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

> 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.

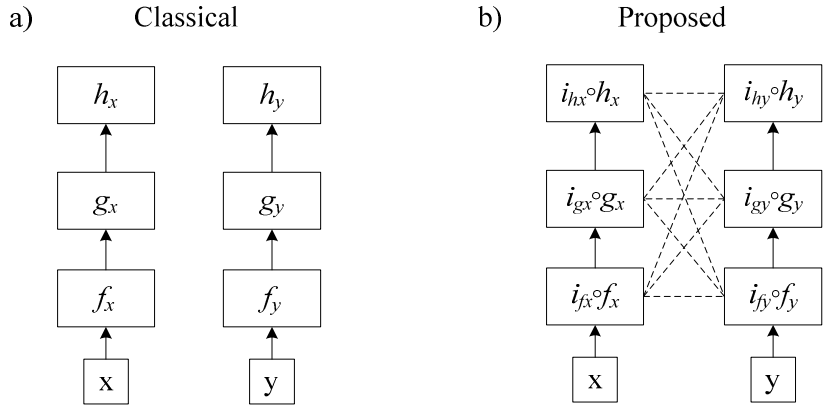

Figure 1.1– Adding an influence network to two preexisting systems. We start in (a) with two pipelined networks that independently compute separate functions. In (b), we compose on each function a corresponding *influence function*, which dynamically modifies its output based on activity at the other 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).

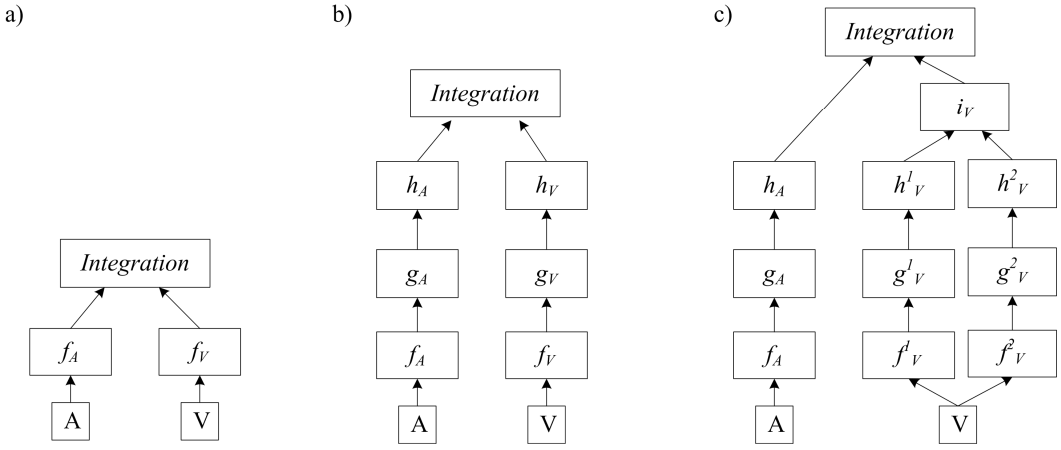

Figure 1.2 – Classical approaches to post-perceptual integration in traditional multimodal systems. Here, auditory (A) and visual (V) inputs pass through specialized unimodal processing pathways and are combined via an integration mechanism, which creates multimodal perceptions by extracting and reconciling 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

