

Cued multimodal learning in infancy: a neuro-computational model

Thomas Hannagan (thom.hannagan@gmail.com)

Laboratoire de Psychologie Cognitive, CNRS and Aix-Marseille University
3, place Victor Hugo, 13331 Marseille, France

Rachel Wu (r.wu@bbk.ac.uk)

Centre for Brain and Cognitive Development, Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London, WC1E 7HX, UK

Abstract

We introduce a connectionist model of cued multimodal learning in infants. Its architecture is inspired by computational studies coming both from the fields of infant habituation and of visual attention. The model embodies in its simplest form the notion that the attentional system involves competitive networks (Lee et al, 1999). Using this model, we are able to reproduce experimental differences in looking times between cued and non-cued conditions. We then show that differences between social and non-social cues recently observed in 8-month-old infants by Wu and Kirkham (2010) can be explained by the amount of information let through from non-cued locations. We discuss these results and future lines of research on this computational work.

Keywords: Computational model; connectionism; cue-target binding; eye-tracking; cognitive development; social cues; eye gaze; multimodal learning.

Introduction

In a busy multimodal world, infants must parse useful information from a swirl of perceptual events. One way to accomplish this is relying on attention cues to guide them to relevant learning events. Many attention cues can guide infants' attention, but which ones help infants learn what to learn?

Recent work has shown that following social cues can shape learning: Some studies have focused on word mapping (e.g., Gliga & Csibra, 2009; Houston-Price, Plunkett, & Duffy, 2006; Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2006; Yu, Smith, & Pereira, 2008) and learning linguistic structures (Goldstein & Schwade, 2008; Thiessen, Hill, & Saffran, 2005). For example, 15-month-olds are able to follow a turning face to an object, and then map a spoken word onto that object rather than a non-cued object (Houston-Price et al., 2006).

Most relevant to this article, Wu and Kirkham (2010) –hereafter W&K – showed that social cues (e.g., a turning face that used infant-directed speech) produce better spatial learning of audio-visual events than non-social cues (i.e., flashing squares that shift attention to the target location), by 8 months of age. With non-social cues (flashing squares), 8-month-olds learned only cued locations regardless of multimodal information. This study measured infants' gaze behavior when they were presented with

dynamic audio-visual events (i.e., cats moving to a bloop sound and dogs moving to a boing sound) in white frames in the corners of a black background. An object's appearance in a spatial location consistently predicted a location-specific sound. On every familiarization trial, infants were shown identical audio-visual events in two diagonally opposite corners of the screen (i.e., two valid binding locations). To test the effects of attentional cueing on audio-visual learning, either a social (i.e., a real face) or non-social (i.e., colorful flashes) cue shifted infants' attention to one of the two identical events on every trial. For the social cue, a face appeared, spoke to the infant, and turned to one of the lower corners containing an object. For the non-social cue, a red flashing square wrapped around the target frame appeared and disappeared at a regular interval (i.e., flashed continuously) throughout the familiarization trial. During the test trials, only the four blank frames were displayed on the screen while one of the sounds played.

The purpose of this article is to characterise the neural mechanisms at work in infants when they are performing this task, without losing track of the interaction between infants and their environment throughout the task. In other words, the model's outputs (where it is going to "look") should determine its next inputs (what it will "see" next). Previous computational work has dealt with isolated aspects of the paradigm used in W&K. The HAB model (Sirois & Mareschal, 2004) can successfully account for robust non-linearities in infant preferential looking data, using two interacting auto-associator networks that learn under opposite principles. However, HAB neither incorporates multimodal learning nor attentional cueing, and its outputs do not determine its inputs. On the other hand, Mozer and Sitton (1998) proposed a computational model of visual attention that embodies the notion of an attentional "spotlight" and accounts for several cueing effects. In order to prevent interference when multiple objects are processed in a single hierarchical network, the authors used a winner-take-all network (WTA) that acted so as to attend to one input region while filtering the others. Importantly, the amount of information filtered in unattended regions was critical to determine attentional shifts. However, explaining the differences between social and non-social cues in a multimodal learning paradigm such

as used in W&K was beyond the scope of this model, since it was trained exclusively in the visual modality.

