A BIOLOGICALLY INSPIRED NEURAL MODEL OF VISION-LANGUAGE INTEGRATION

Alessio Plebe[*], *Marco Mazzone*[†], *Vivian M. De La Cruz*[*]

Abstract: One crucial step in the construction of the human representation of the world is found at the boundary between two basic stimuli: visual experience and the sounds of language. In the developmental stage when the ability of recognizing objects consolidates, and that of segmenting streams of sounds into familiar chunks emerges, the mind gradually grasps the idea that utterances are related to the visible entities of the world. The model presented here is an attempt to reproduce this process, in its basic form, simulating the visual and auditory pathways, and a portion of the prefrontal cortex putatively responsible for more abstract representations of object classes. Simulations have been performed with the model, using a set of images of 100 real world objects seen from many different viewpoints and waveforms of labels of various classes of objects. Subsequently, categorization processes with and without language are also compared.

Key words: *Categorization, self-organization, visual cortex, population coding*

Received: August 29, 2009
Revised and accepted: May 13, 2011

1. Introduction

One of the primary sources of information on our world arrives in the form of visual input. An initial challenge we are faced with in building a representation of the world in our brain is that of learning which light sensations belong to the same class of entities, despite significant changes in appearance. A great step forward in this learning process is, when the mind grasps the idea that sometimes patterns of sound are used to identify and categorize visible objects, an ability that has been claimed to be the essence of human nature [14]. This event takes place in the brain at a boundary between the visual and language systems and it is the object under

[*] A. Plebe, V. M. De La Cruz
Department of Cognitive Science, v. Concezione 8, 98121 Messina, Italy, E-mail: {alessio.plebe,vdelacruz}@unime.it

[†] M. Mazzone
Laboratory of Cognitive Science, viale Andrea Doria 6, 95125 Catania, Italy, mazzonem@unict.it

investigation in this paper. Our model is an attempt to investigate the emergence of semantic categorization, with a computational model that adheres in varying degrees to the reality of the corresponding computations taking place in the brain. We are well aware that our model includes only a fraction of the complexity of the mechanisms involved in categorization, neglecting modalities, such as touch, motion perception, and sensorimotor coordination. However, we believe that the key processes related to early categorization are included in the model and we compare it with previous modeling attempts.

1.1 Theoretical issues in categorization

Our model embraces the idea that object categorization could emerge spontaneously from the ability to grasp patterns of statistical regularities. Several researchers reject this position, Carey and Spelke, for example, postulate that there must necessarily be innate constraints that have the effect of making certain features more salient as opposed to others so as to narrow the hypotheses space with respect to the kinds of objects to be categorized first [10]. Mandler offers a rather different picture, presenting evidence that the first similarity dimensions employed in categorization processes are indeed extremely general [35] and that features could have a different explanation other than that of conceptual nativism. In the same vein, [54] has reported a large body of psychological evidence that she interprets as showing that we do not need to presuppose special mechanisms dedicated to specific domains in early categorization. Smith suggests that words can contribute to category formation in that they behave as features which co-vary with other language-independent features of objects [53]. In [32], the special role visual shape plays was emphasized, and how it could be crucial for bootstrapping categorization processes.

Though Mandler rejects the idea that innate constraints are needed in order to allow the emergence of categories, she puts emphasis on the distinction between perceptual and conceptual categories. There are two main elements specific to conceptual representations: they concern what objects do rather than what they look like, and are promoted by attentive, active cognitive processes. Both constraints are not explicit in our model, however, the involvement of functional information just requires that such information is added to the mere appearance of objects and can be done all the same through simple detection of regular covariation. The second point implies active processing, but can be conceived simply as a pre-processing phase that moves to the foreground the inputs of our current model.

1.2 Computational approaches

Several attempts have been made to shed light on the issue of how the human mind acquires the mapping between words and categories of objects, by means of computational models. In [46] the emergence of simple conceptual systems in infancy is explored, inside the classical PDP framework [50] with an abstract connectionist model trained with backpropagation. Their model learns categories of birds, fish, flowers and trees, by associating a predefined set of visual features, like "red" or "branches", with a fixed set of attributes, such as "can walk", "is living" and so on. Despite the higher level of abstraction, this model demonstrates important facts, those being that quite general dimensions of similarity can emerge without appealing to either physiological or cognitive constraints, simply as the result of a coherent covariation of features.

The LEX model proposed by [44], based on associative memories, is aimed at exploring a different facet of the problem: the relationship between the phonological form of words and their meaning. In this case no attention is given to what constitutes the source of meaning, such as visual features, instead the focus is on the association with word forms. This model has shown that word learning might behave in analogy with what previous models have suggested about categorization: certain features of both objects and words can be made more salient than others, simply as a consequence of regularities in objects, words and their co-variation. In LEX, the level of abstraction is also very high and the phonological features are completely artificial, without any relation to voicing features.

At the other extreme, [49] developed a system able to directly deal with real stimuli, in learning labels of visual objects. The model was able to segment words from utterances and to associate the proper word with the object, seen with impressive accuracy. The computations implemented in this model, however, are careful combinations of standard image processing and signal processing algorithms, without any relationship to biological brain computations.

Limiting the discussion to vision, several models inspired by realistic brain computations in object recognition and categorization can be found [17, 45]. Recently, several models have also attempted to reproduce, to some extent, the structure of the visual cortex [64, 15, 47, 55], as well as an earlier version of this model [43], on which our work is based. Not many neural models have been proposed for the auditory process [39, 61], and little is yet known about the kind of brain computations that lead to word recognition there. Recently, [27] addressed the important issue of the roles and the interactions between ITC (Inferior Temporal Cortex) and PFC (Pre-Frontal Cortex) in categorization, with a neural model, which was again limited to vision without any relation to words.

As far as we know, this is the first model combining visual and auditory paths by simulated cortical maps. While building on the current state-of-the-art, our attempt is that of exploring a territory that is still relatively unknown, the boundaries of integration between language and vision in the brain.

2. Artificial Maps for Visual and Auditory Cortical Pathways

This section will introduce the model, the rationale for its design, its mathematical background and the details of each component.

2.1 Mathematics for developing artificial cortical maps

One of the challenges faced by the attempts to model complex cognitive tasks is that of deciding how to divide the whole task into modules that correspond to elementary computational functions. The concept of “cortical maps”, originated

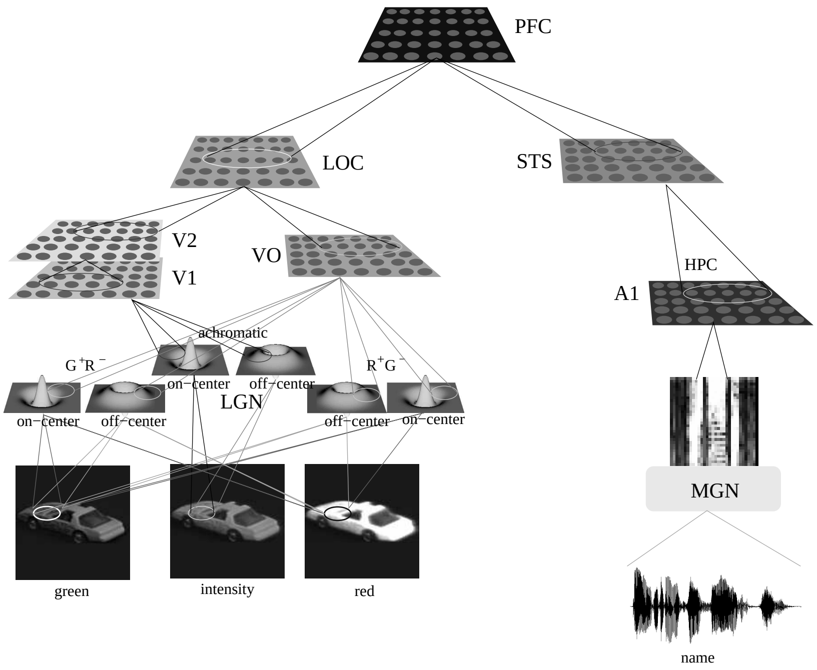

Fig. 1 *Overall scheme of the model. All acronyms are listed in Tab. I.*

by [38], can serve as a useful tool and as a fairly legitimate criterion to use in the task of partitioning the cortex, and has been widely used in vision science [65, 66]. The extent to which the correspondence between modules in our model and cortical maps is faithful is variable. There are computational constraints on the degree of detail in the model, but the most important limitation is the current knowledge availble on the cortical structures that are relevant to this model. Information on these areas continues to be relatively incomplete. As seen in §1.2, there is a large gap between the better knowledge we have of the visual system, and the modest one we have instead of the auditory stream. The boundaries where the two systems meet are even more obscure.

A fruitful mathematical framework for simulating cortical maps from a developmental perspective is the concept of *self-organization* that has been the object of several proposals for artificial neural network schemes. The first implementation was proposed by [62] in models of the development of aspects of the visual system, based exclusively on the local interaction of neurons. Kohonen's SOM (*Self-Organizing Map*) has made the mechanism of self-organization of maps popular thanks to its efficiency and simplicity [30]. However, the winner-takes-all mechanism at the basis of the SOM is a significant departure from the behavior of biological cortical circuits.