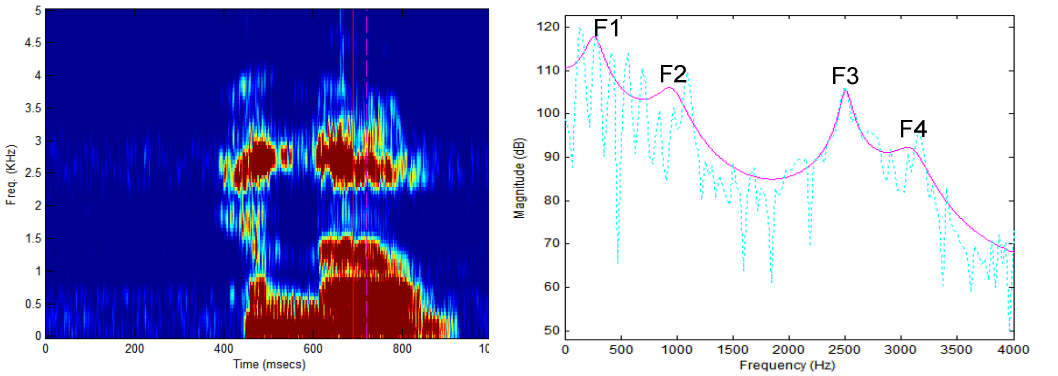

Figure 2.1— On the left is a spectrogram of the author saying, "Hello." The demarcated region (from 690-710ms) marks the middle of phoneme /ow/, corresponding to the middle of the vowel "o" in "hello." The spectrum corresponding to this 20ms window is shown on the right. A 12th order linear predictive 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

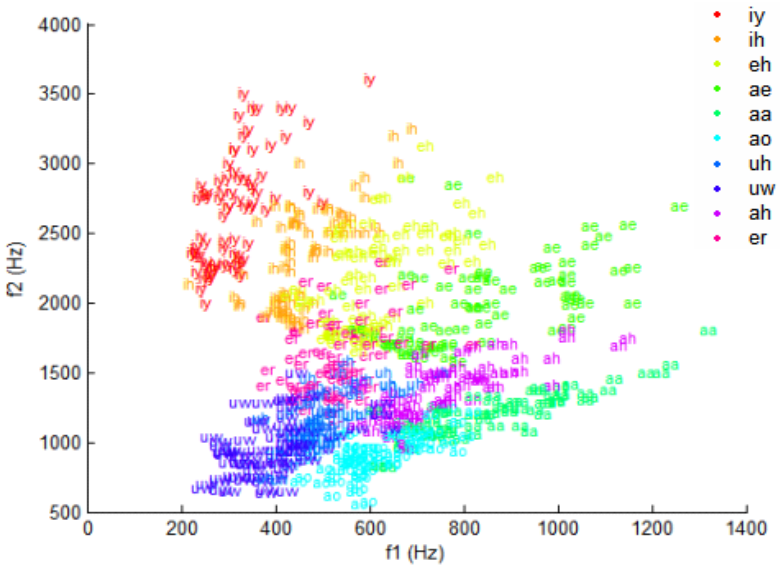

Figure 2.2 – Peterson and Barney Data. On the left is a scatterplot of the first two formants, with different regions labeled by their corresponding vowel categories.

