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 English vowels. They measured the fundamental frequency and first three formants

---

**Figure 1** -- On the left is a spectrogram of the author saying, “Hello.” The demarcated region (from 690-710ms) marks the onset of phoneme /ao/, corresponding to the start of the vowel "o" in “hello.” The spectrum corresponding to this 20ms window is shown on the right. A 12th order 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 formant determination.
(see Figure 1) for each sample and noticed that when plotted in various ways (Figure 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. (viz. 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 literature (e.g., Mase and Pentland 1990, Huang et al. 2003, Potamianos et al.

---

**Figure 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.
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 3 contains images of a speaker voicing different sustained vowels, corresponding to those in Figure 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 4 shows how a speaker’s mouth is represented in this way, with contour data normalized such that a resting mouth configuration

Figure 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.
null (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 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 in Figure 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 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 6) and deriving the kinds of correspondences and segmentations (as shown in

Figure 5 – Labeled scatterplots side-by-side. Formant data is displayed on the left and lip contour data is show on the right. Each plot contains data corresponding to the ten listed vowels in American English.

Figure 6 – Unlabeled data. These are the same data shown above in Figure 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.
Figure 5) without external supervision. This is what we mean by *perceptual grounding* 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. We will examine some of these phenomena in more detail in Chapter 6, but we note here that 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 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 difficulty 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 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 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

---

**Figure 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.
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 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 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 5 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.