A Model of Binding Concepts to Spoken Names

Andrew P. Papliński[1], Lennart Gustafsson[2], and William M. Mount[1]

[1] Monash University, Australia
Andrew.Paplinski@monash.edu
[2] Luleå University of Technology,Sweden
Lennart.Gustafsson@ltu.se

Abstract. We present a preliminary model of binding mental objects to their respective spoken names that aims at mimicking fMRI-tested behaviour. Our model consists of three mutually interconnected association modules which store mental objects, represent their spoken names and bind these to the mental objects, respectively. The auditory information is supplied to the unimodal association auditory module from a sensory or primary auditory module. The mental objects map is created during the learning process by information from the primary objects module. The information exchanged between modules is reduced to a 3-dimensional mental 'label'. It is shown that this highly non-linear dynamic network is able to quickly reconcile spoken names with congruent and incongruous 'thoughts'.

Key words: Multimodal integration, cortical modelling, mental binding

1 Introduction

We present a model of one aspect of processing speech as in the human brain concentrating on the fundamental problem of how our speech related mental activities excite a variety of interconnected cortical areas. The current functional neuroanatomy model of speech processing known as the dual-stream model is presented in [1–4]. This model identifies seven general networks of processing speech information [1]: **Spectrotemporal analysis** is carried out bilaterally in auditory cortices, while the **Phonological network** is responsible for sub-lexical phonological processing and representation. The model then diverges into two broad streams: the articulatory stream, concerned with speech development and production, includes a **Sensorimotor interface** and an **Articulatory network**; the lexical stream, concerned with auditory and speech recognition, comprehension and lexical access, includes the **Lexical interface** which links phonological to semantic information and a **Combinatorial network**, postulated to integrate lexical and articulatory processing. The final network in the model is the **Conceptual network**, a widely distributed system with both posterior perceptual and more anterior cognitive components. All the above networks are asymmetrically located in both hemispheres and are also bi-directionally interconnected, resulting in a rather complicated overall structure. Facing such a complexity we choose to model a small section of the cortical speech processing structure, concentrating on interaction of the phonological, lexical and conceptual networks.

Our model belongs to the class of models using maps that describe, firstly, mapping multidimensional sensory signals into a low-dimensionality cortical representation. We direct the reader's attention to [5] for a discussion on the existence and significance of cortical maps. For a related neurocomputational account of map-based processing and its role in speech comprehension and production, see also [6, 7]. In this paper we study the development of activities over time in modules of our simulated network when thought commands about different mental objects, in this case animals, and their associated spoken names are simultaneously presented. This perceptual or mental binding task is considered central to speech understanding and generation. This case is akin to the fMRI-based study of recognition of spoken words representing animals when subjects had been cross-modally primed for different animals [8]. Although our work has been influenced by many research findings, we will only refer to those that are closest to our approach. We first state that the use of Kohonen self-organizing maps [9] in modeling brain activities is a well-established method. A more recent generalization is the ability to create networks of such maps. Typically, maps and their interconnections are trained using the Kohonen learning law, normalised Hebbian learning, or a combination of both as in [10–13, 6, 7].

In our approach, which follows from our earlier works on multi-modal integration ([14, 15]) self-organized modules form low dimensional labels. These labels are used as afferent signals to up-stream modules, and may represent any type of perceptual, conceptual or lexical 'features', consistent with the neurocomputational modelling efforts of [16].

2 Initial simulation results

The **model** presented here is based upon an hierarchical network of idealised cortical modules involved in the binding of spoken words and names to mental objects. A model, in general, implies a degree of simplification and

abstraction. In this sense, we do not attempt to represent each and every cortical area taking part in perceptual and semantic processing of animal names, in particular, or higher-level language and cognitive processes, in general. Rather our aim is to represent processing in a smaller number of modules or maps aggregating the processes of several cortical areas. Such simplifications allow us to present better how our computational tools and methods can be useful in studying problems related to mental objects or concepts and their spoken or written names.

