Automatic Language Acquisition by an Autonomous Robot

Stephen Levinson and **Kevin Squire** and **Ruei-Sung Lin** and **Matthew McClain**
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801
Email: sel@ifp.uiuc.edu

Abstract

There is no such thing as a disembodied mind. We posit that cognitive development can only occur through interaction with the physical world. To this end, we are developing a robotic platform for the purpose of studying cognition. We suggest that the central component of cognition is a memory which is primarily associative, one where learning occurs as the correlation of events from diverse inputs. We also believe that human-like cognition requires a well-integrated sensory-motor system, to provide these diverse inputs. As implemented in our robot, this system includes binaural hearing, stereo vision, tactile sense, and basic proprioceptive control. On top of these abilities, we are implementing and studying various models of processing, learning and decision making. Our goal is to produce a robot that will learn to carry out simple tasks in response to natural language requests. The robot's understanding of language will be learned concurrently with its other cognitive abilities. We have already developed a robust system and conducted a number of experiments on the way to this goal, some details of which appear in this paper. This is a progress report of what we believe will be a long term project with significant implications.

Introduction

Cognitive development has been studied in various environments–on the playground by the psychologist, under the microscope by the neuro-scientist, and in the armchair by the philosopher. Our study occurs in a robotics lab, where we attempt to embody cognitive models in steel and silicon.

How did we choose this particular habitat? First and foremost, we are scientists and engineers, which immediately suggests forming constructive theories and building things to test them. The particular question we are examining is one of the most fascinating questions that has been asked in the last century: Can machines think?

Alan Turing raised this very question back in 1950. He introduced the idea of a machine engaging in "pure thought" and communicating to the world via teletype writer. As an answer to The Question, he suggested that when the machine's discourse (via teletype) was indistinguishable from a human's, we could say that the machine was thinking. He goes on at the end of the paper to suggest that initially, machines should perhaps learn to compete with men at some purely intellectual task, such as chess, but then, he suddenly presents an alternative approach for creating machine intelligence:

> "It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc." (Turing 1950)

The artificial intelligence community has largely followed the former proposal. We believe the latter holds more promise.

Our research is based on a few fundamental principles. First, we believe that a mind cannot be disembodied–it must interact with the real world. Second, we posit that memory is primarily associative, and that learning is based on the correlation of information from diverse inputs. Third, in humans and higher animals, these two assertions are fulfilled by a complex sensory-motor system. We submit that such a system is necessary for human-like cognition. Finally, we suggest that these ideas provide the basis for a mind which can learn a semantic representation of reality, upon which higher cognition and all linguistic structure is established.

Our experiments are based on these concepts. We are developing a robotic platform with basic sensory-motor capabilities, including binaural hearing, stereo vision, tactile sense, and basic proprioceptive control. On top of this system, we are implementing various processing and learning models, and studying how they contribute to semantic understanding.

Numerous other researchers (Brooks *et al.* 1998; Varshavskaya 2002; Weng, Zhang, & Chen 2003; Stojanov 2001; Fischer & Moratz 2001; Hugues & Drogoul 2001; Cohen, Sutton, & Burns 2002; Fasel *et al.* 2002; Grupen 2003) study cognition using robotics; see (Lungarella & Metta 2003) for a recent survey. One key aspect of our project different from most other work is our focus on language learning and interaction as a basis for higher learning.

The rest of this paper proceeds as follows. In the next section, we will describe the organization and construction of our robotic platform. In the following section, we will discuss the implementation of an associative semantic memory for our robot. We will end with some discussion of our current research on spatial cognition. We wish to point out that