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

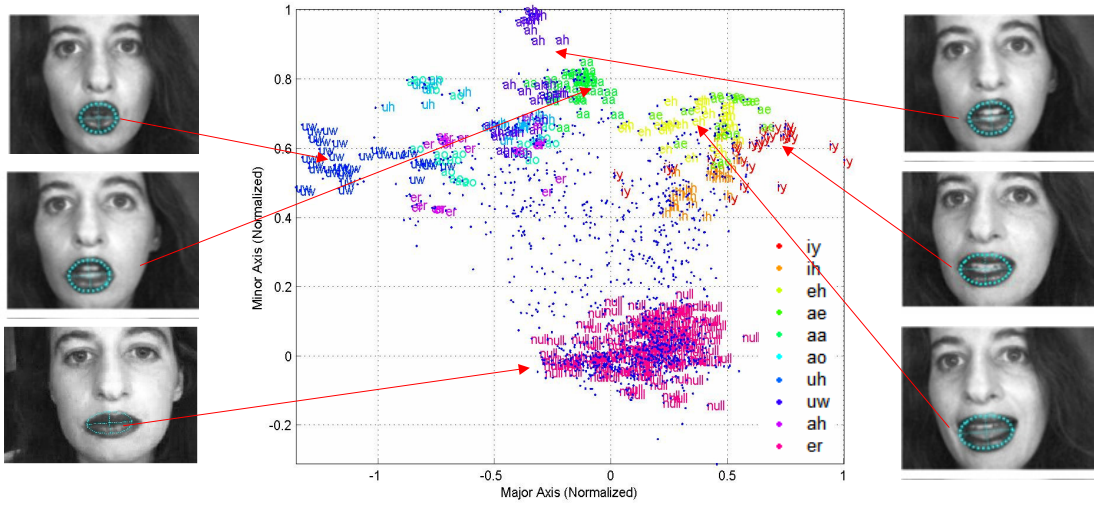

Figure 2.4 -- Modeling lip contours with an ellipse. The scatterplot shows normalized major (x) and minor (y) axes for ellipses corresponding to the same vowels as those in Figure 2.2. In this space, a closed mouth corresponds to a point labeled *null*. Other lip contours can be viewed as offsets from 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

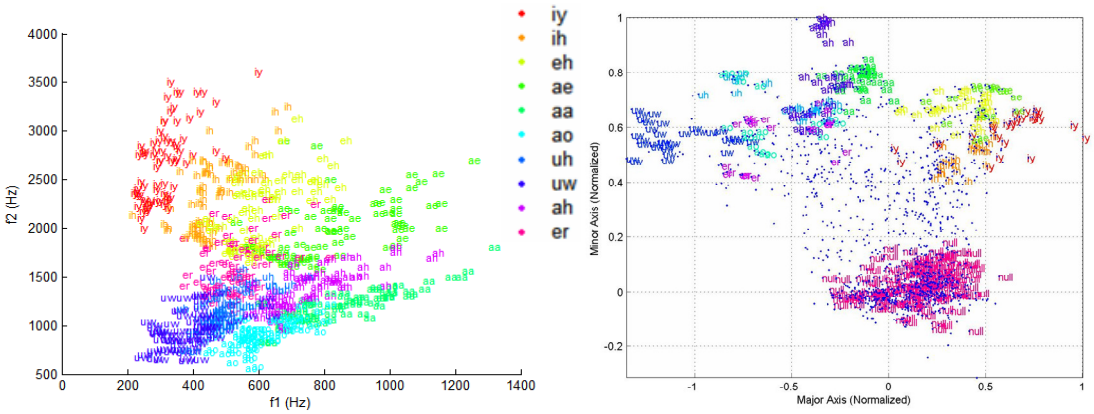

Figure 2.5 – Labeled scatterplots side-by-side. Formant data (from Peterson Barney 1952) is displayed on the left and lip contour data (from the author's wife) is show on the right. Each plot contains data corresponding to the ten listed vowels in American English.

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*

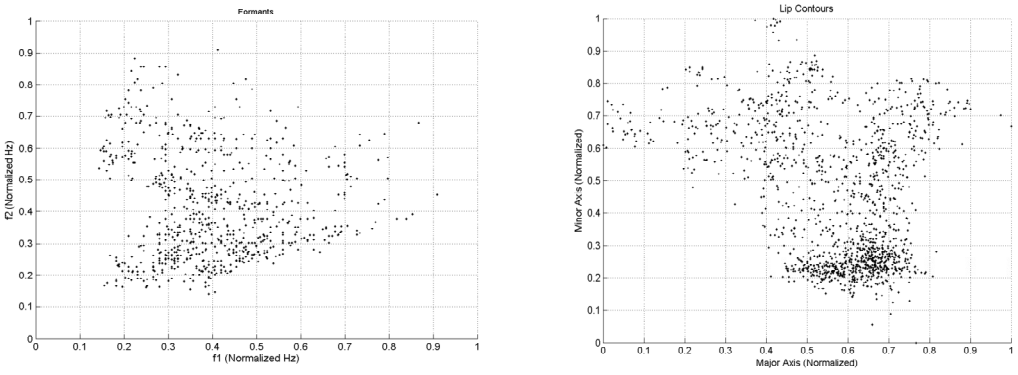

Figure 2.6 – Unlabeled data. These are the same data shown above in Figure 2.5, with the labels removed. This picture is closer to what animals actually encounter in Nature. As above, formants are displayed on the left and lip contours are on the right. Our goal is to learn the categories present in 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