In an attempt to bridge this gap between the two fields we introduce a neuro-computational model that generates a proper sequence of saccades to learn from cued multimodal events. The main challenge in this endeavour was to connect different computational models without producing an intractable model. Our guideline was – in the words of A. Einstein – that the model should be as simple as possible to meet this goal, but not simpler.

Model

The model (illustrated in Figure 1) is essentially an adaptation of Sirois and Mareschal's architecture for infant habituation (Sirois & Mareschal, 2004), combined with Mozer and Sitton's model of visual attention (Mozer & Sitton, 1998). However the model departs from the former in that it is capable of multimodal learning among distractors, and from the latter in that the WTA network is thought to model overt rather than covert attentional shifts. One novel and critical feature of the model is that it is wired in a feedback loop, whereby its last output determines its current inputs. In this way, we can attempt to simulate the processes taking place in the infant's brain as the sequence of visual and audio events unfolds, during training and test trials. Figure 1 bottom and top panels respectively illustrate the W&K experiment and the proposed model, which we now describe in detail.

Simulations begin with the presentation of one of two possible multimodal targets at the model's input level. In W&K, the target events consisted of identical toy animals that moved synchronously at diagonally opposite corners of the screen, while accompanied by a repetitive sound. In the model, these inputs are simplified as patterns of activations distributed over visual and auditory units that remain clamped throughout the trial. There are five sets of N visual input units, each corresponding to an Area of Interest (AOI hereafter) in W&K's eye-tracking study, and a single set of N auditory units (N = 4 in the figure and the simulations). The pattern of activation attributed to the cat toy is presented both in the bottom left and top right visual banks, while another pattern in the center corresponds to the face cue, which in the experiment was presented with the target events during training. The activation pattern corresponding to the sound is presented in the auditory input bank.

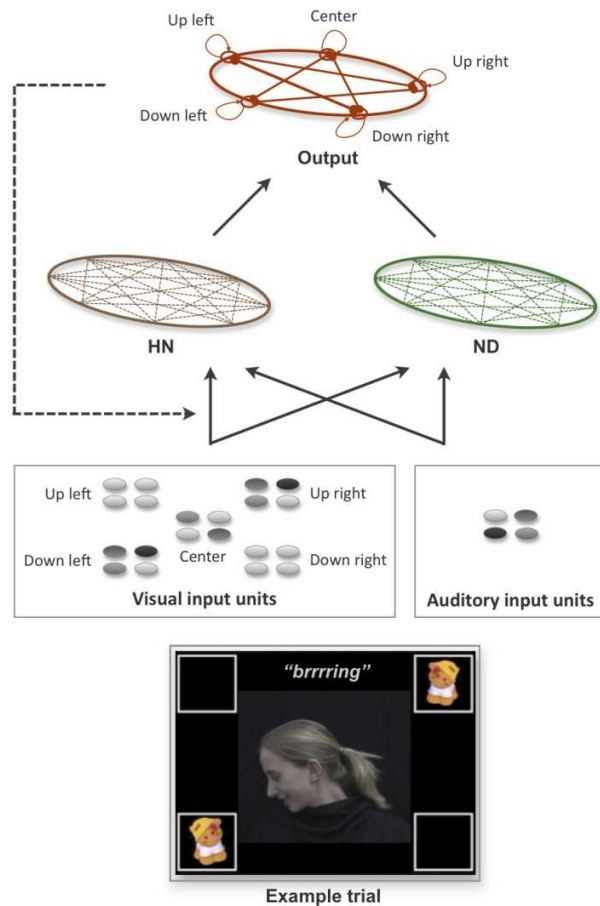


Figure 1: Architecture of the model. Two auto-associator networks are trained to store (left network, Hopfield Network [HN]) or suppress (right network, Novelty Detector [ND]) the activation pattern elicited by some attended part of a multimodal input event (filtered input level). The states to which these networks converge are fed into a winner-take-all network of location units (WTA network, upper network). The winning unit determines the next saccade of the model: which part of the multimodal event will be attended to and which parts will be filtered.