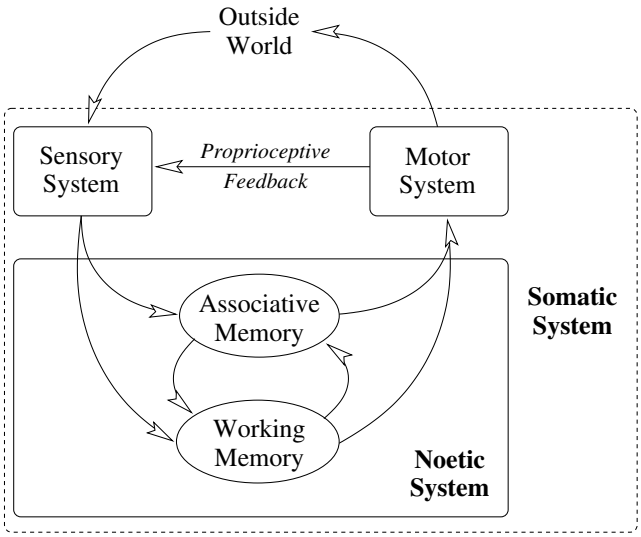

Figure 1: Cognitive Cycle

semantics and spatial understanding are two key elements of language acquistion.

Robot Design and Construction

We have used the basic cognitive cycle depicted in Fig. 1 to guide the design of our robotic system. This simple flow diagram divides cognition into four distinct components: sensory input, long term memory, a working memory, and motor output. The divisions suggested by this model are simple, yet we feel they provide the necessary framework for embodied learning.

As mentioned earlier, interaction with the world through a sensory-motor system is necessary for cognition. Humans perceive the world through the five senses–tactile (touch), gustatory (taste), olfactory (smell), auditory (hearing), and visual (sight). We also perceive information about ourselves, through proprioception (sense of body position and movement) and interoception (internal sensory perception of such things as hunger and body temperature). Ideally, we would like our robot to perceive the world using most of these senses. For now, we have chosen to focus on the senses of sight, sound and touch, with minimal simulation of the others (e.g., proprioception) as needed.

On the other end of the cognitive cycle is a the ability to affect change on the world through a motor system. We can identify two classes of motion actuation in humans that we need to model: (1) movement in the environment; and (2) other articulated body movement (e.g., movement of arms and head, speech).

As shown in Fig. 1, actual cognition takes place between the sensory and motor systems. Taking a cue from psychologists, we model cognition as an associative memory for long-term storage, and a working memory for present information and decision making.

To realize the above system, we chose to work with Arrick Robotics' Trilobot, whose basic anthropomorphic capabilities are complete enough to suit our purposes. In particular,

Figure 2: Robots: Illy, Alan, and Norbert. Alan is an older model Trilobot.

the robot can move freely over level surfaces, can move its head, and can use its arm to manipulate common objects, allowing relatively complex behaviors. For sensory input, the robot has a number of touch and other sensors available.

Starting with this base system, we have heavily augmented the robot's sensory and processing capabilities. We have added cameras and microphones for stereo vision and hearing. A small form-factor computer was installed on-board to collect sensory input from the cameras, microphones, and sensors, and to control the robot. The on-board computer can handle limited processing of the data, but we have also added a wireless transmitter/receiver to transmit sensory data to a distributed network of workstations, where most processing occurs. For this processing, we have developed a robust distributed computing system to manage the actual data transmission and processing, with various modules for sensory data processing, learning, decision making, and control. Finally, because of the additional hardware on-board the robot, we replaced the power supply with a high-capacity sealed lead-acid battery.

We currently have three such robots, whom we have named *Alan*, *Illy* and *Norbert* (see Fig. 2). We are continuing to develop and refine them.

Associative Learning and Memory

Through evolution and in our early childhood development, we first learn to understand the world by associating sensory-motor events and cues (Shanks 1995). Some examples include learning what happens when one touches a hot burner, learning to associate the sight and smell of fire, and learning to associate a word with an event or some other co-occurring cue. In the following sections, we will describe the semantic portion of our associative memory model.

