scripted, interactions. Each of these interactions is subject to what may be deemed the *weakest link* effect, in which every modality must receive input *just so*, i.e., exactly as expected, in order for the overall system to function. The addition of new modalities tends to weaken, rather than strengthen interactive systems, as additional inputs simply offer new opportunities for interpretive errors and combinatorially increase the complexity of fusing perceptions. From a biological perspective, *this is exactly the opposite of what one should expect*. Additional modalities should strengthen perceptual systems by capitalizing on the inherent multimodal nature of events in the real world and the mutual reinforcement between interconnected sensory channels. Modern multimodal systems suffer an unfortunate fate predicted by von Neumann (1956), where adding additional components reduces a system's inherent stability.

Research on perceptual computing has focused almost entirely on unimodal perception: the isolated analysis of auditory, linguistic, visual, haptic, and to a lesser degree biometric data. It seems to put the proverbial cart before the horse to ponder how information from different modalities can be merged while the perceptory mechanisms in the sensory channels are themselves largely unknown. Is it not paradoxical to suggest we should or even could study integration without thoroughly understanding the individual systems to be integrated? Nonetheless, that is the course taken here. We argue that while trying to understand the processing performed within individual sensory channels, we must simultaneously ask how their intermediary results and final products are merged into an integrated perceptual system. We believe that because perceptual systems within a given species coevolved to interoperate, *compatibility pressures existed on their choices of internal representations and processing mechanisms*. In other words, to explain the types of intersensory influences that have been discovered experimentally, disparate perceptual mechanisms must have some degree of overall representational and algorithmic compatibility that makes this influence possible.

Our approach is gestalt, not only from a Gibsonian (1986) perspective, but also because there are few, if any, known examples of complex unimodal sensory systems evolving in isolation. Even the relatively simple perceptual mechanisms in paramecium (Stein and Meredith 1994, Chapter 2) and sponges (Mackie and Singla 1983) have substantial cross-sensory interactions. It seems that perceptual interoperation is a prerequisite for the development of complex perceptual systems. Thus, rather than study any single perceptual system *in depth* – the traditional approach – we prefer to study them *in breadth*, by elucidating and analyzing interactions between different sensory systems. We believe these interactions make possible the co-evolution that leads to complex perceptual mechanisms, without which, they would not otherwise arise. This approach is similar in spirit to the work of (Atkeson et al. 2000, Brooks et al. 1998, Cheng and Kuniyoshi 2000, Ferrell 1996, and Sandini et al. 1997). Although they are primarily concerned with motor coordination, there is a common biological inspiration and long-term goal to use knowledge of human and animal neurophysiology to design more sophisticated artificial systems.

6.4 Our Direct Motivation

The examples for our discussion of multimodal interactions will be drawn from a previous research project known as the Intelligent Room, as described in (Coen 1998, 1999). The Intelligent Room provided a framework for examining fundamental issues in multimodal integration – most importantly for this thesis: *how can independently* *designed perceptual systems be made to work together*? This question proved surprisingly difficult to answer. To see why, we now examine two simple multimodal interactions that were implemented in the room. Much of this section previously appeared in (Coen 2001).

The Intelligent Room had multiple perceptual user interfaces to connect with some of the ordinary human-level events going on within it. The goal of the room was to support people engaged in everyday, traditionally non-computational activity in both work and leisure contexts. The room contained nine video cameras, three of which it could actively steer, several microphones, and a large number of computer-interfaced devices. Its computational infrastructure (Coen et al. 1999), consisting of over 120 software agents running on a network of ten workstations, was housed in an adjacent laboratory to enhance the room's appeal as a *naturally* interactive space.

Before exploring how novel intersensory influences might have been incorporated into a system such as the Intelligent Room, we first examine a traditional and explicit multimodal interaction, with the resolution of a deictic reference, e.g., use of a word such as *this* in referring to a member of some class of objects. Suppose, for example, someone inside the room said, "Computer, dim *this* lamp." The room used its ability to track its occupants, in conjunction with a map of its own layout, to dim the lamp closest to the speaker when the verbal command was uttered. This kind of interaction was implemented with a simple, post-perceptual integration mechanism that reconciled location information obtained from the person tracker with the output of a speech recognition system. Here, multimodal integration of positional and speech information allowed straightforward disambiguation of the deictic lamp reference. Given the simplicity of this example, it seems far from obvious that a more complex integration mechanism might have been called for. To motivate a more involved treatment, we start by examining some of the problems with current approaches to perceptual computing.

Despite many recent and significant advances, computer vision and speech understanding, along with many other perceptual research areas (Picard 1997, Oviatt 2002) are still infant sciences. The non-trivial perceptual components of multimodal systems are therefore never perfect and are subject to a wide variety of failure modes. For example, the Intelligent Room sometimes "lost" people while visually tracking them, due to occlusion, coincidental color matches between fore and background objects, unfavorable lighting conditions, etc. Perceptual components can also dictate sets of environmental conditions that are required for proper operation. For vision systems, these may include assumptions about brightness levels, object size, image backgrounds, scene complexity, pose, etc. Although the particular failure modes of computational perceptual systems varies with them individually, one may assume that under a wide variety of conditions any one of them may temporarily stop working as desired. How these systems manifest this undesired operation is itself highly idiosyncratic. Some may simply provide no information, for example, a speech recognition system confused by a foreign accent. Far more troublesome are those that continue to operate as if nothing were amiss but simply provide incorrect data, such as a vision-based tracking system that mistakes a floor lamp for a person and reports that he is standing remarkably still.

Unimodal systems also suffer from what might be deemed *perceptual impoverishment*. They implement single, highly specific perceptual capabilities, such as locating people within a room, using individual (or far less often, a small number of) recognition paradigms, such as looking for their faces. *They are oblivious to phenomena outside of their perceptual scope, even if these phenomena are indicative of events or conditions they are intended to recognize*.

That perceptual systems have a variety of failure modes is not confined to their artificial instantiations. Biological systems also display a wide range of pathological conditions, many of which are so engrained that they are difficult to notice. These include limitations in innate sensory capability, as with visual blind spots on the human retina, and limited resources while processing sensory input, as with our linguistic difficultly understanding nested embedded clauses (Miller and Chomsky 1963). *Cross-modal perceptual mechanisms have evolved to cope both with innate physiological limitations and specific environmental constraints*. Stein and Meredith (1994) argue for the evolutionary advantages of overlapping and reinforcing sensory abilities; they reduce dependence on specific environmental conditions and reliance on unique perceptual pathways and thereby provide clear survival advantages. The facial vision of the blind discussed earlier is an extreme example of this kind of reinforcement. Generally, the influences are more subtle, never bringing themselves to our attention, but demonstrably a fundamental component of perception nonetheless.

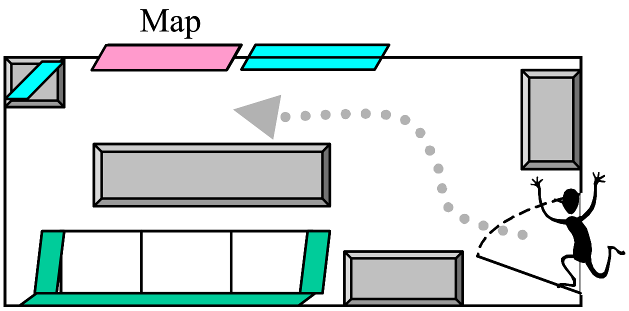
Given their role in biological systems, one might assume cross-modal influences could provide similar benefit in artificial systems such as an Intelligent Room. How then should we go about incorporating them? Answering this question is a two-step process. Because the Intelligent Room was an engineered as opposed to an evolved system, we would first needed to explicitly find potential synergies between its modalities that could have been exploited. Ideally, these influences would be learned, an approach the influence network model enables, but here we examine how this would be done in the absence of a such a model, where the synergies must be identified manually. Once identified, these synergies must then somehow be incorporated (i.e., reverse engineered) into the overall system. At the time, this emerged as the primary obstacle to integrating cross-modal influences into the Intelligent Room and more generally, into other types of engineered interactive systems. Reverse-engineering intersensory influences into systems not designed with them in mind can be convoluted at best and impossible at worst.

6.4.1 Two Examples of Cross-Modal Influence

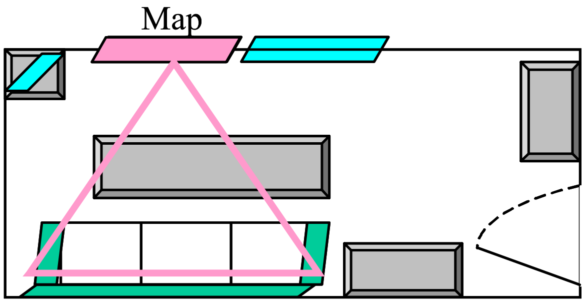
To examine these issues in more detail, we start with the following empirical and complementary observations:

1) People tend to talk about objects they are near. (Figure 6.6a)
2) People tend to be near objects they talk about. (Figure 6.6b)

These heuristics reflect a relationship between a person's location and what that person is referring to when he speaks; knowing something about one of them provides some degree of information about the other. For example, someone walking up to a video display of a map is potentially likely to speak about the map, as in Figure 6a; here, person location



a)



b)

Figure 6.6a (top) – People talk about objects they are near. Someone approaching a projected display showing a map, for example, is more likely to make a geographical utterance. Here, location information can augment speech recognition.