In our simulation experiments we connect four main modules representing four cortical areas, arranged as shown in Figure 1. These areas store, represent and perform the following functions

- Mental Objects map (MO) which stores mental concepts, i.e., combination of perceptual features, of 30 animals, including a tiger as shown in Figure 1.
- Sensory Auditory map (SA) modelling aspects of primary auditory function.
- Unimodal Association auditory map (UA) which performs sub-lexical processing of auditory information and projects the spoken names to the bi-modal association map.
- 'Bimodal' association map (M+A) where perceivable mental objects and their corresponding spoken names are lexically bound together

Before we describe technical details of the structure and functions of the simulation network, let us imagine, with reference to Figure 1, that an fMRI scan has been taken when a human subject has been thinking about and listening to a spoken name 'tiger'. Our simulation model consists of only four cortical-like areas that respond

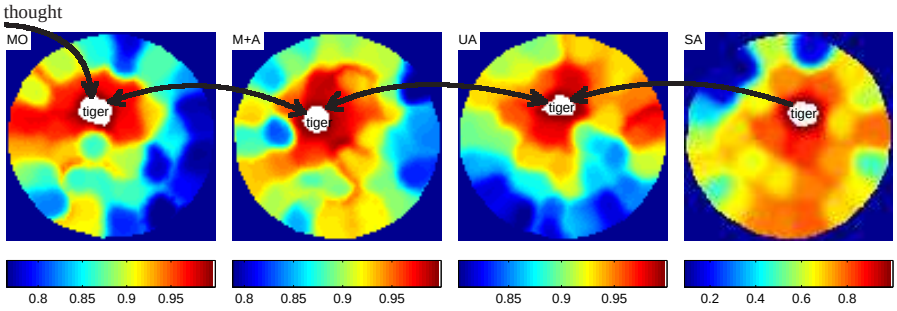

Fig. 1. Interconnection of four simulated cortical maps: MO – Mental Objects map, M+A – Bimodal association map: mental objects and auditory names, UA – Unimodal association Auditory map, SA – Sensory Auditory map

to this stimulus. The spoken word excites cortical patches representing the mental object 'tiger' in its four manifestations as shown in Figure 1. Note that there is a feedforward path from low-level sensory area (SA) through the unimodal association area (UA) to the bimodal binding area (M+A). Similarly a forward pathway exists from the 'conceptual' mental objects area (MO) to M+A. These direct feedforward paths assure that the binding process is rapid – at least in the case of congruous thoughts and inputs. In the case of incongruity between the mental object and the spoken word (I *think* 'horse', you *say* 'tiger') a feedback loop is activated from the M+A area back to the to MO and UA areas. We discuss details and provide examples of activations of the feedforward and feedback paths in the following sections. We also consider the crucial question of what information is actually passed between the maps.

3 Modules and Maps

Each module consists of a number of artificial neuronal units randomly located in a circular area. Relative positions of 'neurons' inside the circle are described by the position matrix $\mathbf{V}$. The total number of neurons is selected in proportion to the number of objects represented by the module and in our case, having 30 objects, the number of neurons is approximately $30 \times 30 = 900$, which is the number of rows in the weight matrices $\mathbf{W}$ characterizing the synaptic strengths in each module.

Conceptually we have two types of modules: sensory modules and association modules. Sensory modules operate on a relatively large number of afferent signals and produce a low-dimensionality efferent signal labeling the sensory object. This label information consists of a three-dimensional vector encoding the relative location of the excited neuron within a given cortical patch, supplemented by the post-synaptic activity level of this neuron. Such labels form a universal representation of all information passed between object maps within the hierarchical model. Association modules, in turn use these labels as their afferent signals and produce efferent labels encoding positions of activated neuronal patches within these maps. This is a fundamental property of our model.

In Figure 3, which we will discuss in detail later, we present an example of three maps, namely, Mental Objects, MO, Unimodal association Auditory, UA, and Bimodal association, U+A maps. In the maps, neuronal positions are marked with the yellow dots and the map area is tesselated with respect to the peaks of neuronal

activities for each stimulus. We can recognize the patch for 'tiger' and compare it with the activity surface for the same stimulus in Figure 1. Note that the tessellation pattern depicted in Figure 3 is rather nominal and an approximation to the specific cortical activity map as in Figure 1.

