Temporal, Environmental, and Social Constraints of Word-Referent Learning in Young Infants: A NeuroRobotic Model of Multimodal Habituation

Richard Veale, Paul Schermerhorn, and Matthias Scheutz

***Abstract*—Infants are able to adaptively associate auditory stimuli with visual stimuli even in their first year of life, as demonstrated by multimodal habituation studies. Different from language acquisition during later developmental stages, this adaptive learning in young infants is temporary and still very much stimulus-driven. Hence, temporal aspects of environmental and social factors figure crucially in the formation of pre-lexical multimodal associations. Study of these associations can offer important clues regarding how semantics are bootstrapped in real-world embodied infants.**

In this paper, we present a neuroanatomically-based embodied computational model of multimodal habituation to explore the temporal and social constraints on the learning observed in very young infants. In particular, the model is able to explain empirical results showing that auditory word stimuli must be presented synchronously with visual stimulus movement for the two to be associated.

***Index Terms*—artificial intelligence, cognitive science, embodied cognition, neural model, developmental robotics**

I. Introduction

THE human cognitive system is remarkable for many reasons, but one of the most fascinating aspects is its ability to acquire and use language. Regardless of whether one believes in innate grammar, language dispositions, or language-related structures in the brain, the fact remains that humans have to learn words and their meanings. While learners in later developmental stages are able to use language itself to scaffold the acquisition of new words and new concepts, building on their previously-acquired linguistic and conceptual capabilities to learn words with increasingly abstract meanings that are far removed from the directly observable, early learners do not have the luxury of relying on such developed structures. In fact, very early learners typically lack knowledge about *both* words *and* referents, with the result that learning is restricted to words whose meanings are closely tied to perceptions. In addition, early learners' brains are still immature and neural structures believed to underlie higher-level cognition do not yet take part in the perception-action loop.

Because the information processing system of young infants is simpler than that of the mature adult, we can use it as the basis for an investigation of how much of the language learning problem is "offloaded" onto simple neural circuits, the body, and the environment. Even the simple control architecture instantiated in a two-month old infant has the necessary categorization and processing faculties to adaptively learn word-referent associations, as demonstrated by its ability to habituate to multimodal audio-visual stimuli presented by a human parent. This ability suggests that a large portion of complex cognitive processing previously thought to be necessary for early language learning in a complex, unstructured environment can be more realistically explained by much simpler circuits as long as they are situated in an appropriately socially scaffolded environment.

This paper describes a robotic model of these early neural circuits and looks at the behavior of the robotic system as temporal properties of the multimodal environment created by the human parent are varied.

II. Background: Word-Referent Association in Infants

Research in developmental psychology has investigated different phases of infant language acquisition, which critically depend on the infant's developmental stage. A commonly accepted view is that the co-occurrence of word and object ought to be sufficient for word learning [1]. Based on this idea, it has been suggested, for example, that infants learn referents of words based on a large number of encounters with words and referents across heterogeneous situations, "cross-situational statistics" [2]); this statistical learning approach is also supported by parent and infant behaviors in unstructured play in 12-20 month old toddlers [3], [4].

While older infants are able to use cross-situational statistics, complex social cues such as eye gaze, or linguistic explanations to determine perceivable referents, younger infants do not have these capabilities. Rather, a very young infant relies primarily on the properties of the stimulus and the environment to guide its cognitive system into states that benefit learning. Hence, the social environment, as established and controlled by the caregiver, provides an important constraining factor that can facilitate (or inhibit) learning of word-referent associations in infants. Empirical results show that parents are sensitive to the cognitive limitations of infants and automatically tailor their teaching behavior to the cognitive needs of infants at different ages [5], [6].

Studies examining multimodal learning in young infants have helped to tease out capabilities and constraints of these stimulus-driven learners over a wide range of stimulus types ([7] provides an extensive overview and discussion). For example, investigations into the conditions under which infants

The authors are members of the Human-Robot Interaction Laboratory of the Cognitive Science Program at Indiana University, Bloomington.

Veale was supported by an NSF IGERT fellowship.

can associate multimodal stimuli has shown that even at two days of age, neonates can habituate to a visual/auditory stimulus pair if the teacher can time the utterance of the word presentation so that it temporally (and spatially) co-occurs with the perception of the referent object [8]. This suggests that even at birth, structures capable of multimodal integration and association must be present and functional.

Another body of research shows that very young infants are also able to reliably habituate to pairs of stimuli between modalities (even when not spatially/temporally co-located) if there are properties that are common between the modalities. This effect is shown to exist when the stimuli are naturalistic (e.g., a mouth moving with a sound at 4-5 months old [9], [10], [11], or physical properties of objects and the sound they make while moving at 3-5 months, [12], [13]), or even if they are arbitrary (e.g., objects of different colors/shapes impacting surfaces with different sounds at 7 months [14], [15]). It seems possible then that the ability to detect redundant properties of relations across modalities also allows for other properties of the multimodal stimuli to become associated, setting the stage for the acquisition of (arbitrary) word-referent relationships. Related studies, especially those by Gogate and colleagues [16], [17], [18], [19] have examined the more difficult association of *arbitrary* auditory stimuli with *arbitrary* visual stimuli in infants from a range of age groups (2-14 months). It is the results of these studies with the youngest of these infants that we will focus on in this paper.

In these infant studies, the experimental setup is meant to expose what happens during a normal parent-child interaction in which the parent "shows" the child an object and names it (Figure 1). The preferential looking paradigm [20][1] is used to examine whether there is a net change in looking time for a visual stimulus (the "shown" object) after it is presented in different manipulated conditions. A simple example involves presenting a visual stimulus V along with an auditory stimulus A (a "word") several times (the "habituation" phase), and then testing whether showing V with A results in significantly different looking time than when V is presented with another, different word B. It is inferred that any change in the looking time between a test with the co-habituated word A, and the test with the other word B, must be caused by some change in the infant that is specific to the stimuli V and A.[2]

The studies reported above demonstrate that infants are learning, even at an early age, but it is important to be clear about exactly what they are actually learning. The infants are *not* learning language or even word-referent associations as they are found in adults (more-or-less permanent lexical entries which can be recruited in a wide range of language situations). The behaviors described above result from a phylogenically older type of adaptive learning called *habituation*. Habituation is a behavioral phenomenon that causes an agent's reaction to a stimulus to weaken if that stimulus is presented in the same modality repeatedly (think of how after wearing sunglasses for a while one stops consciously feeling their contacts on the skin, or how one learns to filter out construction noises in the background). Even though it is not the same long-term lexical acquisition as is present in adults, this early adaptive learning is informative for several reasons. It demonstrates the stimulus-selective processing capacities of infants at this early age–processing capacities which will also be available for learning entries in a long-term lexicon. It has also been demonstrated that habituation learning is a kind of category learning, and not just learning of a particular stimulus token. Infants were able to habituate to several exemplar stimuli and then able to generalize the learned exemplars to a prototypical (but never previously experienced) stimulus [25].

In the infant experiments, the interesting behavior is not habituation itself. Infant experiments will often test whether a *multi*modal stimulus (defined as the *conjunction* of stimuli from two different modalities) can be habituated. If the habituation effect is observed only when the two stimuli are presented in conjunction, but not when either one is presented in isolation, then an "association" must have formed between those two specific unimodal stimuli. Even though the association does not last (disappearing within days/weeks), this association still constitutes learning. How these early associations are related to the permanent acquisition of a lexicon in later years is unknown, but a plausible explanation will be offered for how it would be accomplished in the model presented in this paper.

Although any difference in looking time is taken to indicate an association difference, typically the overall looking time in the "associated" condition is expected to be *less* than in the "control" condition (because the infant has been habituated to that pairing and thus it no longer evokes a "novel" response). In the experimental setup used by Gogate et al., there were two conditions under which subjects could be habituated, which differed in the relative timing of presentation between the visual and object stimuli in each individual presentation episode of the habituation phase. In these experiments, the visual stimuli were visible to the infant the entire time; it was periods of motion of the visual stimulus that constituted "showing" episodes (the hand holding the object waved it around).

The important empirical finding that we target for explanation using an embodied computational model is that learning in the above experiments only occurred during the period in which the auditory and visual stimuli were presented "synchronously" during the habituation phase, and not when they are presented "asynchronously". Following Gogate et al., we can categorize each presentation episode into one of these two groups using two criteria: *onset lag* (L_n) and *offset lag* (L_f) (see Figure 2). Onset lag is the mean time between the onset of the speech act and the onset of movement of the

[1] Under the preferential looking paradigm, learning between (multimodal) stimuli is detected by observing the mean "looking time", which is the amount of time the subject's eyes are pointed at the general region containing the stimulus of interest. The measure is then the ratio of the time spent looking at the stimulus to the trial time.

[2] The logic behind this inference of why pairs of multimodal stimuli can influence looking time is that signals from both modalities are integrated in a region which influences looking time. Candidates for region(s) that actually perform this function have been proposed in the neuroscience literature (some of these are discussed in Section III). There is also behavioral evidence for this assumption (e.g., manipulating the type of auditory stimulus can influence how long infants look at a static visual stimulus [21]). In fact, there is evidence suggesting that at these early ages, senses are not even differentiated yet, i.e., the associations are performed via a network that is not structured to differentiate inputs between modalities [22], [23], [24], [14].

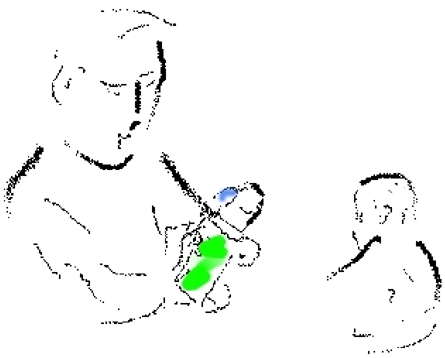

Fig. 1. A parent interacting with a child, "showing" the child a toy while saying, e.g., "bai".

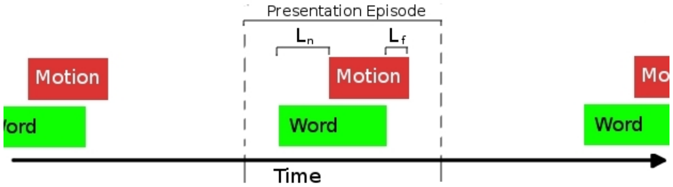

Fig. 2. Timing of stimulus presentations with onset lag and offset lags labeled. Word and object would not overlap (much) in asynchronous condition.

visual stimulus, whereas offset lag is the mean time between the offset of the speech act and the offset of the visual stimulus movement. In the asynchronous condition, L_n and L_f are greater than in the synchronous condition, which effectively results in the word being presented more "between" object motion episodes than "within" them.

The effect of the synchrony/asynchrony condition on whether or not learning occurs is very robust and occurs with a variety of constraints in a wide age range of infants, from as young as 2 months [17]. Yet, there is currently no biologically realistic neural model that explains exactly how real-time multimodal processes affect the formation of word-referent associations in infants when the relative timing of those processes can be characterized more precisely than just being "synchronous" or "asynchronous".

III. Towards a Computational Biologically Plausible Model of Word-Referent Association Learning in Human Infants

The aim of this research is to develop a biologically plausible computational model of the essential processes underlying multimodal habituation. In particular, we are interested in modelling the properties of word-referent association learning observed in 2-month old infants. The model should be as faithful as possible to the functionality *and* the neural architecture of the infant brain. Most importantly, it should replicate the ability of infants to neurally "encode" raw auditory and visual streams, their ability to shift focus of attention using movement (of the eyes, head, body, etc.), and their ability to associate multimodal (in this case, auditory and visual) stimuli.