Figure 6.6b (bottom) – People are near objects they talk about. Someone speaking about the contents of a video display is likely to be located somewhere (delineated by the triangle) from which the display is viewable. Here, speech information can augment person tracking.

data can inform a speech model. Conversely, someone who speaks about a displayed map is likely to be in a position to see it, as illustrated in Figure 6b; here, speech data can inform a location model. Of course, it is easy to imagine situations where these heuristics are wrong. Nonetheless, in our experience they are empirically valid, so how could we have incorporated influences based on them into a system such as an Intelligent Room?

Mechanistically, we might imagine the person tracking system exchanging information with the speech recognition system. For example, the tracking system might provide *hints* to the speech recognition system to preferentially expect utterances involving objects the person is near, such as a displayed map. Conversely, we could also imagine that the speech recognition system would send hints to the person tracking system to be especially observant looking for someone in indicated sections of the room, based on what that person is referring to in his speech.

This seems reasonable until we try to build a system that works this way. There are both representational and algorithmic stumbling blocks that make this conceptually straightforward cross-modal information sharing difficult to implement. These are due not only to post-perceptual architectural integration, but also to how the perceptual are themselves typically created for use in post-perceptual systems. We first examine issues of representational compatibility, the interlingua used to represent shared information, and then address how the systems could incorporate *hints* they receive in this interlingua into their algorithmic models.

Consider a person tracking system that provides the coordinates of people within a room in real-time, relative to some real-world origin; the system outputs the actual locations of the room's occupants. We will refer to the tracking system in the Intelligent Room as a representative example of other such systems (e.g., Wren et al. 1997, Gross et al. 2000). Its only inputs were stereo video cameras and its sole output were sets of *(x,y,z)* tuples representing occupants' centroid head coordinates, which were generated at 20Hz. Contrast this with the room's speech recognition system, which was based upon the Java Speech API (Sun 2001) using IBM's ViaVoice platform and was typical of similar hidden Markov model based spoken language systems (Jelinek 1997). Its inputs were audio voice signals and a formal linguistic model of expected utterances, which were represented as probabilistically weighted context free grammars. Its outputs were sets of parse trees representing what was heard, along with the system's confidence levels that the spoken utterances were interpreted correctly.

How then should these two systems have exchanged information? It does not seem plausible from an engineering perspective, whether in natural or artificial systems, to have provided each modality with access to the internal representations of the other. Thus, we do not expect that a tracking system would know anything about linguistic models nor we do expect the language system would be skilled in spatial reasoning and representation. Even were we to suppose the speech recognition system *could* somehow represent spatial coordinates, the example in Figure 6b above involves *regions* of space, not isolated point coordinates. From an external point of view, it is not obvious how the tracking system internally represents regions, *presuming it even has that capability in the* *first place*. The complementary example of how the tracking system might refer to classes of linguistic utterances is similarly convoluted.

Unfortunately, even were this interlingua problem easily solved and the subsystems had a common language for representing information, the way perceptual subsystems are generally implemented would make incorporation of cross-modal data difficult or impossible. For example, in the case of a person tracking system, the real-world body coordinates are generated via three-dimensional spatial reconstruction based on correspondences between sets of image coordinates. The common techniques for computing the reconstructed coordinates, such as neural networks or fit polynomials, are in a sense closed – once the appropriate coordinate transform has been learned, there is generally no way to bias the transformation in favor of particular points or spatial regions. Thus, there is no way to incorporate the influence, even if the systems had a common way of encoding it. Here again, the complementary situation with influencing speech recognition from a tracking system can be similarly intractable. For example, not all linguistic recognition models support dynamic, preferential weightings for classes of commonly themed utterances. So, even if the tracking system could somehow communicate what the speech recognition system should expect to hear, the speech recognition system might not be able to do anything useful with this information.

We see that not only are the individual modal representations incompatible, the perceptual algorithms (i.e., the computations that occur in the sensory pipelines) are incompatible as well. This comes as no surprise given that these systems were engineered primarily for unimodal applications. Unlike natural perceptual systems within an individual species, artificial perceptual systems do not co-evolve, and therefore, have had no evolutionary pressure to force representational and algorithmic compatibility. These engineered systems are intended to be data sources feeding into other systems, such as the ones performing multimodal integration, that are intended to be data sinks. There is no reason to expect that these perceptual subsystems would or even could directly interoperate.

6.5 Speculation on an Intersensory Model

For disparate perceptual systems to interact, there must be some agreement – implicit or explicit – as to how they represent and process perceptual data. The goal of this section is to motivate and briefly outline a model of intersensory perception that formalizes this notion of "agreement." The intent of the model is to allow us: (1) to formalize an understanding of cross-modal phenomena; and (2) to design artificial systems that exploit intersensory phenomena to increase the scope of their perceptual capabilities. It provides the theoretical framework for evaluating influence networks and other approaches to intersensory integration, by describing essential phenomena that any integration scheme must somehow address. In this sense, it also provides for the development of a competency test for the influence network approach described above.

During the past century, research in intersensory processing has proceeded along two main avenues. The first has been the design of experimental scenarios that reveal what are usually hidden cross-modal influences at the behavioral, and with adult humans, sometimes the conscious level. The second has been based on intrusive physiologic examinations in animals that elucidate the cellular (in primitive organisms) and the neurological (in encephalized species) substrates of intersensory function. Thus, a wide range of phenomenological cross-modal interactions has been described and some of the physiology behind them has been identified. What has not been addressed in a general framework are the processes enabling intersensory influence and their implications to our understanding of the individual senses themselves.

This chapter started by examining evidence that our senses are more interdependent than they seem on the surface. Contrary to the understanding arrived at by introspection, perceptions are the products of complex interactions between our sensory channels. What, however, are the contents of these channels? Returning to a question we asked in Chapter 1, how exactly should we understand what a *sense* is? The common notion might be that a *sense* is the perceptual ability to interpret impressions received via a single sensory organ. For example, vision is the ability to interpret waves of light impacting upon the retinas. Viewed this way, each sense is defined by a wide gamut of sensitivities and abilities tied to a particular biological organ. This view, however, is misleading. It appears instead that each sense is composed of a multitude of capabilities, some of which may be absent in particular individuals (e.g., loss of color perception or depth sensitivity in visual perception) and many of which operate independently. The most persuasive case studies are drawn from experiments with the blind; even though they cannot see, their eyes can still respond to visual cues. For example, people with cerebral cortex injuries rendering them perceptually blind are able to direct their eyes towards spots of light, even though they cannot consciously see these spots and think they are simply guessing (Poppel et al. 1981). Their loss of visual perception has not hindered other unconscious visual processes, which have also been demonstrated in the normally sighted (Wolfe 1983).

The blind can even experience visual cross-modal phenomena. For example, a congenitally blind child has been shown to have convergence of the eyes in response to approaching sounds (Butterworth 1981), even though this convergence has no practical benefit. The existence of independent visual processes has led to a more complex view of visual perception and makes it difficult to view vision as a monolithic capability; it seems instead to be an assortment of somewhat independent processes that have a single sensory organ, the eye, in common. For this reason, although the modes (i.e., perceptual primitives) in multimodal perception are generally understood to correspond to different gross senses, e.g., vision, audition, proprioception, etc., we have found a finer grained definition more useful, in which the different perceptual capabilities of each sensory organ are treated as the individual modes, rather than taking the abstract function of that organ as a whole. Aside from providing a clearer engineering viewpoint, this perspective allows us to make explicit what we will call the cross-modal influences between different perceptual pathways that start from the same sensory organ and would traditionally be considered to be within a single modality.

Given this minimalist view of unimodal perception, we now outline the three types of cross-modal influence to be covered by the model. This type of categorization has not been made in the multisensory literature, but it is well motivated by both phenomenological and neurological data. Although the influences listed below appear to cover a variety of different phenomena, we will argue their primary distinction is temporal and a single mechanism could account for many such influences within an artificial system. Our categories of cross-modal influence are *cueing*, *mutual influence*, and *resolution*:

1. *Cueing* occurs when a sensory system is biased in some way *before* a perceptual event due to cross-modal influences. The most studied form of cueing is *priming* (Meyer and Schvaneveldt 1971), which is an increase in the speed or accuracy of a perceptual decision task, based on previous exposure to some of its content. Although not necessarily intersensory, many priming experiments are conducted cross-modally.[4] Cueing is distinguished from priming because cueing removes any notion of perceptual correctness – influences that simply bias perceptual interpretation are included – and because cueing covers inhibitory, not just excitatory, influences. Thus, *accommodation* experiments (Kohler 1964, Blake et al. 1981), in which prolonged or repeated exposures to sensory stimuli lead to modification in baseline perceptual responses, are also examples of cueing. Kohler's work is perhaps the best known example of these experiments, in which subjects gradually acclimate to the effects of wearing distortive spectacles, e.g., ones in which the left halves of the lens are blue and right halves are yellow. Cueing is important for artificial perceptual systems because they rarely have any notion of how what they should expect to perceive changes over time and with contextual circumstances.