The objects or spoken names are 'placed' in the relevant cortical area during the learning process. Once the learning is completed, the cortical area responds with an activity pattern characteristic to each stimulus. The operation of a cortical module is functionally equivalent to mapping the higher dimensionality input space to a three-dimensional output space.

It is assumed that during normal operation each association module performs a static non-linear mapping of the form:

$$\mathbf{y_{MA}} = g(\mathbf{x_{MA}}) \ , \quad \mathbf{x_{MA}} = [\mathbf{y_{MO}}, \mathbf{y_{UA}}]$$

where $\mathbf{x}$ and $\mathbf{y}$ represent input and output signals, respectively and $g(\cdot)$ describes the Winner-Takes-All function of $\mathbf{W_{MA}x_{MA}}$.

4 The structure and operations of the network

The modules as described above are connected into the network presented in Figure 2. The top row consists

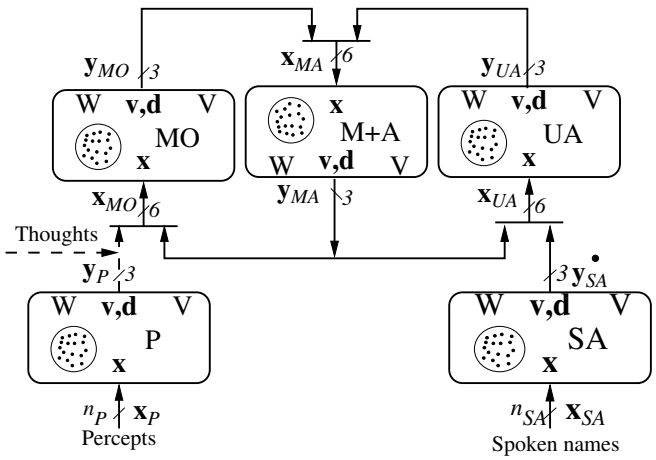

Fig. 2. The four-plus-one map network

of three mutually interconnected association modules with the 'bimodal' integration module (M+A) binding mental objects from the MO module to their spoken names from the UA module.

The auditory sensory module SA operates upon a pre-processed representation of the spoken animal names. The names have been obtained from the Merriam-Webster online dictionary [17]. During pre-processing each name was converted into 36 mel-cepstral frequency coefficients using the VoiceBox toolbox [18] for MATLAB [19]. For each waveform we use three windows overlapping by 50% each window producing 12 mel-cepstral coefficients. The output of the sensory auditory (SA) module is a three-dimensional vector $\mathbf{y_{SA}}$ indicating the relative position of the maximum excitation neuron and its relative postsynaptic strength.

To create the map of mental objects, MO, we first categorize selected animals in the additional primary objects module, P. During the learning process this module is supplied with a feature vector $\mathbf{x_P}$ describing the animals. Following training, the object categorizing module P is disconnected and the signal $\mathbf{y_P}$ is interpreted as **thought commands** used for recalling mental objects stored with the module MO.

Adding feedback to a highly non-linear and complex network is always a challenge. For the bi-modal association module we can write the following formal dynamic equation:

$$\mathbf{y_{MA}}(t+1) = \mathcal{M}(\mathbf{y_P}(t), \mathbf{y_{SA}}(t), \mathbf{y_{MA}}(t)) \tag{1}$$

however this does not help apart from emphasising the recurrent and non-linear nature of the network. Nonetheless from the simulation perspective it is pleasing to note that the trained network settles immediately if we apply known stimuli congruent with the thoughts. In other words, if the labels in the network are known and congruent, the network quickly converges to a stable state.

