Learning to Sing Like a Bird: Self-Supervised Acquisition of Birdsong

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 new framework for *self-supervised* sensorimotor learning. We demonstrate this framework with a system that learns to mimic a zebra finch, directly modeled on the dynamics of how male fledglings acquire birdsong from their fathers. Our system first listens to the song of an adult finch. By listening to its own initially nascent attempts at mimicry through an articulatory synthesizer, the system organizes motor maps generating its vocalizations. Our approach is founded on the notion of *cross-modal clustering*, introduced in (Coen 2005, 2006a), and is unusual for its recursive reuse of perceptual mechanisms in developing motor control. In this paper, we outline this framework, present its results on the unsupervised acquisition of birdsong, and discuss other potential applications.

Introduction

This paper presents a novel computational architecture for sensorimotor learning. In our framework, an agent acquires motor control by learning to imitate the observed, learned behaviors of an external party, such as a parent. The approach presented here is *self-supervised*, meaning no supervisory (corrective) signal is provided to the agent nor are any statistical models describing the expected inputs (or desired outputs) presumed. Instead, the system *supervises its own learning* by observing the effects of its activities and correlating them with the learned behaviors of an external agent.

A power of this mechanism is that it can learn mimicry, a basic form of behavioral learning in which 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 (Meltzoff and Prinz 2002). The developmental notion of *programming by example* is ubiquitous in the animal kingdom (Galef 1988), and this work is a step toward providing this capability to artificial sensorimotor systems.

Furthermore, imitative learning is thought to be among the most powerful developmental mechanisms available both to natural creatures (Thorndike 1898, Piaget 1971) and to artificial ones (Dautenhahn and Nehaniv 2002). The benefits it provides in perceptual and sensorimotor domains are significant because engineered approaches tend to be ad hoc and error prone (Coen 2001); additionally, in sensorimotor learning *we generally have no adequate models to specify the desired input/output behaviors for our systems*. Often, the most transparent and concise specification of a set of desired actions is obtained through observations of another entity performing them, rather than through abstract formalisms.

Our approach recursively applies the perceptual grounding framework of *cross-modal clustering* (Coen 2005, 2006a). This algorithm learns categories in datasets non-parametrically and without assuming any underlying input distributions. It does so by exploiting spatiotemporal redundancies between different – but perceptually overlapping – sensory modalities to learn the number and structure of events they jointly perceive. We briefly motivate and review cross-modal clustering below.

In this paper, we extend this framework by treating the motor component of sensorimotor learning *as if* it were a perceptual problem. This is surprising because one might suppose that motor activity is fundamentally different than perception. However, we take the perspective that motor control can be seen as perception *backwards*. We imagine that – in a notion reminiscent of a Cartesian theater (Dennett 1991) or a Global Workspace (Baars 1997) – 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

Figure 1 – An adult male zebra finch (*Taeniopygia guttata*). Zebra finches are small, social songbirds and are extremely popular for studying neural, physiological, evolutionary, social, and developmental aspects of birdsong acquisition.

In *Proceedings of the Twenty Second National Conference on Artificial Intelligence* (AAAI'07). Vancouver, British Columbia. 2007.

perceptual grounding framework described in (Coen 2006a, Chapters 2 & 3) can then *cross-modally cluster* these internal and external perceptions with respect to one another. The insight behind this approach is that *an agent can develop motor control by learning to mimic the events it has previously learned through unsupervised observations of its world.*

We demonstrate sensorimotor learning in this framework with an artificial system that learns to sing like a fledgling zebra finch. Our system first listens to the song of an adult male zebra finch; it cross-modally clusters this input to learn *songemes*, primitive units of bird song that we propose as avian analogs of phonemes. It then uses a vocalization synthesizer to generate its own nascent birdsong, guided by random exploratory motor behaviors. By simultaneously listening to itself sing, the system organizes its motor maps by cross-modally clustering them with respect to the previously acquired perceptual maps. During this process, the fact that the motor data were derived internally from innate exploratory behaviors rather than from external perceptual events is irrelevant. By treating the motor data as if they were derived perceptually from the outside world, the system thereby learns to reproduce the same sounds to which it was previously exposed.

Our goal in this paper is to demonstrate that motor learning can be accomplished through recursive perceptual grounding. In other words, we show that perceptual and motor learning are unexpectedly similar processes and can be achieved within a common mathematical framework. This surprising result also suggests an approach to grounding higher level cognitive development, by continually reapplying this technique of internal perception, as suggested by the metalinguistic theories of Lakoff (1987).