2. *Mutual influence* occurs when a set of modalities interact in a dynamic feedback process *during* percept formation. A wide range of interactions, particularly between the vestibular, visual, auditory, and proprioceptive systems have been studied, many of these investigated in depth because of the unusual effects of gravity on perception to pilots and astronauts. For example, during high acceleration takeoffs, pilots can experience the sensation that their body is tilted backwards and their instrument panel is rising too quickly (Graybiel 1952). They must learn to ignore these perceptions, because reacting to them could be life threatening. In influences that have a visual component, it is not uncommon for the visual input to dominate the joint perceptual interpretation (Warren at al. 1981), as with the "ventriloquism effect" (Howard and Templeton 1966), where a

[4] A mechanism accounting for priming, *spreading activation* (Collins and Loftus 1975, Collins and Quillian 1969), has been very influential in the development of the theory of semantic networks (e.g., Fahlman 1979) and in later perceptual theories, such as *sequence seeking* (Ullman 1996).

ventriloquist's voice appears to emanate from a dummy's mouth because its lips are moving rather than his. However, the dominant modality in mutual influence is generally the one perceived most strongly (Welch and Warren 1986), and circumstances can uniquely exaggerate any modality to the point of perceptual dominance. The McGurk effect discussed earlier is an unusual example of mutual influence where neither involved modality is dominant. Evidence for mutual influence interactions has been found neurologically; in the thalamus, superior colliculus, and cerebral cortex, neurons respond to and integrate multisensory information in a time frame of 10s of milliseconds following stimulus presentation. Therefore, components of multisensory processing occur long before perceptual and cognitive effects even begin (Meredith 2001). Mutual influence is important for artificial perceptual systems because it allows unimodal systems to overcome perceptual impoverishment by supplementing their sensory input with indicative cross-modal data.

3. *Resolution* occurs when a potentially ambiguous sensory input is crosschecked among parallel sensory input streams and potentially changed *after* percept formation. Resolution influences are reflexive and reactionary, as in those demonstrated in nociception (i.e., pain response) in rodents (Stein and Dixon 1978, Aury et al. 1991). Resolution influences specifically do not involve ratiocination; they are confined to perceptual channels and are therefore unconscious and automatic. There is no direct awareness of the influence itself and it is not subject to willful control. Many post-perceptual intersensory influences do not meet these criteria, including *explicit memory tests* (Jacoby 1983) and cross-modal matching, and will not be considered here. Resolution is important for artificial perceptual systems because it can increase or decrease confidence in perceptions based on cross-modal agreement or conflict. Modern interactive systems rarely have any way to self-validate their own operation or to adjust internal mechanisms (e.g., threshold values) without external supervision. Biological systems have self-supervised plasticity, which provides a dynamic, adaptive fluidity unknown in artificial perception.

How might these influences be effected in a perceptual system? Even though the perceptual channels are themselves highly specialized, we argue that the mechanisms behind intersensory influence can be amodal to a far greater extent – there is less of a need for them to be dedicated to specific perceptual modes. This position at first glance is also hard to support. Mechanistic specialization is so commonplace in biological systems, one can make the case that general-purpose, non-specific mechanisms need justification more so than do ad hoc ones (Gazzaniga 2000, Chomsky 2000). Why then suppose there is an amodal component to intersensory influence?

This question is somewhat misleading. The brain, does, in fact, use amodal codings in perception, the most well known of these being the spatial representations in the sensory and motor maps of the superior colliculus, discussed in more detail below. The superior colliculus is the only known region of the brain where auditory information is represented spatially, as opposed to tonotopically, and one may argue that the extraordinary specialization in this case is actually in the neurological mechanisms that determine spatial localization based on interaural time delays. This elaborate calculation allows representation of auditory information in the superior colliculus' common, amodal coordinate system. *It is precisely the mechanistic perceptual specialization here that allows a general-purpose, amodal system, such as the superior colliculus to exist.* This can be seen identically with other specialized perceptory capabilities unique to given species that provide spatial localization, such as echolocation (bats), whisker displacement (rats and mice), acute audition (owls), infrared detection (rattlesnakes), and electroreception (fish) (Stein and Meredith, Chapter 6). Perceptions in each of these animals are represented in common, amodal coordinate systems and support the view that perceptual specialization in no way precludes the existence of amodal representations.

One may also more tentatively propose strong evolutionary advantages to amodal intersensory perception. It allows incorporation of newly evolved or modified perceptual capabilities without requiring the development of specialized mechanisms that enable their participation in an intersensory perceptual system. More generally, were the mechanisms for intersensory influence between each pair of modalities uniquely specified, there would be $O(n^2)$ of them required for a set of n modalities. General-purpose amodal mechanisms for effecting cross-modal influence would simplify this integration problem. For this reason, one can take the position that regardless of how Nature proceeds, from an engineering perspective, the benefit provided by amodally effected intersensory influence is clear. A general purpose mechanism for incorporating individual sensory components into integrated, artificial perceptual systems would be very useful when building them. This raises the question, however, of how does Nature proceed? What is known about neural substrates of intersensory influence?

Consider the effect of touching someone and having his eyes turn to determine the source of the stimulus. This is an example of cross-modal influence – the position of the *touch* determines the *foveation* of the eyes – but it is different from the interaction described in the McGurk effect above. The touch scenario leads to behavioral effects that center the stimulus with respect to the body and peripheral sensory organs. In contrast, the McGurk effect is an example of sensory influence solely within perceptual channels and has no behavioral component. Is this difference important?

Motor influences – i.e., ones that cause attentive and orientation behaviors – are by far the more understood of the two. The primary neurological substrate behind them is the *superior colliculus* (or *optic tectum* in non-mammalian vertebrates), a small region of the brain that produces signals that cross-modally orient peripheral sensory organs based on sensory input. The superior colliculus contains layered, topographic sensory and motor maps that share a common rostral-caudal vs. medial-lateral coordinate system. The maps so far elucidated are in register; that is, co-located positions in the real world – in the sensory case representing derived locations of perceptual inputs and in the motor case representing peripheral sensory organ motor coordinates that focus on those regions – are essentially vertically overlapping. The actual mechanisms that use these maps to effect intersensory motor influence are currently unknown – variants on spreading vertical activation are suspected – but there is little doubt the maps' organization is a fundamental component of that mechanism.

Far less is known neurologically about *semantic* influences – i.e., ones that have internal effects confined to perceptual channels. The superior colliculus itself has been directly approachable from a research perspective because the brain has dedicated inner space, namely, the tissue of the topographic maps, to representing the outer space of the real-world. The representation is both isomorphic and perspicacious and has made the superior colliculus uniquely amenable to study. The perceptual, as opposed to spatial, representations of the senses are far more elusive because they are specialized to the individual modalities and the organs that perceive them. For example, the auditory system is organized at both the thalamic and cortical levels according to sound frequency whereas the retina organizes visual input spatiotopically (in retinocentric coordinates) and maintains this representation into the visual cortex.

Although many of their neuroanatomical interconnections have been elucidated, how these systems actually share non-spatial perceptual information during intersensory influence is unknown. The approach in this thesis has been to introduce the slice data structure in Chapter 3, which was inspired by the role of cortical topographic maps in equivalently organizing both sensory and motor information in animals. Slices are inherently amodal and independent of the features they represent. This, for example, enables them to freely share information between sensory and motor systems in learning birdsong, without requiring the development of new formalisms or representations.

By treating slices as state spaces, we can model activations in these maps that correspond to dynamic sensory and motor processes. This allows us to incorporate the temporal intersensory influences described above. For example, in our model, *cueing* corresponds to pre-activation of a region within one slice through the prior activation of another. *Mutual influence* corresponds to two slices cooperatively cross-activating each another while they are in the midst of perceiving. These temporal interactions also allow us to use lower dimensional representations for modeling sensory and motor data because they in essence trade space (or dimensionality) for time. In other words, our data need not be completely separable if we can find can find ways of avoiding ambiguity during the temporal window of percept formation. Thus, we believe that cross-modal influence not only provides perceptual stability, it makes sensory systems more computationally efficient in doing so.

6.6 Summary

This chapter has examined the connections between the computational framework presented in this thesis and perception in biological systems. Our goal was to motivate our approach and to provide a context for evaluating related work in multimodal integration. We also examined the representational and algorithmic challenges in engineering biologically realistic perceptual systems. These systems do not co-evolve and may be grossly incompatible with one another, even when external relationships between their perceptions are readily apparent. Finally, we speculated on a number of theoretical issues in intersensory perception and examined how the work in this thesis addresses them.

Chapter 7

Conclusions

The primary contribution of this thesis has been to identify how artificial systems can use cross-modal correspondences to learn without supervision. Building systems that develop by interacting with the world around them has been a dream of artificial intelligence researchers since the days of Turing, and this thesis marks a step forward towards that goal. This work has also demonstrated that a biologically-inspired approach can help answer what are historically difficult computational problems, for example, how to cluster non-parametric data corresponding to an unknown number of categories and how to learn motor control through self-observation.

We have also identified cross-modal cooperation as a central problem in theoretical and applied artificial intelligence. In doing so, we have introduced three key formalisms:

1. ***Slices***, which represent and share information amodally between different subsystems. They can be used to model sensory and motor data and have wide applications for modeling static and dynamic inputs.

2. ***Cross-modal clustering****,* which learns categories in slices without supervision, based on how they co-occur with other slices.

3. ***Influence networks***, which spread influence among different slices and provide perceptual and interpretative stability.