The model includes analogues of relevant parts of the human auditory system, containing a *cochlear model* to extract low-level features of sound and an auditory circuit intended to represent a column of the primary auditory cortex A1. It also instantiates a plausible visual system, using a *saliency map* to represent the neural response of relevant regions of the LGN and visual cortices and the superior colliculus to control attention (eye location). A second recurrent circuit (the "integration circuit") is included that receives primary inputs from V1 retinotopic visual feature representations, and also receives projections from the auditory circuit. Based on the differential development of brain areas and connectivity at 2 months of age and current understanding of the substrate of (unimodal) habituation, evidence points to the early-developing perirhinal/entorhinal corticies in the hippocampal structure as a convincing candidate for the location of this integration/association of multimodal stimulus information.

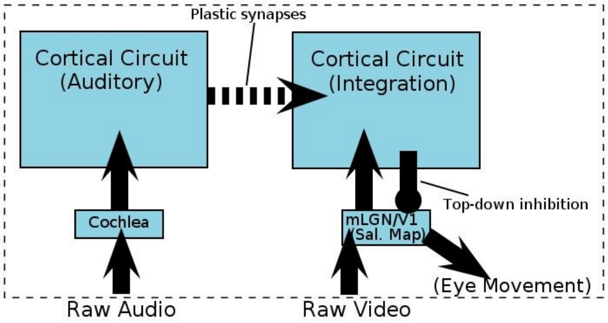

Fig. 3. A sketch showing the major architectural pieces of the model.

Unique encoding of temporally-extended unimodal stimuli is accomplished by the two recurrent circuits: the auditory circuit for auditory stimuli, and the integration circuit for visual stimuli (and as a second level for auditory stimuli). Each of these circuits can differentiate between unimodal stimuli in its respective modality because many different feature combinations from sensors are injected into different subsets of the circuits' constituent neurons. In addition, the recurrent connections within these circuits ensure that the state of a circuit at a given point is to some extent influenced by its previous state (i.e. it integrates some information over time). The model learns conjunctions of auditory-visual stimuli as a result of changes in connections between neurons in the auditory circuit and neurons in the integration circuit. The synapses are sensitive to the timing of activity in the neurons that they connect; their strength changes dependent on the degree to which activity in the two circuits is synchronized. This results in association learning because it is essentially establishing a linear mapping (weighting) between the two encoding spaces.

The elements included in the model were carefully chosen such that their combination will display several key characteristics that we believe must be accounted for by any model that purports to faithfully represent (both functionally and mechanistically) the behavior of very young infants. The most important characteristic is agnosticism towards explicit categories. There are no neurons in the system that are wired or intended for a pre-determined purpose, or to represent a specific thing, except where biological analogs exist in

Fig. 4. The Nao robot platform on which the computational model was developed.

infants.[3] In addition to being biologically valid, this constraint ensures that we make no assumptions regarding *uni*modal cognitive development in infants—we are agnostic regarding the formation of independent, modality-specific categories. All behaviors that arise from the model will be explainable without recourse to modality-specific categories. This is crucial, since neurological evidence for young infants suggests that it is unlikely that the brain differentiates between stimuli from different modalities, so they may not have access to such categories [14].

In addition to the exclusive use of biologically plausible neural mechanisms in the architecture, we also ensure that results obtained using the computational model are consistent with infant behavior by adhering to two important real-world constraints: the nature of the visual and auditory stimuli, and real-world human-robot interactions analogous to parent-infant interactions. Hence, auditory and visual stimuli are fed to the model in the format provided by the microphone and camera sensors without pre-processing.

The robot can interact with human teachers in the same way that infants do in real-time, thus allowing the systematic exploration of the phenomenon of interest: the effect on the model of different social timing behaviors exhibited by the parent.

The robot implementation will allow us to gather evidence for a conjecture: that the empirical results found in developmental psychology about the kinds of multimodal association learning possible in early stages of human development can be explained solely in terms of very simple and minimal "reactive" stimulus responses in an infant's sensory processing and integration areas. The following subsections provide the design rationales (and neurological support) for each of the major subsystems of the model and describe details of how they were implemented computationally so that they can be run in real-time on the physical robot (Figure 4).

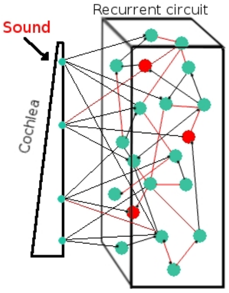

Fig. 5. Visualization of the auditory network, from where sound enters the system and is passed through the cochlear model (left), producing activity in output channels that are injected into the recurrent circuit (right).

A. Auditory Processing

Raw audio streams are converted into neural activations by a cochlear model ([26], [27]) which effectively applies band-pass filters and transformations to the sound wave to approximate the firing activity of a set of neural channels along the cochlea. These cochlear channel neurons project synapses into a recurrent neural circuit intended to model primary auditory cortex (or at least a single column in it, see Figure 5). This architecture was chosen based on evidence (e.g., from [28]) that such a circuit instantiates a fading-memory filter which contains sufficient unique information of the correct type to independently train spoken word classifiers. The circuit accomplishes this by reliably entering different, yet consistent, state-space trajectories in response to different real-time stimulus streams. In other words, if the cochlear model converts the sound streams of two different spoken words into sufficiently different patterns of firing over 500 ms in its various channel neurons, then the recurrent circuit's response should (consistently) differ between the two input stimuli as well.[4] The same paradigm will be invoked to propose a plausible account of how to handle complex stimulus responses in the visual modality as well.

Implementation details: The cochlear model described in [27][5] was modified and updated to run in real time. Parameter defaults are retained (exceptions: $decimationFactor$=16 and $stepFactor$ =0.5) producing 41 output channels which encode on each simulation step the probability that a spike occurs in that channel on that step (1 ms for the experiments below). The auditory recurrent circuit is implemented with few changes using parameters drawn from [28] (based on empirical recordings from rat somatosensory cortex), and is comprised of a $15 \times 3 \times 3$ column of current-based leaky integrate-and-fire (LIF) neurons (for a total of 135 neurons, 20% of which are randomly chosen to be inhibitory), whose membrane potential's (V_m) dynamics are described by the

[3] Examples include the lowest-level feature-tuned neurons in the cochlea and the saliency map. Also, prototypical color selective neurons were hard-wired for the vision system implementation used in the experiments in order to simplify the analysis of learning.

[4] Mathematically, the circuit does this by mapping its input onto a much higher-dimensional vector space (i.e., the super-dimensional system represented by the configuration of all the parameters of the neural circuit over time) such that different inputs will occupy regions of this high-dimensional space in different orders; see [28] for details.

[5] C source available from http://www.slaney.org/malcolm/pubs.html

following equation:

$$\frac{\partial V_m}{\partial t} = \frac{-(V_m - V_{rest}) + R_m \cdot (I_{bg} + I_{syn})}{\tau_m} \tag{1}$$

where R_m is the membrane resistance (uniformly 1.0 MΩ for all neurons), I_{bg} the background current (uniformly 13.5 mV for all neurons) and I_{syn} the total current impinging from afferent synapses. The $-V_m$ represents the leakage term, causing the membrane potential to decay exponentially with time constant τ_m (uniformly 30 ms for all neurons). The resting potential of the membrane V_{rest} is assumed to be 0 mV. When the membrane potential of the neuron exceeds a threshold V_{thresh} (uniformly 15.0 mV), the neuron is assumed to fire an action potential. The membrane potential is reset to V_{reset} (uniformly 13.5 mV) and the neuron enters a refractory period during which the dynamics of the neuron model are frozen. The refractory period of excitatory neurons is 3 ms (inhibitory neurons 2 ms). I_{syn} at a given time is equal to the sum of the post-synaptic responses (PSR) of excitatory afferent synapses minus the sum of the PSRs of inhibitory afferent synapses.

The 41 channels from the cochlear model diverge to inject current directly into a randomly selected 30% of circuit neurons via static synapses (i.e., the current is added directly to I_{syn}). The amplitude A of each of these input synapses was drawn from a Gaussian distribution with mean A_{mean}=18.0 when the post-synaptic neuron was excitatory and A_{mean}=9.0 when it was inhibitory. The standard deviation was chosen to be 100% of the mean. Negative weights were set appropriately from a uniform distribution between $0.001 \cdot A_{mean}$ and $2 \cdot A_{mean}$.

The output of each channel of the cochlear model was linearly scaled by a constant value (200.0) to normalize it to the correct magnitude and range of the circuit neurons. Neurons within the auditory recurrent circuit are probabilistically connected based on a function of their Euclidean distance that gives priority to local connections. Specifically, the probability that a synapse exists between neurons at 3-D points a and b is $C \cdot e^{(-D(a,b)/\lambda)^2}$, where λ is a global parameter controlling the density of connections (=2.0), $D(\cdot)$ is the Euclidean distance function, and C is a parameter to modulate the probability of a synapse depending on properties of the connected neurons. In our case, $C = 0.3$ if a is an excitatory neuron and b is an excitatory neuron (EE), $C = 0.2$ for excitatory and inhibitory neurons (EI), $C = 0.4$ for inhibitory and excitatory neurons (IE), and $C = 0.1$ for two inhibitory neurons (II).

Synapses have transmission delays of 2 ms (i.e., from presynaptic neuron firing, it takes 2 ms for the action potential to impact the synapse and have an effect on the post-synaptic neuron) in the case of excitatory-excitatory neurons and 1 ms otherwise, and are modelled as exponential-decay synapses. The dynamics of the post-synaptic response q_{psr} of a synapse is thus:

$$\frac{\partial q_{psr}}{\partial t} = \frac{-q_{psr}}{\tau_{syn}} \tag{2}$$

where τ_{syn} is the time constant of the synapse (τ_{syn} = 3.0 ms for excitatory synapses and 6.0 ms for inhibitory synapses).

The synaptic dynamics of intra-circuit synapses are modeled according to the UDF model proposed in [29], [30] using the parameters described in [28]. In this model, the arrival of a spike k (represented as a Dirac-delta function, $\delta(t)$) after an interspike interval Δ_{k-1} induces an increase in the post-synaptic charge of amplitude A_k:

$$A_k = w \cdot u_k \cdot R_k \tag{3}$$

$$u_k = U + u_{k-1}(1 - U)e^{-\Delta_{k-1}/F} \tag{4}$$

$$R_k = 1 + (R_{k-1} - u_{k-1}R_{k-1} - 1)e^{-\Delta_{k-1}/D} \tag{5}$$

where w is the weight of the synapse (synaptic efficacy), u_k and R_k are hidden dynamic variables maintaining the facilitatory and depressionary tendencies of the short-term plasticity of the synapse, and U, D and F are the parameters modulating local synaptic use, time constant of depression (in seconds), and time constant of facilitation (in seconds), respectively. Initially, $R_k = 1$ and $u_k = U$. The parameter triples (U,D,F) were selected for each synapse depending on the type of neurons that were connected, i.e., EE, EI, IE, and II, and were drawn from a Gaussian distribution with means (0.5, 1.1s, 0.05s) for EE, (0.05, 0.125s, 0.120s) for EI, (0.25, 0.7s, 0.02s) for IE, and (0.32, 0.144s, 0.06s) for II (the standard deviation was 50% of the respective means in all cases). Negative results were redrawn from a uniform distribution between 0.001 of the mean to double the mean. The weights w of the synapses were drawn from Gamma distributions with means 0.3 (for EE), 0.6 (for EI), 0.19 (for IE), and 0.19 (for II); SD of 100% was used for each mean, with negative results redrawn from a uniform distribution as described above. Simulation of the model is performed via Euler integration with dt=1 ms. While closed-form solutions to the dynamics of each of the neuron/synapse/STDP models exist, a closed-form solution for the entire network does not, and must be numerically approximated.