Associative Semantic Learning Using a Cascade of Hidden Markov Models

One necessary condition for effective communication between two people (or even a person and a robot) is that they share a similar model of the world—that they can experience the world in similar ways and understand a similar set of concepts. One way of posing the problem of robotic language acquisition, then, is that a robot should to learn a model of the world through verbal interactions with a hu-

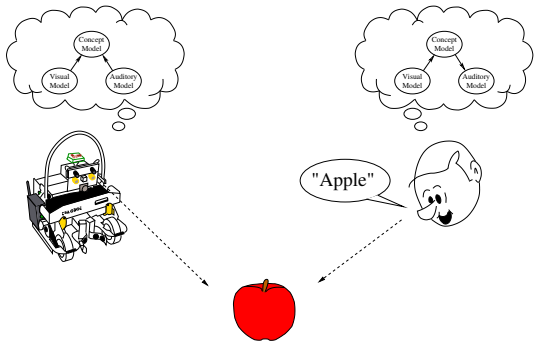

Figure 3: Associative learning of the word "apple."

man, as suggested by Fig. 3. This figure shows an interaction between two subjects, a boy and a robot, each with his own cognitive model of the world. The immediate goal of the robot is to learn the cognitive model the boy is using to comprehend the immediate environment. As the robot learns the boy's model of the world, it can use that model when making decisions.

For our purposes, we will assume that the boy's cognitive model contains the boy's semantic knowledge of the world, encoded as *concepts*. As suggested earlier, our understanding of the world is intimately related to our senses, and so, as suggested by Fig. 4, concepts such as *apple* are learned and recognized through corresponding sensory inputs and other related knowledge. In order to form concepts similar to the boy's, the robot must therefore learn associations among related sensory inputs from multiple senses.

Fig. 5 shows this idea in an abstract, simplified manner. Each submodel in this figure is a classifier. Classes in the visual model correspond to aspects of the different objects in its environment, including such things as colors, shapes, textures, or types of motion. Classes in the audio model represent unique audio cues, including speech. For the concept model, classes represent concepts and are formed from frequently co-occurring audio and visual inputs. Learning in all submodels is unsupervised.

As an example, suppose that the world contains apples and oranges. As the robot explores, its visual model will be presented repeatedly with features from the various objects (such as, shape and color information), and will form classes corresponding to the visual aspects of apples and oranges. Additionally, auditory features from words or sounds

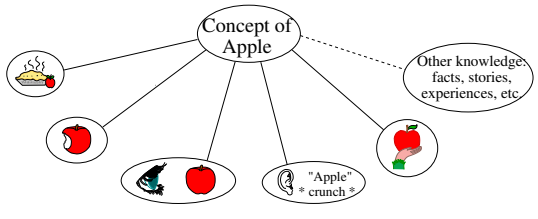

Figure 4: The concept of *apple*.

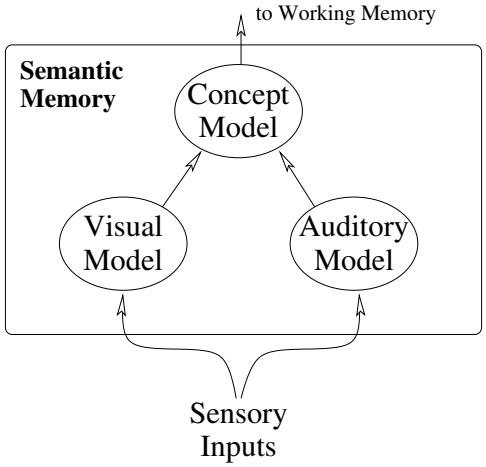

Figure 5: Visual/auditory concept hierarchy.

corresponding to objects and environment will be presented to the auditory model, and it will form classes corresponding to commonly repeated sounds or words, such as "apple", "round", or (perhaps) non-speech related sounds. The classifications of the audio and visual inputs by these models will be presented to the concept model, which will form classes for commonly co-occuring inputs, corresponding to, e.g., shapes or object names.

In our work, we have implemented model in Fig. 5 using a cascade of hidden Markov models (HMMs), the topology of which is shown in Fig. 6. We describe this implementation below.