Next, it is important to motivate these input assumptions. In many computational studies of multimodal learning (Althaus & Mareschal, in press; Mayor & Plunkett, 2010), input patterns are derived from actual pixelated images and Mel-scale filtered voices, whereas our inputs are simple arbitrary patterns in the spirit of the HAB model (Sirois & Mareschal, 2004). In addition (and at odds with the dynamical nature of the actual stimuli), our input patterns are randomly generated only once at the beginning of the simulation, and they remain clamped for every trial. These choices were made considering that the actual similarity between representations and the representational changes elicited by moving stimuli were not thought to be essential in the simulated experiment. Rather, our computational model focuses on understanding which information is being sent forward to associative structures, and on testing the nature of the attentional mechanisms involved.

Indeed, not all visual inputs are forwarded to the associative networks: we assume that some attentional filtering is exerted by the WTA network (dynamics explained in the next section). Every time a saccade is made, this filtering lets information about the attended AOI pass through undisturbed, whereas in other AOIs only a fraction of the activation is forwarded. This filtering mechanism and the WTA network that produces it come from Mozer and Sitton's model of visual attention (Mozer and Sitton, 1998), except for the default amount of filtering exerted on unattended regions which was of 90% in Mozer and Sitton, compared to 50% in our model. This difference reflects the fact that attentional systems are subject to cortical maturation (Johnson, 1990), although its precise value was arbitrary and needs to be investigated further. The filter only operates on visual inputs, and it is initialized in a state that depends on the cue and target condition. At the beginning of a trial, central patterns are less likely to be filtered, following experimental data showing that babies are more likely to look at the center (because of the attention getter that was just presented centrally). Filtered and unfiltered inputs are then forwarded to the auto-associator networks.

Auditory and visual patterns then arrive in the core of the model, which consists of two auto-associator networks: the Hopfield network (HN in Figure 1) and the novelty detector (ND in Figure 1). This dual system comes from the HAB model (Sirois and Mareschal, 2004), and like HAB, this is the only part of our model that learns by modifying connection weights during every cycle in each trial of the training phases. HN and ND are fully connected networks of 6N units each, with small connection weights initially generated at random. Each network is presented with full auditory and filtered visual patterns. However, the networks differ in the associative learning rule they use: whereas HN uses Hebbian learning to strengthen connections between active units, ND uses anti-Hebbian learning to decrease these same connections. Over the course of training, HN comes to memorize the patterns it was exposed to by virtue of repeated auto-associations between coactive parts, so much so that eventually presentation of a part (for instance the audio part) is sufficient to retrieve the entire trained pattern. Meanwhile ND progressively learns to suppress the activation elicited by the patterns it is being trained with, so that eventually trained patterns are perfectly suppressed and new patterns produce large activities; they are, in this sense, detected. Finally HN and ND do not gate each other's inputs and outputs, as they do in HAB, but rather the visual units in each network send their activation forward to the WTA network.

The WTA network (Figure 1, top network) is the structure of the model that determines where it will "look" next. It is a standard winner-take-all network (as in Mozer and Sitton, 1998) of five units, one for each AOI. WTA units increase their own activity by way of auto-excitation, and also receive activation from units of the same AOI in HN and ND. Critically, WTA units are wired so as to compete with one another via inhibitory connections. The net effect of this entire set-up is that the unit that receives the most input will build activation faster and win the competition, by which

we mean that its activity crosses a .95 threshold and triggers an ocular saccade to the corresponding AOI. Triggering a saccade in the model means changing the filter's values so as to change the flow of information from input to auto-associator networks. Consistent with the phenomenon of inhibition of return that can last for several seconds (Klein, 2000), we suppress activation in the winning unit until the next saccade is made, which favors foraging of the visual scene.

Simulations

Procedure

The simulations procedure followed the dynamic spatial indexing paradigm used in W&K. After checking that each sub-network (HN, ND and WTA) was operational, 20 models were generated, similar to the average number of infants in each of the three conditions. Models were generated at random and thus differed in their input representations and initial connection weights. Each model was trained over four familiarization blocks, where one block contained six trials of target events (three trials per event type). Target events were randomized, but the same target could not be presented for more than two trials in a row. A trial was limited to 10 cycles, during any of which the connection weights in HN and ND were updated. Testing took place at the end of a block, and consisted of two trials, where the auditory pattern for each target event was presented alone for 10 cycles. Mean proportional looking times and standard errors were then calculated from output saccades to the five AOIs.

We simulated 4 cueing conditions: No Cue, Social Cue (70% filter), Square Cue (70% filter) and Social Cue (90% filter).