In order to improve performance of the network, particularly during the learning phase, we project all n-dimensional vectors onto a surface of a n+1-dimensional unity hypersphere. Working with unity vectors makes it easy to compare them by calculating relevant inner products. This allows us to use a simplified dot-product learning law [9]. The sensory-level maps, P and SA, are trained first, independently of each other, producing maps similar to that presented in Figure 3. The learning parameters are carefully selected so that learning

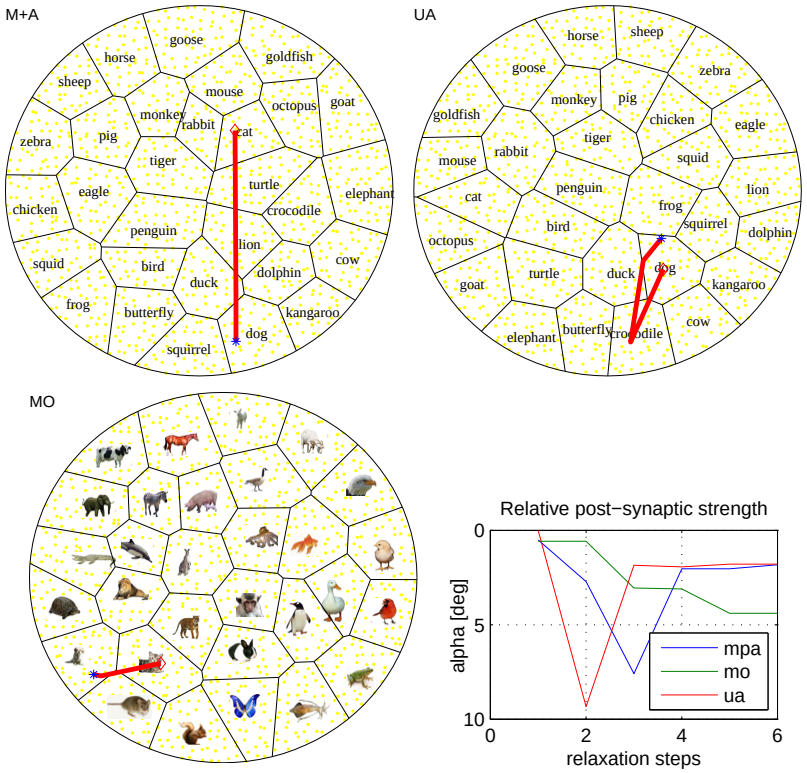

Fig. 3. Incongruent, but similar objects (*think* 'cat', *hear* 'dog'). The red line is the trajectory of winners. The starting and terminal points are marked with '⋄' and '*', respectively. Relative post-synapic strength is measured as $\cos\alpha$

proceeds relatively quickly through the competition and cooperation phases (see [9] for details). After a few hundred epochs the maps are typically fully developed. The three association modules forming the recurrent part of the network are then trained together. This time, after each learning step we perform several relaxation steps running the network as in eqn (1), until all efferent signals in the network are constant. There is a limit on the number of iterations imposed that is important in the initial learning stages. Once the maps are fully developed, the network stabilizes quickly after a small number of iterations depending on congruence between the thoughts and auditory inputs.

5 Simulating incongruent thoughts and names

Now we can conduct number of tests exposing the network to congruent, noisy and incongruent stimuli. When we apply auditory sensory inputs congruent with internal 'thoughts' the network quickly activates corresponding patches across cortical areas as in Figure 1. In our previous works we tested responses of similar networks [14, 15] to congruent, but noisy stimuli (phonemes and letters). Here we present the results for incongruent 'thoughts' and objects' spoken names. In general we have identified three basic types of behaviour depending on the level of similarities between the incongruent objects and related thoughts.

First we consider the situation when the objects are similar, but names are dissimilar e.g. *think* 'cat', *hear* 'dog'. The results of such simulation are presented in Figure 3. In this simulation only three association modules: MO, UA and M+A, which are interconnected by the feedback loop (see Figure 2) actively participate. At the starting point, the MO module *thinks* 'cat', whereas the auditory unimodal association module UA *hears* the word 'dog'. The bimodal association map, M+A is arbitrarily initialized with a mental object 'cat'. This initial state is marked on three maps in Figure 3 with the '⋄'. Since these positions have been learned during the training procedure, the relative post-synaptic strengths are initially at their maximum value, as seen in the plot in Figure 3.

