Self-Supervised Acquisition of Vowels in American English

Michael H. Coen
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street
Cambridge, MA 02139
mhcoen@csail.mit.edu

Abstract
This paper presents a self-supervised framework for perceptual learning based upon correlations in different sensory modalities. We demonstrate this with a system that has learned the vowel structure of American English – i.e., the number of vowels and their phonetic descriptions – by simultaneously watching and listening to someone speak. It is highly non-parametric, knowing neither the number of vowels nor their input distributions in advance, and it has no prior linguistic knowledge. This work is the first example of unsupervised phonetic acquisition of which we are aware, outside of that done by human infants. This system is based on the cross-modal clustering framework introduced by [4], which has been significantly enhanced here. This paper presents our results and focuses on the mathematical framework that enables this type of intersensory self-supervised learning.

Introduction
This paper presents 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 show that sensory systems can be perceptually grounded by mutually bootstrapping off each other.

As a demonstration, 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 [4] 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 output of this system is displayed in Figure 1. The goal of this paper is to elaborate upon these results and outline the framework through which they were obtained.

Our approach to perceptual grounding has been to mathematically formalize an insight in Aristotle's De Anima [1], that differences in the world are only detectable because different senses perceive the same world events differently. This implies both that sensory systems need 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. This insight was the basis for the cross-modal clustering framework in [4], which is the foundation for this work and is significantly enhanced here. This approach has been motivated by recent results in the cognitive and neurosciences [13,2,12] detailing the extraordinary degree of interaction between modalities during ordinary perception. These biological motivations are discussed at length in [3]. We believe 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. This is an important problem in computer science, cognitive science, and neuroscience.

We proceed by first defining what is meant by the word "sense." We then introduce our application domain and discuss why perceptual grounding is a difficult problem. Finally, we present our enhancements to cross-modal clustering and demonstrate how the main results in this paper were obtained. We note that the figures in this paper are most easily viewed in color.

**What Is a "Sense?"**

We have 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. 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 [15], thus demonstrating they are phenomenologically independent.

**Problem Statement**

Our demonstration for perceptual grounding has been inspired by the classic study of Peterson and Barney [10], who studied recognition of spoken vowels (monophthongs) in English according to their formant frequencies. (An explanation of formant frequencies is contained in Figure 2.) Their observation that formant space could be approximately partitioned for vowel identification, as in Figure 3, was among the earliest approaches to spectral-based speech understanding. The corresponding classification problem remains a popular application for machine learning, e.g., [6].

It is well known that acoustically ambiguous sounds tend to have visually unambiguous features. 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, as shown in Figure 4. The images are taken from a mouth tracking system written by the author, where the mouth position is modeled by the major and minor axes of an ellipse fit onto the speaker's lips.

In Figure 5A, we examine formant and lip data side-by-side, in color-coded, labeled scatterplots over the same set of 10 vowels in American English. We note that ambiguous regions in one mode tend to be unambiguous in the other and vice versa. It is easy to see how this type of intersensory disambiguation could enhance speech recognition, which is a well-studied computational problem [11].
We are interested here, however, in a more fundamental problem: how do sensory systems learn to segment their inputs to begin with? In the color-coded plots in Figure 5A, 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 5B 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 5B) and deriving the kinds of correspondences and segmentations (as shown in Figure 5A) 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.

**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 are unknown. Clustering algorithms generally make strong assumptions about one or both of these and when faced with nonparametric, distribution-free data, algorithmic clustering techniques tend not be robust [7,14].

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.

**The Simplest Complex Example**

We proceed by means of an example. 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 in Figure 6, 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 (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
input data to define the distance metric within each mode.

Our approach
We would like to assemble the clusters within each slice
into larger regions that represent actual perceptual
categories present in the input data. Consider the colored
regions in Figure 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. We proceed
by formulating a metric that minimizes the distance
between codebook regions that are actually within the
same perceptual region and maximizes the distance
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.

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, where connections
between regions are strengthened as they are
simultaneously active. The result of this process is
illustrated in Figure 9, where the slices 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.

Hebbian Projections
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 \( C_A = \{ p_1, p_2, \ldots, p_n \} \) and \( C_B = \{ q_1, q_2, \ldots, q_n \} \), with
cluster centroids \( p, q \in \mathbb{R}^n \). We define the Hebbian
projection of a \( p \in C_A \) onto mode \( M_B \):

\[ p \rightarrow M_B(p) \]

\[ p \rightarrow \text{closest point in } M_B \text{ to } p \]
A Hebbian projection is simply a conditional spatial probability distribution that lets us know what mode $M_d$ probabilistically "looks" like when a region $p_i$ is active in co-occurring mode $M_s$. This is visualized in Figure 10.

We can equivalently define a Hebbian projection for a region $r \subseteq M_s$ constructed out of a subset of its codebook clusters $C_r = \{p_{1r}, p_{2r}, ... , p_{nr}\} \subseteq C_s$:

$$\hat{H}_s^A(p_i) = \left[ \Pr(q_1 | p_i), \Pr(q_2 | p_i), ..., \Pr(q_n | p_i) \right]$$

where the infimum is taken over all joint distributions $J$ on $x$ and $y$ with marginals respectively. In this paper we assume that $d$, the metric on $\Omega$, is Euclidean.

$D_{OTO}$ is a novel metric called the one-to-many distance. Let $f$ and $g$ be the density functions of $\mu$ and $\nu$ respectively. Then the one-to-many distance between $\mu$ and $\nu$ is:

$$D_{OTO}(\mu, \nu) = \int \sqrt{f(x) \cdot D_W(x, \nu) \cdot dx} = \int \int f(x) \cdot g(y) \cdot d(x, y) \cdot dydx = \int g(y) \cdot D_W(\mu, y) \cdot dy = D_{OTO}(\nu, \mu)$$

Further details of these metrics, including their definitions over discrete distributions and their computational complexities, are contained in [5].

For the results below, we replace the cross-modal distance metric in [4] with Similarity distance $D_s$ and use the same cross-modal clustering algorithm.

**Experimental Results**

To learn the vowel structure of American English, data was gathered according to the same pronunciation protocol employed by [10]. Each vowel was spoken within the context of an English word beginning with [h] and ending with [d]; for example, /æ/ 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, using a 30ms FFT window and a 12th order LPC model. Lip contours were extracted using the system described above. Time-stamped formant and lip contour data were fed into slices in an implementation of the work in [4], using the Similarity distance defined above. We note this
implementation was used to generate most of the figures in this paper, which represent actual system outputs. The results of this experiment are shown in Figures 1 and 12. This is the first unsupervised acquisition of human phonetic data of which we are aware. The work of de Sa [6] has studied unsupervised cross-modal refinement of perceptual boundaries, but it requires that the number of categories be known in advance. We note also there is a vast literature on unsupervised clustering techniques, but these generally make strong assumptions about the data being clustering or they have no corresponding notion of correctness associated with their results. The intersensory approach taken here is entirely non-parametric and makes no a priori assumptions about underlying distributions or the number of clusters being represented.

Acknowledgements

The author is indebted to Whitman Richards, Howard Shrobe, Patrick Winston, and Robert Berwick for encouragement and feedback. This work is sponsored by AFRL under contract #FA8750-05-2-0274. Thanks to the DARPA/IPTO BICA program and to AFRL. The author also thanks the anonymous reviewers for their highly insightful comments.

References