Hidden Markov Models. An HMM φ is a discrete-time stochastic process with two components, $\{X_n, Y_n\}$, where (i) $\{X_n\}$ is a finite-state Markov chain, and (ii) given $\{X_n\}$, $\{Y_n\}$ is a sequence of conditionally independent random variables. The conditional distribution of Y_k depends on $\{X_n\}$ only through X_k. The name *hidden Markov model*

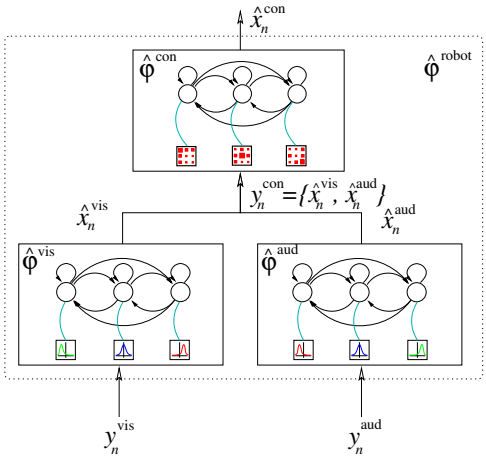

Figure 6: Cascade of Hidden Markov Models.

arises from the assumption that $\{X_n\}$ is not observable, and so its statistics can only be ascertained from $\{Y_n\}$.

Compared to a traditional classifier, we note that (1) the states of Markov chain $\{X_n\}$ correspond to the classes in a classifier, and (2) the observations $\{Y_n\}$, which are considered to be independent in most traditional classifiers, here are only conditionally independent when conditioned on the current state. Therefore, by assuming that the state sequence is a Markov chain, an HMM has advantages over other classifiers in that it takes into account time dependencies in the data.

An important aspect of learning in the robot is that it should be incremental and adaptive. Therefore, we are using the recursive maximum-likelihood estimation (RMLE) algorithm described in (Krishnamurthy & Moore 1993) to train the HMMs running on our robots.

Cascade of HMMs. As mentioned above, Fig. 6 shows the topology of our cascade model, using HMMs for the individual auditory, visual, and concept submodels. Formally, assume that our robot's model of the world is a cascade model $\hat{\boldsymbol{\varphi}}^{\mathrm{robot}} = \{\hat{\boldsymbol{\varphi}}^{\mathrm{con}}, \hat{\boldsymbol{\varphi}}^{\mathrm{aud}}, \hat{\boldsymbol{\varphi}}^{\mathrm{vis}}\}$, where $\hat{\boldsymbol{\varphi}}^{\mathrm{aud}}$ and $\hat{\boldsymbol{\varphi}}^{\mathrm{vis}}$ are auditory and visual HMMs, respectively, and $\hat{\boldsymbol{\varphi}}^{\mathrm{con}}$ is a concept HMM. As described above, the auditory HMM learns classes for audio feature inputs, the visual HMM learns classes for visual features, and the concept HMM learns classes for (co-occuring) audio and visual classifications. At the moment, learning and classification for each model is independent of the other models.

It is important to note that, for each submodel in the cascade, we are not using multiple right-to-left HMMs, as is common in many applications. Instead, we use a single fully connected HMM where each state of the model represents one class of a classifier.

Depending on the state of the robot and the availability and types of features available to the models, not all models will necessarily be active at any given time. There are three modes that the cascade model runs in:

1. Audio-only mode. In this mode, the visual HMM, $\hat{\boldsymbol{\varphi}}^{\mathrm{vis}}$, is not active. The audio HMM, $\hat{\boldsymbol{\varphi}}^{\mathrm{aud}}$, classifies incoming audio features, and the concept HMM, $\hat{\boldsymbol{\varphi}}^{\mathrm{con}}$, accepts as input and classifies the audio state. Because we are using an online update, learning can occur in the auditory HMM, but not in the visual or concept HMM. In this mode, the concept HMM can also produce a corresponding visual *target* state.