Cross-Modal Clustering

We briefly review cross-modal clustering, which is a biologically motivated algorithm for learning categories based upon correlations in different sensory modalities. This approach was inspired by the notion that the sensory information gathered by animals is inherently redundant (Richards 1980). This redundancy can enable learning without requiring explicit teaching (as in supervised learning) or statistical modeling (as in unsupervised learning). In Nature, these are frequently unavailable and yet animals learn anyway. In other words, redundancy allows animals to supervise their own learning, hence the designation of *self-supervised* learning.

For example, (Coen 2006b) has shown cross-modal clustering can learn the vowel structure of American English by simultaneously watching and listening to someone speak. It learns the number of vowels and their formant structure without any *a priori* linguistic knowledge and receives no information outside of its perceptual channels. One may contrast this with the large number of clustering algorithms that require harsh

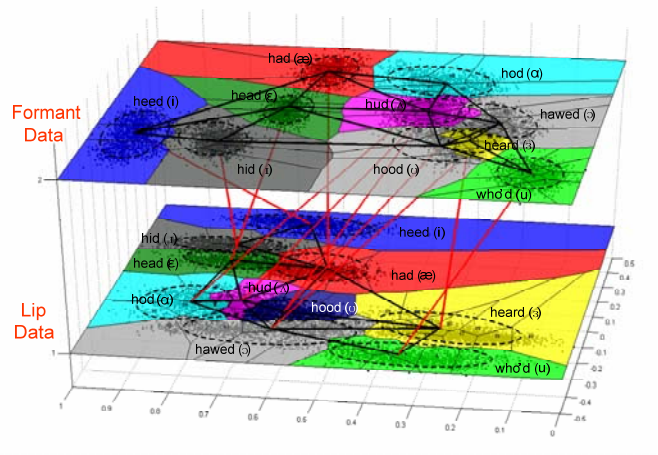

Figure 2 – Perceptual grounding through cross-modal clustering. This figure shows we can learn the number and structure of vowels in American English by simultaneously watching and listening to someone speak. Auditory formant data is displayed on top and visual lip data – corresponding to major and minor axes of ellipses fit on the speaker's 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. This figure is taken from (Coen 2006b).

simplifying assumptions (e.g., the clusters are Gaussian) or careful setting of parameters requiring advance knowledge of the data (see the surveys in Jain et al. 1999, Weiss 1999). In contrast, cross-modal clustering can learn categories in nonparametric data, *where we know neither the number of represented categories nor their input distributions.* However, it requires that we have multiple modalities that simultaneously view events in the world; in other words, each modality possesses a unique viewpoint of the same world events.

The cross-modal clustering algorithm precedes by hyperclustering the inputs within each dataset into a large number of Voronoi regions: this means we construct a Voronoi partitioning of each dataset with a far larger number of clusters than we assume exists in the data. For each Voronoi region, the algorithm calculates its projection onto another modality, based on simple temporal co-occurrences. These projections are thus spatial probability distributions onto the multi-dimensional metric space defined by each dataset. By comparing the distances between these projected distributions, the algorithm merges Voronoi regions (within a dataset) that have "similar" projections onto other modalities; in this way, it generates larger perceptual regions. In essence, it merges regions within one modality that appear conditionally identical from the perspective of other modalities.

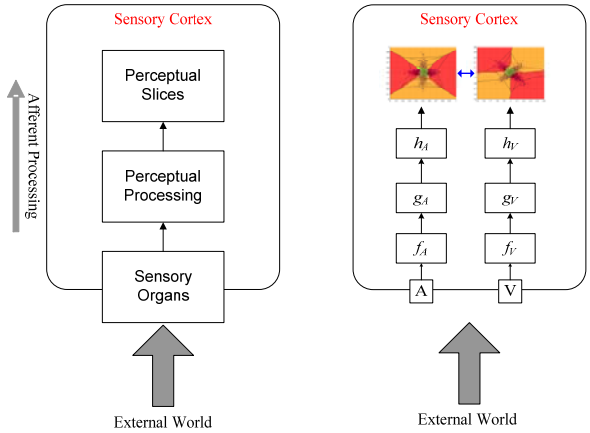

Figure 3 – An abstract model of sensory processing in our framework. A schematic view is shown on the left, which is expanded upon in the example of the right. Events in the world are detected by sensory organs, here labeled A and V, representing auditory and visual receptors. These are fed into processing pipelines shown here by the composition of functional units. The features extracted from these pipelines are fed into slices, which are then cross-modally clustered with respect to one another.