In our model, we use the LISSOM architecture (*Laterally Interconnected Synergetically Self-Organizing Map*) [52], which implements flexible and modifiable lateral connections of both excitatory and inhibitory types. The basic equation of the LISSOM describes the activation level x_i of a neuron i at a certain time step k:

$$x_i^{(k)} = f\left(\frac{\gamma_{\mathrm{A}}}{1+\gamma_{\mathrm{N}}\mathbf{U}\cdot\mathbf{v}_{r_{\mathrm{A}},i}} \mathbf{a}_{r_{\mathrm{A}},i}\cdot\mathbf{v}_{r_{\mathrm{A}},i} + \gamma_{\mathrm{E}}\mathbf{e}_{r_{\mathrm{E}},i}\cdot\mathbf{x}^{(k-1)}_{r_{\mathrm{E}},i} - \gamma_{\mathrm{I}}\mathbf{i}_{r_{\mathrm{I}},i}\cdot\mathbf{x}^{(k-1)}_{r_{\mathrm{I}},i} \right), \quad (1)$$

where vector $\mathbf{v}_{r_{\mathrm{A}},i}$ is composed by afferent to unit i in a circular radius r_{A}, the vectors $\mathbf{x}^{(k-1)}_{r_{\mathrm{E}},i}$ and $\mathbf{x}^{(k-1)}_{r_{\mathrm{I}},i}$ are the activations of all neurons in the map, where a lateral connection exists with neuron i of an excitatory or inhibitory type, respectively. Their fields are circular areas of radius, respectively, r_{E}, r_{I}. Vectors $\mathbf{e}_i$ and $\mathbf{i}_i$ are composed by all connection strengths of the excitatory or inhibitory neurons projecting to i. The scalars γ_{X}, γ_{E}, and γ_{I}, are constants modulating the contribution of afferents.

The scalar γ_{N} controls the setting of the push-pull effect in the afferent weights, allowing inhibitory effects without negative weight values. Mathematically, it represents dividing the response from the excitatory weights by the response from a uniform disc of inhibitory weights over the receptive field of neuron i. In equation (1) and all the following ones the operation $\mathbf{x}\cdot\mathbf{y}$ are the product of vectors $\mathbf{x}$ and $\mathbf{y}$. Vector $\mathbf{U}$ is just a vector of 1s of the same dimension of $\mathbf{x}_i$. The function f can be any monotonic nonlinear continuous growing function limited between 0 and 1. For computational economy, it has been implemented as a piecewise linear approximation of the sigmoid function, with a typical lower threshold of 0.1, and an upper threshold of 0.7.

The final activation value of the neurons is assessed after a certain settling time K, typically about 10 time steps. All connection strengths adapt according to the general Hebbian principle, and include a normalization mechanism that counterbalances the overall increase of connections of the pure Hebbian rule. The equations are the following:

$$\begin{aligned}
\Delta\mathbf{a}_{r_{\mathrm{A}},i} &= \frac{\mathbf{a}_{r_{\mathrm{A}},i}+\eta_{\mathrm{A}}x_i\mathbf{v}_{r_{\mathrm{A}},i}}{\|\mathbf{a}_{r_{\mathrm{A}},i}+\eta_{\mathrm{A}}x_i\mathbf{v}_{r_{\mathrm{A}},i}\|} - \mathbf{a}_{r_{\mathrm{A}},i}, && (2)\\
\Delta\mathbf{e}_{r_{\mathrm{E}},i} &= \frac{\mathbf{e}_{r_{\mathrm{E}},i}+\eta_{\mathrm{E}}x_i\mathbf{x}_{r_{\mathrm{E}},i}}{\|\mathbf{a}_{r_{\mathrm{E}},i}+\eta_{\mathrm{E}}x_i\mathbf{x}_{r_{\mathrm{E}},i}\|} - \mathbf{e}_{r_{\mathrm{E}},i}, && (3)\\
\Delta\mathbf{i}_{r_{\mathrm{I}},i} &= \frac{\mathbf{i}_{r_{\mathrm{I}},i}+\eta_{\mathrm{I}}x_i\mathbf{x}_{r_{\mathrm{I}},i}}{\|\mathbf{i}_{r_{\mathrm{I}},i}+\eta_{\mathrm{I}}x_i\mathbf{x}_{r_{\mathrm{I}},i}\|} - \mathbf{i}_{r_{\mathrm{I}},i}, && (4)
\end{aligned}$$

where $\eta_{\{\mathrm{A,E,I}\}}$ are the learning rates for the afferent, excitatory, and inhibitory weights.

LISSOM has been adapted as a model for vision [4], with an organization of the components of input as receptive fields. The vector $\mathbf{a}$ is now made up of afferent signals organized in a two-dimensional fashion, and $\mathbf{v}$ can be thought of as a two dimensional function shaping the receptive field. Therefore, using two orthogonal indexes r and c, equation (1) may be rewritten as:

$$x_{r,c}^{(k)} = f\left(\frac{\gamma_{\mathrm{X}}}{1+\gamma_{\mathrm{N}}\mathbf{I}\cdot\mathbf{a}_{r,c}} \gamma_{\mathrm{X}}\mathbf{a}_{r,c}\cdot\mathbf{v}_{r,c} + \gamma_{\mathrm{E}}\mathbf{e}_{r,c}\cdot\mathbf{x}^{(k-1)}_{r,c} - \gamma_{\mathrm{I}}\mathbf{i}_{r,c}\cdot\mathbf{x}^{(k-1)}_{r,c} \right), \quad (5)$$

where $\mathbf{v}_{r,c}$ is now a vector composed by all values in a two-dimensional array included in the circular receptive field projected by the neural element at coordinates

layer		size	$r_{\rm A}$	$r_{\rm E}$	$r_{\rm I}$	$\gamma_{\rm A}$	$\gamma_{\rm E}$	$\gamma_{\rm I}$	$\gamma_{\rm N}$
LGN	Lateral Geniculated Nucleus	144×144	4.7	-	-	-	-	-	-
MGN	Medial Geniculated Nucleus	32 × 32	2.0	-	-	-	-	-	-
V1	Primary Visual Cortex	96 × 96	8.5	1.5	7.0	1.5	1.0	1.0	0.0
V2	Secondary Visual Cortex	30 × 30	7.5	8.5	3.5	50.0	3.2	2.5	0.7
VO	Ventral Occipital	30 × 30	24.5	4.0	8.0	1.8	1.0	1.0	0.0
A1	Auditory Primary Cortex	24 × 24	6.5	1.5	5.0	1.7	0.9	0.9	0.0
LOC	Lateral Occipital Complex	16 × 16	6.5	1.5	3.5	0.5	1.1	1.7	0.0
STS	Superior Temporal Sulcus	16 × 16	2.5	1.5	5.5	1.8	1.0	1.2	0.0
PFC	Pre-Frontal Cortex	24 × 24	2.5	1.5	5.5	1.5	3.2	4.1	0.0

Tab. I *Main parameters of all maps used in the model.*

r, c. There is a topological correspondence between a translation of r, c on the map and the translation of the field in the input array. The input array can be the retina, as well as another LISSOM map, so that complex models can be built using this architecture. In our model the two-dimensional arrangement is extended to sound stimuli in the time-frequency domains.

2.2 Overall structure of the model

An outline of the modules that make up the model is shown in Fig. 1, the main parameters and the names of all the components are in Tab. I. There are two main paths, one for the visual process and another for the auditory channel. The visual path is built upon and extends a previous model of visual object recognition [43]. The visual system encompasses the ventral stream only, the "what" stream used in the classical dichotomy established by [57], which is specialized for object recognition [42]. A similar division has been found in the auditory system as well, with two streams originating in the caudal and rostral parts of the auditory cortex, targeting spatial and non-spatial domains of the frontal lobe [48]. Here again, the "what" stream is the only contribution taken into account by our model.

2.3 Subcortical components

Both visual and auditory paths include thalamic modules. They are not the object of this study and are, therefore, hardwired according to what is known about their functions. Their governing equations are the following:

$$x^{\odot} = f\left(\left(\mathbf{l}_{r_{\rm A}} + \mathbf{m}_{r_{\rm A}}\right) \cdot \left(\mathbf{g}_{r_{\rm A}}^{\sigma_{\rm N}} - \mathbf{g}_{r_{\rm A}}^{\sigma_{\rm W}}\right)\right) \tag{6}$$

$$x^{\circledcirc} = f\left(\left(\mathbf{l}_{r_{\rm A}} + \mathbf{m}_{r_{\rm A}}\right) \cdot \left(\mathbf{g}_{r_{\rm A}}^{\sigma_{\rm W}} - \mathbf{g}_{r_{\rm A}}^{\sigma_{\rm N}}\right)\right) \tag{7}$$

$$x^{\rm R^{+}G^{-}\odot} = f\left(\mathbf{l}_{r_{\rm A}} \cdot \mathbf{g}_{r_{\rm A}}^{\sigma_{\rm N}} - \mathbf{m}_{r_{\rm A}} \mathbf{g}_{r_{\rm A}}^{\sigma_{\rm W}}\right) \tag{8}$$

$$x^{\rm R^{+}G^{-}\circledcirc} = f\left(\mathbf{l}_{r_{\rm A}} \cdot \mathbf{g}_{r_{\rm A}}^{\sigma_{\rm W}} - \mathbf{m}_{r_{\rm A}} \mathbf{g}_{r_{\rm A}}^{\sigma_{\rm N}}\right) \tag{9}$$

