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

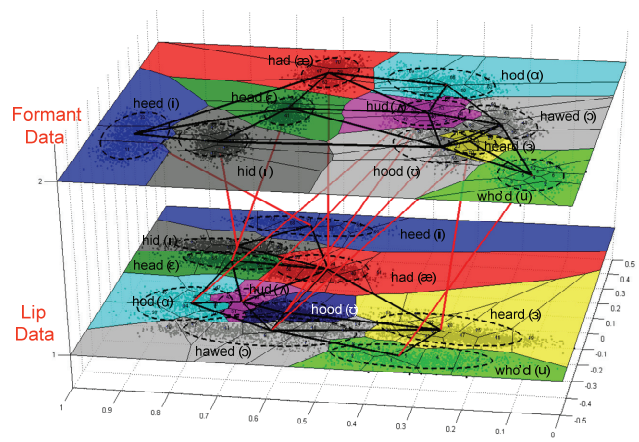

Figure 1 – Mutual bootstrapping through cross-modal clustering. This figure shows we can learn the number and structure of vowels in American English by simultaneously watching and listening to someone speak. Auditory formant data is displayed on top and visual lip data – corresponding to major and minor axes of an ellipse fit on the mouth – is on the bottom. Initially, nothing is known about the events these systems perceive. Cross-modal clustering lets them mutually structure their perceptual representations and thereby learn the event categories that generated their sensory inputs. The region colors show the correspondences obtained from cross-modal clustering. Red lines connect corresponding vowels between the two datasets and black lines show neighboring regions within each dataset. The phonetic labels were manually added to show identity. The data are from a real speaker and were normalized.

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.

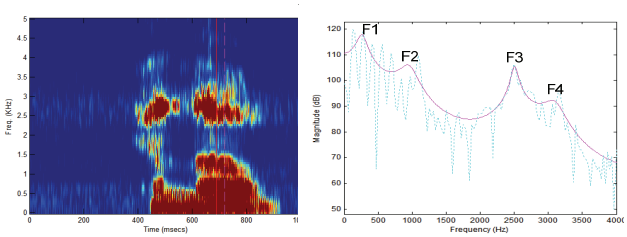

Figure 2 – On the left is a spectrogram of the author saying, "hello." The demarcated region (from 690-710ms) marks the onset of phoneme /ao/, corresponding to the start of the vowel "o" in "hello." The spectrum corresponding to this 20ms window is shown on the right. A 12th order LPC model is shown 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.

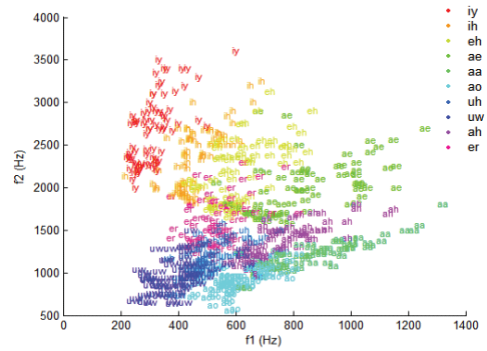

Figure 3 – Peterson and Barney Data. A scatterplot of the first two formants, with different regions labeled by their corresponding vowel categories.

Therefore, we prefer a finer grained approach to perception. From this perspective, intersensory influence can happen between modes within the same sensory system, e.g., entirely within vision, or 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, it treats both cases equivalently.

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].

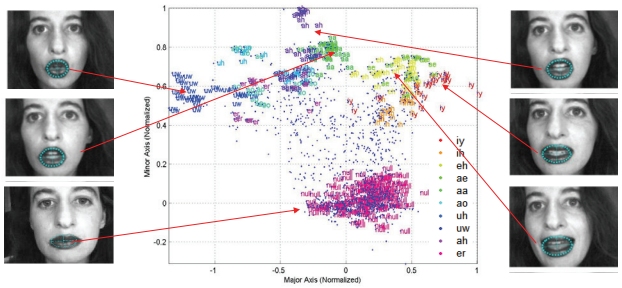

Figure 4 – Modeling lip contours with ellipses. The scatterplot shows normalized major (x) and minor (y) axes for ellipses corresponding to the same vowels as those in Figure 3. In this space, a closed mouth corresponds to a point labeled null. Other lip contours can be viewed as offsets from the null configuration and are shown here segmented by color. These data points were collected from video of this woman speaking.