B. Visual Processing

While auditory stimuli can be separated based on the sequence of onsets, offsets, and strengths of frequencies present in the sound stream, a corresponding method for vision that will convert a sequence of raw camera frames to a unique response is not obvious, even after a review of the neuroscience literature [31]. In humans, it is known that some neurons in the optical ganglia, and subsequently in the LGN, are responsive to certain basic color stimuli, and that there exist columns in visual areas V1 and V3 that are tuned to the orientation (e.g., 45° or 90°) of visual lines that fall within their receptive fields [32], [33]. Additionally, columns responsive to the direction of motion have been reported reliably in area MT [34], and there is evidence for simple color maps in V1 ("blobs", [35]), and even hue-maps representing the continuum of color hues (from "stripes" in V2 [36], and "globs" in V4 [37]). Recent research has attempted to determine the method by which contours are perceived based on combinations of lower-level feature detectors of this type [38].

These empirical findings point to what we will consider a set of basic visual features that can be used to differentiate stimuli. However, these results are based on studies with *adult* animals, while we are interested in *infants*, who are only partway through the developmental trajectory to adulthood. It turns out that many of the above cited visual processing capabilities do not emerge for quite some time postnatally (see [39] for an overview). Hence, they can not possibly contribute to the observed behavior of infants. This permits us to base our model on a much simpler neuroanatomical circuit than the full "adult" visual circuit.

The developmental neuroanatomy literature indicates that at 2 months of age the only pathways by which information can proceed from the retina to the muscles that control eye movement are via a cortico-collicular (corticotectal) route: retina, magnocellular lateral geniculate nucleus (mLGN), V1 deep layers 5/6, superior colliculus (SC) deep layers, brainstem [40], or a thalamic relay via extrastriate area MT: retina, SC superficial layers, pulvinar, V5/MT, SC deep, brainstem [39]. This latter route is associated primarily with processing motion signals. However, since the primary feature dimensions defining objects in our paradigm are not motion-related, we will focus on the former (direct cortico-collicular) route. A retinotopic representation of the visual field is present in layers 5/6 of the primary visual cortex (V1), receiving its inputs from the retina via the magnocellular pathway of the LGN (mLGN). This magnocellular pathway is the "broadband" pathway and mostly encodes information about rod-cells (light intensity-detecting) from the retinal array. It thus encodes very coarse-grained, flicker/intensity information.[6] This retinotopic map in V1 then projects retinotopicically to deep layers of the superior colliculus ([41]), which contains neurons which in turn project to motor control related regions in the brainstem which elicit an eye movement based on the activation distribution in the SC. The genesis of these retinotopic maps is not addressed in this paper, but they are probably mediated by both experientially-based and molecular mechanisms [39].

The circuit described above purports to explain the mechanisms by which the eyes are directed to locations in the environment based on the activation of regions of a retinotopic map in, e.g., V1 and then SC. These activations are based on bottom-up activation (i.e., based on properties of the stimuli in the environment and how the system is built to filter these properties). In reality, all locations in the retinotopic map would be activated in parallel (and so have effects on deeper areas) but for computational simplicity it is assumed that only the "winning" (maximal) region is uninhibited and able to innervate the deeper areas.

The above mechanism to determine the activations of retinotopic regions and to select the highest one is implemented via a *saliency map*. The saliency map allocates higher "activation" to (retinotopic) regions of globally high bottom-up contrast across many different scales and feature types [42]. The theory is that these regions are more likely to contain the most informative stimuli, and useful information to uniquely encode the informative visual stimuli. Thus, directing the eyes to maximally encompass those stimuli will allow quicker and more efficient encoding of the information, allowing the animal to respond appropriately.

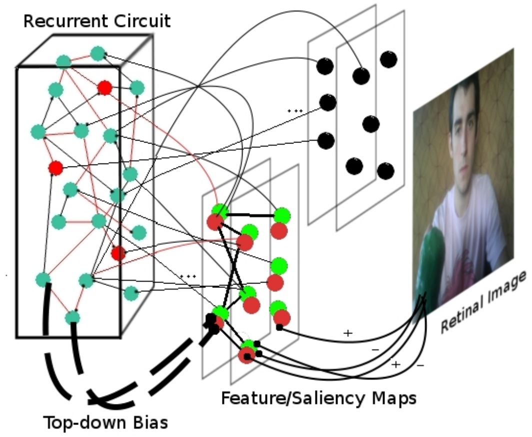

Fig. 6. The pathway and connectivity of the visual system, which also controls bottom-up and top-down biasing of attention. (1) Pixels on the visual field are sampled in combinations, producing neural correlates of saliency map channels. (2) The active neurons in the saliency maps (V1) influence a recurrent neural circuit in combinations that reflect their location and feature-type. (3) Feature-tuned neurons in the saliency map level that correspond to the currently active trajectory-encoding are inhibited, causing those regions to be disadvantaged.

It is informative to look at the behavior of the system when only one feature-type contributes to the encoding of visual stimuli. This is also closer to the situation of an infant. By simplifying the system in this way, one can sidestep complexities of dealing with combinations of too many different feature-types—problems which are not central to the main research questions addressed in this paper.[7] We implement and examine the visual system outlined above, but with the maps and detectors for all feature-types except for color lesioned. Thus, the only dimension along which the system will be able to differentiate between two stimuli is if they are of different (R-G/B-Y) colors. The motion map is also active, but does not influence the higher-level encoding. In other words, only combinations of color-sensitive saliency map neurons project into the integration circuit. However, both motion and color neurons contribute to deep SC projections that drive eye movements. Based on the developed circuit at 2 months postnatal, real infants would only have access to intensity (and thus possibly orientation), and flicker-type information to encode stimuli in V1. Area MT may additionally have other types of motion encoded. How to uniquely represent, e.g., rotation and translation invariant visual stimuli based on these basic feature types (as in the brain) is not well-understood and not the focus of this paper. The problem was side-stepped by using the easy-to-represent feature "color" as the only defining feature of visual stimuli, even though infants of the target age

[6] The parvocellular pathway, which is not developed yet, is believed to be what processes, e.g., color information. These projections into more superficial layers of V1 only become myelinated later in development and so likely do not play a role in control at 2 months of age. In the experiments we will be using color information instead of configurations of these flicker/intensity features because it is more straightforward.

[7] We can do this without affecting the ability of the system to behave in the real world, "up to" its ability to discern visual stimuli. It is similar to the situation of a color-blind man who behaves normally until he comes upon two stimuli, one red and one green, that he cannot discern.

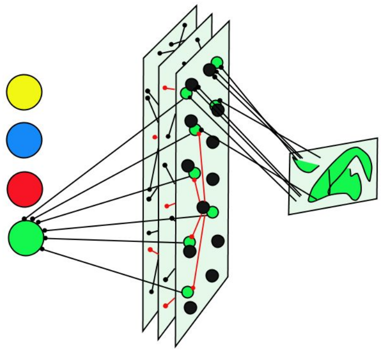

Fig. 7. The architecture of the vision system as implemented. From right to left: Retinal image projects in various combinations onto neurons in saliency maps, called “green” (in green) or “motion” (in black) neurons depending on what the projections they receive implement. Here, green are only feed-forward feature-tuned neurons, whereas motion maps inhibit laterally and within map, thus implementing saliency map. Green-tuned map neurons project onto a neuron in a simple contrived circuit—the green circuit neuron represents “green” because of the combinations of inputs it gets from saliency maps (not because it’s a predetermined green-tuned neuron).

probably can not actually use color.

Even in such a seemingly simple domain as color, neurological evidence cannot support one explanation over all others regarding which representations of colors (i.e., full hue maps, or coarse color opponencies) are engaged in which behaviors [31]. Lacking an empirical basis for constructing the network, we implemented it in the most straightforward manner possible (diagrammed in Figure 7). The “recurrent circuit” comprising the integration region has been simplified and reduced to a four-neuron circuit with no lateral connections, contrived to demonstrate how individual circuit neurons gain identities based on the combination of inputs they received (here the bottom-most neuron is visually “green” because it happens to receive positive inputs from green-tuned V1 neurons, and no connections from other neurons).

Implementation details: The “fully neural” saliency map is approximated by an optimized version (“envision”) of the common implementation of the saliency map ([42]) available under the GPL license.[8] The code was modified to run in real time with top-down biases (i.e., to V1 state). Fixations are determined by a winner-take-all (WTA) network (intended to correspond to SC) implemented as a 2-dimensional array of LIF neurons with a uniform membrane time constant 20 ms, which receives input from the instantaneous saliency map calculation (i.e., V1 state) run on the current visual frame every 20 ms. The input to each neuron in the WTA network is determined by application of a Gaussian filter to the level-4 output of the instantaneous saliency map, normalized to the range $[0, 1]$. The filter’s width is $1/6$ of the pixel-width of the visual field. LIF neurons in the WTA network are modeled without a refractory period, and are assumed to continuously fire while above threshold. The WTA network is fully laterally connected with inhibitory connections (static, linear synapses with amplitude A=0.2). After a WTA neuron reaches threshold and a fixation is executed, a region equivalent to the width of the Gaussian filter is inhibited by a constant signal of strength A=0.3 beginning 300 ms after the onset of firing of the winning WTA neuron and ending 500 ms after the onset of firing. The average color (the only distinguishing feature type) is extracted from the region of the visual frame corresponding to the location of the WTA neuron by averaging the RGB color channel values for all pixels that lie within a circle of radius $1/20$ of the width of the visual field.

The average rgb-space color of the foveated region is converted into R-G/B-Y color space based on the equations used in [42]:

$$R = r - \frac{(g+b)}{2} \tag{6}$$

$$G = g - \frac{(r+b)}{2} \tag{7}$$

$$B = b - \frac{(g+r)}{2} \tag{8}$$

$$Y = r + g - 2(|r-g| + b) \tag{9}$$

LIF neurons (with parameters identical to those in the auditory circuit) tuned to each of the 4 colors are injected directly with the value of the corresponding (R/G/B/Y) color channel, linearly scaled by 75.0 to convert from the color space range to the neural parametal range, with synaptic scaling parameter uniformly A=1.0. These 4 neurons are intended to represent the firing activity of the population of color-tuned neurons in the V1 area with the highest response strength (to save having to represent all regions in the retinotopic map in parallel).

The “integration circuit” is modeled as an additional 4 LIF neurons that receive static input from the 4 V1 color-tuned neurons in a color-preserving manner (connected only to the corresponding color neuron, A=5.0). These parameters produce a constant firing rate of roughly 50Hz in the circuit with sustained input of a pure color.