2. Visual-only mode. In this mode, auditory inputs are ignored, and only the visual and concept HMMs, $\hat{\boldsymbol{\varphi}}^{\mathrm{vis}}$ and $\hat{\boldsymbol{\varphi}}^{\mathrm{con}}$, are active. It is otherwise the same as the audio-only mode.

3. Audio-visual mode. This is the robot's "learning mode." All three HMMs are actively classifying their inputs, and learning can occur in all three simultaneously. Notably, this is the only mode where concepts are learned, since it is the only time that auditory and visual information can be associated together.

Monte-Carlo Simulation. The goal of a Monte-Carlo simulation is to generate data with a known model and attempt to learn the parameters in a new model of similar or identical structure. We ran two such simulations. The cascade model as proposed above is not fully generative, so the source model was modified to be fully generative in order to produce appropriate observations in each simulation.

In both simulations, a source model generated simulated "auditory" and "visual" data using known concept, visual, and auditory models. The generated data streams were presented to a second model running in audio-visual mode. In both simulations, we were able to learn a set of parameters in the concept HMM similar to the set of concept HMM parameters used in the generative model. For details of these simulations, see (Squire 2004).

Robotic Experiments. After validating the cascade model via simulation, we incorporated the model into a robot demonstration. In our scenario, our robot is wandering around a benign environment, and is instinctually motivated to look for "interesting" things. We expect the following behaviors:

1. It will be attracted to objects, especially ones that it has not seen before, or not seen recently; it will "play" with these objects, attempting to first pick them up, then knock them over.
2. It will be attracted by loud noises, turning toward them and assuming, e.g., that someone wants to get its attention.
3. Using the cascade model, it will
   (a) learn to recognize the visual objects in its environment,
   (b) learn to recognize distinct words spoken to it, and
   (c) learn the concepts associated with the various words and objects.
4. Also using our HMM cascade, it will demonstrate that it recognizes these concepts by
   (a) recognizing a word, choosing a corresponding concept, and finding an object which also matches that concept, and
   (b) recognizing an object and saying the name of a concept corresponding to that object.

The behaviors listed in numbers one and two above were first demonstrated by McClain (McClain 2003). The demonstration described here builds on his work and on the work of others, including

- sound source localization research by Li and Levinson (Li & Levinson 2003),
- speech feature extraction and synthesis research by Kleffner (Kleffner 2003), and
- visual feature extraction by Lin (unpublished).

The specific objects we are using in this demonstration are shown in Fig. 7, and the list of words and phrases we say are listed in Table 1. These words were chosen to test the learning of concepts for specifically named objects (such as *cat*) as well as concepts for general categories (such as *animal*).

Figure 7: Objects used in our robot demonstration.

Table 1: List of words used in our robot demonstration.

animal
ball
cat
dog
green ball
red ball

As we were now working with real auditory and visual inputs, we had a number of implementation issues to decide. For our auditory HMM, we created a simple word recognizer. For visual features, we chose color histogram, moment, and width to height ratio. In the case where more than one object was visible, features for each object were fed sequentially into the visual HMM. Since the word HMM ran slower than the visual HMM, the classification output of the auditory model was upsampled to match the classification rate of the faster visual model, and timestamps were used to align the signals.

As this was the first experiment on the robot, the models were small and of fixed size. The auditory model had six classes, the visual model had four classes, and the concept model also had six classes. Prerecorded auditory and visual data was used to do unsupervised initialization of the auditory and visual models using recursive maximum-likelihood estimation. Although not strictly necessary, the concept model was initialized by hand with slight biases toward desired concepts.

Results. Our goal in this experiment was to show that the concept model $\hat{\varphi}^{\text{con}}$ can learn concepts from a set of real inputs. We trained the auditory and visual models off-line using recorded auditory and visual features, respectively. Note that, even though the training occurred off-line, we used recursive maximum-likelihood estimation to learn the model parameters, so this training could be done online.