Nature Does Not Label Its Data

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

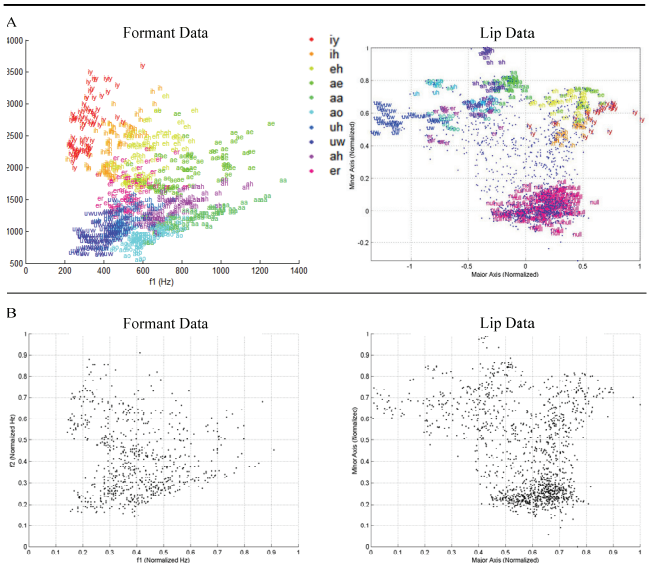

Figure 5A (top): Labeled scatterplots side-by-side. Formant data is displayed on the left and lip contour data is on the right. Each plot contains data corresponding to the ten listed vowels in American English

Figure 5B (bottom): Unlabeled data. These are the same data shown in Figure 5A, 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 shown directly above.

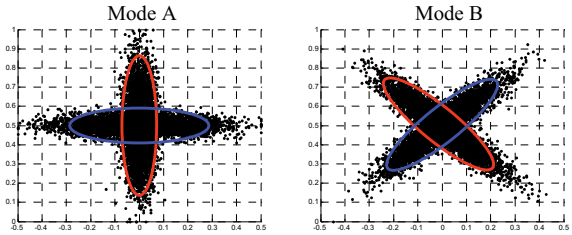

Figure 6 – Two hypothetical co-occurring perceptual modes. Each mode, unbeknownst to itself, receives inputs generated by a simple, overlapping Gaussian mixture model. To make matters more concrete, we might imagine Mode A is a simple auditory system that hears two different events in the world and Mode B is a simple visual system sees those same two events, which are indicated by the red and blue ellipses.

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 spatiotemporal correlations between them.

Defining Slices

Our approach is to represent the modes' perceptual inputs within *slices* [4,5]. Slices are a convenient way to discretely model perceptual inputs (see Figure 7) and are inspired by surface models of cortical tissue. Formally, they are topological manifolds that discretize data within Voronoi partitionings, where the regions' densities have been normalized.

Intuitively, a slice is a codebook [8] 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.

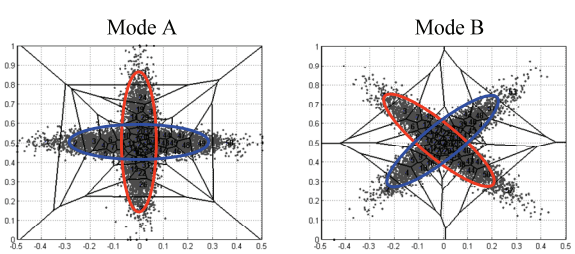

Figure 7 – Slices generated for Modes A and B using the hyperclustering algorithm in [5]. We refer to each Voronoi cluster within a slice as a *codebook* region.

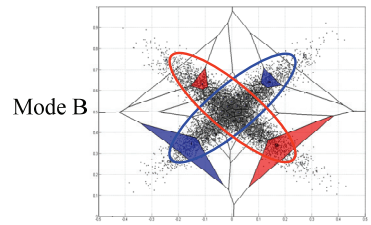

Figure 8 – Combining codebook regions within a slice to construct perceptual regions. We would like to determine that regions within each ellipse are all part of the same perceptual event. Here, for example, the two blue codebook regions (probabilistically) correspond to the blue event and the red regions correspond to the red event.