Importantly, we note that the notion of modality here does not necessarily correspond to an entire sensory system, and cross-modal clustering can be applied to sets of processed data within the same sensory channel. For example, in learning birdsong, we will focus exclusively on auditory features. The representational primitive in cross-modal clustering for representing data is a structure called a *slice*. For example, Figure 2 contains two 2-D slices corresponding to auditory and visual datasets. Slices are manifolds that we will use to represent both sensory and motor maps in this paper. We equivalently use the terms "slice" and "map" in this paper to refer to the same type of object.

A Sensorimotor Architecture

We begin by examining abstract models of innate sensory and motor processing in isolation. Afterwards, we integrate them to enable sensorimotor learning.

A Simple Model of Innate Sensory Perception

Our framework begins with the model of afferent sensory perception outlined in Figure 3, which schematically diagrams an abstract computational sensory cortex. In this model, external events in the world impinge upon sensory organs. These receptors in turn generate perceptual inputs, which feed into specialized perceptual processing channels. A primary outcome of this processing is the extraction of descriptive features, which capture abstract sensory detail. This process occurs in parallel within multiple sensory pipelines, as illustrated in Figure 3 on the right. This hypothetical example shows auditory and visual receptors that provide inputs to their respective perceptual pathways. These channels extract features from their perceptual input streams, which are fed into the *slices* displayed on top and then cross-modally clustered to learn the different events they are capable of distinguishing.

A Simple Model of Innate Motor Activity

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

An abstract model describing the generation of innate efferent motor activity is shown in Figure 4. In a sense, this model is the reverse of the one displayed in Figure 3. Instead of the outside world generating events, we assume an innate generative mechanism stimulates a motor control center. This in turn evokes coordinated activity in a muscle or effector system, leading to the generation of an external event in the world.

In our model, the innate specification of developmental behaviors is represented by a joint probability distribution over a set of parameters governing motor activity. This is

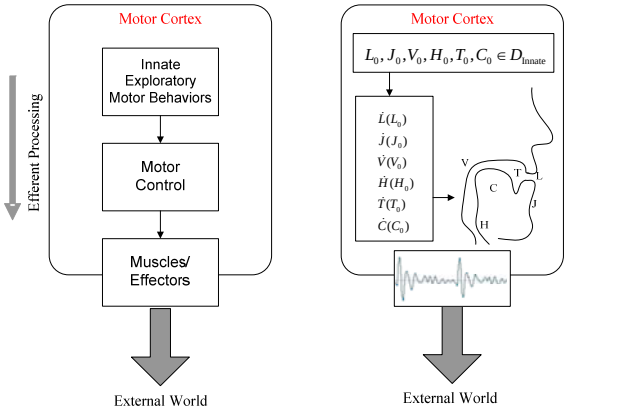

Figure 4 – An abstract model of innate motor activity. A schematic view is shown on the left, which is expanded upon in the example on the right. For illustration, we examine a model of human vocal articulation. This is parameterized by articulator positions at the lips (L), tongue tip (T), jaw (J), tongue center (C), velum (V), and hyoid (H). Motor control corresponds to a set of state equations governing the constrained movement of these articulators over some time period. Parameters describing this movement are selected from some assumed innate distribution on the top right.