$$x^{\rm G^{+}R^{-}\odot} = f\left(\mathbf{m}_{r_{\rm A}} \cdot \mathbf{g}_{r_{\rm A}}^{\sigma_{\rm N}} - \mathbf{l}_{r_{\rm A}} \mathbf{g}_{r_{\rm A}}^{\sigma_{\rm W}}\right) \tag{10}$$

$$x^{\mathrm{G}^{+}\mathrm{R}^{-}\circledcirc} = f\left(\mathbf{m}_{r_{\mathrm{A}}} \cdot \mathbf{g}_{r_{\mathrm{A}}}^{\sigma_{\mathrm{W}}} - \mathbf{l}_{r_{\mathrm{A}}} \mathbf{g}_{r_{\mathrm{A}}}^{\sigma_{\mathrm{N}}}\right) \tag{11}$$

$$x_{\tau,\omega}^{MGN} = \left| \sum_{t=t_0}^{t_M} v(t) w(t-\tau) e^{-j\omega t} \right|^2, \tag{12}$$

where x is the activation of a neuron, without indices for clarity, except in (12), where the horizontal dimension τ is time, and the vertical dimension ω is frequency. Function $w(\cdot)$ in (12) is a short-term temporal window that performs a spectrogram-like response similar to that given by the combination of cochlear and MGN processes [9]. In the visual path, there are two distinct pathways, one achromatic processed by equations (6), (7), and another which is sensitive to colors, limited here to medium and long wavelengths. The equations are: (8), (9), (10), (11). The symbol $\odot$ refers to on-center receptive fields, and symbol $\circledcirc$ to off-center receptive fields. The profile of all visual receptive fields is given by differences of two Gaussians $\mathbf{g}^{\sigma_{\mathrm{N}}}$ and $\mathbf{g}^{\sigma_{\mathrm{W}}}$, with standard deviations $\sigma_{\mathrm{N}} < \sigma_{\mathrm{W}}$. This is an approximation of the combined contribution of ganglion cells and LGN [16]. The color-sensitive units combine on/off-center response with color opponency, as shown in Fig. 1, for example units $x^{\mathrm{R}^{+}\mathrm{G}^{-}\odot}$ will have an increase in activation if red light is shown in the center of the receptive field and/or green light if it is shown in the surround.

2.4 Lower cortical maps

The lower cortical maps included in the model are V1, V2 and VO, the primary, secondary, and ventral-occipital visual areas, and A1, the auditory primary cortex. Their equations are the following:

$$\begin{aligned} x^{\mathrm{V1}} = f\Bigg(& \gamma_{\mathrm{A}}^{\mathrm{V1}} \left(\mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{V1}\leftarrow\odot} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{\odot} + \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{V1}\leftarrow\circledcirc} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{\circledcirc}\right) + \\ & + \gamma_{\mathrm{E}}^{\mathrm{V1}} \mathbf{e}_{r_{\mathrm{E}}}^{\mathrm{V1}} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{\mathrm{V1}} - \gamma_{\mathrm{I}}^{\mathrm{V1}} \mathbf{i}_{r_{\mathrm{I}}}^{\mathrm{V1}} \cdot \mathbf{x}_{r_{\mathrm{I}}}^{\mathrm{V1}} \Bigg) \end{aligned} \tag{13}$$

$$x^{\mathrm{V2}} = f\left(\gamma_{\mathrm{A}}^{\mathrm{V2}} \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{V2}\leftarrow V1} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{V1} + \gamma_{\mathrm{E}}^{\mathrm{V2}} \mathbf{e}_{r_{\mathrm{E}}}^{\mathrm{V2}} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{\mathrm{V2}} - \gamma_{\mathrm{I}}^{\mathrm{V2}} \mathbf{i}_{r_{\mathrm{I}}}^{\mathrm{V2}} \cdot \mathbf{x}_{r_{\mathrm{I}}}^{\mathrm{V2}}\right) \tag{14}$$

$$\begin{aligned} x^{\mathrm{VO}} = f\Bigg(& \gamma_{\mathrm{A}}^{\mathrm{VO}} \Bigg(\mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{VO}\leftarrow\mathrm{R}^{+}\mathrm{G}^{-}\odot} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{\mathrm{R}^{+}\mathrm{G}^{-}\odot} + \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{VO}\leftarrow\mathrm{R}^{+}\mathrm{G}^{-}\circledcirc} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{\mathrm{R}^{+}\mathrm{G}^{-}\circledcirc} + \\ & \qquad + \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{VO}\leftarrow\mathrm{G}^{+}\mathrm{R}^{-}\odot} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{\mathrm{G}^{+}\mathrm{R}^{-}\odot} + \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{VO}\leftarrow\mathrm{G}^{+}\mathrm{R}^{-}\circledcirc} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{\mathrm{G}^{+}\mathrm{R}^{-}\circledcirc}\Bigg) + \\ & + \gamma_{\mathrm{E}}^{\mathrm{VO}} \mathbf{e}_{r_{\mathrm{E}}}^{\mathrm{VO}} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{\mathrm{VO}} - \gamma_{\mathrm{I}}^{\mathrm{VO}} \mathbf{i}_{r_{\mathrm{I}}}^{\mathrm{VO}} \cdot \mathbf{x}_{r_{\mathrm{I}}}^{\mathrm{VO}} \Bigg) \end{aligned} \tag{15}$$

$$x^{\mathrm{A1}} = f\left(\gamma_{\mathrm{A}}^{\mathrm{A1}} \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{A1}\leftarrow MGN} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{MGN} + \gamma_{\mathrm{E}}^{\mathrm{A1}} \mathbf{e}_{r_{\mathrm{E}}}^{\mathrm{A1}} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{\mathrm{A1}} - \gamma_{\mathrm{I}}^{\mathrm{A1}} \mathbf{i}_{r_{\mathrm{I}}}^{\mathrm{A1}} \cdot \mathbf{x}_{r_{\mathrm{I}}}^{\mathrm{A1}}\right) \tag{16}$$

In all the equations, the activation x has to be taken as the activation of a generic i-th neuron of that level, and all receptive fields have to be intended as referring to that neuron, the index i has been omitted for clarity. For the same reason, the indication of the radius r of the circular receptive fields has not superscripted with the indication of the layer, for example in equation (14) all r_{A} should be read $r_{\mathrm{A}}^{\mathrm{V2}}$, as the radii are different in each map.

In the visual path, the use of still pictures justifies the lack of area V3 whose main role is the processing of motion information [19]. In exploiting the modularity of the model, enacted by the correspondence with cortical maps, a simplification has been introduced by way of the separation of the processing of shape and color. Shape is elaborated through V1 and V2, by equations (13) and (14), and the processing of color is entrusted to VO, with equation (15).

There is evidence, in fact, that suggests that in the visual system no segregation of functions, such as shape or color, processing takes place, and that almost all visual cortical maps cooperate in analysing form, color, motion and stereo information [58, 51]. On the other hand, it is clear that visual areas are not equally involved in all aspects of object recognition. It is possible to identify specialization in one main function in certain maps. This is the case in what has been called the color center area by [69, 70] who named it "V4". We are using the more general name of VO (Ventral Occipital), given by [66]. V1 is the well-known primary visual cortex, the most studied part of the brain [24, 25]. One of its main functions is the organization of the map into domains of orientation tuned neurons [8, 59], which are fundamental for early shape analysis, our model discards the contributions of V1 to all other processes. The main projection from V1 is to its immediately anterior area, V2, which is as large as V1, but by far much less understood. A general and shared idea is that V2 is responsible for shape analysis at a level of complexity and scale that is larger than that of V1 [29, 26, 2].

Very little is known about the computational organization of the auditory primary cortex as compared to the early visual maps just discussed [34]. Our model discards binaural interaction and preserves the main connectivity from single cochlear signals in the Medial Geniculate nucleus to A1. A large body of evidence points to an organization of A1 with a fundamental dependency on sound frequencies along one cortical dimension and a distribution of neural responses to temporal properties [3, 37, 68], model A1 has frequency and time as map dimensions.

Two additional notes on the connectivities in the model. The parameter $\gamma_{mathrmN}$ that allows inhibitory action at the afferent level is, in fact, an additional flexibility of the model, which is not always necessary, because the intracortical lateral inhibitory effect often suffices in developing organizations in the maps which are close enough to biological reality. As can be seen in Tab. I, it has been used only for the V2 component. This version of the model does not include backprojections, this is a limitation since the visual path and the auditory path as well are characterized by significant amounts of projections from higher areas down to lower areas. A further limitation is that of segregating modalities into separate paths, due to the fact that in each perceptual path, neurons are actually responding, to some extents, to other modalities as well [1].