We demonstrated the power of cross-modal cooperation and clustering by learning the vowel structure of American English, simply by watching and listening to someone speak. This is the first self-supervised computational approach to this problem of which we are aware. We further demonstrated that this framework can equivalently be applied to sensorimotor learning, by acquiring birdsong according to the developmental stages of a juvenile zebra finch. In this, we have shown that this work can be applied recursively to learn higher-level knowledge. In the example given in Chapter 5, we recursively grounded motor control in self-acquired sensory categories, thereby demonstrating afferent and efferent equivalence in sensorimotor learning, which in itself is a very surprising result. It further suggests how to carry this work forward towards more complex, higher-level learning tasks, as we discuss below.

7.1 Future Work

Our immediate plans are to apply the birdsong learning framework in Chapter 5 towards learning human babbling and the entire phonetic set of American English. Early protolinguistic behavior in humans is not a well studied nor understood phenomenon. We intend to create a system that learns to babble at the level of an eight month old child, in a developmentally realistic way. We believe examining the earliest intellectual development in people provides a path to understanding fundamental issues in knowledge acquisition and to building a new generation of intelligent machines. Along these lines, we plan to extend this work to hierarchical clustering and address issues in concept formation based on perceptual grounding. We would particularly like to apply the ideas of Lakoff (1987) for grounding internal mental activity using external, real world metaphors and observations, which is a basic feature of human cognition.

We would also like to embed this framework in a robotic platform or an interactive environment, to provide a real-time environment for visual and motor development.

We are also interested in non-perceptual applications of the work presented here. The mathematical framework in this thesis can be applied to problems with similar structures. For example, by treating the messages passed between human or software agents as perceptual events, we can categorize both the messages and the agents passing them. This has applications both in software engineering and in understanding cells of interacting people. It also is applicable to learning how to detect events within distributed sensor networks. This work also has applications in separating non-ergodic Markov chains into their ergodic components, which has applications in probability theory and statistical physics.

Chapter 8

References

1. Agin, G.J. Representation and description of curved objects. Ph.D. Thesis, Stanford University, Stanford, CA. 1972.

2. Alcock, J. Animal Behavior: An Evolutionary Approach, 7th edition. Sinauer Associates, Inc., Sunderland, Massachusetts. 2001.

3. Aleksandrovsky, B., Whitson, J., Andes, G., Lynch, G. Granger, R. Novel Speech Processing Mechanism Derived from Auditory Neocortical Circuit Analysis. In Proceedings of ICSLP'96. 1996.

4. Amari, S. Topographic organization of nerve fields. *Bulletin of Mathematical Biology*, 42:339-364. 1980.

5. Arbib, M.A., Schemas for the temporal organization of behavior. *Human Neurobiology*, 4, 63-72. 1985

6. Aristotle. De Anima. 350 BCE. Translated by Tancred, H.L. Penguin Classics. London. 1987.

7. Atkeson C.G., Hale J, Pollick F., Riley M., Kotosaka S., Schaal S., Shibata T., Tevatia G., Vijayakumar S., Ude A., Kawato M. Using humanoid robots to study human behavior. *IEEE Intelligent Systems*, Special Issue on Humanoid Robotics, 46-56. 2000.

8. Auroy P., Irthum B., Woda A. Oral nociceptive activity in the rat superior colliculus. *Brain Res*. 549:275-284. 1991.

9. Bangalore, S. and Johnston, M. Integrating multimodal language processing with speech recognition. In *ICSLP-2000*, vol.2, 126-129. 2000.

10. Bartlett, M.S. Face Image Analysis by Unsupervised Learning. Kluwer. MA. 2001.

11. Becker, S. A computational principle for hippocampal learning and neurogenesis. Hippocampus 15(6):722-738. 2005.

12. Becker, S. and Hinton, G. E., Spatial coherence as an internal teacher for a neural network. In Backpropagation: Theory, Architectures, and Applications, Y. Chauvin and D. Rumelhart (eds),. Lawrence Erlbaum. Hillsdale, N.J. 1995.

13. Bednar, J.A., Choe, Y., De Paula, J., Miikkulainen, R., Provost, J., and Tversky, T. Modeling Cortical Maps with Topographica, *Neurocomputing*. 58—60 pp. 1129-1135. 2004.

14. Bellman, R. *Adaptive Control Processes*. Princeton University Press, 1961.

15. Bender, R. E. The Consquest of Deafness (3rd Ed.). Danville, Il: Interstate Publishers. 1981.

16. Binford, T. O., Visual perception by computer. In the *Proceedings of the IEEE Systems Science and Cybernetics Conference*. Miami, FL. *IEEE*, New York. 1971.

17. Bishop, C.M., Neural Networks for Pattern Recognition. Oxford University Press, 1995.

18. Blake, R., Sobel, K. V., and Gilroy, L. A., Visual Motion Retards Alternations between conflicting Perceptual Interpretations. *Neuron*, (39):1–20, August, 2003.

19. Bloedel, JR, Ebner, TJ, and Wise, SP (eds.), Acquisition of Motor Behavior in Vertebrates, MIT Press: Cambridge, MA 1996.

20. Blum, A., and Mitchell, T. Combining Labeled and Unlabeled Data with Co-Training. With Tom Mitchell. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92--100, 1998.

21. Bobick, A.; Intille, S.; Davis, J.; Baird, F.; Pinhanez, C.; Campbell, L.; Ivanov, Y.; Schütte, A.; and Wilson, A. The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. M.I.T. Media Laboratory Perceptual Computing Section 398. 1996.

22. Boersma, P., and Weenink, D., PRAAT, a system for doing phonetics by computer. (Version 4.3.14) [Computer program]. Glot International 5(9/10): 341-345.http://www.praat.org/. May 26, 2005.

23. Bolt, R. A., Put-That-There: Voice and gesture at the graphics interface. *Computer Graphics*. Vol. 14, No. 3. pp. 262 – 270. July, 1980.

24. Brainard, M.S. and Doupe, A.J. Auditory feedback in learning and maintenance of vocal behaviour. *Nat Rev Neurosci*, 2000.

25. Brenowitz, E.A., Margoliash, D., Nordeen, K.W. (Eds.) Special issue on birdsong. *Journal of Neurobiology*. Volume 33, Issue 5, pp 495-709. 1997.

26. Brooke N. M., & Scott S. D. PCA image coding schemes and visual speech intelligibility. *Proc. Institute of Acoustics*, **16**(5), 123–129. 1994.

27. Brooks, R., A Robust Layered Control System For A Mobile Robot. *IEEE Journal of Robotics and Automation.* 2:14-23. March, 1986.

28. Brooks, R. Intelligence without Representation, *Artificial Intelligence*. 47:139–159. 1991a.

29. Brooks, R. Intelligence without Reason. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence* (IJCAI-91). Sydney, Australia. 1991b.

30. Brooks, R.A., C. Breazeal (Ferrell), R. Irie, C. Kemp, M. Marjanovic, B. Scassellati and M. Williamson, Alternate Essences of Intelligence. *In Proceedings of The Fifteenth National Conference on Artificial Intelligence.* (AAAI98). Madison, Wisconsin. 1998.

31. Bruner, J.S. and Postman, L. On the Perception of Incongruity: A Paradigm. In *Journal of Personality*, *18*, 206-223. 1949.

32. Bushara, K.O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., and Hallett, M., Neural correlates of cross-modal binding. *Nature Neuroscience* 6, 190-195. 1 Feb 2003.

33. Butterworth, G. The origins of auditory-visual perception and visual proprioception in human development. In *Intersensory Perception and Sensory Integration*, R.D. Walk and L.H. Pick, Jr. (Eds.) New York. Plenum. 1981.

34. Cadez, I. and Smyth, P. Probabilistic Clustering using Hierarchical Models. 1999.

35. Calvert, A.G., Spence, C., and Stein, B.E. The Handbook of Multisensory Processes. Bradford Books. 2004

36. Calvert, G.A., Bullmore, E., Brammer, M.J., Campbell, R., Iversen, S.D., Woodruff, P., McGuire, P., Williams, S., and David, A.S., Silent lipreading activates the auditory cortex. *Science*, 276, 593-596. 1997.

37. Calvert, GA., Campbell R., and Brammer, MJ., Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal. *Curr. Biol.*, 2000.

38. Camarillo D.B., Krummel T.M., Salisbury J.K., Robotic technology in surgery: past, present, and future. *Am J Surg*. 188:2S-15S. 2004.

39. Cheng, G., and Kuniyoshi**,** Y. Complex Continuous Meaningful Humanoid Interaction: A Multi Sensory-Cue Based Approach *Proc. of IEEE International Conference on Robotics and Automation* (ICRA 2000), pp.2235-2242, San Francisco, USA, April 24-28, 2000.

40. Cherry, E. C., Some experiments on the recognition of speech with one and two ears. *J. Acoust. Soc. Am*. 25, 975–979. 17, 227–246. 1953.

41. Chomsky, N. Language and the Brain. Quandaries and Prospects. Hans-Lukas Teuber Lecture. Massachusetts Institute of Technology. October 13, 2000.

42. Chomsky, N., and Halle, M. *The sound pattern of English.* New York: Harper and Row. 1968.

43. Citti, G. and Sarti, A. A cortical based model of perceptual completion in the roto-translation space. In *Proceeding of the Workshop on Second Order Subelliptic Equations and Applications.* Cortona. 2003.

44. Clark B, and Graybiel A., Factors Contributing to the Delay in the Perception of the Oculogravic Illusion. *Am J Psychol 79*:377-88. 1966.