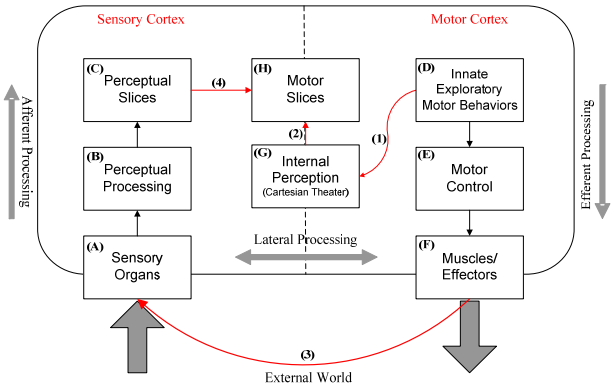

Figure 5 – An integrated sensory motor framework. We connect the isolated sensory and motor systems with the addition of a Cartesian theater (G), which receives data via (1), corresponding to innate exploratory behaviors generated in (D). These data are fed into motor slices (H) via (2). These exploratory behaviors also trigger motor activity via the efferent pathways in (E) and (F). Most importantly, the system is able to perceive its own actions, as shown by (3). These inputs feed into the afferent sensory system, where features are extracted and fed into perceptual slices (C). We thereby learn the Hebbian probabilities between the perceptual slices (C) and motor slices (H), which describe the generation of these perceptions. In the final step, we cross-modally cluster the motor slices (H) with respect to the perceptual slices (C); we thereby learn the motor categories that generate previously acquired sensory categories learned when the system was perceptually grounded.

motivated by the specification of a motor program as a descriptive parameterization. Alternatively, one could assume the existence of a set of deterministic motor schemas, corresponding to predetermined patterns of activity.

An Integrated Architecture

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

In our model, internal perception occurs through the addition of a Cartesian theater (Dennett 1991), so named because it provides a platform for internal observation. Pursuing this philosophical metaphor, the homunculus in our theater will be replaced by cross-modal clustering. We may therefore use the notion of a Cartesian theater without engendering its associated dualist criticisms. We argue that internal perception is a useful framework for higher level cognitive bootstrapping, where cross-modal clustering replaces an internal observer and any notions of "intentionality" are attributed to innate phylogenetic structures and tendencies. Our integrated sensorimotor framework is shown in Figure 5.

Most importantly, the system observes its own actions. Innately generated events impinge upon the sensory organs and are fed into the sensory apparatus on the left. Features extracted from these data are fed into sensory slices (C). This process thereby creates the co-occurrence linkages used by cross-modal clustering between the sensory slices (C) and the motor slices (D), which correspond to conditional spatial probability distributions on regions within their manifolds.

We point out that slices are what may be deemed

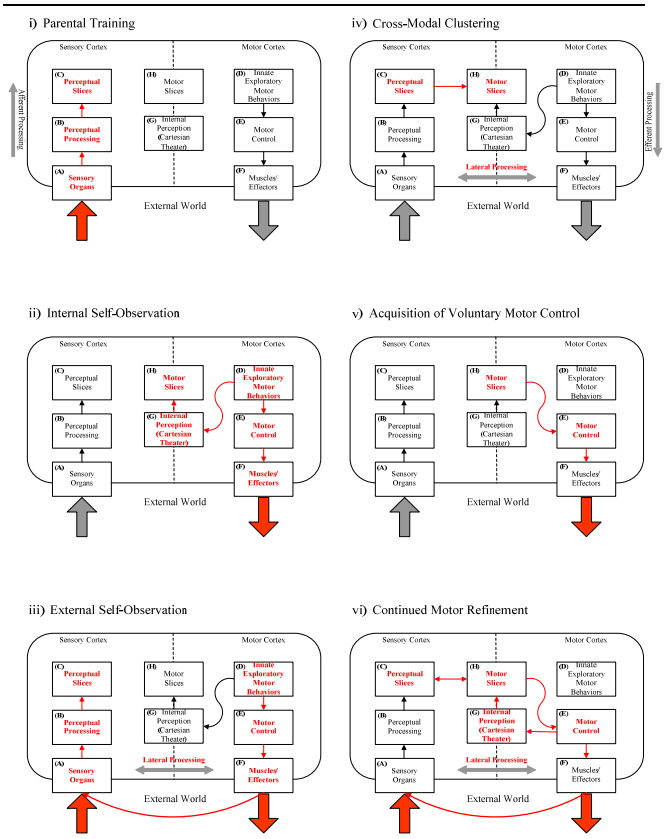

Figure 6 – Developmental stages in our model. i) The juvenile acquires perceptual structures from its parent. ii) Motor acts are observed internally through a Cartesian Theater. iii) The effects of motor acts are observed externally through perceptual channels. iv) Motor slices are cross-modally clustered with respect to perceptual slices. The juvenile thereby learns how to generate the events it learned in stage (i). v) Random exploratory behaviors are disconnected and motor slices take over the generation of motor activity. The juvenile is now able to intentionally generate the sensory events acquired from its parent. vi) Internal perception can be used subsequently in non-juveniles to refine motor control.