2.5 Higher cortical maps

The equations for the higher cortical maps in the model are the following:

$$\begin{aligned} x^{\mathrm{LOC}} = f\Bigg(& \gamma_{\mathrm{A}}^{\mathrm{LOC}} \left(\mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{LOC} \leftarrow V2} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{V2} + \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{LOC} \leftarrow VO} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{VO} \right) + \\ & + \gamma_{\mathrm{E}}^{\mathrm{LOC}} \mathbf{e}_{r_{\mathrm{E}}}^{\mathrm{LOC}} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{\mathrm{LOC}} - \gamma_{\mathrm{I}}^{\mathrm{LOC}} \mathbf{i}_{r_{\mathrm{I}}}^{\mathrm{LOC}} \cdot \mathbf{x}_{r_{\mathrm{I}}}^{\mathrm{LOC}} \Bigg) \end{aligned} \tag{17}$$

$$x^{\mathrm{STS}} = f\left(\gamma_{\mathrm{A}}^{\mathrm{STS}} \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{STS}\leftarrow A1} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{A1} + \gamma_{\mathrm{E}}^{\mathrm{STS}} \mathbf{e}_{r_{\mathrm{E}}}^{\mathrm{STS}} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{\mathrm{STS}} - \gamma_{\mathrm{I}}^{\mathrm{STS}} \mathbf{i}_{r_{\mathrm{I}}}^{\mathrm{STS}} \cdot \mathbf{x}_{r_{\mathrm{I}}}^{\mathrm{STS}}\right) \tag{18}$$

$$\begin{aligned} x^{\mathrm{PFC}} = f\Bigg(&\gamma_{\mathrm{A}}^{\mathrm{PFC}} \left(\mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{PFC}\leftarrow LOC} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{LOC} + \mathbf{a}_{r_{\mathrm{A}}}^{\mathrm{PFC}\leftarrow STS} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{STS}\right) + \\ &+ \gamma_{\mathrm{E}}^{\mathrm{PFC}} \mathbf{e}_{r_{\mathrm{E}}}^{\mathrm{PFC}} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{\mathrm{PFC}} - \gamma_{\mathrm{I}}^{\mathrm{PFC}} \mathbf{i}_{r_{\mathrm{I}}}^{\mathrm{PFC}} \cdot \mathbf{x}_{r_{\mathrm{I}}}^{\mathrm{PFC}}\Bigg). \end{aligned} \tag{19}$$

Equation (17) describes the convergence of shape and color processing paths into the LOC model map, corresponding to the area in the human cortex thought to be crucial for the task of recognition in vision, located anterior to Brodmann's area 19, near the lateral occipital sulcus, extending into the posterior and mid fusiform gyrus and occipital-temporal sulcus, with an overall surface size similar to that of V1. Perhaps one of the most important ideas concerning this area, and one that has obtained a certain amount of consensus, is that it is involved in visual behavior in which recognition is the main task [21, 22, 63, 31, 67].

In the auditory stream, equation (18) states the projection from A1 to STS. This is the model's correlate of a region in the cortical ventral auditory stream on which there is accumulating evidence and a convergence of opinion on its role in representing and processing phonological information [5, 6, 33, 23].

The model map where the ventral visual path and the auditory path meet, equation (19), is PFC. As pointed out in §1.2, there are actually several areas where visual and auditory signals converge and more than one area is activated in categorization tasks. The reason for calling the model map PFC is that it is the upper map in this system hierarchy, where categorization is expected to express the best possible abstraction and integration of all available object information. This is a role which is ascribed to biological PFC as well. Needless to say, biological PFC deals with larger sets of object information than those used in the model, and these would include for example, the functional information pertaining to what objects can be used for.

From the model equation it can be seen that all radii being of the unit projections constant, its algorithmic complexity is linear using as dimension the number of units. The time of one recall from stimuli to PFC is of about 400 milliseconds on an AMD Athlon 7750, with 1350 KHz clock and 512 KB cache size.

3. Resulting Functions in the Maps

In this section, we will describe how the model has been trained in this experiment and report on the functions developed in all the lower cortical maps of the model. We will then discuss the various aspects of categorization that emerged in the higher maps.

3.1 Simulation of inner and environmental experiences

The model has been exposed to a variety of stimuli, at different stages of its development, that to some extent parallel periods of human development from the prenatal stage to that of early language acquisition. Initially, only V1, VO and A1 maps are allowed to modify their synaptic weights by equations (2), (3), and

(4). The stimuli presented to V1 and VO are synthetic random blobs that mimic waves of spontaneous retinal activity that are known to play a fundamental role in the ontogenesis of the visual system [36, 28, 56, 20, 11]. Blobs presented to V1 are elongated along random directions to stimulate orientation selectivity. Blobs to VO are circular, with constant hues, and random size, position, and intensity. The A1 map is exposed to short trains of waves sweeping linearly around a central frequency. Time durations, central frequencies and sweeping intervals are changed randomly. The next period of development involves V2 and STS maps as well. The visual stimuli comprises pairs of elongated blobs with a coinciding end point, to enhance the experience of patterns that are slightly more complex than lines, such as corners. The auditory stimuli are synthesized waves of the 7200 most common English words (from `http://www.bckelk.uklinux.net/menu.html`), with length of 3 to 10 characters. All words are converted from text to waves using *Festival* software [7], with cepstral order 64 and a unified time window of 2.3 seconds. In the development stage that corresponds to that just after eye opening, natural images are used. In order to include the primary and most realistic difficulty in recognition, which is the identification of an object under different views, the COIL-100 collection has been used [40], where for each of the 100 objects, 72 different views are available. In most experiments, unless otherwise stated, only 8 views per object have been used during the learning phase of the model and all 72 views are used in the testing phases.

The last stage of the experiment simulates events in which an object is viewed and a label corresponding to its basic category is heard simultaneously. The 100 objects have been grouped manually into 38 categories. Certain categories, such as `cup` or `medicine`, have 5 exemplars in the object collection, while others, such as `telephone`, have only one exemplar. Each category word is converted from text to waves using the `en1` "Roger" male voice, and the `us1` female American speaker in the *Festival* software. Both male and female utterances are duplicated at standard and slower speeds, using the 1.3 value of the `Duration_Stretch` parameter in *Festival.* Examples of the stimuli used can be seen in Fig. 2.

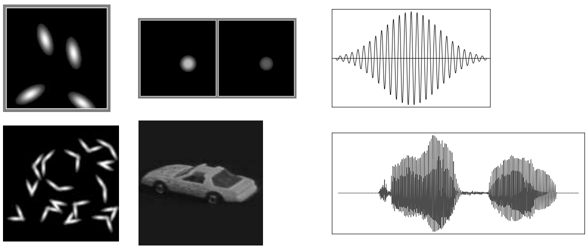

Fig. 2 *Example of stimuli to the model. From the left, elongated blobs input to V1, hue circular blobs for VO, wave trains for A1, couple of blobs for V2, real images for the visual path, and word waves for the auditory path.*

3.2 Emergence of organization in the lower maps

At the end of development, different organizations are found in the lower maps that enable the performance of processes that are essential to recognition, and that are similar to those found in corresponding brain areas. These are mentioned in §2.4. The V1 map in the model is organized according to orientation selectivity, with responsiveness of neurons to oriented segments arranged over repeated patterns of gradually changing orientations, broken by few discontinuities, as shown in Fig. 3. This ordering resembles the one known to be found in biological primary cortex [8, 59].

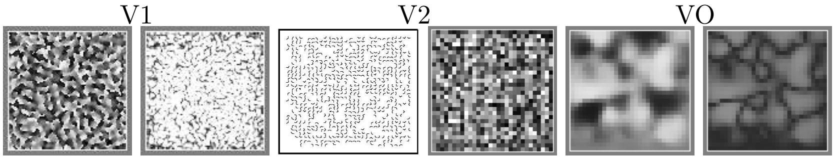

Fig. 3 *Organization of neural responses in the lower maps of the model. The two maps on the left show orientation selectivity in V1. Of these, the map on the left shows the preferred orientation for each neuron, coded in gray level from black→horizontal to white→vertical. The map on the right contains the degree of neuronal selectivity in a gray scale from black→equally responsive to all orientations, to white→highly selective to a single orientation. The two middle maps show angle selectivity in V2. The map on the left is overlaid with patterns inducing peak responses in the neurons, the map on the far right is the degree of selectivity. The pair of maps on the far right represent the hue-responding domains in VO. The map on the left encodes the preferred hue, from black→red to white→green, the map on the left is again the degree of selectivity.*

In the VO map of the model, most neurons respond to specific hues regardless of intensity. This is one of the basic features of color processing. Color constancy is crucial in object recognition and is known to develop somewhere between two and four months of age [13]. The organization in the model's VO is shown in Fig. 3. The kind of mapping found in A1 is typically tonotopic and it encodes the dimensions of frequency and time sequences in a sound pattern. This is known to be the main ordering of neurons in biological A1 [60]. The main organization found in the V2 map is responsiveness to angles, especially in the 60 and 150 degrees range, this is shown in Fig. 3. This kind of selectivity is one of the major phenomena recently discovered in biological V2 [26, 2]. We refer to other published works for further details on the functions that emerged in V1 and VO [43] and V2 [41].

3.3 Categorization in the higher maps

In the scope of this experiment the most interesting results are found in the three higher maps. It is here where we expect forms of categorization to take place that concern both visual and word forms, and mostly conceptual categorization that intergrates the two.

A common method of analysis has been carried out for all the upper maps to reveal possible distributions of neural activation as population coding of categories.