45. Clarkson, P. and Moreno, P.J. On the use of support vector machines for phonetic classification. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99).* 1999.

46. Clemins, P.J. and Johnson, M.T. Application Of Speech Recognition To African Elephant (Loxodonta Africana) Vocalizations. In Proceedings of the *International Conference on Acoustics, Speech, and Signal Processing*. Vol. I p. 487. 2003.

47. Coen, M.H. Self-supervised acquisition of vowels in American English. To appear in *Proceedings of the Twenty First National Conference on Artificial Intelligence.* (*AAAI-06*) Boston, MA. 2006.

48. Coen, M.H. Cross-modal clustering. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*. (*AAAI-05*) Pittsburg, Pennsylvania. 2005.

49. Coen, M.H. Multimodal interaction: a biological view. In *Proceedings of 17th International Joint Conference on Artificial Intelligence*. (*IJCAI-01*) Seattle, Washington. 2001.

50. Coen, M.H., and Wilson, K. Learning Spatial Event Models from Multiple-Camera Perspectives in an Intelligent Room. In *Proceedings of MANSE'99*. Dublin, Ireland. 1999.

51. Coen, M.H, Phillips, B., Warshawsky, N., Weisman, L., Peters, S., Gajos, K., and Finin, P. Meeting the computational needs of intelligent environments: The Metaglue System. In *Proceedings of MANSE'99*. Dublin, Ireland. 1999.

52. Coen, M.H., The Future of Human-Computer Interaction, or How I Learned to Stop Worrying and Love My Intelligent Room, *IEEE Intelligent Systems*, vol. 14, no. 2, Mar./Apr., pp. 8—19. 1999.

53. Coen, M.H. Design Principles for Intelligent Environments. In *Proceedings of The Fifteenth National Conference on Artificial Intelligence.* (*AAAI-98*). Madison, Wisconsin. 1998.

54. Cohen, P. R., Johnston, M., McGee, D., Smith, I. Oviatt, S., Pittman, J., Chen, L., and Clow, J. QuickSet: Multimodal interaction for simulation set-up and control. 1997, *Proceedings of the Applied Natural Language Conference,* Association for Computational Linguistics. 1997.

55. Collins, A. M. & Loftus, E. F., A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407-428. 1975.

56. Collins, A. M. & Quillian, M. R., Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240-247. 1969.

57. Cotzin, M. and Dallenbach, K., Facial Vision: the role of pitch and loudness in the perception of obstacles by the blind. *The American Journal of Psychology*, 63: 485-515. 1950.

58. Cytowic, R. E., Synesthesia: A Union of Senses. New York. Springer-Verlag. 1989.

59. Dale AM, Fischl B, and Sereno MI., Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*. Feb; 9(2):179-94. 1999.

60. Dantzig, G.B. *Application of the simplex method to a transportation problem*. In Activity Analysis of Production and Allocation, Koopmans, T.C. (Ed.) pp. 339--347. Wiley, New York, 1951.

61. Darrell, T., Gordon, G., Harville, M., and Woodfill, J., Integrated person tracking using stereo, color, and pattern detection, *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR '98),* pp. 601-609, Santa Barbara, June, 1998.

62. Dautenhahn, K. and Nehaniv, C. L. (eds.), Imitation in Animals and Artifacts. MIT Press: London, 2002.

63. de Marken, C. Unsupervised Language Acquisition, PhD Thesis. Massachusetts Institute of Technology. Cambridge, MA. 1996.

64. de Sa, V.R. *Unsupervised Classification Learning from Cross-Modal Environmental Structure*. Doctoral Dissertation, Department of Computer Science, University of Rochester. 1994.

65. de Sa, V.R., & Ballard, D. Category Learning through Multi-Modality Sensing. In *Neural Computation* 10(5). 1998.

66. Dempster, A., Laird, N. and Rubin, D., Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1):1–38, 1977.

67. Dennett, D.C. Consciousness Explained, Little, Brown and Company. Boston, MA. 1991.

68. Dennett, D. C., and Haugeland, J. Intentionality. The Oxford Companion to the Mind. Gregory, R.L. (Ed.)., Oxford University Press. 1987.

69. Dey, A.K., Salber, D. and Abowd, G.D. A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. *Human-Computer Interaction (HCI) Journal*, Volume 16 (2-4), pp. 97-166. 2001.

70. Dobbins, A. C., Jeo, R., and Allman, J., Absence of spike frequency adaptation during binocular rivalry [Abstract]. *Society for Neuroscience Abstracts, 21.* 1995.

71. Donoho, D.L., and Grimes, C., Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *PNAS*; vol. 100; no. 10;5591-5596. May 13, 2003.

72. Elbert, T., Pantev, C., Wienbruch, C., Rockstroh, B., and Taub, E. Increased Cortical Representation of the Fingers of the Left Hand in String Players, *Science*: 270: 305-307. 1995.

73. Ernst, M.O., and Banks, M.S., Humans integrate visual and haptic information in a statistically optimal fashion. *Nature.* 415, 429-433; doi: 10.1038/415429a. 24 January 2002.

74. Fahlman, S.E. NETL: A System for Representing and Using Real-World Knowledge, MIT Press, Cambridge MA. 1979.

75. Ferrell, C. Orientation behavior using registered topographic maps. In *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior (SAB-96).* Society of Adaptive Behavior. 1996.

76. Finney, E.M., Fine, I., and Dobkins, K.R., Visual stimuli activate auditory cortex in the deaf. *Nature Neuroscience* 4, 1171-1173, 2001.

77. Fischl B., Sereno, M.I. and Dale, A.M., Cortical Surface-Based Analysis. II: Inflation, Flattening, and a Surface-Based Coordinate System. *Neuroimage*. 9(2):195-207. 1999.

78. Fisher, J. and Darrell, T., Signal-Level Audio Video Fusion Using Information Theory. *Proceedings of Workshop on Perceptive User Interfaces*. 2001.

79. Fitzgibbon, A., Pilu, M., and Fisher, R.B., "Direct Least Square Fitting of Ellipses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 476-480, May, 1999.

80. Fitzpatrick, R. and McCloskey, D.I., Proprioceptive, visual and vestibular thresholds for the perception of sway during standing in humans. *The Journal of Physiology*, Vol 478, Issue 1 pp. 173-186, 1994.

81. Flynn, A., Redundant sensors for mobile robot navigation, M.S. Thesis, Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA, July 1985.

82. Frank, A. On Kuhn's Hungarian Method - a tribute from Hungary. *Naval Research & Logistics*. 52. 2005.

83. Frisby, J. P., and Davies, I. R. L., Is the Haptic Müller–Lyer a Visual Phenomenon? *Nature.* 231, 463-465. 18 Jun 1971.

84. Galef, B. G., Imitation in animals: History, definition, and interpretation of data from the psychological laboratory. In: Social learning: Psychological and biological perspectives, eds. T. R. Zentall & B. G. Galef, Jr. Erlbaum. 1988.

85. Garnicia, O. K., Some prosodic and paralinguistic features of speech to young children. In C. E. Snow and C. A. Ferguson (Eds.) Talking to children. Cambridge University Press. 1977.

86. Gazzaniga, M. (Ed.) The New Cognitive Neurosciences. MIT Press. 2nd edition. Cambridge, MA. 2000.

87. Gerstner, W and Kistler, W.M., Spiking Neuron Models. Single Neurons, Populations, Plasticity Cambridge University Press. 2002.

88. Gibbon, J. Scalar expectancy theory and Weber's law in animal timing. *Psychol. Review* 84(3):279-325. 1977.

89. Gibbs, A.L and Su, F.E. On choosing and bounding probability metrics. *International Statistical Review*, vol. 70, number 3, 419-435. 2002

90. Gibson, J.J. The Perception of the Visual World. Boston, Houghton Mifflin. 1950.

91. Gibson, J.J. The Ecological Approach to Visual Perception. Lawrence Earlbaum Associates. Hillsdale, N.J. 1987.

92. Gold, B. and Morgan, N. *Speech and Audio Signal Processing*. Wiley Press, New York. 2000.

93. Graybiel A. Oculogravic Illusion. Archives of Ophthalmology (Chicago, IL) 48:605-15. 1952.

94. Gross, R., Yang, J., and Waibel, A. Face Recognition in a meeting room. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000

95. Grunfeld, E.A., Okada, T, Jauregui-Renaud, K., and Bronstein, A.M., The effect of habituation and plane of rotation on vestibular perceptual responses. *J Vestib Res*. 10(4-5):193-200. 2000.

96. Haesler, S., Wada, K., Nshdejan, A., Morrisey, E.E., Lints, T., Jarvis, E.D., and Scharff, C. FoxP2 Expression in Avian Vocal Learners and Non-Learners. *J. Neurosci*. 24: 3164-3175. 2004

97. Hall, A.C. and Moschovakis, A. eds., The superior colliculus: new approaches for studying sensorimotor integration. (Boca Raton: CRC Press) 2004.

98. Hamill, N.J., McGinn, M.D. and Horowitz, J.M., Characteristics of auditory brainstem responses in ground squirrels. *Journal of Comparative Physiology* B: Volume 159, Number 2. Pages: 159 – 165. March 1989.