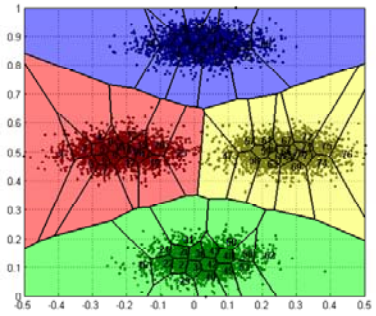

Figure 7 – A hypothetical sensory system that has learned four events in the world. These are acquired through cross-modal clustering, using the framework in the previous chapter. For simplicity, only a single sensory mode is illustrated here.

agnostic data structures – they neither "know" nor "care" what type of data they contain. We can therefore cross-modally cluster the motor slices (D), based on the categories acquired during the perceptual grounding of the sensory slices (C). We note that this is a one-way process. In other words, we fix the sensory categories and only cluster the motor data. We thereby learn motor categories that correspond to previously acquired perceptual categories. The developmental stages of this model are independently illustrated in Figure 6.

An Example

Let us consider a simple example. Consider the situation illustrated in Figure 7, in which a hypothetical sensory system that has learned four events in the world. These are acquired through cross-modal clustering, using the framework described above. For simplicity, only a single sensory mode is illustrated here. This corresponds to stage (i) in Figure 6.

The second and third stages of our model correspond to a system's observation of its own innate, exploratory motor activity, as illustrated in Figure 8. Although reflexive behaviors are phylogenetically selected in animals to satisfy their individual motor requirements (Tinbergen 1951), in artificial systems, we must specify how these innate behaviors are generated. While it may often be reasonable to design exploratory behaviors that are predetermined to satisfy a set of motor goals, we examine a generic strategy here. Our goal is simply to explore a motor space and in doing so, simultaneously observe the effects internally through the Cartesian theater and externally through normal perceptual channels. Consider, for example, the problem of generating pairs of exploratory parameters *(x,y)* in a hypothetical motor system. We have found it useful to select these parameters and thereby explore motor spaces according to an Archimedean spiral. In this case, the *internal perception* of this motor activity might be represented by the slice on left in Figure 8.

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

We see this phenomenon in the slice on the right in Figure 8, which visualizes perception of the motor activity generated by the slice on the left. For the purpose of this example, we generated a non-bijective mapping between the motor parameters and the perceptual features of events they generate in order to illustrate the degree to which corresponding motor and perceptual slices may appear incongruous. Thus, even though these two slices may represent the same set of percepts abstractly, they should not be expected to bear any superficial resemblance to each other.

Figure 9 illustrates the outcome of the cross-modal clustering phase (iv) in Figure 6. In this example, the system learned four events during parental training, illustrated on the left by the four colored distributions. By

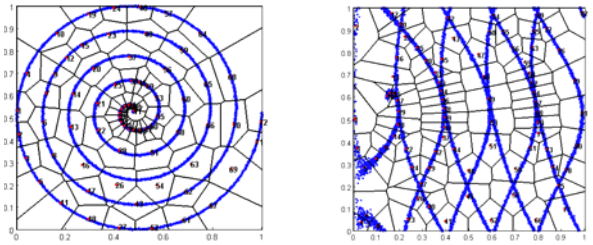

Figure 8 – On the left: Internal perception of exploratory motor behavior corresponding to an Archimedean spiral. These data correspond to the parameters used to generate motor activity. **On the right:** External perception of exploratory motor behavior. This slice perceives the events generated by the motor activity described by on the left. These data correspond to perceptual features describing sensory observations.

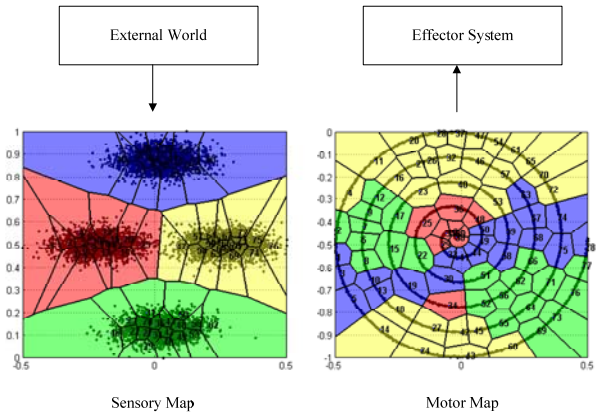

Figure 9 – Acquisition of voluntary motor control. Regions in the motor map on the right are now labeled with the perceptual events they generate in the sensory map.