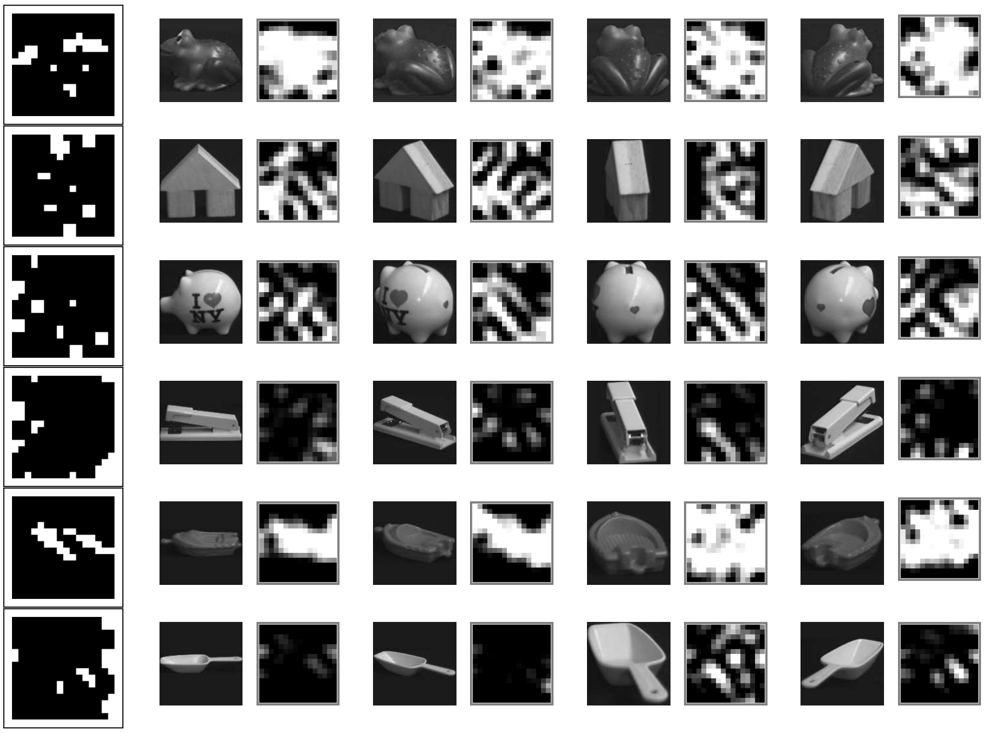

Fig. 4 *Examples of population coding of objects in the LOC map. In each row the map on the far left displays the coding neurons. The following images in the row are samples of views of the same object and the corresponding response patterns in LOC.*

Let us introduce the following function:

$$x_i(s) : S \in \mathcal{S} \to \mathbb{R}^+; \quad s \in S \in \mathcal{S}, \tag{20}$$

that gives the activation x of a generic neuron i in one of the higher maps in response to the presentation of the stimulus s to the system. This stimulus is an instance of a class S, belonging to the set of all classes of stimuli $\mathcal{S}$ available in the experiment. For a class $S \in \mathcal{S}$ we can define the two sets:

$$X_{S,i} = \{x_i(s_j) : s_j \in S\}; \quad \overline{X}_{S,i} = \{x_i(s_j) : s_j \in S' \neq S \in \mathcal{S}\}. \tag{21}$$

We can then associate with the class S a set of neurons in the map by ranking it with the following function:

$$r(S,i) = \frac{\mu_{X_{S,i}} - \mu_{\overline{X_{S,i}}}}{\sqrt{\frac{\sigma_{X_{S,i}}}{|X_{S,i}|} + \frac{\sigma_{\overline{X_{S,i}}}}{|\overline{X}_{S,i}|}}}, \tag{22}$$

where μ is the average and σ the standard deviation of the values in the two sets, and $|\cdot|$ is the cardinality of a set. Now the following relation can be established as the population code of a class S:

$$p(S) : \mathcal{S} \to \{\langle i_1, i_2, \cdots, i_M\rangle : r(S,i_1) > r(S,i_2) > \cdots > r(S,i_M)\}, \tag{23}$$

# of views	discriminative power	
	average over all objects	standard deviation
4	0.624	0.325
8	0.647	0.314
18	0.653	0.323

Tab. II *Model LOC discriminative power of individual objects by population coding. As a comparative figure, discrimination by chance would be 0.01.*

where M is a given constant, typically one order of magnitude smaller than the number of neurons in the map. The population code $p(S)$ computed with (23) can be used to classify a stimulus s in an expected category:

$$c(s) = \arg\max_{S \in \mathcal{S}} \left\{ \sum_{j=1\cdots M} \alpha^j x_{p(S)_j}(s) \right\}, \tag{24}$$

where $p(S)_j$ denotes the j-th element in the ordered set $p(S)$ and α is a constant that is close, but smaller, than one. It is possible to evaluate how the population code of a map is effective in discriminating a category S by measuring the fraction of hits in classifying stimuli belonging to that category:

$$a(S) = \frac{|\{s :\ s \in S \wedge c(s) = S\}|}{|S|}. \tag{25}$$

3.3.1 LOC map

It is possible to begin evaluating the type of representation established in LOC by using as stimuli view o of an object O, therefore specializing the equation (20) as:

$$x_i^{\mathrm{LOC}}(o) : O \in \mathcal{O} \to \mathbb{R}^+; \quad o \in O \in \mathcal{O}, \tag{26}$$

where $x_i^{\mathrm{LOC}}(o)$ is computed by (17), when the image o is presented to the visual path. In this case, a category S of (20) is just an individual object whose instances can be ascribed to the same entity, or mistaken for a different entity, if a specific point of view does not resemble the typical view of that object. Samples of population coding of individual objects are shown in Fig. 4. As can be seen from the comparison of the population coding map with LOC responses to different views of the same object, most coding neurons exhibit invariant responses. Invariance, an ability to recognize known objects despite large changes in their appearance on the sensory surface, is the most important property to fulfill the requirement for an object-recognition cortical area. It has been identified in human LOC by several studies [22, 63, 31, 67].

Although the term invariance is commonly used in vision, a better term would be "degree of tolerance" with respect to classes of changes in appearance of the same object, since no cortical map contains neurons with responses that are absolutely invariant to transformations of the same object [12]. There is also psychophysical evidence of inabilities to recognize objects viewed from less common perspectives, even in adult humans [18]. In the model's LOC the supposed coding by population of individual objects is efficient in discriminating most of the views for many objects, although not in an absolute way. By means of equation (25), adapted to stimuli conditions described in (26), the discrimination power of LOC, i.e. the fraction of samples correctly classified in their category, has been quantitatively assessed, and is shown in Tab. II.

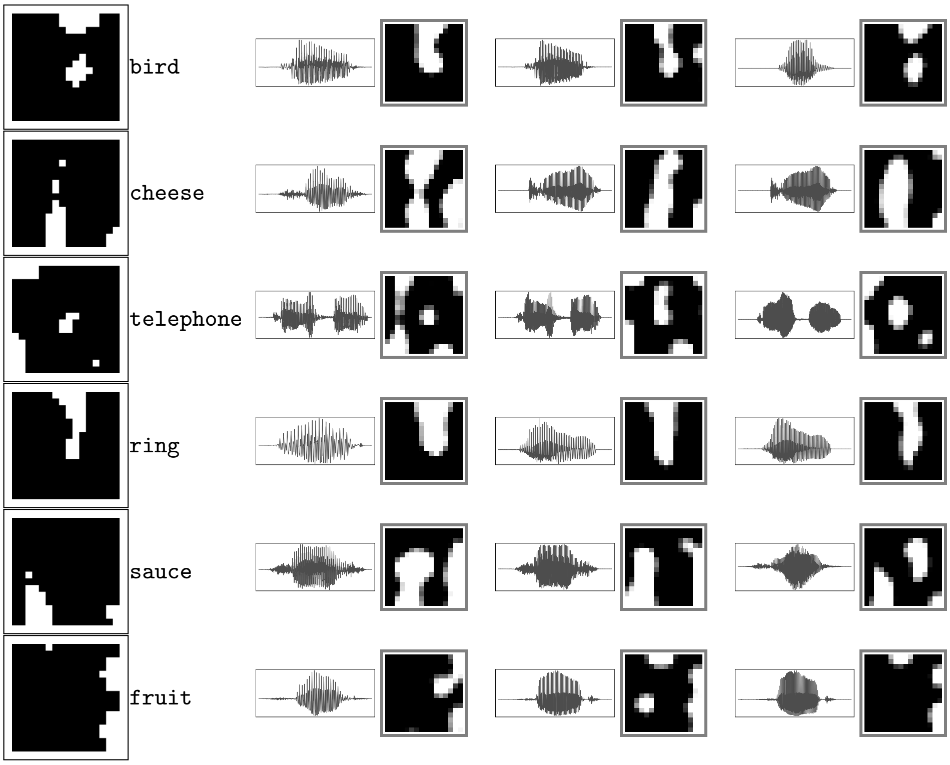

Fig. 5 *Examples of population coding of word sounds in the STS map. In each row the leftmost map displays the coding neurons. The following images in the row are samples of sound of the same label and corresponding response patterns in STS.*

3.3.2 STS map

In the STS map it is possible to evaluate the population coding of labels, as spoken in four different voices, therefore using stimuli that specialize equation (20) as:

$$x_i^{\mathrm{STS}}(n) : N \in \mathcal{N} \to \mathbb{R}; \quad n \in N \in \mathcal{N}, \tag{27}$$

where $x_i^{\mathrm{STS}}(n)$ is computed by (18), when the sound n is presented to the auditory path. There are 38 classes N of sounds in $\mathcal{N}$ corresponding to the labels of 38 object

voices	discriminative power	
	average over all names	standard deviation
female	0.882	0.242
male	0.895	0.234
both	0.658	0.300

Tab. III *Model STS discriminative power of words heard by different voices. As a comparative figure, discrimination by chance would be 0.026.*

voices	discriminative power	
	average over all categories	standard deviation
female	0.878	0.220
male	0.895	0.167
both	0.695	0.240

Tab. IV *Model PFC accuracy in discriminating linguistic categories by population coding. As a comparative figure, discrimination by chance would be 0.026.*

categories. Samples of population coding of labels are shown in Fig. 5. In Tab. III the accuracy of STS in discriminating between the different labels is shown. The development of STS has been experimented using three different sets of stimuli: female voices only, male voices only, or using the full set of voices. Listening to voices of a single gender facilitates the identification of names, as can be observed in children.