C. The Integration Region

We postulate that the perirhinal/entorhinal cortex could play the role of multimodal integrator for stimulus quality, and parahippocampal/entorhinal for spatial. The perirhinal cortex receives afferent projections from both V1 and from primary auditory cortex A1 and is well-developed and active even prenatally [43]. Additionally, the hippocampus and the surrounding “old” cortices (parahippocampal, entorhinal, perirhinal) have been shown to be involved in habituation and preferential looking based on novelty detection [44], [45], [46]. Synaptic plasticity (STDP) mechanisms have been shown to be active within these cortical areas, providing an explanation for how associations could be learned between some multimodal stimuli but not others [47]. Finally, it has been shown that extensive recurrent connections project back from perirhinal/parahippocampal cortex to the (primary) sensory areas V1 (and A1) [48], [49]. Also, stimulus or location-selective neurons have been shown to exist in inferior temporal cortex (the rhinal sulcus) [50] and visual regions [51], which show interesting biased behavior during simple memory tasks. In particular, this bias is manifest in that neurons which are selective to a particular stimulus S will actually become inhibited about 100 ms after stimulus presentation

[8] http://ilab.usc.edu/bu/

if the stimulus they select for is not the "target" stimulus. Meanwhile, the target stimulus will stay maximally active until finally an eye movement is made (a further 100 ms later). This is evidence that top-down (or lateral) inhibition can actually be modulating the magnitude of a representative response. In the case of the above studies the top-down inhibition was memory of which stimulus is the "target" (via an unspecified mechanism which maintains state for a short time—i.e., short-term memory). In our case it will be top-down inhibition of whichever perirhinal "multimodal representation" is most active. In normal, unimodal visual habituation this would be caused by a straightforward growth in afferent connections to the representative neural population/system in perirhinal cortex. Note that the same unimodal (e.g., V1) input would evoke a stronger perirhinal response, and thus increased feedback inhibition to itself via the feedback connections. This inhibition would decrease the activation of that location in V1, and then in SC, thus reducing the probability of looking to that location.

In this paper the perirhinal visual "representations" are kept simple (just colors); only the weights of the multimodal input coming from auditory cortex are learned. The modification of these weights has the effect of modulating the bias that a given auditory response will have on a given visual representation (in our case, canonical color). This way multimodal habituation is modeled without addressing the effects of unimodal visual (or auditory) habituation, which are assumed to be constant for all stimuli. Thus, the whole system will serve as a mechanistic "novelty filter" for *conjunctions* of audio and visual stimuli, since there is no mechanism in the system to "learn" visual categories (only auditory categories), and the only mechanism to direct the eyes and orient towards a stimulus is via the visual system. Yet, it can habituate to multimodal stimuli because the inhibition to the retinotopic map which directs eye movement increases as A and V are associated, causing less looking time to V only in the simultaneous presence of A.

Neuroanatomically, the connections from sensory areas *to* the parahippocampal cortex are probably not the synapses whose plasticity is responsible for habituation. Rather, it is more probably connections between neurons within the perirhinal/entorhinal cortex that play this role, because it would be difficult to get sufficient complexity in the sensory-perirhinal synapses to recognize complex stimuli to which people can habituate. Thus, in reality, it would be the case that the primary sensory areas would feed with (mostly static) projections into different subsets of the recurrent circuit representing the peri/entorhinal cortex, and then it would be the weights between constituent neurons of this recurrent circuit that would be plastic. For both computational and conceptual simplicity, we have removed one of these two locations of change. This leaves the qualitative mechanism of the habituation in place, while making later analysis and understanding of the system simpler and paving the way for future mechanistic additions. However, we found that the model manifests some undesirable properties, possibly as a result of the simplification, which we will discuss in section V.

D. Association Mechanism: Spike-Time Dependent Plasticity

Spike-time dependent plasticity (STDP) is a mechanism whereby the efficacy of a synapse is modulated dependent on the relative timing of pre- and post-synaptic spikes [52]. Learning in the neural model is accomplished via the plasticity of synapses between the recurrent auditory circuits ("auditory cortex") and the integration region (peri/entorhinal cortex) that receives inputs in static configurations from the saliency map's (i.e., V1's) output.

STDP is a mechanism well-suited for multimodal integration because the potentiation of synapses will only occur where pre- and post-synaptic neurons are firing in a synchronous pattern—in particular when the action potential from the pre-synaptic neuron's firing reaches the synapse before the post-synaptic neuron fires. Depression of the synapse will tend to occur if the synapses are out of synchrony, or if they randomly switch between preceding and trailing one another.

Thus, increase in weight will be more likely to happen when stimuli are presented synchronously. We hypothesize that the increased weight will allow the auditory circuit to "entrain" the integration circuit, biasing it towards continuing to fire in synchrony, allowing more learning to occur. Firing in synchrony can also result in increased information flow between circuits, since action potentials will be less likely to arrive when the neuron is least ready to integrate them (i.e., in its refractory period).

Implementation details: Static exponential decay synapses connect every excitatory neuron in the auditory circuit with each of the 4 integration region circuit neurons. STDP is modeled for each of these synapses between the two input maps using the STDP model described in [53], with $\tau_+ = \tau_-$ = 20 ms and A_+=0.005 and A_-=0.006. In this model, the change in synaptic efficacy (weight) of each of N afferent synapses to a post-synaptic neuron is modeled according to $N + 1$ functions. One function, $M(t)$ is used to decrease synaptic strength. The others, $P_a(t)$ (for $a = 0, 1, 2...N - 1$) are used to increase synaptic strength. All of these functions $P_a(t)$ and $M(t)$ are initially zero, and decay exponentially, i.e.

$$\frac{\partial M}{\partial t} = \frac{-M}{\tau_-} \tag{10}$$

$$\frac{\partial P_a}{\partial t} = \frac{-P_a}{\tau_+} \tag{11}$$

Every time the postsynaptic neuron fires, $M(t)$ is *decremented* by A_-. When synapse a receives a presynaptic action potential at time t, its weight is modified: $w_a = w_a + M(t)w_{max}$. w_{max} is the maximal weight parameter, representing physical constraints on synaptic transmission efficiency per synapse, globally set at $w_{max} = 3.0$ for the experiments. The value of w_a is clipped if it would be greater than w_{max} or less than zero.

$P_a(t)$ is *incremented* by A_+ every time synapse a receives an action potential. When the post-synaptic neuron fires at a time t, w_a is modified: $w_a = w_a + P_a(t)w_{max}$. Again, w_a is clipped if it would be greater than w_{max} or less than zero. Initial weights for all synapses are zero so that the net

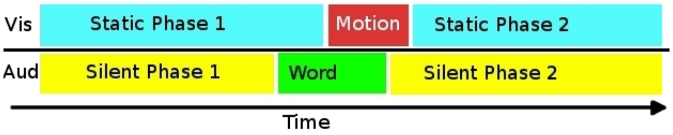

Fig. 8. The time line for a single learning episode. Phases in both auditory (A) and visual (V) modalities are shown side-by-side. In this case the word is slightly longer than, and has a negative offset to, the object motion.

change can be used as a direct learning metric in the empirical experiments.

IV. Evaluating the Robotic Model in Learning Trials

The measure of success for the robotic model evaluation is how closely the observed behavior of the robot implementation matches the infant experiment results described in Section II. If the robotic model is able to produce behaviors consistent with the infant studies, i.e., if it shows evidence of learning word-referent associations in the synchronous condition but *not* in the asynchronous condition, then we have succeeded in demonstrating that the proposed model is plausible, given the additional measures taken to construct it based on current neurological evidence. Note that given the experimental paradigm used for infants and the available psychological results, it is not possible to do more detailed model fitting than these two data points.

A. The Role of Synchrony

For the evaluations, we use a design based on the model employed by Gogate and colleagues [16], [17]. In particular, we examine single episodes of word-object presentations to determine the effects of various parameter combinations on learning performance; a single experimental run consists of a human experimenter waving an object in the robot's visual field and speaking a word at a given offset from the onset of the object stimulus (see Figure 8). The conditions are defined on two dimensions: *word length* (two distinct words presented by the speaker) and *word offset* relative to object motion (i.e., the difference between the temporal onset of the word stimulus and the temporal onset of the object stimulus).

The model was implemented using our distributed robot infrastructure *ADE* (the *Agent Development Environment* [54]). The ADE development model encapsulates major functional components as *ADE servers* (e.g., the vision server that provides frames to the saliency map, or the server used to control the Nao robot's movement). ADE provides facilities for establishing and maintaining connections between servers and invoking actions in other (possibly remote) servers, allowing the system architect to focus on implementing the functional capabilities of the robot. In this case, the learning model itself is implemented as multiple threads in a single ADE server that calls on other ADE servers (e.g., vision) as needed.

During a teaching episode, the teacher faces the robot, monopolizing much of its visual field. The teacher holds the visual stimulus (a green ball that takes up roughly $1/16$ of the visual field) so that it is fully visible during the whole episode.

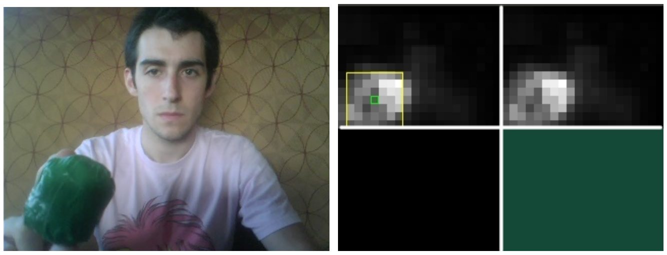

Fig. 9. Typical frame from visual stimulus motion, and saliency map visualization. The average color of the foveated region is indicated in the bottom-right sub-box; the upper-left is the final output, and the yellow square indicates the winning area (i.e., currently being foveated).

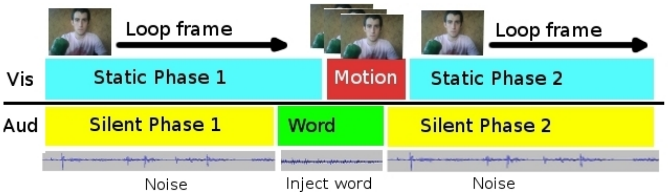

Fig. 10. The time line for an "unsupervised" learning episode. The video fed during the "static" phases is a loop of the frames immediately preceding and succeeding the motion phase, and the audio fed during the "silent" phases is background noise.

The ball is held still except during the object stimulus motion, during which the teacher waves it back and forth (primarily lateral movement no more than half the ball's size in the visual field). The auditory stimulus is spoken at the appropriate time (for the given offset), and the performance (i.e., the degree to which an association is established, which is proportional to the decrease in probability of looking to the object in the future) is recorded.

Note that it is impossible for any human to consistently present the stimuli with the degree of precision required for comparisons of very subtly differing (e.g., by 40 ms) conditions. Moreover, it will be very difficult to eliminate all sources of noise during the "non-stimulus" portions of a trial (e.g., variations in background audio or slight movements of the object, which could bias the system, or even lead to erroneous associations under certain circumstances). While the system is fairly robust to such noise, particularly over multiple episodes, it is useful to isolate the effects of the stimuli from the influence of noise while at the same time ensuring consistent, precise variations of each parameter.

The computational implementation affords us the flexibility to control those factors, as we have direct control over the inputs and how they are presented. We eliminate the sources of imprecision: instead of having an experimenter repeatedly present the stimuli to the robot, we construct audio and video input streams of the episode and feed these directly to the robot architecture for processing in real time, just as if the data were arriving by microphone and camera.