In all conditions, the information from the attended location was entirely sent forward. However, in the no-cue condition, models were initialized with unattended filters set to 50%, meaning that only 50% of activation from unattended locations could propagate to the associative systems. In contrast, social and non-social cue conditions had more stringent filters for unattended locations (either 70% or 90% depending on the cue and the hypothesis being tested), meaning that less information from these locations was let through. Apart from the manipulation of this single parameter for the purposes of hypothesis testing, exactly the same set of parameters was used for all models and for all conditions (an exhaustive list of simulation parameters is not specified here for lack of space, but is available upon request to the first author).

Results

Statistical tests of significance were not available at the time of submission, and here we only report mean data as well as standard errors. We believe this is sufficient for the purpose of showing that the model exhibits a pattern of results consistent with the differences observed in W&K with or without cues, and between types of cues.

Cued versus non-cued learning

Over the four blocks in W&K’s No Cue condition, infants were equally likely to look at all locations when presented with the auditory cue. In particular, the authors failed to find any significant advantage of lower locations (labeled “cued” in Figure 2, for consistency with other conditions) over upper locations (labeled “non-cued”) that could have accounted for a bias in the other conditions. This finding is mirrored in our simulations, where cued and non-cued locations are indistinguishable. However, the model was slightly more likely to look at the center than at any other locations.

In contrast, when multimodal training events were cued in W&K’s study, infants looked significantly more at cued locations (in the Square condition) or cued correct locations (in the Social condition, last two blocks) during test trials. The middle right and bottom right graphs in Figure 2 show the same advantage in the model for cued locations over non-cued locations.

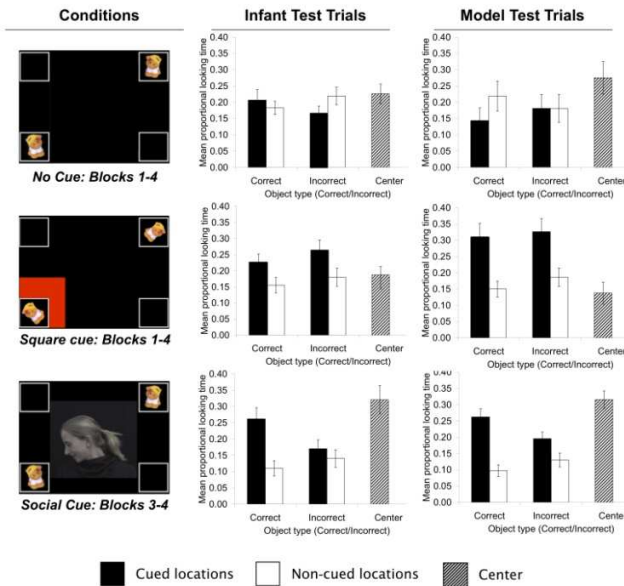


Figure 2: Mean proportional looking times for the model (right) and for infants (center), with the corresponding typical stimuli used in each experiment during training (left screenshots, note that no visual stimuli were provided during test) in No Cue, Square, and Social conditions (resp. top, middle and bottom panels). Courtesy, Wu & Kirkham (2010).

Differences between cues

The main finding from W&K was that different cues produced different types of learning. What we might call “shallow learning” was observed in the Square condition, where infants looked preferentially at locations that had been cued during training (in Figure 2, middle, black bars were superior to white bars) but without associating a location to a sound (black bars are of equal height). On the contrary, “deep learning” was observed in the

Social condition, but only in the last two blocks, where infants looked significantly more at the correct cued location than at any other peripheral location (in Figure 2, bottom, the correct black bar is largely superior both to the incorrect black bar and to the white bars).

Table 1: Proportional looking times (Means and SE) for Infants and Model in the simulated conditions.