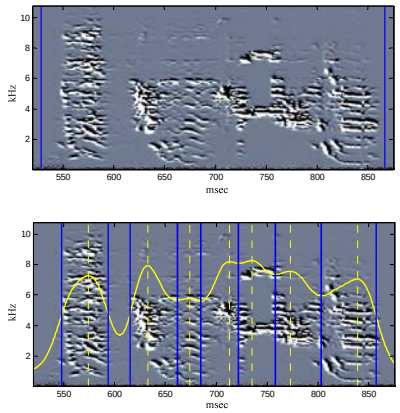

Figure 10 – Partitioning a single syllable into songemes, viewed as spectral derivatives. The song syllable on top is partitioned into songemes on the bottom, which captures its fine structure. The songeme partitioning is computed by finding the peaks in the smoothed log(power) of the signal between 860 and 8600Hz, corresponding to the expected vocalization range of a zebra finch. The songeme boundaries are determined by finding the local minima adjacent to these peaks.

following its innate exploratory pattern and simultaneously observing the effects of its own actions, the system is able to cross-modally cluster its motor map to generate the events to which it was exposed during parental training.

It is important to note that most motor activities are not discrete, discontinuous phenomena. Thus, rather than select individual points in a motor map to trigger behaviors, it is far more plausible to imagine a system "moving" through a motor map during a time period corresponding to sustained activity. One may wish to therefore incorporate additional constraints into motor systems, for example, to minimize energy, avoid perceptual ambiguity, or maximize stability.

Learning Birdsong

We now demonstrate the use of this framework for self-supervised learning of birdsong. Our presentation focuses on song learning in the zebra finch, a popular species for studying oscine songbird vocal production. Specifically, we examine a system that learns to imitate an adult zebra finch in a developmentally realistic way, modeled on the dynamics of how male juvenile finches learn birdsong from their fathers (Marler 1997, Tchernichovski 2000, Nottebohm 2005). Each male essentially learns its father's song with minor, idiosyncratic variations.

Zebra finches are unique for the noisy spectral quality of their songs, which are distinct from the whistled, tonal characteristics of most other songbirds, such as sparrows or canaries. We can see this harmonic complexity in the song snippet in Figure 10. The complexity of their vocalizations, along with a range of behavioral and neurological similarities, has prompted many researchers to propose studying song learning in zebra finches as a model for understanding speech development in humans. Towards this end, we introduce the novel notion of a *songeme*, which we propose as avian analog to phonemes. Songemes are defined in Figure 10 and are motivated by the dynamics of syringeal vocalization in oscine songbirds.

Most importantly, songemes give us a way to break complex bird song into its constituent, generative components. It is these primitive components that are acquired during our system's development.

To train our system, we are indebted to Tchernichovski (2005), who provided approximately 5,000 samples of a single bird's song to act as its "father." After partitioning these into songemes, we extracted streams of acoustic features using a customized version of the "Sound Analysis for Matlab" software package (Saar 2005). For each songeme, we averaged the feature values within it to obtain a compact acoustic description (Figure 11). These values were cross-modally clustered within an assembly of 11 slices, as shown in Figure 12. Although these types of interconnections are likely phylogenetically determined in nature, a more sophisticated artificial system might employ techniques to automatically select interconnections between acoustic features based on mutual information. However, given our task here is to demonstrate acquisition