Image sequences were recorded under realistic conditions in the native frame rate (30 Hz) and format (320x240 raster) of the onboard camera (Figure 9, left). Timestamps for each incoming frame were logged to allow the accurate reconstruction of the stream, including realistic variation due to system

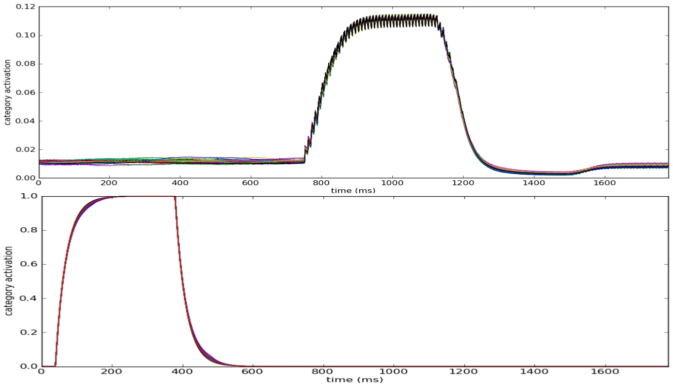

Fig. 11. Mean activations of the green-tuned neuron (top) and LSM word-probe neuron (bottom) for each offset in the 420 ms word condition. The word activation is the integrated output of a readout neuron, described in section V

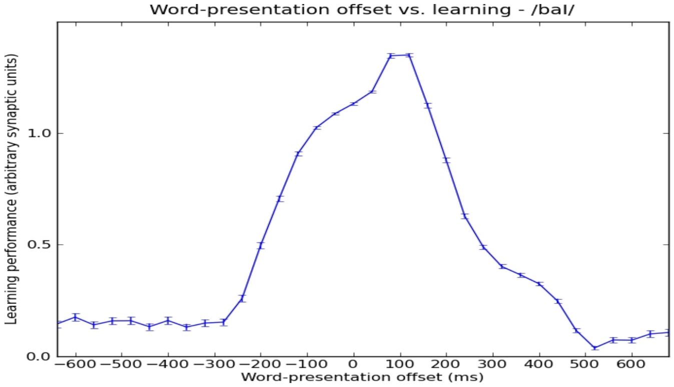

Fig. 12. Mean learning performance for 420 ms (/baI/) word trials (with standard error bars).

load, etc. Extraneous motion is eliminated from the pre- and post-stimulus phases by repeating the frame immediately preceding the motion onset for the entire pre-stimulus phase and repeating the frame immediately succeeding the end of motion for the entire post-stimulus phase. The robot architecture does not distinguish these "static" frames, so the effect is of the teacher holding the ball perfectly still. The audio stream is constructed in much the same way: the word itself is recorded and combined with recordings of "clean" background noise from the same environment for the pre- and post-stimulus phases. Figure 10 shows the progression of a learning trial conducted using these "idealized" input streams.

The lengths of the three phases of the video input are held constant for all trials (1600 ms pre-motion, 360 ms motion, 1200 ms post-motion). Hence, the motion stimulus onset is always 1600 ms, and the word-motion offset variation is introduced by manipulating word onset in the audio stream. The offset was varied from -680 ms to +720 ms, with a granularity of 40 ms, for a total of 35 offsets (and audio stream permutations). Two different word stimuli (/baI/ and /da/) were tested, with lengths of 420 ms and 250 ms, respectively, making non-overlapping presentations of the word stimulus possible both before and after the motion stimulus and providing examples in which the word length is longer and shorter than the motion length.

As noted above, the robot model runs in exactly the same way with the prepared input as it would with dynamic input; the only adjustment to the ADE servers is to allow sensory input from the recorded streams (e.g., a group of image frames instead of a camera). The learning module acquires frames in the same way whether they are live or recorded. The learning component processes the inputs in real time in multiple threads subject to operating system process scheduling, which introduces a small amount of variability in the phase lengths; the error is controlled by aborting and restarting trials in which the threads became more than 10 ms out of step. Figure 11 shows the mean output from the visual (top) and auditory (bottom) subsystems of the model for each of the offset values examined in the 420 ms word condition. The curves for each modality appear very similar across all offsets, and one-way ANOVAs for each modality with *offset condition* as the independent variable and *activation* as the dependent variable confirm that the differences are insignificant for both video (F(34,62265)=.222,p=1.0) and audio (F(34,62265)=.0001,p=1.0); we take this as verification that the architecture achieves the consistency desired and the results presented below are not attributable to input bias.

450 trials were conducted for each word/offset combination. To avoid possible effects of a preferred fixation length and to ensure that the model is not always in the same fixation state at the onset of the motion stimulus, the model is reset to its base state at a random point in the first 400 ms of each trial, which is well before the onset of either stimulus, and hence does not *directly* affect the resultant activity of either. The state of the model is logged at regular intervals throughout each trial, and each trial's outcome is recorded when it completes. To measure learning, instead of employing an analogue to the "switch" test used in the infant experiments, we calculate the net change in synaptic efficacy of the plastic synapses that run between the auditory recurrent circuit and the integration circuit neuron that corresponds to (in our case) the color of the stimulus.[9]

Figures 12 and 13 show for each of the two words the mean learning measured in 450 trials of each offset. Aside from minor differences in the curves (attributable to the two words' distinct activation patterns), both words exhibit the same general pattern: a range of word-motion offsets in which association strength is low (attributable to chance/noise), followed by a range in which learning is observed, and then another range of low association. In each case, the range in which learning occurs is roughly centered around the point at which the co-occurrence from the activations from

[9] This metric most accurately encodes the "strength" of the association between the auditory stimulus response and the visual stimulus response. It corresponds to the strength of the "bias" caused on the visual circuit by the auditory circuit when it is activated by the same auditory stimulus. It can thus be easily converted to a bias in looking times by having activation in the visual circuit inhibit the saliency of regions containing corresponding features. In the "same" case, activation of the associated auditory stimulus will push the vision circuit towards the trajectory that would correspond to the visual stimulus actually being present, inhibiting the saliency of the corresponding regions and resulting in lower probability of fixation and shorter fixations on those regions. Activation of a non-associated auditory stimulus ("control" condition), on the other hand, would not inject meaningful activation, and thus attention would not be allocated significantly differently than normal.

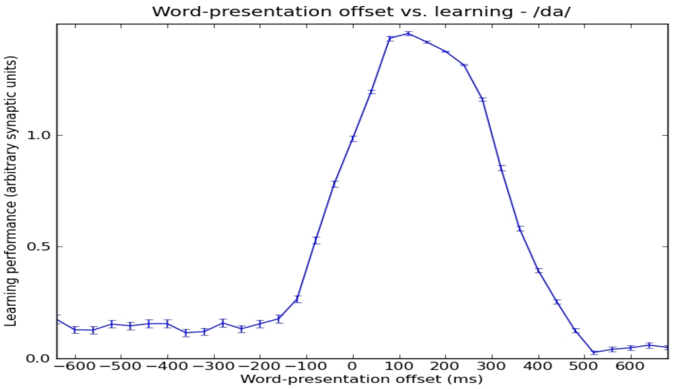

Fig. 13. Mean learning performance for 250 ms (/da/) word trials (with standard error bars).

the two sources is greatest. Note that this does not occur when the offset Δt=0 ms (i.e., when the presentations start simultaneously), but rather when Δt=120 ms, which is the peak for both words. This delay corresponds directly to the mean attention shift "lag" (an effect of the inhibition-of-return mechanism of the saliency map implementation) and, is thus due to the particular implementation of the robot model; in other implementations the peak could occur earlier or later.

Significant learning is observed only for offsets that allow a period of overlap in the two subsystems' activations. When $\Delta t < 0ms$, word length determines the amount of overlap (i.e., how much of the word activation phase remains when the color activation begins). Hence, the range of offsets for which learning occurs is wider and present for more negative offsets (see Figure 12). Conversely, when $\Delta t > 0ms$, the determinant of co-occurrence is motion length (i.e., how much of the word activation phase occurs before the word activation ends). The result is that learning trails off at roughly the same offset (520 ms) for both words.

A 2-way 2x35 ANOVA with *word* and *offset* as independent variables and *learning performance* as the dependent variable finds that both word (F(1,31430)=5.442,p=.02) and offset (F(34,31430)=2456.569,p <.001) are significant main effects. The main effect on word is obviously expected given the difference of the employed words. The more interesting main effect of offset indicates that learning performance does improve above the baseline in the region of overlap. Moreover, there is a highly significant 2-way interaction (F(34,31430)=250.138,p <.001) resulting from the different possible overlaps associated with different word lengths: longer words have more potential for overlap with the visual stimulus. Post-hoc analysis confirms that the average learning performance of every offset in the no-overlap ranges ($[-640, -240]$ and $[480, 720]$) is significantly lower than the average learning performance for every offset in ($[-200, 440]$), the region of overlap (p <.01, Tukey's HSD). In addition, for offsets in $[-240, 0]$, learning performance is greater for the 420 ms stimulus than for the 250 ms stimulus, while the reverse is true for offsets in $[80, 360]$ (p <.01, Tukey's HSD).

These results strongly indicate that stimulus co-occurrence (i.e., simultaneity) plays a key role in learning multimodal associations. Other effects are clearly present in the results (e.g., the kurtosis seen in Figure 12), but these are entirely unsurprising given the constraints imposed by our design goals (cf. Section III). The model is designed to be biologically plausible in both its constituent parts and their interactions; some variation is inevitable from one trial to the next, and characteristics of the neural architecture prevent the system from instantaneously tracking the input (e.g., the attention shift delay). In addition, the input, although "unnaturally" consistent across learning episodes, is taken from a real-world interaction and contains much of the noise that is normally found in data from audio and video sources. Finally, the two stimulus words are different (and therefore generate different activation patterns in the auditory circuit). However, these other effects are relatively small compared to the influence of overlap; co-activation appears to be the strongest indicator of learning performance.

B. Functional Category Learning

The evaluation above demonstrates that learning in the model scales with the amount of synchrony overlap between the auditory/visual stimuli. Further experiments were necessary to verify that the model components and learning mechanism are actually capable of categorization of stimuli into "habituated" stimuli ("same") condition and "novel" stimuli ("switch") condition, as infants are. Gogate and colleagues [17] determined that at 2 months of age, while infants were capable of habituating to one multimodal stimulus at a time (only if the multimodal stimuli were presented synchronously) they were unable to simultaneously habituate to two multimodal stimuli even if the stimuli were presented in synchrony.

To investigate the behavior of the model in similar circumstances, we simulated habituation of the model to multimodal word-color pairings and then tested the response of the model when presented with the habituated word versus a novel word. If the model was successfully able to extract information specific to the habituated word stimulus and associate it with the visual color stimulus, then the co-habituated (integration region) color neuron's activation should increase more when the "same" word is presented than when a non-co-habituated "switch" word is presented. Since both the "same" and "switch" word share a neural substrate (the auditory cortex), both will necessary excite the integration region neuron to some extent as a result of the changes in synaptic strengths that resulted from the co-habituation to the "same" word. However, simultaneous experience of the visual stimulus and the word stimulus should have caused weight changes that select specifically for properties of the auditory circuit's response to the "same" word, whereas no such learning will have happened for any non-co-experienced word stimuli. Thus, the channels between the auditory circuit and the integration region neuron should be set up in a way that selects for co-incidence of the "same" word and the visual stimulus, but not necessarily for co-incidence of the visual stimulus with any other word stimulus.