99. Hebb, D.O. 1949. *The Organization of Behaviour*. John Wiley & Sons, New York.

100. Heil, P. and Neubauer ,H., A unifying basis of auditory thresholds based on temporal summation. *PNAS*, Vol. 100, no. 10. May 13, 2003.

101. Held, R. Shifts in binaural localization after prolonged exposures to atypical combinations of stimuli. *Am. J. Psychol*. 68L526-266. 1955.

102. Helmholtz, H. v. *Handbook of Physiological Optics*. 1856. as reprinted. in James P.C. Southall. (Ed.) 2000.

103. Hennecke M. E., Prasad K. V., & Stork D. G. Using deformable templates to infer visual speech dynamics. In *Proc. 28th Annual Asilomar Conf. on Signals, Systems and Computers*. 1994.

104. Hinton, G.E. and Sejnowski, T.J. Learning and relearning in Boltzman machines. In *Parallel Distributed Processing*, Rumelhart, D.E. and McClelland, J.L. (eds), Volume 1:Foundations. MIT Press, Cambridge, MA. 1986.

105. Holbrook, A., and Fairbanks, G. Diphthong Formants and their Movements. *J. Speech Hear. Res*. 5, 38-58. 1962

106. Honkela, T. and Hyvärinen, A. Linguistic Feature Extraction using Independent Component Analysis. In *Proceedings of IJCNN 2004, Intern. Joint Conf. on Neural Networks*, pp. 279-284. Budapest, Hungary, 25-29 July 2004.

107. Horn, B.K.P. Shape from Shading. MIT Artificial Intelligence Laboratory Technical Report. AITR 232. November 1970.

108. Hoshino, O. and Kuroiwa, K., Echo sound detection in the inferior colliculus for human echolocation. *Neurocomputing*: *An International Journal Special Issue*, 38-40, 1289-1296. 2001.

109. Howard, I. P. & Templeton, W. B. *Human Spatial Orientation*, Wiley, New York. 1966.

110. Hsu, W.H., and Ray, S.R., Construction of recurrent mixture models for time series classification. *International Joint Conference on Neural Networks* (IJCNN'99), 1999.

111. Huang, X, Acero, A., and Hon, H.W. Spoken Language Processing. Prentice Hall, New Jersey. 2001.

112. Hubel, D. H. and Wiesel, T. N., *J. Physiol.,* London. 160, 106−154. 1962.

113. Hughes, J. W. The threshold of audition for short periods of stimulation, *Proc. R. Soc. London Ser. B* 133, 486-490. 1946.

114. Intille, S. S., Larson, K., and Tapia, E. M., Designing and evaluating technology for independent aging in the home, in *Proceedings of the International Conference on Aging, Disability and Independence*. 2003.

115. J. Kleinberg. An Impossibility Theorem for Clustering. *Advances in Neural Information Processing Systems (NIPS)* Whistler, British Columbia, Canada. 2002.

116. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E. Adaptive mixtures of local experts. *Neural Computation,* 3, 79-87. 1991.

117. Jacoby, L.L. Remembering the data: analyzing interactive processes in reading. *Journal of Verbal learning and Verbal Behaviour, 22,* 485-508. 1983.

118. James, W. 1890. Principles of Psychology. Vol. 2. Dover. New York. 1955.

119. Jelinek, F. Statistical Methods for Speech Recognition. MIT Press. Cambridge, MA. 1997.

120. Johnson-Laird, P.N. Mental models: Towards a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press. 1983.

121. Kaas J.H., and Hackett T.A. Subdivisions of auditory cortex and processing streams in primates. In *PNAS.* Oct 24;97(22):11793-9. 2000.

122. Kaas, J. The Reorganization of Sensory and Motor Maps after Injury in Adult Mammals, in M. Gazzaniga (ed.), The New Cognitive Neurosciences, 2nd ed. (Cambridge: MIT Press: 223-236). 2000.

123. Kardar, M. and Zee, A. Information optimization in coupled audio-visual cortical maps. In *PNAS* 99: 15894-15897. 2002

124. Kautau, A. Classification of the Peterson & Barney vowels using Weka. Technical Report. UCSD. 2002.

125. Keele S.W. and Summers, J. J., The Structure of Motor Programs. In Motor Control – Issues and Trends. Stelmach, G. E., (Ed.) Academic Press. San Diego. 1976.

126. Klee, V., and Minty, G. J., *How good is the simplex algorithm*?, in Inequalitites, III. Shisha, O. (Ed.), Academic Press, New York, pp. 159-175. 1972.

127. Kogan, J.A. and Margoliash, D. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. *Journal of the Acoustic Society of America*. Vol. 103, No. 4, April 1998.

128. Kohler, I. The formation and transformation of the perceptual world. *Psychological Issues* 3(4):1-173. 1964.

129. Kohonen, T. Self-Organization and Associative Memory. Springer-Verlag, Berlin. 1984.

130. Kuhl, P.K. Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50, pp. 93-107. 1991.

131. Kuhn, H. The Hungarian method for the assignment problem. *Naval Res. Logist. Q.* 2:83—97. 1955.

132. Kumar, V. Towards Trainable Man-machine Interfaces: Combining Top-down Constraints with Bottom-up Learning in Facial Analysis. Ph.D. Thesis. Department of Brain and Cognitive Sciences. Massachusetts Institute of Technology. Cambridge, MA. 2002.

133. Lakoff, G. Women, Fire, and Dangerous Things: What Categories Reveal About the Mind. University of Chicago Press, 1987.

134. Leopold, D. A., Wilke, M., Maier, A. and Logothetis, N.K. Stable perception of visually ambiguous patterns. *Nature*. Volume 5, Number 6, pp 605 – 609. June 2002.

135. Levina, E., and Bickel P.J., The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics. *Proceedings of ICCV*, Vancouver, Canada, pp. 251-256. 2001.

136. Lewkowicz, D.J. and Lickliter, R. (eds.) The Development of Intersensory Perception. Lawrence Erlbaum Associations. Hillsdale, N.J. 1994.

137. Lippmann, R.P. Speech recognition by machines and humans. *Speech Communication* 22, 1-15. 1987.

138. Lloyd, S. P., `Least squares quantization in PCM,' *IEEE Trans. Inform. Theory*, vol. 28, pp. 129-137, 1982.

139. López, M.E., Barea, R., Bergasa, L.M. and Escudero, M.S., Visually Augmented POMDP for Indoor Robot Navigation (Spain) *From Proceedings of* Applied Informatics. 2003.

140. MacKay, D.H.J. Information Theory, Inference, and Learning Algorithms. Cambridge University Press. 2003.

141. Mackey, M.C., and Glass, L. Oscillation and Chaos in Physiological Control Systems. *Science*,

142. Mackie, G.O. and Singla, C.L. Studies on Hexactinellid Sponges. I. Histology of Rhabdocalyptus dawsoni (Lambe, 1873). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, Vol. 301, No. 1107. (Jul. 5, 1983), pp. 365-400. 1983.

143. Marler, P. Three models of song learning: Evidence from behavior. *Journal of Neurobiology*. Volume 33, Issue 5 , Pages 501 – 516. 1997.

144. Marr, D. Vision. WH Freeman, San Francisco, 1982.

145. Mase, K., and Pentland, A. Automatic Lipreading by Computer. *Trans. IEEE.*, vol. J73-D-II, No. 6, pp. 796-803, June 1990.

146. Massaro, D.W. The fuzzy logical model of speech perception: A framework for research and theory. In Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp.79-82). Ohmsha Ltd, Tokyo. 1992.

147. Massaro, D.W. and Cohen, M.M. Fuzzy logical model bimodal emotion perception: Comment on The perception of emotions by ear and by eye by de Gelder and Vroomen *Cognition and Emotion, 14(3),* 313-320. 2000.

148. Massaro, D.W., Cohen, M.M., Gesi, A. and Heredia, R. Bimodal Speech Perception: An Examination across Languages. *Journal of Phonetics, 21*, 445-478. 1993.

149. Massie, T. M. and Salisbury, J. K.. *The PHANToM Haptic Interface: A Device for Probing Virtual Objects*. In ASME Haptic Interfaces for Virtual Environment and Teleoperator Systems 1994, Dynamic Systems and Control 1994, volume 1, pages 295--301, Nov. 1994.

150. Massone, L., and Bizzi, E., On the Role of Input Representations in Sensorimotor Mapping, *Proc. IJCNN*, Washington DC, 1:173-176. 1990.

151. Mataric, M.J., Studying the Role of Embodiment in Cognition. *Cybernetics and Systems* 28(6):457-470. 1997.

152. Maunsell, J.H.R, Nealy, T.A., Sclar, G., and DePriest, D.D. Representation of extraretinal information in mokey visual cortex. In *Neural mechanisms of visual perception. Proceedings of the Second Retina Research Foundation Symposium*, 14-15 April 1989. D.M Lam and C.D. Gilbert (eds). Woodlands, Texas. Portfolio Pub. Co. 1989.

153. McCallum, A., and Nigam, K. Employing EM in Pool-Based Active Learning for Text Classification. In *Proceedings of ICML-98, 15th International Conference on Machine Learning*. 1998.

154. McGraw, M. B., The Neuromuscular Maturation of the Human Infant. New York: Institute of Child Development. 1939.

155. McGurk, H., and MacDonald, J. Hearing lips and seeing voices. *Nature*. 264:746-748. 1976.

156. Mellinger, D.K. and Clark. C.W. Bioacoustic transient detection by image convolution. *Journal of the Acoustical Society of America*. -- Volume 93, Issue 4, p. 2358. April 1993.