3.3.3 PFC map

The PFC map is where the visual and the linguistic information meet, the population coding is the computed specializing equation (20) as:

$$x_i^{\mathrm{PFC}}(c) : C \in \mathcal{C} \to \mathbb{R}; \quad c = \langle o, n \rangle \in C = \left(\{\epsilon\} \cup \bigcup_{O \in \mathcal{O}_C} O \right) \times \left(\{\epsilon\} \cup N_C \right), \quad (28)$$

where $x_i^{\mathrm{PFC}}(c)$ is computed by (19), when the sound n is presented to the auditory path and the object o to the visual path. The 38 object categories introduce a partition in the set of objects $\mathcal{O}$, such that all sets of views in the partition $O \in \mathcal{O}_C$ are of objects of that category C. N_C is the set of utterances naming category C. Note that the empty sample ϵ is included, for experiments in which only a single modality is presented, $c = \langle o, \epsilon \rangle$ in the case of the visual modality only, and $c = \langle \epsilon, n \rangle$ in the case of linguistic input only. The development of PFC followed that of all the lower hierarchies in the model and has been replicated, as in the case of STS, for single gender voices, and for the full set. The objects have been presented using 8 views, of the 72 available, for each. Tab. IV displays the accuracy achieved at the end of the development, as for STS, the discrimination is quite more accurate when words are spoken by a person of a single gender. Nevertheless, the model achieves a remarkable ability of categorizing objects taking into account their names, even when thrown into a two gender speaking environment. All the results described next refer to the full voice set experiment. Samples of population coding and actual activations are shown in Fig. 5. As can be seen, both the population coding and map activations do not preserve any explicit trace of visual features. The activations of neurons seem to code the categories in a rather abstract way, as well as the peculiarities of the single samples. For each pair of waves and images the activations in the PFC map exhibit an important overlap with the population coding of the corresponding category, with contingent additional or lacking activations denoting the specific view and the specific voice.

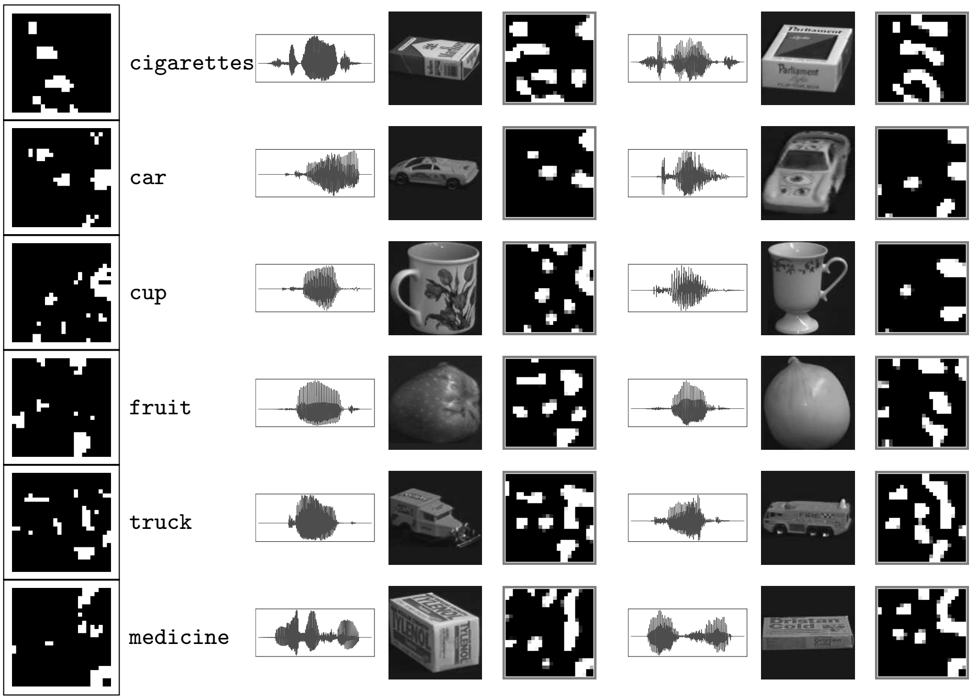

Fig. 6 *Examples of population coding in the PFC map. In each row the leftmost map displays the coding neurons. Following images in the row are samples of a sound of the same name, a view of an object of the relevant class, and the corresponding response patterns in PFC. The first sample always has a male voice sound and the second sample a female voice.*

The PFC map is a high level abstract center, where multiple modalities converge. In the specific case of this model, it is the place of convergence of the visual and linguistic modalities. By supplying the model with partial inputs of the two types $c = \langle o, \epsilon \rangle$ and $c = \langle \epsilon, n \rangle$, it is possible to derive population codings of PFC as being unimodal by always using equations (23), (25), and thus investigate the separate effects of the two modalities. Several of these modal codings are compared in Tab. 7 with the full multimodal map. There is a general spontaneous partitioning

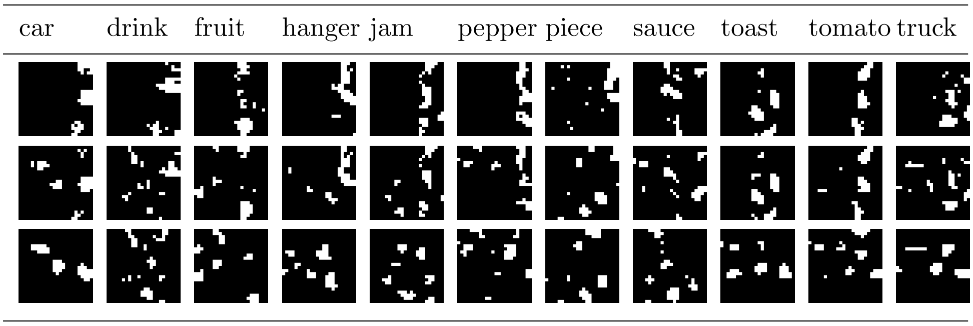

Fig. 7 *Examples in the PFC map of modal and multimodal population coding. Images in the upper row are coding maps for the auditory mode only, in the lower row we find the visual mode only, and the middle row is the full multimodal population coding of the categories.*

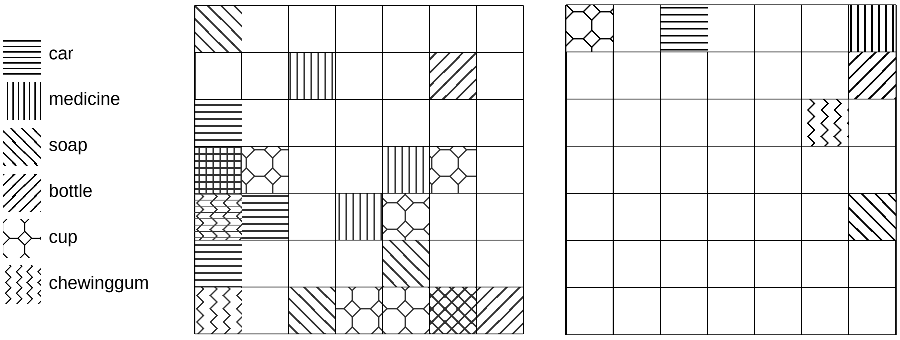

Fig. 8 *Comparison of categorization in LOC and PFC maps of the model, for few classes of objects. The two graphs are SOM clusterings of the LOC map on the left, and the PFC map on the right.*

in this map, with linguistic information more clustered on the right side and visual on the left, but with large overlaps. In several cases, the multimodal coding looks like a combination of the visual and linguistic representations, as in the case of `car`, while others, like `cigarettes` or `toast`, are more influenced by the linguistic representation. In the case of car, it could be due to the larger variety of visual samples inside the same category `car`.