Two sets of stimuli were prepared to test the multimodal category-learning ability of the network. The same stimuli used in the synchrony experiments /baI/ and /da/ were compared, and also /baI/ was compared with a reversed version of itself /iab/ as an easy way to control for length. For each pair of stimuli, in two different sets of experiments each stimulus was used as both the habituated stimulus and as the novel stimulus in turn. Thus, 4 conditions were run, comparing (/baI/ habituated, /da/ novel), (/da/ habituated, /baI/ novel), (/baI/ habituated, /iab/ novel), and (/iab/ habituated, /baI/ novel). Within each condition, a novel model was generated randomly according to the process and parameters described in Section III, habituated to the target stimulus over 10 repetitions of the target word stimulus while simultaneously stimulating the color-responsive neuron in the integration region at 200Hz beginning at word onset and ending at word offset. 100 ms of audio/visual silence was inserted between each word repetition during which the network was updated but there was no input. Finally, both the habituated word and the non-habituated word were presented in turn (with 100 ms of separation) while hyperpolarizing the integration circuit color-responsive neuron so that it would not fire an action potential (this effectively prevented synapses from being modified during the test phase). The net change in membrane potential of the color-responsive neuron over the duration of each of the "same"/"switch" words was recorded. To measure the change between each of the 10 familiarization presentations, the test was actually administered between *every* stimulus repetition, and not just at the end of the 10 presentations. It was also administered once before any familiarization presentations to generate a baseline value (which was always zero because of the initial weight distribution).

In order to account for differing word-lengths and acoustic power (translating to increased pre-synaptic activity and thus increased post-synaptic activation) between words, the net change over the length of the word was divided by the total number of pre-synaptic spikes recorded over the duration of the word stimulus presentation. This is necessary because the auditory network is input-driven (i.e., it has no passive spiking activity without input), and so words with differing power can significantly change the amount of energy present in the system to begin with. This is a shortcoming that is recognized in the literature (solutions, such as synaptic tuning, self-scaling, etc., have been proposed) and that we plan to reconcile in the future (see discussion below). The number of presynaptic spikes simultaneously represents both the length of the word and the average power per unit time, and is thus an ideal normalization measure for our purposes. 500 trials were conducted for each condition (each with 10 repetitions for habituation and 11 tests of each word). Figures 14 and 15 plot the mean difference between the normalized *increase* in activation of the color-responsive integration neuron in response to each word stimulus being presented separately.

If the system is able to learn the correct categorization, it is expected that the difference be always positive, as this indicates that the co-habituated ("same") word is able to elicit more energy *per spike* than a novel ("switch") word which was never co-experienced with the visual color stimulus. Both plots show that this is indeed the case for all combinations of the stimuli tested. Paired one-sided T-tests for each of the 4 conditions and 11 repetitions (for each of the 44 possibilities, $t(499) > 3.75$, $p < 0.00001$) show that the observed difference is significantly positive for every case.

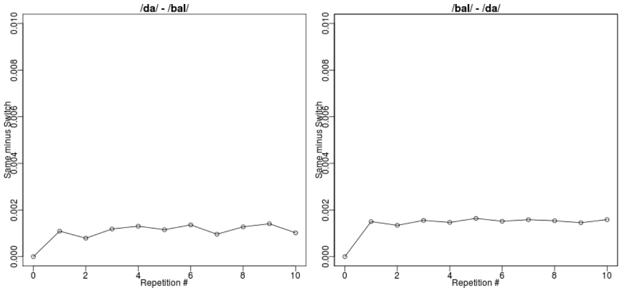

Fig. 14. Mean postsynaptic injected current difference *per spike* for stimuli /da/ and /baI/.

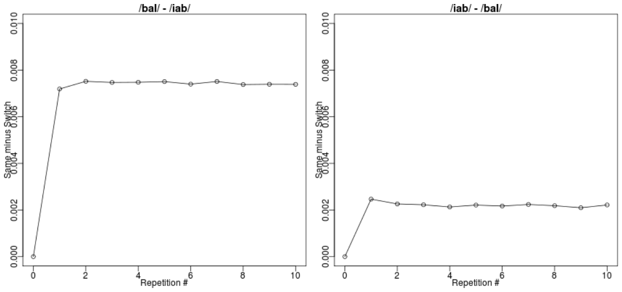

Fig. 15. Mean postsynaptic injected current difference *per spike* for stimuli /baI/ and /iab/.

The results demonstrate that with repeated synchronous presentation of the audio/visual stimulus pair, the model's dynamics are such that the strengths of weights will be organized to transfer more energy per spike for the corresponding auditory stimulus than for a novel auditory stimulus. However, it also seems that a single encounter with the stimulus is sufficient for the model to extract all the information necessary to maximally select for that stimulus. This is possibly due to learning rate-related parameters in the model being too high, or the visual stimuli representations being excessively simple. However, it also raises an important issue, which is that stimuli are continuous and temporally extended. To an infant, what constitutes a single "stimulus encounter" is not well understood. For instance, how much effect does exposing an infant to only a portion of an auditory stimulus have? Empirical studies do not address this issue, treating stimulus presentations as discrete and atomic. This is the first attempt to bridge the explanatory gap which exists between mathematical models of habituation (e.g. [55], [56]) which measure stimuli in arbitrary dimensions ("presentations") and the actual real-world stimuli and motor motions which produce the empirical measurements that lead to those models.

There is some evidence that infants' systems could be interpreted as experiencing the world as atomic, discrete periods of stimulus presentations. This is based on the observation that infants enter periods of "attention" (alertness/arousal, usually measured in terms of deccelerated heart-rate), during which they are better able to learn. In the extreme, when not in an alert state, the infant is not able to habituate. It has been shown that presenting a stimulus to infants for 5 seconds during the "sustained attention" phase results in the same amount of learning that presenting a stimulus for 20 seconds (at a random time) normally would [57]. In neonates, alert states are endogenously generated and cyclic, similar to the circadian rhythm. As the infant develops, it becomes possible to elicit alert states using engaging external stimuli [58], [59]. However, other research shows that infants actually *recognize* stimuli less well when not aroused. This is taken to imply that at least part of the observed benefit of arousal is that it facilitates the *processing* of stimuli, instead of (just) facilitating learning [58], [59].

If interpreted as recruitment of low-level functions that facilitate learning and/or encoding/processing, then several mechanisms could be postulated involving modulations to the overall system by arousal state that move the system into states receptive to learning. The identity of the biological structures and their functions will need to be addressed in future models as they obviously play a major role related to stimulus processing and learning in infants. Arousal is thought to be elicited functionally by ascending (subcortical) pathways from the brainstem reticular formation. These target common limbic and cortical regions, including those that figure centrally in the present model such as the parahippocampal structures. The functional effects of the release of these (extracellular/horomonal) neurotransmitters (noradrenaline, acetylcholine, seratonin, and dopamine) associated with these ascending pathways probably play a large role in learning and memory, and will need to be addressed in future model iterations.

For example, we can hypothesize based on neurobiological and computational modelling evidence that the arousal functions described are involved in the generation of beneficial passive/baseline circuit activity, bias of afferent input over intracortical connections, and even increased synaptic plasticity (role of ACh is reviewed in [60], for reward-based plasticity see [61], [62]). Such baseline activity could also prove to be a solution to the strong dependence of the current model on stimulus energy, and the bias towards afferent inputs could have the effect of tightening cortical representations of stimulus input for each modality, producing higher-fidelity associations.

C. Discussion of Results/Model Evaluation

Based on the model evaluations presented above, it is clear that several improvements can be made to future models to account for additional properties of recognition memory and multimodal habituation as they occur in human infants.[10] Many of these seem to stem from oversimplification of the target biological circuit, which is not necessarily negative, since our goal was to determine a minimal functional circuit. While it turned out that the proposed circuit was able to demonstrate some aspects of infant multimodal habituation, it was unable to account for others. An example of this is the current model's dependence on low-level stimulus energy. Some desirable properties that are expected to alleviate the dependence of results on stimulus power include: baseline network activity, improved synaptic learning mechanisms, and more processing of stimuli before they reach the plastic synapses. The addition of an additional cortical layer to model the internal dynamics of the integration region (in our case peri/entorhinal cortices) is expected to solve this last problem, and in addition will bring the model closer to the actual anatomy. With baseline activity the amount of energy in the network would stay relatively constant, but the distribution of energy would be perturbed by incoming stimuli. This would reduce the drastic change in network response to different stimuli, as is observed in the current input-driven network. It is known that the human cortices demonstrate baseline activity, especially during periods of alertness elicited by ACh diffusion. Integration of similar horomonally-induced changes is a possible solution to this problem, though balancing complex models to have constant baseline activity is not a trivial problem. Finally, the Song et al model of STDP implemented in the current model is one of the simplest computational models of STDP and only takes into account the relative timing of every pre-post spike pair. It does not take into account recent data and models which recognize the modulatory effects of spike triples or quadruples (e.g., [63], [64]) which have been shown to better approximate the functions defined by pre-synaptic input [65].

V. General Discussion

The computational model of infant word-referent learning presented here includes biologically plausible neural components of the relevant portions of the auditory/visual sensory, attentional control, and sensory integration systems of human infants. The robotic implementation of the model was tested using an experimental paradigm taken directly from the infant learning literature. The results demonstrate that the robot's responses to manipulations of the relative timing of the presentation of auditory/visual stimuli are consistent with those of young human infants in early language learning studies.

Most notably, our model is capable of learning multimodal categories without needing to (explicitly) learn unimodal categories. Agnosticism with respect to explicit categories sets this model apart from previous models of word-referent association (addressed in section VI), which construe it as the association between independent representations abstracted away from the temporally extended processes that activate them. However, the results described here demonstrate that it is precisely the description at sub-stimulus temporal timeframes that can explain the synchrony/asynchrony phenomena in young infants. In our model, stimuli are only uniquely identified by the particular temporally-extended trajectory of circuit state space through which a recurrent circuit travels in response to the stimulus being presented. Thus, our explanation of the mechanism of association learning must operate only on the temporally local state of the system. A consequence of this is that the circuit cannot "represent" input for more than about 30 *milliseconds* (the time constant of the neural membranes). This does not seem to be enough to, e.g., recognize whole words (or even phonemes) and associate them with visual objects at a category level. The fact that the model can accomplish robust association learning under these extreme constraints is perhaps

[10] The authors would like to thank the two anonymous reviewers for their constructive comments which enabled these additional avenues for improvement of the model.

its most interesting and novel contribution. It demonstrates that the multimodal associations can be accounted for by very simple mechanisms that are driven entirely by perceptual input in very simple circuits. It also explains the reason behind the synchrony and environmental variables that are observed to constrain the learning behavior.

This is not to say that unimodal categories never form, or do not exist at all. Indeed, the trajectory of state-space that the circuit takes in response to a stimulus can be thought of as an "implicit" category representation. It is possible to extract the equivalent of an "explicit" category representation by taking a specific weighted projection of the neural state. This projection will respond only when it is likely that the sequence of states the circuit travels through corresponds to that of the stimulus the projection classifies. The projection is an example of "readouts" presented in [28], and we used one to represent "word activation".[11]

The weights learned via STDP that project into the integration circuit can be viewed as a special type of "readout", specifically one that represents the existence or non-existence of a category in the auditory circuit *in terms of* the integration circuit's state-space. It is as if specialized categories whose only purpose is to "inform" the integration circuit (and then bias the visual circuit, to bias looking) have been learned. These categories can be seen as similar to "action-based representations" in that they only encode the category in the way necessary for it to perform its role in producing system behavior (in our case, biasing looking behavior). Since the post-synaptic vision circuit activation effectively served as a supervisor (but only during synchronous presentations!), this is a satisfyingly plausible conceptualization of what is happening.

Another interesting hypothesis is that the development of more abstract, time-invariant categories (like readout neurons) can explain later phases in infant development, such as the development of unimodal categories. The association between these time-invariant categories could then be referenced to explain later lexicon development (i.e., permanent word-referent association).