157. Meltzoff, A.N. and Moore, M.K. Imitation of facial and manual gestures by human neonates. *Science* 198:75-78. 1977.

158. Meltzoff, A. N., and Prinz, W., The imitative mind: Development, evolution, and brain bases. Cambridge, England: Cambridge University Press. 2002.

159. Metcalfe, J.S., . Chang, T.Y., Chen, L.C., McDowell, K., Jeka, J.J., and Clark, J. E. Development of somatosensory-motor integration: An event-related analysis of infant posture in the first year of independent walking. *Developmental Psychobiology* 46(1) 19-35. 2005.

160. Meyer, D. E., and Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* 90:227--235.

161. Miller, G. and Chomsky, N. Finitary models of language users. In Luce, R.; Bush, R. and Galanter, E. (eds.) *Handbook of Mathematical Psychology, Vol 2.* New York: Wiley. 419-93. 1963.

162. Miller, D. B. The acoustic basis of mate recognition by female zebra finches (Taeniopygia guttata). *Anim. Behav*. 27, 376−380. 1979.

163. Minnen, D., Starner, T., Ward, J.A. ,Lukowicz, P., and Troester, G., Recognizing and Discovering Human Actions from On-Body Sensor Data ICME 2005, Amsterdam, NL, July 6-8, 2005.

164. Minsky, M., A Framework for Representing Knowledge. Reprinted in *The Psychology of Computer Visio*n, Winston, P.H. (Ed.), McGraw Hill, 1975.

165. Mitchell, T., Machine Learning. McGraw Hill, 1997.

166. Moody, J. and Darken, C. J. Fast learning in networks of locally tuned processing units. *Neural Computation*, 1:281-294. 1989.

167. Moran, T.P., and Dourish, R., Introduction to This Special Issue* on Context-Aware Computing. *Human-Computer Interaction*, Volume 16, 2001.

168. Müller, C. M. and Leppelsack, H. J., Feature extraction and tonotopic organization in the avian auditory forebrain. *Experimental Brain Research* (Historical Archive), Volume 59, Issue 3, pp. 587 – 599, Aug, 1985.

169. Mumford, S. Laws in Nature. London, England: Routledge. 2004.

170. Naeaetaenen, R., Tervaniemi, M., and E Sussman, P., Primitive intelligence in the auditory cortex. *Trends Neurosci*. 2001.

171. Nakayama, K. and Silverman, G. H., Serial and parallel processing of visual feature conjunctions. *Nature 320*, 264–265. 1986.

172. Nefian, A., Liang, L., Pi, X., Xiaoxiang, L., Mao, C. and Murphy, K. *ICASSP '02 (IEEE Int'l Conf on Acoustics, Speech and Signal Proc.)*, 2:2013--2016. 2002

173. Newell A. and Rosenbloom P.S., Mechanisms of skill acquisition and the law of practice. In: Cognitive skills and their acquisition (Anderson JR, Ed.), pp. 1–51. Hillsdale, NJ: Erlbaum. 1981.

174. Newell, A., & Simon, H.A. Computer science as empirical inquiry: Symbols and search. Communications of the Association for Computing Machinery, 19(3), 113-126. 1976.

175. Nilson, N. (Ed.), Shakey the Robot. SRI A.I. Center Technical Note 323. April, 1984.

176. Nordeen KW, Nordeen EJ. Auditory feedback is necessary for the maintenance of stereotyped song in adult zebra finches. *Behav Neural Biol*. Jan;57(1):58-66. 1992.

177. Ohnishi T, Matsuda H, Asada T, Aruga M, Hirakata M, Nishikawa M, Katoh A, Imabayashi E. Functional anatomy of musical perception in musicians. *Cereb Cortex*. 11(8):754-60. 2001.

178. Ölveczky B.P., Andalman A.S., Fee M.S. Vocal Experimentation in the Juvenile Songbird Requires a Basal Ganglia Circuit. PLoS Biol 3(5): e153. 2005

179. Oviatt, S. Multimodal interfaces, The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications, Lawrence Erlbaum Associates, Inc., Mahwah, NJ. 2002

180. Pearson, P. and Kingdom, F.A.A., On the interference of task-irrelevant hue variation on texture segmentation. *Perception*, 30, 559-569. 2001.

181. Peterson, G.E. and Barney, H.L. Control methods used in a study of the vowels. *J.Acoust.Soc.Am.* 24, 175-184. 1952.

182. Piaget, J. Construction of reality in the child. London: Routledge & Kegan Paul, 1954.

183. Piaget, J. Genetic Epistemology. W. W. Norton & Co. Scranton, Pennsylvania. 1971.

184. Picard, R. Affective Computing. MIT Press. Cambridge, MA. 1997

185. Pierce, W.D., and Cheney, C.D. Behavior Analysis and Learning. 3rd edition. Lawrence Erlbaum Associates. NJ. 2003.

186. Poppel, E. A hierarchical model of temporal perception. *Trends in Cognitive Sciences,* Volume 1, Number 2, pp. 56-61(6). May 1997.

187. Poppel, E., Held, R., and Frost. D. Residual visual function after brain wounds involving the central visual pathways in man. *Nature*, 243, 295-296. 1973.

188. Potamianos, G., Neti, C., Luettin, J., and Matthews, I. Audio-Visual Automatic Speech Recognition: An Overview. In: *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press (In Press), 2004.

189. Ratnanather, J. T., Barta, P. E., Honeycutt, N. A., Lee, N. G., Morris, H. M., Dziorny, A. C., Hurdal, M. K., Pearlson, G. D., and Miller, M. I. Dynamic programming generation of boundaries of local coordinatized submanifolds in the neocortex: application to the Planum Temporale, *NeuroImage*, vol. 20, pp. 359-377. 2003.

190. Remagnino, P., Foresti, G., and Ellis, T.J. (Eds.), Ambient Intelligence a Novel Approach. Springer-Verlag New York Inc. 2004.

191. Richards, W. (Ed.) Natural Computation. Cambridge, MA. The MIT Press. 1988.

192. Rodriguez, A., Whitson, J., Granger, R. Derivation and analysis of basic computational operations of thalamocortical circuits. *Journal of Cognitive Neuroscience*, 16: 856-877. 2004.

193. Rubin, P., Baer, T. and Mermelstein, P., An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70, 321-328. 1981.

194. Rubner, Y., Tomasi, C., and Guibas, L. J., A Metric for Distributions with Applications to Image Databases. *Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India*, ,pp. 59-66. January 1998.

195. Russell, B. Theory of Knowledge: The 1913 Manuscript, London, Boston, Sydney: George Allen and Unwin, 1984.

196. Ryle, G. The Concept of Mind. Chicago: The University of Chicago Press. Chicago, IL. 1949.

197. Saar, S. Sound analysis in Matlab. http://ofer.sci.ccny.cuny.edu/html/sam.html. 2005.

198. Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O., Lu, S., and Simola, J. Seeing speech: Visual information from lip movements modified activity in the human auditory cortex. *Neurosci. Lett.* 127:141-145. 1991.

199. Sandini G., Metta G. and Konczak J. Human Sensori-motor Development and Artificial System. In *Proceedings of AIR&IHAS '97*, Japan. 1997.

200. Schmidt, R. A., Motor control and learning, 2nd edition. Human Kinetics Publishers. 1988.

201. Shimojo, S., and Shams, L. Sensory modalities are not separate modalities: plasticity and interactions. Current Opinion in Neurobiology. 11:505-509. 2001.

202. Slaney, M. A Critique of Pure Audition. In *Proceedings of Computational Auditory Scene Analysis Workshop*. International Joint Conference on Artificial Intelligence, Montreal, Canada. 1995.

203. Slater, P. J. B., Eales, L. A., & Clayton, N. S. Song learning in zebra finches: Progress and prospects. *Advances in the Study of Behavior*, 18, 1–34. 1988.

204. Smolensky, P. . Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. Cambridge, MA: MIT Press/Bradford Books. 194-281. 1986.

205. Starner, T., Wearable Computing and Contextual Awareness. Ph.D. Thesis Massachusetts Institute of Technology. Cambridge, MA. 1999.

206. Stauffer, C. Perceptual Data Mining. Ph.D. Thesis, Massachusetts Institute of Technology. Cambridge, MA. 2002.

207. Stein, B.E. and Dixon, J.E. Superior colliculus neurons respond to noxious stimuli. *Brain Research*. 158:65-73. 1978.

208. Stein, B.E., and Meredith, M. A. 1994. The Merging of the Senses. Cambridge, MA. MIT Press.

209. Stein, R.B. The frequency of nerve action potentials generated by applied currents. *Proc. R. Soc. Lond B. Biol. Sci.*, 1967.

210. Stevens, S.S., On the psychophysical law. *Psychol. Review* 64(3):153-181, 1957.

211. Still, S., and Bialek, W. How many clusters? An information theoretic perspective, *Neural Computation*. 16:2483-2506. 2004.

212. Stork, D.G., and Hennecke, M. Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques", *Proc. of the Second Int. Conf. on Auto. Face and Gesture Recog.* Killington, VT pp. xvi--xxvi. 1996.