Condition	Infants Mean (SE)	Model Mean (SE)
No Cue		
Cued, correct	0.21 (.03)	0.14(0.04)
Non cued, correct	0.18 (.02)	0.22(0.05)
Cued, incorrect	0.17 (.02)	0.18(0.04)
Non Cued, incorrect	0.22 (.03)	0.18(0.04)
Central	0.23 (.03)	0.27(0.05)
Square		
Cued, correct	0.23 (.03)	0.31 (.04)
Non cued, correct	0.15 (.03)	0.11 (.02)
Cued, incorrect	0.26 (.03)	0.33 (.04)
Non Cued, incorrect	0.18 (.03)	0.11 (.03)
Central	0.19 (.03)	0.14 (.03)
Social (90%)		
<i>Blocks 1 & 2</i>		
Cued, correct	0.15 (.03)	0.24 (.02)
Non cued, correct	0.18 (.03)	0.14 (.02)
Cued, incorrect	0.25 (.03)	0.23 (.02)
Non Cued, incorrect	0.16 (.03)	0.06 (.02)
Central	0.25 (.03)	0.33 (.03)
<i>Blocks 3 & 4</i>		
Cued, correct	0.26 (.04)	0.26 (.03)
Non cued, correct	0.11 (.02)	0.10 (.02)
Cued, incorrect	0.17 (.03)	0.20 (.02)
Non Cued, incorrect	0.14 (.03)	0.13 (.02)
Central	0.32 (.04)	0.32 (.03)
Social (70%)		
<i>Blocks 1 & 2</i>		
Cued, correct	-	0.19 (.04)
Non cued, correct	-	0.16 (.03)
Cued, incorrect	-	0.17 (.04)
Non Cued, incorrect	-	0.14 (.02)
Central	-	0.33 (.04)
<i>Blocks 3 & 4</i>		
Cued, correct	-	0.15 (.04)
Non cued, correct	-	0.14 (.03)
Cued, incorrect	-	0.18 (.04)
Non Cued, incorrect	-	0.15 (.03)
Central	-	0.37 (.04)

This behavior is to be contrasted with the looking times observed in the “Social 70%” condition, which acts as a control for our hypothesis that social cueing increases attentional filtering (as shown in Table 1, these simulations do not have a counterpart in infant data). Indeed if the improvement in learning for the Social 90% relative to the Square condition was not due to the increased

filter but rather to the central presence of a visual stimulus, then the same improvement should be expected when the filter is lowered down to the value used for the Square condition. On the contrary, Table 1 shows that no preference for cued object locations was apparent in the Social 70% condition, and cued locations were only marginally superior to non-cued locations.

The “Social 90%” entry in Table 1 is divided in blocks 1&2 and blocks 3&4, to be compared to the block analysis carried out in W&K. We see that the model can reproduce the same late but deep learning effect: it is more likely to look at the correct cued location only in the last two blocks, thereby showing a learning curve. The agreement between infant and model on blocks 3&4 is illustrated in Figure 2, bottom panel. However, note that in the first two blocks, the model exhibits the same pattern of results as in the Square condition (preferential looking for both cued locations, in equal proportion), whereas surprisingly infants appear to look at cued incorrect locations.

General Discussion

Although a true understanding of this model can only be achieved through a detailed enquiry into training saccades and the mechanisms behind them, here we wish to provide the reader with elements of explanation that might shed some light on our main results.

Explaining the impact of cueing

Cueing in the model is achieved by letting through more activation from the location that is being cued, than would normally be allowed. That is, if the model is “looking”, say, at the upper right location while the lower right location is cued, 70% activation from the lower right is forwarded to the associative areas rather than the usual 50% when there is no cue.

This simple mechanism means first that, in the auto-associator networks, some learning will occur for cued locations even if the model actually never “looks” at them, and second, that the model will in fact be biased to look at these cued locations. This is because the increase of activity drives the HN auto-associator into a state that resembles more and more the cued location, so that the corresponding unit in the WTA network would be fed more activation and would tend to win the competition more often. As training proceeds, these two effects reinforce each other and help the model associate auditory patterns to cued objects, which explains how it is able to account for experimental differences between cued and non-cued conditions. However this mechanism alone cannot explain why the model fails to distinguish between cued correct and cued incorrect locations in the Square condition and succeeds only in the last two blocks of the Social condition. Instead, with only this mechanism, the model treats all cues equally.

Explaining social cues versus square cues

Following a proposal in the attentional literature, we tested the hypothesis that the superior learning observed with social cues

resulted from a kind of narrowing of the infant’s receptive fields (Laeng et al, 2010). While maintaining the original cueing mechanism, this narrowing was modeled by more stringent filters for every other location than the fixated and the cued locations (that could possibly differ). Instead of the usual 50%, only 30% activation would be forwarded in the Square condition, against 10% in the Social (90%) condition.