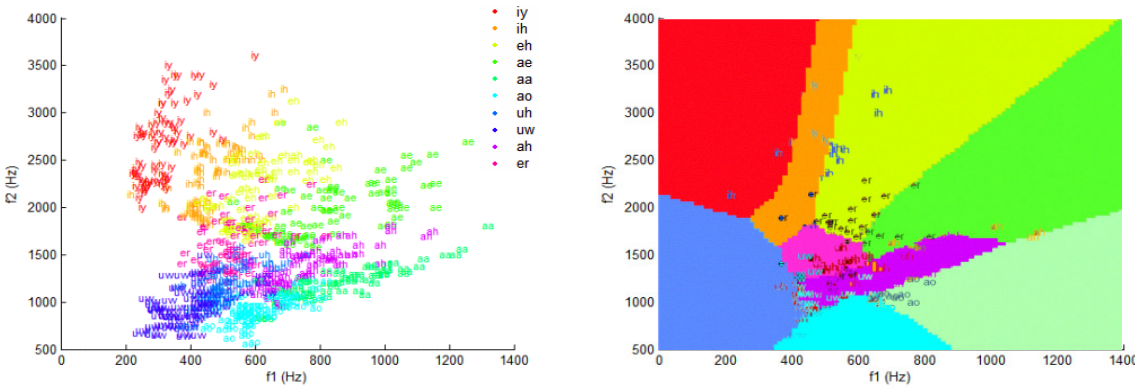

Figure 2.7 – On the left is a scatterplot of the first two formants, with different regions labeled by their corresponding vowel categories. The output of a backpropagation neural network trained on this data is shown on the right and displays decision boundaries and misclassified points. The misclassification 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

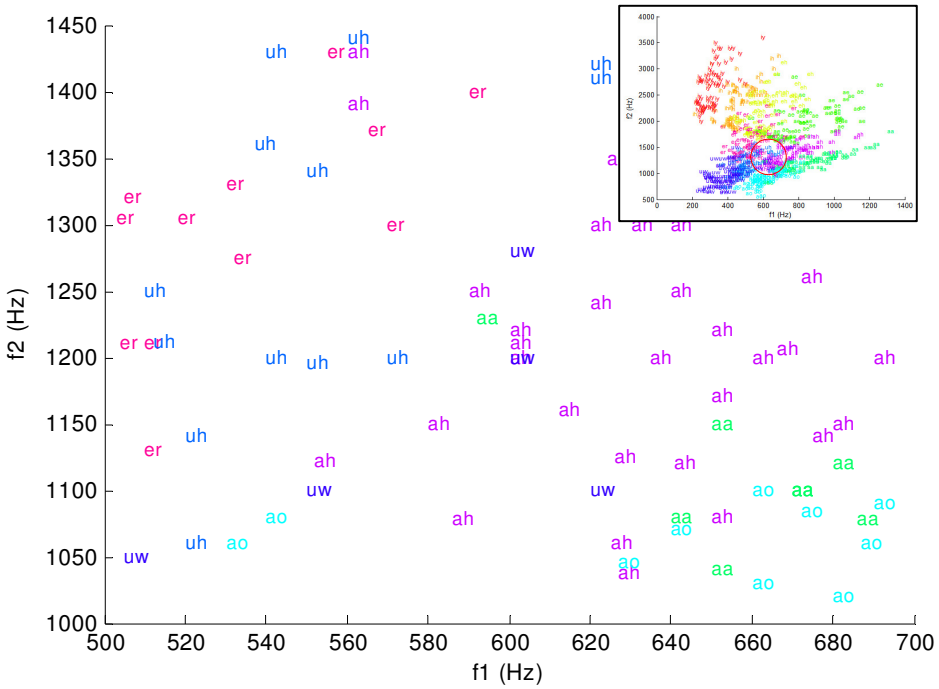

Figure 2.8 – Focusing on one of many ambiguous regions in the Peterson-Barney dataset. Due to a confluence of factors described in the text, the data in these regions are not obviously separable.

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.