3.3.4 Spontaneous and linguistic categories

Representations in PFC are the result of the tension between the force of naming and the regular similarities in visual appearance, in shaping categories. It is interesting to compare the categories developed in PFC with those emerged in LOC, where language has no effect. For this purpose, a Self Organizing Map [30] has been used to cluster the output of LOC and PFC for all object samples. Both SOM maps are 7 × 7 to allow for 38 categories. In Fig. 8 some classes of objects

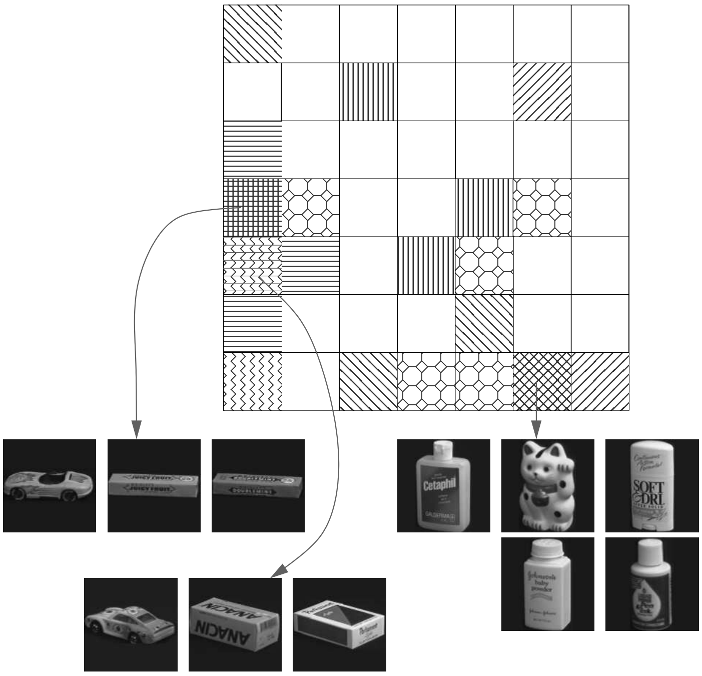

Fig. 9 *Examples of non linguistic categories in SOM clusters of LOC maps that overlap with linguistic categories in PFC.*

are compared in the two SOM maps. As expected, in the case of PFC there is a correspondence of SOM elements with linguistic classes of objects, while in the case of LOC they are spread along multiple SOM elements, with overlaps as well. The content of some LOC "natural" classes that include overlaps of linguistic classes are shown in Fig. 9. The two cells (far left column), where the linguistic `car` category is spread, share objects of categories `chewing gum` (upper cell) and `medicine`, `cigarettes` (lower cell). In both cases, there are visual similarities in the presence of labels with color contrast, in yellow and light green. There is no possibility of paying special attention to the wheels as elements in building a specific concept of `car`, without perceptual knowledge of how `wheels` are essential to cars. The other cell in the lower row shares categories `soap` and `bottle`, probably because of their shape and labeling. More surprising is the presence of an object class `kitten`. But on the other hand, it might not be too surprising, if you take into account that it displays a similar upright shape and pattern of colors (dominant white background), and that nothing in the model induces the "face" specificity in the visual recognition process.

3.3.5 Visual imagery elicited by labels

It is possible to supply the model with input of type $c = \langle \epsilon, \breve{n} \rangle$, with missing visual input, and check what visual imagery has been elicited by the word sounds $\breve{n}$. For this purpose, a view $\widehat{o}$ of an object is selected by the following:

$$\widehat{o} = \operatorname*{arg\,min}_{o_i \in \bigcup O,\, n_{o_i} \in N_{o_i}} \left\{ \sum_{j=1\cdots M} \alpha^j \left(x^{\mathrm{PFC}}_{p(\langle\epsilon,\breve{n}\rangle)_j}(\langle\epsilon,\breve{n}\rangle) - x^{\mathrm{PFC}}_{p(\langle\epsilon,\breve{n}\rangle)_j}(\langle o_i, n_{o_i}\rangle) \right)^2 \right\}, \qquad (29)$$

where $p(S)_j$ denotes the j-th element in the ordered set $p(S)$, as in (24), the function $x^{\mathrm{PFC}}(\cdot)$ is that defined by (28), and N_{o_i} is the set of label sounds referring to the category of object o_i. In Fig. 10 several examples of images elicited by labels are shown, uttered by female and male voices. Often, all possible voices elicit the same view of an object that is a sort of a prototype for its class. Again, as in the condition of spontaneous categorization, `kitten` is misunderstood, confirming that face-like cues have no special effect on the model. The source of confusion for the sound `dog` is likely to be phonetic due to the similarity with `jug`, the category to which the elicited objects belong.

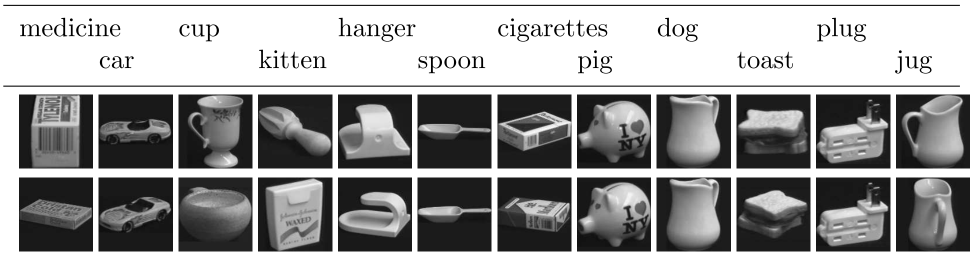

Fig. 10 *Images elicited in PFC by voices of a category word, female voices in the first row, male voices in the second row.*

We should stress that this ability of model PFC of evoking one modality with the other is limited compared to the rich forms that are possible in the human brain; this is due to the obviously extreme simplfications of the model with respect to a real brain, discussed in §2.4.

4. Conclusions

The model described here demonstrates that the perceptual experience of light and sound is enough for a first emergence of concepts of visual objects and their labels, with sounds being those uttered by people naming objects, and light being the vision of those objects. Clearly, the grasping of a name in this model should be interpreted as the pure association with a class of objects that share some visual feature. There is nothing implemented here that would suggest any other aspect of labels or names that would enrich their concepts in brain development. For example, `medicine` is only a sound that induces one to accept some commonalities between boxes with certain colored labels. In the real world, this name would, of course, acquire the meaning of something one eats to treat illness, but only when language has been grasped to the point of becoming the main source for the creation of networks of associations between meanings. This can never happen in the model because it is not implemented. The step implemented here, however, is not marginal at all, in that it is a first and difficult step that is necessary in order to grasp the idea that sounds, spoken by other humans, might cluster categories in the world. For this reason, our model allows us to address human symbolic ability as such.

A strength in the model is the attempt to adhere to the chain of computational processes taking place in the brain, especially in the cerebral cortex, and to rely on fundamental plasticity mechanisms that allow functions to emerge from sensorial experience. The model benefits from this effort to different degrees due to the different levels of knowledge we have of the involved cortical processes. Much more is known about visual recognition, and this explains why this path is much more detailed in the model.

We conceive the model described here as an important step in a long-term project of developing a brain-informed simulation of phenomena in the early acquisition of language, in connection with visual experience. Its novelty is not just in demonstrating the emergence of concepts from the association of visual information and sounds, that has been already simulated in previous work, but rather in its showing how this could happen in neural structures that are similar to the cortical architecture of the human brain.

References

[1] Allman B. L., Keniston L. P., Meredith M. A.: Not just for bimodal neurons anymore: The contribution of unimodal neurons to cortical multisensory processing. Brain Topography, **21**, 2009, pp. 157–167.

[2] Anzai A., Peng X., Essen D. C. V.: Neurons in monkey visual area V2 encode combinations of orientations. Nature Neuroscience, **10**, 2007, pp. 1313–1321.

[3] Atzori M., Lei S., Evans D. I. P., Kanold P. O., Phillips-Tansey E., McIntyre O., McBain C. J.: Differential synaptic processing separates stationary from transient inputs to the auditory cortex. Neural Networks, **4**, 2001, pp. 1230–1237.

[4] Bednar J. A.: Learning to See: Genetic and Environmental Influences on Visual Development. PhD thesis, University of Texas at Austin, Tech. Report AI-TR-02-294, 2002.

[5] Belin P., Zatorre R. J., Lafaille P., Ahad P., Pike B.: Voice-selective areas in human auditory cortex. Cognitive Brain Research, 403, 2000, pp. 309–312.

[6] Belin P., Fecteau S., Bédard C.: Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences, 8, 2004, pp. 129–135.

[7] Black A. W., Taylor P. A.: The festival speech synthesis system: System documentation. Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Edinburgh, UK, 1997.

[8] Blasdel G. G.: Orientation selectivity, preference, and continuity in monkey striate cortex. Journal of Neuroscience, 12, 1992, pp. 3139–3161.

[9] Brown M. C.: Audition. In: L. R. Squire, F. Bloom, S. McConnell, J. Roberts, N. Spitzer, and M. Zigmond, editors, Fundamental Neuroscience, Academic Press, New York, 2003, pp. 699–726.

[10] Carey S., Spelke E.: Science and core knowledge. Journal of Philosophy of Science, 63, 1996, pp. 515–533.

[11] Chapman B., Stryker M. P., Bonhoeffer T.: Development of orientation preference maps in ferret primary visual cortex. Journal of Neuroscience, 16, 1996, pp. 6443–6453.

[12] Cox D. D., Meier P., Oertelt N., DiCarlo J. J.: ‘Breaking’ position-invariant object recognition. Nature Neuroscience, 8, 2005, pp. 1145–1147.

[13] Dannemiller J. L.: A test of color constancy in 9- and 20-weeks-old human infants following simulated illuminant changes. Developmental Psychology, 25, 1989, pp. 171–184.

[14] Deacon T. W.: Symbolic Species: The Co-Evolution of Language and the Brain. Norton, W. W. & Company, Inc., New York, 1998.