Although the performance of the computational model is encouraging, there are areas where improvements can be made. The empirical results suggest some unforeseen effects of the model implementation. These should be addressed since they may cause the behavior of the model to diverge from the behavior of the modeled infant in model extensions.

One such effect is a result of the adopted association mechanism, STDP. Because *every* pre- or post-synaptic spike results in depression or potentiation of the synaptic weight, denser spiking activity will cause much faster weight change. The circuits receive input directly from world-facing feature detectors, so input activity is only lightly filtered and thus may reflect physical properties of the stimulus itself. In addition, the model of synaptic dynamics that is implemented in the recurrent circuit synapses makes it likely that the onset of a stimulus will evoke high firing rates, and it will take a finite amount of time before the firing stabilizes. This makes it more likely that weight change will occur during the more active portion of the stimulus presentation, which is not something that we took into account. In the end, the distribution of firing rates during the course of a stimulus can affect the amount of learning that occurs, in addition to stimulus overlap/synchrony. Whether this is also the case with infants cannot be determined decisively, since it is not possible to sufficiently control the timing of their responses to stimuli nor to accurately measure the "amount" of learning that resulted from one particular auditory-visual stimulus presentation. However, it is certainly a phenomenon that could have a significant effect on the performance space of learning of the associations, and the exact implications should be explored in future studies.

VI. Related Work

The use of robots in the study of language learning is not unique to the present study. Sugita and Tani [66] proposed a neural model that could learn grammatical sentences and their meanings via association with referent action sequences. Their approach to language learning focuses on motor experience, in particular on sequences of motor actions and/or proprioceptive feedback from them. The model was implemented on a robot, which was then required to learn the arbitrary relation between linguistic expressions and their meanings based on supervised co-occurrence between the motor modality and the linguistic modality. Thus, while their proposed model takes advantage of the co-activation phenomenon investigated above, they do not investigate how the temporal relationship between the two modalities affect learning. Moreover, the Sugita and Tani model is intended to demonstrate a principle without being constrained to fit any specific empirical data, unlike the model described here, which was developed specifically to model multimodal information integration and association learning as it occurs in early infants who are engaged in interaction involving "showing" and naming behavior by a parent.

A robotic model for the more specific task of word-referent association learning was proposed by Roy and Pentland [67]. Similar to the model presented in this paper, their model operates directly on unprocessed visual and auditory inputs and relies on the co-occurrence of multimodal stimuli to generate associations. However, there are important differences between the two models. For one, the Roy and Pentland model was not explicitly designed with the goal of biological plausibility or with the intent to model the behavior of a specific age of infant. The model associates strings of phoneme probabilities generated by a pre-trained recurrent neural network ("words") with pre-trained visual objects (defined by histograms of local shapes from various angles of the unoccluded object). It does this by storing instances of co-occurrences as "prototypes" in a "long-term memory" list, and then saving them as "lexical items" if the mutual co-information between the word and object is high enough. A "short-term-memory" decides when to introduce prototypes into LTM based on how often phoneme

[11] The 135-dimensional state of the auditory circuit had to be visualized, so a readout neuron was trained via linear regression to respond positively when the circuit state corresponded to the word being present, and negatively otherwise. The instantaneous activation can be low-pass filtered with an exponential decay kernel (time constant 30 ms) to obtain the continuous function that is plotted in Figure 11.

sub-sequences co-occur with the visual categories. In contrast, our computational model is specifically based on empirical and neuroscientific results about the abilities of young infants. It does not require any pre-training or pre-processing of auditory or visual streams, nor does it rely on abstract structures such as long-term or short-term memory buffers. In fact, our model makes no assumptions about the (pre-)existence of auditory or visual categories (not even of phonemes), as it is not even possible for the model to explicitly learn or extract "category" representations (e.g., of words) in the way required for the Roy and Pentland model. All parameters of the network specific to each modality (i.e., the connections from the feature-tuned neurons to the recurrent circuits, and synapses within the recurrent circuits) are determined randomly at the beginning, and do not change.

Another proposal by Rolf et al. [68] uses a robotic model to investigate the distinguishing characteristics of child-directed versus adult-directed teaching, and in particular the performance of different models/parameterizations of attention allocation. They found that child-directed teaching tended to involve more synchrony than adult-directed teaching. As such, attentional models that are tuned to this synchrony will tend to direct gaze more beneficially for infants who need this synchrony to learn the associations. Our model complements that work by focusing on the effect of varying degrees of synchrony on learning itself. Thus, it explains the crucial importance for learning of both a correct attentional system (that will focus on stimuli with high synchrony) *and* a caretaker sensitive to the constraints of the system (i.e., who will tend to present multimodal stimuli synchronously). It also explains how learning breaks down when one of these criteria is not fully met (e.g., in a situation where the caretaker is presenting stimuli in an adult-directed fashion—not reliably maintaining synchrony). In sum, our model highlights the fact that the word-referent learning problem in young infants is really dependent on the combination of the infant's internal configuration (e.g., attentional system) and the correct social action of the caretaker sensitive to the young learner's needs for synchrony. It shows what will happen when the latter criterion breaks down to different degrees, and offers a platform to investigate what will happen when the former is not parameterized correctly.

VII. Conclusion

We presented an embodied cognitive model of word-referent association learning based on a biologically plausible neural architecture. Implemented on a robot, the model demonstrates young human infant-like word-referent association behavior in response to real-world stimuli in a learning task taken from developmental psychology studies of learning in young infants. Confirming empirical findings, it shows sensitivity to the same timing constraints as young human infants: multimodal stimuli need to be presented simultaneously for the looking-time biasing effects of habituation to occur.

The model and evaluation task were explicitly designed to replicate the critical architectural, environmental, and social aspects of word-learning in infants. Great care was taken to ensure that the architecture is consistent with current accounts of infant cognition in psychology and neuroscience. Realistic processing of environmental cues (i.e., auditory and visual stimuli) is ensured by using unprocessed sensory stimuli that are directly fed into the relevant architecture components. Successful learning in the model relies on social interaction, and the experimental paradigm explicitly allowed us to vary the parameters of the interaction to determine their effects on learning performance.

While the model is interesting in itself, it has also enabled us to run empirical trials with a high level of stimulus timing control (which is impossible with real infants) to reveal what learning occurs for different levels of synchrony between the visual and auditory stimuli. We found that the shape of this space is succinctly approximated when framed as the "co-activation" of the neural response to the visual and auditory stimuli respectively, i.e., the amount that they overlap. This supports previous hypotheses which suggested that synchrony and temporal contiguity were necessary for multimodal habituation learning, but also provides a mechanistic explanation for *why* these conditions must be met.

The empirical results also explain why parents adopt different behaviors when teaching children the names of things compared to when they do the same for adults, highlighting the crucial role that the parent's environmental scaffolding plays in the early development of language in human infants. Moreover, the results highlight how learning mechanisms at play in infants at these early stages of development differ significantly from those in older infants who have more complex cognitive structures to rely on.

A logical next step for the model development would thus be to expand the model and use it to explain the change in language capabilities of infants as they enter subsequent developmental stages and additional cortical regions develop and become active. For instance, the model may be able to provide insight into the question of how symbolic categories develop and become associated to form lexical entries later in life, and what role the early multimodal associations play in that development.

It will also be beneficial to explore how including the ability to manipulate objects, as seen in slightly older infants, affects learning. Older infants demonstrate a tendency to bring attended objects closer to their faces, which seems to result in self-initiated feedback loops that configure the environment in a way that is beneficial to the learning problem. This would be an analogue to the parent-initiated scaffolding presented in this paper, and demonstrate how developed cognitive abilities will actually begin to take over for the work that must be done by an external social interactor for very young infants.

Finally, physical motion (of the infant's eyes and head) serves both to reorganize the perceptual field and to provide cues to the caregiver regarding the infant's learning processes. This can result in interesting feedback loops and couplings between the teacher and learner which may have effects on the dynamics of the neural model. Interactive experiments with real humans teaching the robot could expose the effects of these head and eye motions. While the model has been tested on what is theoretically the full range of possible behaviors that could be adopted by human teachers, it would be interesting to see whether distinct teaching strategies exist and investigate how close people actually get to presenting stimuli at the optimal time for the learner, especially when the learner is a robot.

Acknowledgment

The authors would like to extend special thanks to Xuan Wu for his help in preparing the experiments and in the data analysis.

References

[1] E. Bates and B. MacWhinney, "Competition, variation, and language learning," in *Mechanisms of language acquisition*, B. MacWhinney, Ed. Psychology Press, 1987.

[2] C. Yu and D. H. Ballard, "Understanding human behaviors based on eye-head-hand coordination," in *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*. London, UK: Springer-Verlag, 2002, pp. 611–619.

[3] L. B. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, pp. 1558–1568, 2008.

[4] C. Yu and L. B. Smith, "Rapid word learning under uncertainty via cross-situational statistics," *Psychological Science*, vol. 18, pp. 414–420, 2007.

[5] L. J. Gogate, L. E. Bahrick, and J. D. Watson, "A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures," *Child Development*, vol. 71, no. 4, pp. 878–894, 2000.

[6] P. Zukow-Goldring, "A social ecological realist approach to the emergence of the lexicon: educating attention to amodal invariants in gesture and speech," in *Evolving explanations of development: Ecological approaches to organism-environment systems*, C. Dent-Read and P. Zukow-Goldring, Eds. American Psychological Association, 1997.

[7] L. J. Gogate, A. S. Walker-Andrews, , and L. E. Bahrick, "The intersensory origins of word comprehension: an ecological-dynamic systems view," *Developmental Science*, vol. 4, no. 1, pp. 1–37, 2001.

[8] E. B. Alan Slater and M. Badenoch, "Intermodal perception at birth: Newborn infants' memory for arbitrary auditoryvisual pairings," *Early Development and Parenting*, vol. 6, pp. 99–104, 1997.

[9] P. K. Kuhl and A. N. Meltzoff, "The bimodal perception of speech in infancy," *Science*, vol. 218, pp. 1138–1141, December 1982.

[10] ——, "The intermodal representation of speech in infants," *Infant Behavior and Development*, vol. 7, pp. 361–381, 1984.

[11] ——, "Speech as an intermodal object of perception," in *Minnesota symposia on child psychology: Vol. 20. The development of perception*, A. Yonas, Ed. Hillsdale, NJ: Erlbaum, 1988.

[12] L. E. Bahrick, "Infants' perception of substance and temporal synchrony in multimodal events," *Infant Behavior and Development*, vol. 6, no. 4, pp. 429–451, 1983.

[13] ——, "Infants' perceptual differentiation of amodal and modality-specific audio-visual relations," *Journal of Experimental Child Psychology*, vol. 53, no. 42, pp. 180–199, 1992.

[14] R. Lickliter and L. E. Bahrick, "Perceptual development and the origins of multisensory responsiveness," in *The Handbook of Multisensory Processes*, G. A. Calvert, C. Spence, and B. E. Stein, Eds. MIT Press, 2002.

[15] L. E. Bahrick, "The development of infants' sensitivity to arbitrary intermodal relations," *Ecological Psychology*, vol. 6, no. 2, pp. 111–123, 1994.

[16] L. J. Gogate and L. E. Bahrick, "Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven–month–old infants," *Journal of Experimental Child Psychology*, vol. 69, pp. 133–149, 1998.

[17] L. J. Gogate, C. G. Prince, and D. J. Matatyaho, "Two–month–old infants sensitivity to changes in arbitrary syllable-object pairings: The role of temporal synchrony," *Journal of Experimental Child Psychology*, vol. 35, no. 2, pp. 508–519, 2009.