We will refer to each individual cluster within a slice as a *codebook region*, and will define the non-Euclidean distance metric between them below.

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 $C_A = \{p_1, p_2, ..., p_a\}$ and $C_B = \{q_1, q_2, ..., q_b\}$, with cluster centroids $p_i, q_j \in \mathbb{R}^N$. We define the *Hebbian projection* of a $p_i \in C_A$ onto mode M_B:

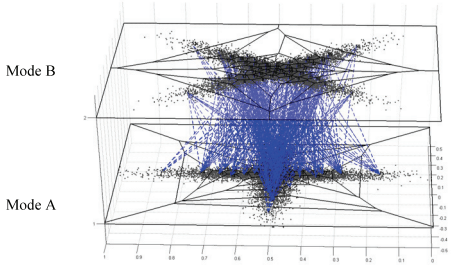

Figure 9 – Viewing Hebbian linkages between two different slices. The slices have been vertically stacked here to make the correspondences clearer. The blue lines indicate that two codebook regions temporally co-occur with each other. Note that these connections are weighted based on their strengths, which are not visually represented here, and that these weights are additionally asymmetric between each pair of connected regions.

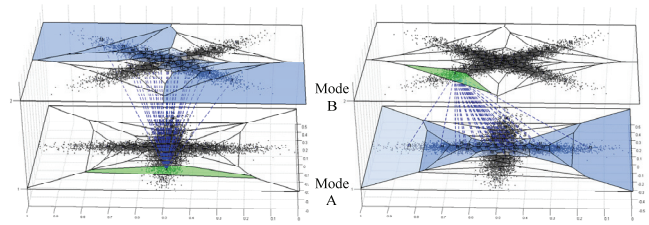

Figure 10 – Visualizations of Hebbian projections. On the left, we project from a cluster p_i in Mode A onto Mode B. The dotted lines correspond to Hebbian linkages and the blue shading in each cluster q_i in Mode B is proportional to $\Pr(q_i|p_i)$. A Hebbian projection lets us know what Mode B probabilistically "looks" like when some prototype in Mode A is active. On the right, we see a projection from a cluster in Mode B onto Mode A.

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

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 10.

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

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

A Cross-Modal Distance Metric

We use the Hebbian projections defined in the previous section to define the distance between 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 determine the distance between codebook regions by comparing their Hebbian projections onto co-occurring slices. This process is illustrated in Figure 11.

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.

Instead, we use the notion of *Similarity distance* defined in [5], which measures the similarity between distributions on a metric space. Let μ and ν be distributions on state space $\Omega = \mathbb{R}^n$, corresponding to Hebbian projections. The Similarity distance between μ and ν is:

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

Here, D_W is the Kantorovich-Wasserstein distance [9]:

$$D_W(\mu,\nu) = \inf_J \{d(x,y) : L(x) = \mu,\ L(y) = \nu\}$$

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 Ω, is Euclidean.

D_{OTM} is a novel metric called the *one-to-many distance*. Let f and g be the density functions of μ and ν respectively. Then the one-to-many distance between μ and ν is:

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

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, /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, 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