[15] Deco G., Rolls E.: A neurodynamical cortical model of visual attention and invariant object recognition. Vision Research, 44, 2004, pp. 621–642.

[16] Dowling J. E.: The Retina: An Approachable Part of the Brain. Cambridge University Press, Cambridge (UK), 1987.

[17] Edelman S., Duvdevani-Bar S.: A model of visual recognition and categorization. Philosophical transactions of the Royal Society of London, 352, 1997, pp. 1191–1202.

[18] Farah M. J.: Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision. MIT Press, Cambridge (MA), 1990.

[19] Gegenfurtner K. R., Kiper D. C., Levitt J. B.: Functional properties of neurons in macaque area V3. Journal of Neurophysiology, 77, 1997, pp. 1906–1923.

[20] Gödecke I., Bonhoeffer T.: Development of identical orientation maps for two eyes without common visual experience. Nature, 379, 1996, pp. 251–254.

[21] Grill-Spector K., Kushnir T., Edelman S., Itzchak Y., Malach R.: Cue-invariant activation in object-related areas in the human occipital lobe. Neuron, 21, 1998, pp. 191–202.

[22] Grill-Spector K., Kushnir T., Edelman S., Avidan-Carmel G., Itzchak Y., Malach R.: Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron, 24, 1999, pp. 187–203.

[23] Hickok G., Poeppel D.: The cortical organization of speech processing. Nature Reviews Neuroscience, 8, 2007, pp. 393–402.

[24] Hubel D., Wiesel T.: Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. Journal of Physiology, 160, 1962, pp. 106–154.

[25] Hubel D. H., Wiesel T. N.: Brain and Visual Perception: The Story of a 25-Year Collaboration. Oxford University Press, Oxford (UK), 1959.

[26] Ito M., Komatsu H.: Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. Journal of Neuroscience, 24, 2004, pp. 3313–3324.

[27] Kashimori Y., Ichinose Y., Fujita K.: A functional role of interaction between IT cortex and PF cortex in visual categorization task. Neurocomputing, 70, 2007, pp. 1813–1818.

[28] Katz L., Shatz C.: Synaptic activity and the construction of cortical circuits. Science, 274, 1996, pp. 1133–1138.

[29] Kobatake E., Tanaka K.: Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. Journal of Neurophysiology, 71, 1994, pp. 856–867.

[30] Kohonen T.: Self-Organizing Maps. Springer-Verlag, Berlin, 1995.

[31] Kourtzi Z., Erb M., Grodd W., Bülthoff H. H.: Representation of the perceived 3-d object shape in the human lateral occipital complex. Cerebral Cortex, 13, 2003, pp. 911–920.

[32] Landau B., Smith L. B., Jones S.: The importance of shape in early lexical learning. Cognitive Development, 3, 1988, pp. 299–321.

[33] Liebenthal E., Binder J. R., Spitzer S. M., Possing E. T., Medler D. A.: Neural substrates of phonemic perception. Cerebral Cortex, 15, 2005, pp. 1621–1631.

[34] Linden J. F., Schreiner C. E.: Columnar transformations in auditory cortex? A comparison to visual and somatosensory cortices. Cerebral Cortex, 13, 2006, pp. 83–89.

[35] Mandler J. M.: The Foundations of Mind. Oxford University Press, Oxford (UK), 2004.

[36] Mastronarde D. N.: Correlated firing of retinal ganglion cells: I. spontaneously active inputs in X- and Y-cells. Journal of Neuroscience, 14, 1983, pp. 409–441.

[37] Miller L. M., Escab M. A., Read H. L., Schreiner C. E.: Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. Journal of Neurophysiology, 87, 2002, pp. 516–527.

[38] Mountcastle V.: Modality and topographic properties of single neurons in cats somatic sensory cortex. Journal of Neurophysiology, 20, 1957, pp. 408–434.

[39] Näger C., Storck J., Deco G.: Speech recognition with spiking neurons and dynamic synapses: a model motivated by the human auditory pathway. Neurocomputing, 44–46, 2002, pp. 937–942.

[40] Nayar S., Murase H.: Visual learning and recognition of 3-d object by appearence. International Journal of Computer Vision, 14, 1995, pp. 5–24.

[41] Plebe A.: A model of angle selectivity development in visual area V2. Neurocomputing, 70, 2007, pp. 2060–2066.

[42] Plebe A.: The ventral visual path: Moving beyond V1 with computational models. In T. A. Portocello and R. B. Velloti, editors, Visual Cortex: New Research, Nova Science Publishers, New York, 2008, pp. 97–160.

[43] Plebe A., Domenella R. G.: Object recognition by artificial cortical maps. Neural Networks, 20, 2007, pp. 763–780.

[44] Regier T.: The emergence of words: Attentional learning in form and meaning. Cognitive Science, 29, 2005, pp. 819–865.

[45] Riesenhuber M., Poggio T.: Models of object recognition. Nature Neuroscience, 3, 2000, pp. 1199–1204.

[46] Rogers T. T., McClelland J. L.: Semantic Cognition - A Parallel Distributed Processing Approach. MIT Press, Cambridge (MA), 2006.

[47] Rolls E. T., Stringer S. M.: Invariant visual object recognition: A model, with lighting invariance. Journal of Physiology – Paris, 100, 2006, pp. 43–62.

[48] Romanski L. M., Goldman-Rakic P. S.: Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nature Neuroscience, 2, 1999, pp. 1131–1136.

[49] Roy D., Pentland A.: Learning words from sights and sounds: a computational model. Cognitive Science, 26, 2002, pp. 113–146.

[50] Rumelhart D. E., McClelland J. L., editors. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge (MA), 1986.

[51] Schiller P. H.: On the specificity of neurons and visual areas. Behavioural Brain Research, 76, 1996, pp. 21–35.

[52] Sirosh J., Miikkulainen R.: Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. Neural Computation, 9, 1997, pp. 577–594.

[53] Smith L. B.: Children's noun learning: How general learning processes make specialized learning mechanisms. In B. MacWhinney, editor, The Emergence of Language. Lawrence Erlbaum Associates, Mahwah (NJ), second edition, 1999.

[54] Smith L. B.: How domain-general processes may create domain-specific biases. In M. Bowerman and S. Levinson, editors, Language Acquisition and Conceptual Development. Cambridge University Press, Cambridge (UK), 2001.

[55] Taylor N. R., Hartley M., Taylor J. G.: Coding of objects in low-level visual cortical areas. In W. Duch, J. Kacprzyk, E. Oja, and S. Zadrony, editors, Artificial Neural Networks – ICANN '05. 15th International Conference Proceedings, Berlin, 2005. Springer-Verlag, pp. 57–63.

[56] Thompson I.: Cortical development: A role for spontaneous activity? Current Biology, 7, 1997, pp. 324–326.

[57] Ungerleider L., Mishkin M.: Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, Analysis of visual behavior, MIT Press, Cambridge (MA), 1982, pp. 549–586.

[58] Van Essen D. C., DeYoe E. A.: Concurrent processing in the primate visual cortex. In M. S. Gazzaniga, editor, The Cognitive Neurosciences. MIT Press, Cambridge (MA), 1994.

[59] Vanduffel W., Tootell R. B., Schoups A. A., Orban G. A.: The organization of orientation selectivity throughout the macaque visual cortex. Cerebral Cortex, 12, 2002, pp. 647–662.

[60] Verkindt C., Bertrand O., Echallier F., Pernier J.: Tonotopic organization of the human auditory cortex: N100 topography and multiple dipole model analysis. Electroencephalography and Clinical Neurophisiology, 96, 1995, pp. 143–156.

[61] Volkmer M.: A pulsed neural network model of spectro-temporal receptive fields and population coding in auditory cortex. Neural Computing, 3, 2004, pp. 177–193.

[62] von der Malsburg C.: Self-organization of orientation sensitive cells in the striate cortex. Kybernetic, 14, 1973, pp. 85–100.

[63] Vuilleumier P., Henson R. N., Driver J., Dolan R. J.: Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. Nature Neuroscience, 5, 2002, pp. 491–499.

[64] Wallis G., Rolls E.: Invariant face and object recognition in the visual system. Progress in Neurobiology, 51, 1997, pp. 167–194.

[65] Wandell B. A.: Computational neuroimaging of human visual cortex. Annual Review of Neuroscience, 10, 1999, pp. 145–173.

[66] Wandell B. A., Brewer A. A., Dougher R. F.: Visual field map clusters in human cortex. Philosophical Transactions of the Royal Society of London, 360, 2005, pp. 693–707.

[67] Weigelt S., Kourtzi Z., Kohler A., Singer W., Muckli L.: The cortical representation of objects rotating in depth. Journal of Neuroscience, 27, 2007, pp. 3864–3874.

[68] Winer J. A., Miller L. M., Lee C. C., Schreiner C. E.: Auditory thalamocortical transformation: structure and function. Neuron, 28, 2005, pp. 255–263.

[69] Zeki S.: Colour coding in the cerebral cortex: The reaction of cells in monkey visual cortex to wavelenghts and colours. Neuroscience, 9, 1983a, pp. 741–765.

[70] Zeki S.: Colour coding in the cerebral cortex: the responses of wavelength-selective and colour-coded cells in monkey visual cortex to changes in wavelength composition. Neuroscience, 9, 1983b, pp. 767–781.