[18] L. J. Gogate, "Learning of syllable-object relations by preverbal infants: The role of temporal synchrony and syllable distinctiveness," *Journal of Experimental Child Psychology*, vol. 105, pp. 178–197, 2010.

[19] J. F. Werker, L. B. Cohen, V. L. Lloyd, M. Casasola, and C. L. Stager, "Acquisition of word-object associations by 14-month-old infants," *Developmental Psychology*, vol. 34, no. 6, pp. 1289–1309, 1998.

[20] R. M. Golinkoff, K. Hirsh-Pasek, K. M. Cauley, and L. Gordon, "The eyes have it: lexical and syntactic comprehension in a new paradigm." *Journal of child language*, vol. 14, no. 1, pp. 23–45, February 1987.

[21] P. S. Kaplan and M. J. Owren, "Dishabituation of visual attention in 4-month-olds by infant-directed frequency sweeps," *Infant Behavior and Development*, vol. 17, no. 4, pp. 347–358, 1994.

[22] R. Lickliter, "Premature visual experience facilitates visual responsiveness in bobwhite quail neonates," *Developmental Psychobiology*, vol. 23, pp. 15 – 17, 1990.

[23] ——, "Premature visual experience facilitates visual responsiveness in bobwhite quail neonates," *Infant Behavior and Development*, vol. 13, no. 4, pp. 487 – 496, 1990.

[24] R. Lickliter and J. Stoumbos, "Enhanced prenatal auditory experience facilitates species-specific visual responsiveness in bobwhite quail chicks (colinus virginianus)," *Journal of Comparative Psychology*, vol. 105, no. 1, pp. 89 – 94, 1991.

[25] P. C. Quinn, "The categorical representation of visual pattern information by young infants," *Cognition*, vol. 27, pp. 145–179, 1987.

[26] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82*, May 1982, pp. 1282–1285.

[27] M. Slaney, "Lyon's cochlear model," Apple Computer Inc. Cupertino, Ca., Tech. Rep. 13, 1998.

[28] W. Maass, T. Natschlager, and H. Markram, "Real–time computing without stable states: A new framework for neural computation based on perturbations," *Neural Computation*, vol. 14, no. 11, pp. 2531–2560, 2002.

[29] H. Markram, Y. Wang, and M. Tsodyks, "Differential signaling via the same axon of neocortical pyramidal neurons," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 9, April 1998, pp. 5323–5328.

[30] R. Legenstein, C. Naeger, and W. Maass, "What can a neuron learn with spike-timing-dependent plasticity?" *Neural Comput.*, vol. 17, no. 11, pp. 2337–2382, 2005.

[31] S. G. Solomon and P. Lennie, "The machinery of colour vision," *Nature Reviews Neuroscience*, vol. 8, pp. 276–286, 2007.

[32] D. L. Adams and S. Zeki, "Functional organization of macaque v3 for stereoscopic depth," *Journal of Neurophysiology*, vol. 86, pp. 2195–2203, 2001.

[33] M. S. Livingstone, "Mechanisms of direction selectivity in macaque v1," *Neuron*, vol. 20, pp. 509–526, 1998.

[34] T. D. Albright, R. Desimone, and C. G. Gross, "Columnar orgnization of directionally selective cells in visual area mt of the macaque," *Journal of Neurophysiology*, vol. 54, no. 1, pp. 16–31, 1984.

[35] K. R. Gegenfurtner, "Cortical mechanisms of color vision," *Nature Reviews Neuroscience*, vol. 4, pp. 563–572, 2003.

[36] Y. Xiao, Y. Wang, and D. J. Felleman, "A spatially organized representation of colour in macaque cortical area v2," *Nature*, vol. 421, pp. 535–539, 2003.

[37] B. R. Conway and D. Y. Tsao, "Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex," *PNAS*, vol. 106, no. 42, pp. 18 034–18 039, 2009.

[38] T. N. Mundhenk and L. Itti, "Computational modeling and exploration of contour integration for visual saliency," *Biological Cybernetics*, vol. 93, pp. 188–212, 2005.

[39] J. Bourne, "Unravelling the development of the visual cortex: implications for plasticity and repair," *Journal of Anatomy*, vol. 217, pp. 449–468, 2010.

[40] M. H. Johnson, "Cortical maturation and the development of visual attention in early infancy," *J. Cognitive Neuroscience*, vol. 2, no. 2, pp. 81–95, 1990.

[41] A. King, J. Schnupp, S. Carlile, A. Smith, and I. Thompson, "The development of topographically-aligned maps of visual and auditory space in the superior colliculus," *Progressive Brain Research*, vol. 112, pp. 335–350, 1996.

[42] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[43] L. Seress, "Morphological changes in human hippocampal formation from midgestation to early childhood," in *Handbook of developmental cognitive neuroscience*, C. A. Nelson and M. Luciana, Eds. MIT, 2001, pp. 45–58.

[44] J. Bachevalier, M. Brickson, and C. Hagger, "Limbic-dependent recognition memory in monkeys develops early in infancy," *NeuroReport*, vol. 4, pp. 77–80, 1993.

[45] O. Pascalis and J. Bachevalier, “Neonatal aspiration lesions of the hippocampal formation impair visual recognition memory when assesed by paired-comparison task but not by delayed nonmatching-to-sample task,” *Hippocampus*, vol. 6, pp. 609–616, 1999.

[46] C. Nelson, “The ontogeny of human memory: a cognitive neuroscience perspective,” *Developmental Psychology*, vol. 31, no. 5, pp. 723–738, 1995.

[47] J. Haas, T. Nowotny, and H. Abarbanel, “Spike-timing-dependent plasticity of inhibitory synapses in the entorhinal cortex,” *Journal of Neurophysiology*, vol. 96, pp. 3305–3313, 2006.

[48] M. Witter, H. Groenewegen, F. L. de Silva, and A. Lohman, “Functional organization of the extrinsic and intrinsic circuitry of the parahippocampal region,” *Progress in Neurobiology*, vol. 33, pp. 161–253, 1989.

[49] M. Mishkin, W. Suzuki, D. Gadian, and F. Vargha-Khadem, “Hierarchical organization of cognitive memory,” *Philosophical transactions of the Royal Society of London B*, vol. 352, pp. 1461–1467, 1997.

[50] R. Desimone, “Neural mechanisms for visual memory and their role in attention,” *PNAS*, vol. 93, no. 24, pp. 13 494–13 499, 1996.

[51] ——, “Visual attention mediated by biased competition in extrastriate visual cortex,” *Philosophical transactions of the Royal Society of London B*, vol. 353, pp. 1245–1255, 1998.

[52] G. Bi and M. Poo, “Synaptic modifications by correlated activity: Hebbs postulate revisited,” *Ann. Rev. Neurosci.*, vol. 24, pp. 139–166, 2001.

[53] S. Song, K. D. Miller, and L. F. Abbott, “Competitive hebbian learning through spike-timing-dependent synaptic plasticity,” *Nature Neuroscience*, vol. 3, no. 9, pp. 919–926, September 2000.

[54] M. Scheutz, “ADE - steps towards a distributed development and runtime environment for complex robotic agent architectures,” *Applied Artificial Intelligence*, vol. 20, no. 4-5, pp. 275–304, 2006.

[55] G. Shoner and E. Thelen, “Using dynamic field theory to rethink infant habituation,” *Psychological Review*, vol. 113, no. 2, pp. 273–299, 2006.

[56] S. Sirois and D. Mareschal, “An interacting systems model of infant habituation,” *Journal of Cognitive Neuroscience*, vol. 16, no. 8, pp. 1352–1362, 2004.

[57] J. E. Richards, “Effects of attention on infants’ preference for briefly exposed visual stimuli in the paired-comparison recognition-memory paradigm,” *Developmental Psychology*, vol. 33, no. 1, pp. 22–31, 1997.

[58] J. Colombo, “The development of visual attention in infancy,” *Annual Review of Psychology*, vol. 52, pp. 337–367, 2001.

[59] L. Hainline, “Summary and commentary: Eye movements, attention, and development,” in *Cognitive Neuroscience of Attention: A developmental perspective*, J. E. Richards, Ed. Lawrence Earlbaum, 1998, pp. 163–178.

[60] M. E. Hasselmo, “The role of acetylcholine in learning and memory,” *Current opinion in neurobiology*, vol. 16, pp. 710–715, 2006.

[61] R. Legenstein, D. Pecevski, and W. Maass, “A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback,” *PLoS Computational Biology*, vol. 4, no. 10, pp. 1–27, 2008.

[62] ——, “Theoretical analysis of learning with reward-modulated spike-timing-dependent plasticity,” *Advances in Neural Information Processing Systems*, vol. 20, pp. 881–888, 2008.

[63] R. C. Froemke and Y. Dan, “Spike-timing-dependent synaptic modification induced by natural spike trains,” *Nature*, vol. 416, pp. 433–438, 2002.

[64] A. Morrison, A. Aertsen, and M. Diesmann, “Spike-timing-dependent plasticity in balanced random networks,” *Neural Computation*, vol. 19, pp. 1437–1467, 2007.

[65] R. Gutig, R. Aharonov, S. Rotter, and H. Sompolinsky, “Learning input correlations through nonlinear temporally asymmetric hebbian plasticity,” *The Journal of Neuroscience*, vol. 23, no. 9, pp. 3697–3714, 2003.

[66] Sugita and J. Tani, “Learning semantic combinatoriality from the interaction between linguistic and behavioral processes,” *Adaptive Behavior*, vol. 13, no. 1, pp. 33–52, 2005.

[67] D. Roy and A. Pentland, “Learning words from sights and sounds: a computational model,” *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

[68] M. Rolf, M. Hanheide, and K. Rohlfing, “Attention via synchrony: Making use of multimodal cues in social learning,” *Autonomous Mental Development, IEEE Transactions on*, vol. 1, no. 1, pp. 55–67, May 2009.

Richard Veale Richard Veale received degrees in computer science (B.S., 2008) and philosophy (B.A., 2008) from Ursinus College (Pennsylvania, USA). He is currently a Ph.D. student in the Cognitive and Computer Science joint degree program at Indiana University, Bloomington. His research interests include neural and developmental approaches to understanding how the environment and body scaffold traditionally “cognitive” learning behaviors, and how these findings can be used to build more intelligent robots.

Paul Schermerhorn Paul Schermerhorn received a degree in philosophy (M.A. 1999) from Northern Illinois University (DeKalb, IL) and degrees in computer science (M.S. 2002, Ph.D. 2006) from the University of Notre Dame (Notre Dame, IN). Paul is currently a research scientist in the Human-Robot Interaction Laboratory of the Cognitive Science Program at Indiana University Bloomington. His research interests include affective robot control systems and agent-based modeling of social interactions in biological agents.

Matthias Scheutz Matthias Scheutz received degrees in philosophy (M.A. 1989, Ph.D. 1995) and formal logic (M.S. 1993) from the University of Vienna and in computer engineering (M.S. 1993) from the Vienna University of Technology (1993) in Austria. He also received the joint Ph.D. in cognitive science and computer science from Indiana University in 1999. Matthias is currently an associate professor of computer and cognitive science in the Department of Computer Science at Tufts University. He has over 100 peer-reviewed publications in artificial intelligence, artificial life, agent-based computing, natural language processing, cognitive modeling, robotics, human-robot interaction and foundations of cognitive science. His current research and teaching interests include multi-scale agent-based models of social behavior and complex cognitive and affective robots with natural language capabilities for natural human-robot interaction.