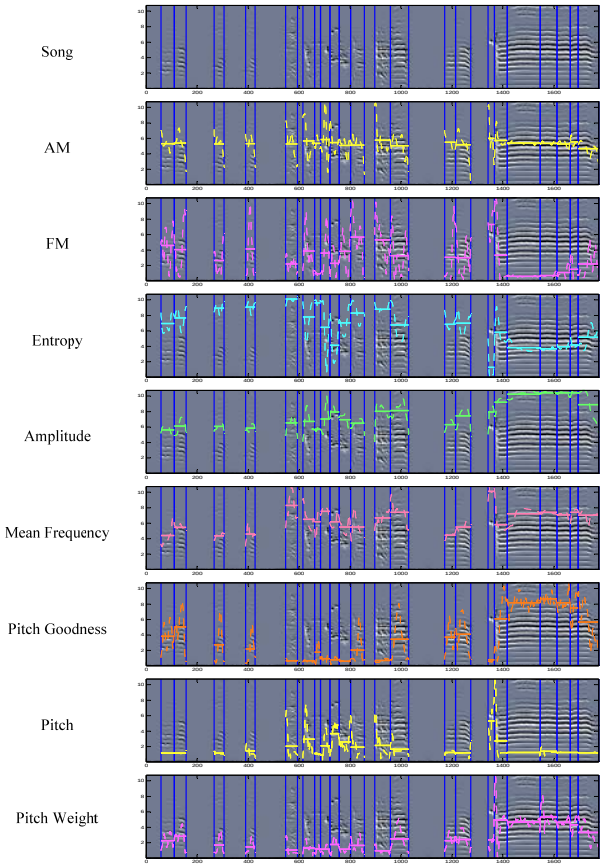

Figure 11 – Feature extraction for a single zebra finch song partitioned into songemes. The solid line within each songeme shows the mean value for the corresponding feature.

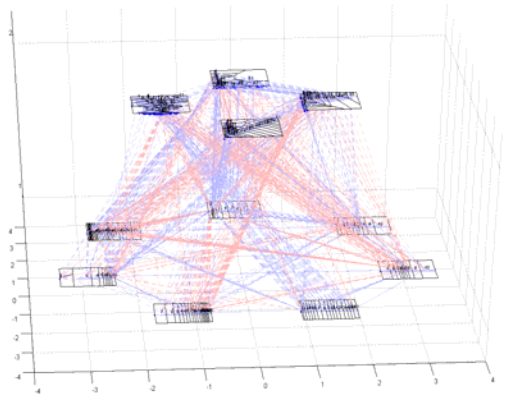

Figure 12 – Slices for birdsong learning. On the bottom, one dimensional slices feed songeme feature values into the two dimensional slices on the top. The colored lines represent learned Hebbian linkages. The slices are then cross-modally clustered to learn songeme categories. This perceptually grounds the system with respect to its "parent's" song.

of sensorimotor control using perceptual mechanisms, automating the feature selection and their interconnections was not of great concern. Nonetheless, it should be acknowledged that a fair amount of manual effort went into architecting the birdsong learner illustrated in Figure 12, which then was cross-modally clustered based on the input song of its parent.

We examine two of these slices in detail in Figure 14, which are among the most interesting of our results from a scientific perspective.

Birdsong Synthesis

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

We refer to the nascent activity of the system as babbling, and some examples of increasingly complex babbling are shown in Figure 13. These demonstrate the system's acquisition of harmonic complexity in response to auditory feedback. The initial babbling corresponds to

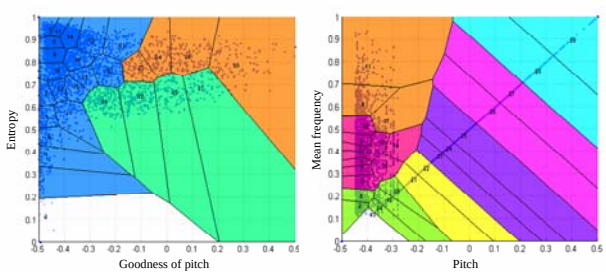

Figure 14– Two cross-modally clustered zebra finch slices. Among the most interesting of our results, we can interpret the left slice as demonstrating the system has learned there are three different types of parent vocalizations: 1) the blue region (far left) corresponds to chaotic sounds. This is similar to fricative speech in humans; 2) the green region (bottom) corresponds to pure tones, such as whistles; and 3) the orange region (top) corresponds to harmonic sounds, such as in the distance call. The right slice shows the system has learned the pitch structure for seven different component vocalizations used to form songemes.

uninformed, innate motor behavior. As the system simultaneously listens to its own outputs while "watching" the internal generation of motor activity, it learns which regions of its motor maps are responsible for providing harmonic complexity, thereby matching this feature of its father's song through cross-modal clustering. Each feature is acquired in turn, until the complete set of songemes is learned. This differential acquisition of acoustic features is characteristic of zebra finch song development.