The net effect of this assumption is to minimize interference in HN: the network is equally biased to attend to the cued location in the Square and the Social condition (as in W&K), but only in the latter can it associate precisely the cued visual information to the auditory pattern. In the former condition, the unfiltered activation that comes from the non-cued location gets involved in the association, so that during test trials part of the activation pattern for the non-cued correct location is retrieved, which disturbs the WTA network.

Role of different parts of the model in this account

In this account it would appear that the best part is played by the HN network, while ND appears to have no explanatory power. This is not so, but its role is obscured by the fact that in this model ND is much more active in early phases of training. When training begins, ND has not yet learned how to suppress activation for training input patterns and thus through ND every unfiltered piece of information can contribute to the competition in the WTA. As training unfolds however, ND learns to suppress activation for known patterns, thereby ensuring that unfiltered information cannot use this path anymore to drive the model’s output. This difference between early and late training might be the reason for the learning effect observed in the “Social 90%” condition, although this cannot explain why the same effect was not found in the Square condition.

Size of auditory input

One unexpected clue to understanding the network that might be of significance is the size of the auditory input pattern. The tuning phase of the network revealed that large auditory formats were detrimental to the model’s learning capacity, while the best performance was obtained when it was equal to N (the size of one set of visual units). The reason for this is as follows. Auto-associator networks are known to be very sensitive to the correlation between the patterns to be stored, and this is especially true of the kind of local, incremental learning rule used in HAB and in this model. When the patterns of activity that are to be memorized are too close from one another, as they are when the auditory units vastly outnumber the set of active visual units, interference occurs, and the network can converge to wild configuration states, often called “spurious attractors” (Hopfield et al, 1983). Therefore, limiting auditory inputs to the same format as a single visual location makes multimodal patterns more different to one another and makes for better learning. It would be interesting to investigate how this prediction of the network could be tested in the lab.

Conclusion

We have presented a neuro-computational model that builds on two successful predecessors coming from different fields of cognitive science. The model can account for new infant data involving cued multimodal learning in the presence of distractors. In particular, we have found a candidate mechanism that might underlie largely observed differences between social and non-social cues in infancy. This mechanism holds that infants make use of more stringent attentional filters when they are exposed to social cues than to non-social cues.

Prospects

Future research should aim to better understand how the network behaves, presumably by tracking down the evolution of proportional looking times as training unfolds. In the long term, the model could also be improved by strengthening its links to the brain. For instance Sirois and Mareschal related HN and ND to the cortex and the hippocampus, respectively, and the model might be improved by reinstating the interaction that was originally present between these two systems in HAB. More generally the cortex, hippocampus and superior colliculus obviously all perform more than one function that might well be relevant in this model, for instance coding for auditory maps in the case of the colliculus (King et al, 1996), or input recoding (Levy et al, 2005) and interleaved learning (McClelland et al, 1995) in the case of the hippocampus. A model that could recode input patterns for better storage and present them repeatedly to the infant during less active periods could offer new perspectives into how infants succeed universally to learn what to learn.

Acknowledgments

We thank Denis Mareschal, Sylvain Sirois, Dan Yurovsky, Chen Yu, Shohei Hidaka, Joshen Triesch and Bruno Laeng for their help and useful remarks.

References

Althaus, N. and Mareschal, D. (in press). Early language as multimodal learning. Proceedings of the 12th Neural Computation and Psychology Workshop.

Cleveland, A., Schug, M., & Striano, T. (2007). Joint attention and object learning in 5- and 7-month-old infants. *Infant and Child Development*, 16(3), 295-306.

Gliga, T., & Csibra, G. (2009). One-year-old infants appreciate the referential nature of deictic gestures and words. *Psychological Science*, 20(3), 347-353.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, 19(5), 515-523.

Hopfield, J. J., Feinstein D. I., & Palmer R. G. (1983). 'Unlearning' has a stabilizing effect in collective memories. *Nature*, 304(5922), 158-159.

Houston-Price, C., Plunkett, K., & Duffy, H. (2006). The use of social and salience cues in early word learning. *Journal of Experimental Child Psychology*, 95(1), 27-55.

Johnson, M. H. (1990). Cortical maturation and the development of visual attention in early infancy. *Journal of Cognitive Neuroscience*, 2, 81-95.