213. Sumby, W.H., and Pollack, I. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26:212-215. 1954.

214. Summerfield, Q. Some preliminaries to a comprehensive account of audio-visual speech perception, in Dodd, B. and Campbell, R., editors, Hearing by Eye: The psychology of lip-reading. Lawrence Erlbaum Associates, Hillsdale NJ. pgs 3-52. 1987.

215. Sun. http://www.javasoft.com/products/java-media/speech/. 2001.

216. Sung, Michael and Pentland, Alex (Sandy)., Minimally-Invasive Physiological Sensing for Human-Aware Interfaces. *HCI International.* 2005.

217. Sussman E., Winkler I., Ritter W., Alho K., and Naatanen, R., Temporal integration of auditory stimulus deviance as reflected by the mismatch negativity. *Neuroscience Letters*, Volume 264, Number 1, 2, pp. 161-164(4). April 1999.

218. Sussman, G. Abelson, H. and Sussman, J. Structure and Interpretation of Computer Programs. MIT Press. Cambridge, MA. 1983.

219. Takeuchi, A. and Amari S-I., Formation of Topographic Maps and Columnar Microstructures in Nerve Fields. 1999.

220. Taskar, B., Segal, E., and Koller, D. Probabilistic Clustering in Relational Data, In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, Seattle, Washington, August 2001.

221. Tchernichovski O, Nottebohm F, Ho C, Pesaran B, Mitra P. A procedure for an automated measurement of song similarity. *Anim Behav* 59: 1167-1176. 2000.

222. Tchernichovski, O. Private communication. 2005.

223. Tenenbaum, J. B., de Silva, V. and Langford, J. C., A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*. 290 (5500): 2319-2323, 22 December 2000.

224. Thelen, E., and Smith, L. A Dynamic Systems Approach to the Development of Cognition and Action. Cambridge, MIT Press. 1998.

225. Thompson, DW. On Growth and Form. New York: Dover Publications. 1917 revised 1942.

226. Thorndike, E. L., Animal intelligence: An experimental study of the associative process in animals. *Psychological Review Monograph* 2(8):551–53. 1898.

227. Thrun, S., Burgard, W., and Fox, D., Probabilistic Robotics. MIT Press. Cambridge, MA. 2005.

228. Tinbergen, N., The Study of Instinct, Oxford University Press, New York, NY. 1951.

229. Treisman, A., Preattentive processing in vision. *Computer Vision, Graphics and Image Processing 31*, 156–177. 1985.

230. Trussell LO. PDF Physiological mechanisms for coding timing in auditory neurons. *Ann. Rev. Physiol.* 61:477-496. 1999.

231. Ullman, Shimon. High-level vision: object recognition and visual cognition. Cambridge. MIT Press. 1996.

232. van der Meer A.L., van der Weel F.R. and Lee D.N., The functional significance of arm movements in neonates. *Science*. 267(5198):693-5. 1995.

233. Vignal, C., Mathevon, N. & Mottin, S. Audience drives male songbird response to partner's voice. *Nature* 430, 448−451. 2004.

234. Von Neumann, J. The Computer and the Brain, *Silliman Lectures Series*, Yale Univ. Press, New Haven, CT. 1958.

235. Waibel, A., Vo, M.T., Duchnowski, P., and Manke, S. Multimodal Interfaces. *Artificial Intelligence Review*. 10:3-4. p299-319. 1996.

236. Wang, F., Ma, Y.F., Zhang, H.J., Li, J.T., A Generic Framework for Semantic Sports Video Analysis Using Dynamic Bayesian Networks, pp. 115-122, 11th International Multimedia Modelling Conference (MMM'05), 2005.

237. Wang, L., Walker, V.E., Sardi, H., Fraser, C. and Jacob, T.J.C., The correlation between physiological and psychological responses to odour stimulation in human subjects. *Clinical Neurophysiology* 113, 542-551. 2002.

238. Warren, D.H., Welch, R.B. and McCarthy, T.J. The role of auditory-visual `*compellingness'* in the ventriloquism effect: Implications for transitivity amongst the spatial senses. *Perception and Psychophysics,* 30(6), pp 557- 564. 1981

239. Watanabe, S. *Pattern Recognition*. Human and Mechanical. John Wiley, New York. 1985

240. Watson, A.B., Temporal sensitivity in *Handbook of perception and human performance*. K. Boff, L. Kaufman and J. Thomas, (Eds.) Volume 1 (A87-33501 14-53). New York, Wiley-Interscience, p. 6-1 to 6-43. 1986.

241. Watts, D., and Strogatz, S. *Collective dynamics of 'small-world' networks*. Nature 393:440-442. 1998.

242. Webb, D. M., and Zhang, J. FoxP2 in Song-Learning Birds and Vocal-Learning Mammals. *Journal of Heredity*. 96: 212-216. 2005.

243. Weiner, N. Cybernetics: On Control and Communication in the Animal and the Machine. John Wiley, 1948.

244. Weiser, Mark. The Computer for the Twenty-First Century. *Scientific American.* 265(3):94—104, September 1991.

245. Weiser, Mark. The world is not a desktop. *Interactions*. pp. 7—8. January, 1994.

246. Welch, R. B., and Warren, D. H. 1986. "Intersensory Interactions." In *Handbook of Perception and Human Performance*, edited by K. R. Boff, L. Kaufman, and J. P. Thomas, chap. 25. New York: Wiley.

247. Wertheimer, M. Laws of Organization in Perceptual Forms. First published as Untersuchungen zur Lehre von der Gestalt II, in *Psycologische Forschung*, *4*, 301-350. Translation published in Ellis, W. *A source book of Gestalt psychology* (pp. 71-88). London: Routledge & Kegan Paul. 1938

248. Wiener, N. Cybernetics: or the Control and Communication in the Animal and the Machine, Cambridge: MIT Press. 1948.

249. Wiggs, C.L., and Martin, A., Properties and mechanisms of perceptual priming. Current Opinion in Neurobiology *Cognitive Neuroscience*, (8):227-233, 1998.

250. Williams, H., and Staples, K. Syllable chunking in zebra finch (Taeniopygia guttata) song. *J Comp Psychol*. 106(3):278-86. 1992.

251. Winston, P.H. Learning Structural Descriptions from Examples. MIT Artificial Intelligence Laboratory Technical Report. AITR-231. September 1970.

252. Witten, I, and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 2005.

253. Wolfe, J.M. Hidden visual processes. *Scientific American, 248(2),* 94-103. 1983.

254. Wolfe, J.M., & Cave, K.R. The psychophysical evidence for a binding problem in human vision. *Neuron, 24*, 11-17. 2000.

255. Wren, C., Azarbayejani, A., Darrell, T., and Pentland, P., "Pfinder: Real-Time Tracking of the Human Body ", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 1997.

256. Wu, Lizhong, Oviatt, Sharon L., Cohen, Philip R., Multimodal Integration -- A Statistical View, *IEEE Transactions on Multimedia*, Vol. 1, No. 4, December 1999, pp. 334-341.

257. Young, S.J. and Woodland, P.C. Hidden Markov Model Toolkit, Entropic Research Laboratory, Washington, DC, 1993.

258. Zann R. The Zebra Finch: A Synthesis of Field and Laboratory Studies. Oxford: Oxford University Press, 1996

259. Ziegler, P. & Marler, P. (Eds.) Special issue: Neurobiology of Birdsong, *Annals of the New York Academy of Science*. 1016: 348–363. 2004.

Appendix 1

Glossary

Amodal – Not relating to any particular sense directly. For example, *intensity* is amodal, because different senses can have their inputs described as being *intense*, whereas *purple* is not an amodal attribute because it applies only to visual inputs.

Artificial perceptual system – A computational system that perceives real-world phenomena. For example, speech recognition and computer vision systems are examples of artificial perceptual systems.

Cross-Modal – Any phenomenon that involves information being shared between seemingly disparate perceptual channels. In this thesis, these channels may fall within the same gross modality, such as *vision*, as long as different *modes* are involved.

Intersensory – Used interchangeably with *cross-modal*.

Modality – A sense or perceptory capability, such as touch, vision, etc. Generally refers to an entire class of such capabilities, such as the *vision* modality, which is comprised of a range of individual visual capabilities, such as *color sensitivity*, *peripheral vision*, *motion detection*, etc., all of which share a common sensory organ.

Mode – A highly specific sense or perceptory capability. Unlike *modality*, modes refer to the specific capabilities of particular senses, e.g., the *sweetness* mode of taste or the *color sensitivity* mode of vision, etc. This definition is elaborated upon in §1.5.

Multimodal – A perceptory phenomenon involving multiple modes simultaneously. For example, *hand clapping* is multimodal, because it can be both seen and heard. For that matter, most real-world events are multimodal, because they generate multiple types of energy simultaneously to which different sensory channels are receptive. Also used to describe *artificial perceptual systems* that are sensitive to multiple modalities; these are typically called *multimodal systems*.

Sense – In this thesis, a *sense* is used equivalently with a *mode*, as defined above. In other words, a sense refers to a highly specific perceptual capability.

System – A computer system unless context indicates biological.

Unimodal – A perceptory phenomenon or quality involving only a single modality. For example, speech recognition systems are generally considered unimodal, because they only perceive spoken language inputs. The designation can be confusing, however, because many unimodal systems have traditional graphical user interfaces (GUI) that provide inputs outside of their perceptual channel. They are nonetheless still deemed unimodal, because GUI inputs are non-perceptual.