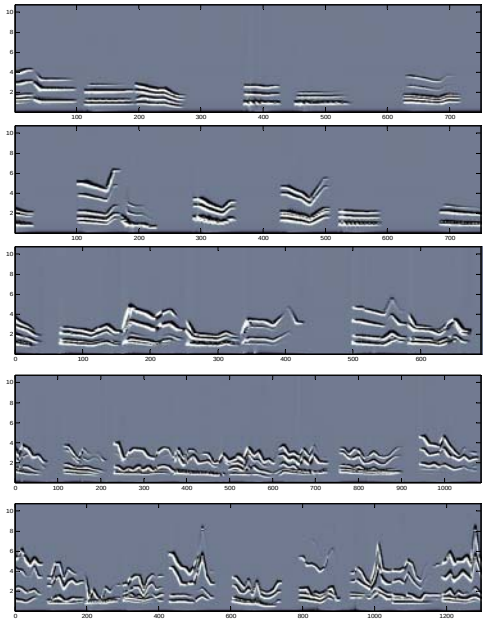

Figure 13 – The temporal evolution of bird babbling in our system. This figure illustrates the acquisition of harmonic complexity due to auditory feedback.

Results

Evaluating the cross-species empirical results presented in this paper is challenging, as the notion of a "correct" answer proves elusive. Spectrograms corresponding to the "father" and "son's" songs are shown in Figure 15. Although the acquired song sounds recognizably like the parent bird's to human ears, does it sounds like a bird to another bird?

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

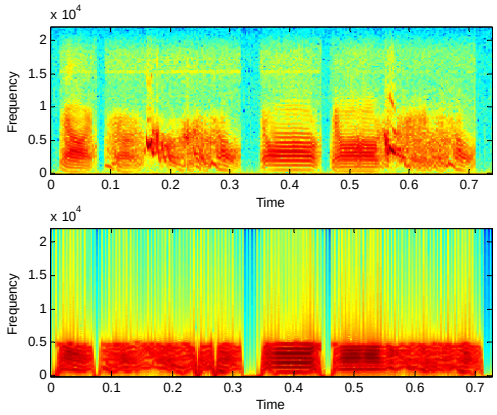

Figure 15 – Empirical evaluation: On the top is a sample of the zebra finch song used as the "parent" for our system. On the bottom is the system's learned imitation, where the acquired songemes have been fit to the template of the parent's song and smoothed. Audio files corresponding to these songs may be obtained by searching for the author's web page via a search engine.

References

1. Baars, B.J. In the Theatre of Consciousness: Global Workspace Theory, A Rigorous Scientific Theory of Consciousness. *Journal of Consciousness Studies* 4. 4: 292-309. 1997.

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

3. Coen, M.H. Multimodal Dynamics: Self-Supervised Learning in Perceptual and Motor Systems. Ph.D. Dissertation. Massachusetts Institute of Technology. 2006a.

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

5. Coen, M.H. Multimodal Integration – A Biological View. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pp. 1417-1424. Seattle, WA, 2001.

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

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

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

9. Jain, A.K., Murty M.N., and Flynn P.J. Data Clustering: A Review, ACM Computing Surveys, Vol 31, No. 3, 264-323. 1999.

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

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

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

13. Nottebohm F The neural basis of birdsong. PLOS BIOL May ;3(5):759-761. 2005

14. Piaget, J. Genetic Epistemology. W. W. Norton & Co. Scranton, Pennsylvania. 1971.

15. Richards, W. (Ed.) Natural Computation. Cambridge, MA. The MIT Press. 1988.

16. Saar, S. Sound analysis in Matlab. http://ofer.sci.ccny.cuny.edu/html/sam.html. 2005.

17. Tchernichovski O, Nottebohm F, Ho C, Pesaran B, Mitra P. A procedure for an automated measurement of song similarity. Anim Behav 59: 1167-1176. 2000.

18. Tchernichovski, O. Private communication. 2005.

19. Thorndike, E. L., Animal intelligence: An experimental study of the associative process in animals. Psychological Review Monograph 2(8):551–53. 1898. 228.

20. Tinbergen, N., The Study of Instinct, Oxford University Press, New York, NY. 1951.

21. Weiss Y. Segmentation using eigenvectors: a unifying view. Proceedings IEEE International Conference on Computer Vision p. 975-982. 1999.

Acknowledgements

The author is indebted to Whitman Richards, Howard Shrobe, Patrick Winston, Robert Berwick, Rodney Brooks, and Gerald Sussman for encouragement and feedback. Deep gratitude is offered to Ofer Tchernichovski for making his birdsong data publicly available in the interest of furthering research. This work was sponsored by AFRL under contract #FA8750-05-2-0274. Thanks to the DARPA/IPTO BICA program and to the AFRL. The author also thanks the anonymous reviewers for their informative and interesting comments.