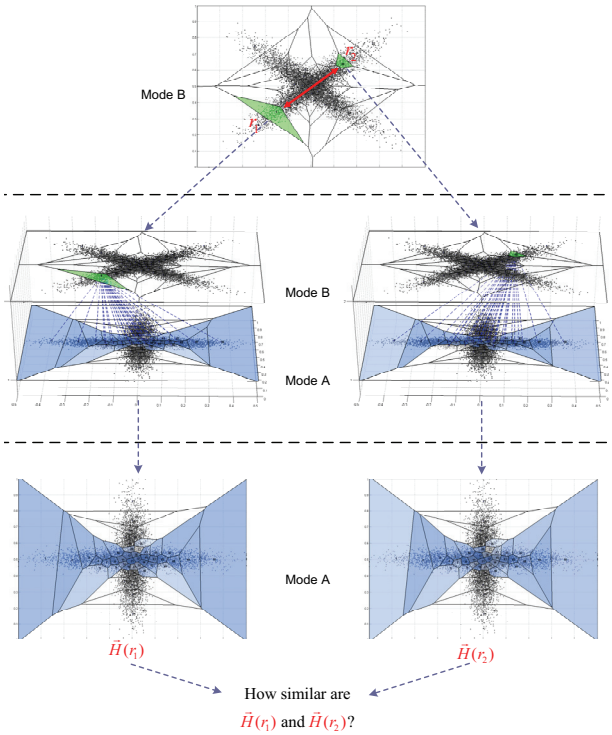

Figure 11 – Our approach to computing distances cross-modally. To determine the distance between codebook regions r_1 and r_2 in Mode B on top, we project them onto a co-occurring modality Mode A as shown in the middle. We then ask how similar their Hebbian projections onto Mode A are, as shown on the bottom. We have thereby transformed our question about distance between regions into a question of similarity between the spatial probability distributions provided by their Hebbian projections. This is computed via their Similarity distance D_S .

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.

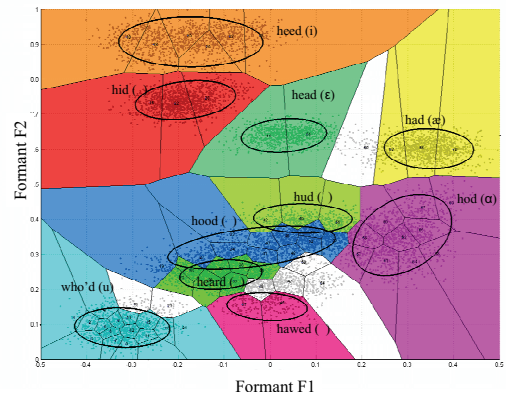

Figure 12 – Self-supervised acquisition of vowels (monophthongs) in American English. The identifying labels were manually added for reference and ellipses were fit onto the regions to aid visualization. Unlabeled regions have ambiguous classifications. All data have been normalized. Note the correspondence between this and the Peterson-Barney data shown in Figure 3.

References

1. Aristotle. De Anima. 350 BCE. Translated by Tancred, H.L. Penguin Classics. London. 1987.
2. Calvert, A.G., Spence, C., and Stein, B.E. The Handbook of Multisensory Processes. Bradford Books. 2004.
3. Coen, M.H. Multimodal interaction: a biological view. In *Proceedings of 17th International Joint Conference on Artificial Intelligence. (IJCAI-01).* Seattle, Washington. 2001.
4. Coen, M.H. Cross-Modal Clustering. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI'05),* pp. 932-937. Pittsburg, PA. 2005.
5. Coen, M. H. Multimodal Dynamics: Self-Supervised Learning in Perceptual and Motor Systems. Ph.D. Thesis. Massachusetts Institute of Technology. 2006.
6. de Sa, V.R. Unsupervised Classification Learning from Cross-Modal Environmental Structure. Doctoral Dissertation, Department of Computer Science, University of Rochester. 1994.
7. Fraley, C. and Raftery, A.E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97, 611-631.
8. Gray, R.M. Vector Quantization. *IEEE ASSP*, pp. 4--29, April 1984.
9. Kantorovich L., On the translocation of masses, C. R. Acad. Sci. URSS (N.S) 37:199-201. 1942. (Republished in *Journal of Mathematical Sciences*, Vol. 133, No. 4, 2006. Translated by A.N. Sobolevskii.)
10. Peterson, G.E. and Barney, H.L. Control methods used in a study of the vowels. *J.Acoust.Soc.Am*. 24, 175-184. 1952.
11. 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. 2004.
12. Spence, C., & Driver, J. (Eds.) . Crossmodal space and crossmodal attention. Oxford, UK: Oxford University Press. 2004.
13. Stein, B.E., and Meredith, M. A. The Merging of the Senses. Cambridge, MA. MIT Press. 1994.
14. Still, S., and Bialek, W. How many clusters? An information theoretic perspective, *Neural Computation*. 16:2483-2506. 2004.
15. Wolfe, J.M. Hidden visual processes. *Scientific American*, 248(2), 94-103. 1983.