Laeng, B., Okubo, M., Saneyoshi, A., & Michimata, C. (in press). Processing spatial relations with different apertures of attention. *Cognitive Science*.

King, A. J., Schnupp, J.W.H., Carlile, S., Smith, A. L., & Thompson, I. D. (1996). The development of topographically-aligned maps of visual and auditory space in the superior colliculus. In B. E. Stein, M. Narita and T. Bando (eds): Extrageniculostriate Mechanisms of Visually Guided Orientation Behavior., *Progress in Brain Research*, 112, pp. 335-50.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, 4(4), 138-147.

Lee, D. K., Itti, L., Koch, C. & Braun, J. (1999). Attention activates winner-take-all competition among visual filters, *Nature Neuroscience*, 2(4), pp. 375-81.

Levy, W. B., Hocking, A. B., & Wu, X. (2005). Interpreting hippocampal function as recoding and forecasting. *Neural Networks*, 18(9), 1242-1264.

Mayor, J., & Plunkett, K. (2010). A neuro-computational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, 117, 1-31.

McClelland, J. L., McNaughton, B. L., & O'Reilly R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419-457.

Mozer, M. C., & Sitton, M. (1998). Computational modeling of spatial attention. In H. Pashler (Ed.), *Attention* (pp. 341-393). East Sussex: Psychology Press Ltd.

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development*, 77(2), 266-280.

Sirois, S., & Mareschal, D. (2004). An interacting systems model of infant habituation. *Journal of Cognitive Neuroscience*, 16, 1352-1362.

Appendix

Here we give some implementation details on how activation propagates and how learning takes place in the model. These equations can also be found in Mozer and Sitton (1998) and Sirois and Mareschal (2004) in a slightly more general form. Note that as in the latter paper, the model's structure allows for a useful abuse of notation that simplifies the exposition of the formula without loss of content. Indeed because each visual and auditory input unit has exactly one counterpart unit in HN and ND, a single index i is used, depending on the context, to refer to one of these three corresponding units.

Activation dynamics

Units in the model use a standard sigmoid function:

$$a_i = \frac{1}{1 + \exp^{-Net_i}}$$

Where Net_i is the net input to unit i , defined as:

$$Net_i = \sum_j w_{ij} a_j$$

For any unit in HN (resp. ND), this net input can be broken into the contribution from the filtered input and the contribution from HN (resp. ND), and it evolves in time like:

$$\forall X \in \{HN, ND\}, \quad Net_i(t+1) = I_i \cdot Filter(WTA_{loc(i)})(t) + \sum_{j \in X} w_{ij} a_j(t)$$

Where I_i indicates unit i at the input level, which is filtered depending on the value of the relevant location unit in the WTA network in the following way:

$$Filter(WTA_{loc(i)}) = \begin{cases} 1, & \text{if location } loc(i) \text{ is attended} \\ \sigma, & \text{if location } loc(i) \text{ is cued} \\ 0, & \text{otherwise} \end{cases}$$

Here $loc(\cdot)$ is simply the fan-in function that maps each unit i in HN and ND to one of the five location units in WTA. A location is said to be attended if the activation of the corresponding WTA unit has crossed a given threshold:

$$loc(i) \text{ is attended} \Leftrightarrow WTA_{loc(i)}(t) > \rho$$

Finally activation in the WTA network is accumulative, which is obtained by taking a convex combination with its previous state :

$$WTA_{loc(i)}(t+1) = \begin{cases} \tau (WTA_{loc(i)}(t)) + (1 - \tau) \sum_{j \in WTA} w_{ij} a_j(t), & \text{if location } loc(i) \text{ is not attended} \\ 0, & \text{if location } loc(i) \text{ is attended} \end{cases}$$

Throughout the simulations and for all networks we used the following values for the parameters: $\tau=0.1$, $\rho=0.95$, and $\sigma=0.5$; 0.3 or 0.1 depending on the conditions.

Learning rules

The learning rules for ND and HN are given respectively by:

$$ND: \Delta w_{ij} = -\lambda_{ND} a_i a_j$$

$$HN: \Delta w_{ij} = \lambda_{HN} (I_i - a_i a_j)$$

Where $\lambda_{ND}=0.8$ and $\lambda_{HN}=0.05$ throughout the simulations.