The concept model was then trained using RMLE during the simulation run. Specifically, the robot would approach and sit in front of an object of interest. The visual model $\hat{\varphi}^{\text{vis}}$ would continuously recognize this object, and the auditory model $\hat{\varphi}^{\text{aud}}$ would recognize words that were spoken into a close-talk microphone. When a word was spoken and recognized, the state $\hat{x}^{\text{aud}}$ of model $\hat{\varphi}^{\text{aud}}$ corresponding to that word and the state $\hat{x}^{\text{vis}}$ of model $\hat{\varphi}^{\text{vis}}$ were presented to the concept model, and the model was updated according to the RMLE algorithm. To speed up training, each input pair was presented 10 times each time the a word was recognized. This process was repeated multiple times for each object as the robot wandered around and played with its toys.

After a short simulation run, we evaluated the performance of the model in two ways. First, when the robot was wandering around its environment, we would get its attention by making a loud sound. The robot would then turn toward the sound and listen. The robot was then told a word that it knew. This would activate a corresponding concept class, and from this class an appropriate visual target would be chosen. The robot would then look for this target object, and if found, approach and play with it. After about 30 minutes of training, the robot had learned to correctly associate the six words with the four objects it knew.

Second, we were able to look directly at the observation HMM for the concept model. In doing so, we discovered that, as expected, each set of concept HMM transition probabilities and observation densities indicated convergence toward a unique set of concepts. For example, the concept *ball* initially corresponded to a visual representation of the red or green ball with probabilities 0.3 and 0.3, and to a visual representation of the cat or dog with probabilities 0.2 and 0.2. The trained values of these states showed a stronger proclivity to all initial biases. Taking the *ball* example again, the final observation probabilities for the red ball and green ball were 0.43 and 0.42, respectively, and the observation probabilities for the visual representations of cat and dog went down accordingly. The same was true for other observation probabilities for both auditory and visual inputs. For details of all of these experiments, see (Squire 2004).

Discussion. The original motivation for this research was to implement, for our robot, an associative memory for learning the symbolic concepts mentioned above. The model is currently implemented and running in our robot, and has worked very well. We have been able to run the model as part of a demonstration, learn concepts from auditory and visual cues in the environment, and use these concepts to make decisions. An important perspective on this simple statement is that our model converting analog inputs to discrete symbols, allowing the robot's controller to make decisions symbolically using discrete representations of the environment. Moreover, these symbols form the basis needed for more complex symbolic manipulation, such as language.

Spatial Cognition

Based the proposed model of associative learning and our previous experiences with robot learning, the next step of our research will be to study the relationship between spatial language and spatial cognition. We are developing a cognitive map learning algorithm for the robot to extract spatial knowledge of the environment from continuous navigation experiences. Our current map learning algorithm is mainly based on visual sensory input, but it does not exclude sensory inputs from other modalities. Unlike current map learning approaches which focus on recovering the geometry of the environment, our cognitive map contains both the geometric structure of the environment and the robot's navigational experiences in this environment. We firmly believe that both types of information are important for spatial cog-

nition.

Once we have built a cognitive map, we can begin to acquire spatial reasoning. An important extension of spatial reasoning is the ability to extend this understanding to other domains. For example, we would like the robots to learn to understand the temporal concepts of "at" (a certain time), "before," and "after" using the spatial concepts of "at" (a certain location), "in front of," and "behind."

Conclusion

Our ultimate goal is nothing less than construction and explanation of a mechanical "mind". While the study of mind has an intrinsic theoretical and philosophical component, the matter cannot be resolved by a thought experiment. Some constructive approach, however crude, is required. We consider our project to be a humble but serious beginning to a long-range research program which has significant technological and social implications.

We have proposed three fundamental hypotheses upon which we believe a constructive cognitive theory should rest. First, manipulation of our mental model of reality is primarily accomplished by storing, fetching and comparing memorized associations. Second, this mental model depends critically on a fully integrated sensory-motor periphery. Third, the dominant structure of language is semantics. We have proposed to test these hypotheses through the vehicle of an autonomous intelligent robot, trained in a reinforcement paradigm.

On the basis of these hypotheses, our robot has already acquired a number of important abilities and behaviors. It can localize sound sources, and learn how to characterize those sounds. It can autonomously explore its environment in a robust manner, and can learn to visually recognize and play with objects it finds. It has also begun to learn concepts by recognizing the correlation among speech and visual objects. Underlying all of these behaviors is a robust communications framework allowing the various system components to interact and run concurrently.

We are now at a critical juncture in experiments at which simple behaviors are transformed into complex ones. We believe this complexity will arise from the interaction of numerous simpler components. Although our ultimate goal is still far off, we have made some progress defining the function and interaction of these components, and obtained very encouraging results.

Our work is quite challenging and ambitious, and perhaps controversial. Yet we feel that our experiments are technically feasible and potentially of great practical value if successful. Most importantly, however, in our best scientific and technical judgment, when a mechanical mind is eventually constructed, it will much more closely resemble the ideas expressed herein than the mainstream ideas being pursued so vigorously at the present.

References

Brooks, R.; Breazeal, C.; Marjanovic, M.; Scassellati, B.; and Williamson, M. 1998. The Cog project: Building a humanoid robot. In Nehaniv, C., ed., *Computation for Metaphors, Analogy and Agents*. Berlin: Springer-Verlag. 52–87.

Cohen, P. R.; Sutton, C.; and Burns, B. 2002. Learning effects of robot actions using temporal associations. In *Proc. 2nd Int. Conf. on Development and Learning*, 96–101.

Fasel, I.; Deák, G. O.; Triesch, J.; and Movellan, J. 2002. Combining embodied models and empirical research for understanding the development of shared attention. In *Proc. 2nd Int. Conf. on Development and Learning*, 21–27.

Fischer, K., and Moratz, R. 2001. From communicative strategies to cognitive modelling. In *Proc. 1st Int. Workshop on Epigenetic Robotics*.

Grupen, R. A. 2003. A developmental organization for robot behavior. In *Proc. 3rd Int. Workshop on Epigenetic Robotics*.

Hugues, L., and Drogoul, A. 2001. Shaping of robot behaviors by demonstration. In *Proc. 1st Int. Workshop on Epigenetic Robotics*.

Kleffner, M. 2003. A method of automatic speech imitation via warped linear prediction. M.S. thesis, University of Illinois at Urbana-Champaign.

Krishnamurthy, V., and Moore, J. B. 1993. On-line estimation of hidden Markov model parameters based on the Kullback-Leiber information measure. *IEEE Trans. Signal Processing* 41(8):2557–2573.

Li, D., and Levinson, S. E. 2003. A Bayes-rule based hierarchical system for binaural sound source localization. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*.

Lungarella, M., and Metta, G. 2003. Beyond gazing, pointing, and reaching: A survey of developmental robotics. In *Proc. 3rd Int. Workshop on Epigenetic Robotics*.

McClain, M. 2003. The role of exploration in language acquisition for an autonomous robot. M.S. thesis, University of Illinois at Urbana-Champaign.

Shanks, D. R. 1995. *The Psychology of Associative Learning*. Cambridge: Cambridge University Press.

Squire, K. 2004. *A Robotic Framework for Semantic Learning*. Ph.d. dissertation, University of Illinois at Urbana-Champaign. (forthcoming).

Stojanov, G. 2001. Petitagé: A case study in developmental robotics. In *Proc. 1st Int. Workshop on Epigenetic Robotics*.

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59:433–460.

Varshavskaya, P. 2002. Behavior-based early language development on a humanoid robot. In *Proc. 2nd Int. Workshop on Epigenetic Robotics*.

Weng, J.; Zhang, Y.; and Chen, Y. 2003. Developing early senses about the world: 'Object permanence' and visuoauditory real-time learning. In *Proc. INNS/IEEE Int. Joint Conf. Neural Networks*, volume 4, 2710–2715.