Due to the corrective feedback through the network, the perceptual strengths at each of the maps varies from the initial points, but after just five steps, settles to a stable pattern of activity. The MO and M+A maps confabulates from their initial 'cat' position to a position in the 'dog' area, with a post-synaptic strength that indicates that the level of confidence has been reduced (see the plot in Figure 3) because the thought 'cat' is still being supplied. Similarly, the auditory UA map shifts its response from the original 'dog' positions to modified positions within the same patch, showing slightly reduced confidence levels. This is a good result in two respects, processing speed and correctness; the maps quickly negotiate the initial confusion and select a sensible, "predominantly 'dog' " solution.

The space limitation allows for only a brief description of two other types of results. Consider now the *think* 'frog', *hear* 'dog' situation when the objects are dissimilar but the names are acoustically similar. In this

case the network quickly settles to a solution when the thought prevails: the UA map moves from the initial 'dog' position into the 'frog' area. Finally, when both, objects and names are dissimilar, the maps can oscillate between two positions until the thought is changed to coincide with the name.

We hope that the above examples clearly demonstrate the power of presented simulation model.

Conclusion

We presented a model for binding spoken names to mental objects and provide preliminary results to demonstrate how this can potentially emulate several realistic cognitive behaviours related to speech and language processing. These include the virtual immediacy by which object names are perceptually bound to perceivable objects during typical human activities such as reading and listening, cognitive delays typically experienced when resolving conflicts with partially incongruous, or perhaps misheard auditory information and even the rivalrous sequences of perception or oscillating mental activity that can result when exogenous stimuli (evidence of the senses) contradict endogenous patterns of thought.

In the interest of maintaining structural simplicity of the model, several assumptions have been made and in the case of the mental objects map, the distinction between perceptual modalities, conceptual categories and semantic relationships have been somewhat blurred. This simplification has distinct advantages in terms of computational efficiency, allowing conceptual information to be encoded as arbitrary lists of features and object qualities. In a more comprehensive and biologically realistic model, the auxiliary 'P' module could be divided into specific sensory modalities or sub-modalities, used to represent visual, tactile, spatial or other modally-grounded perceptual features applicable to mental object categories (such as animals). The representation of the mental objects would then involve a multi-modal integration of such features and lexical binding of these to their associated written or spoken names.

References

1. Hickok, G., Poeppel, D.: The cortical organization of speech processing. Nature Rev. Neurosci. **82** (2007) 393–402
2. Binder, J., Desai, R., Graves, W., Conan, L.: Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex **19** (2009) 2767–2796
3. Price, C.: The anatomy of language: a review of 100 fMRI studies published in 2009. Ann. N.Y. Acad. Sci. **1191** (2010) 62–88
4. Rauschecker, J.P., Scott, S.K.: Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neurosci. **12**(6) (2009) 718–724
5. Schreiner, C.E., Winer, J.A.: Auditory cortex mapmaking: Principles, projections, and plasticity. Neuron **56**(2) (2007) 356–365
6. Li, P., Zhao, X., MacWhinney, B.: Dynamic self-organization and early lexical development in children. Neuron **31** (2007) 581–612
7. Mayor, J., Plunkett, K.: A neurocomputational account of taxonomic responding and fast mapping in early word learning. Psychol Rev. **117**(1) (2010) 1–31
8. Noppeney, U., Josephs, O., Hocking, J., Price, C., Friston, K.: The effect of prior visual information on recognition of speech and sounds. Cerebral Cortex **18** (2008) 598–609
9. Kohonen, T.: Self-Organising Maps. 3rd edn. Springer-Verlag, Berlin (2001)
10. Miikkulainen, R.: Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. Brain and Language (1997) 334–366
11. Sit, Y.F., Miikkulainen, R.: Computational predictions on the receptive fields and organization of V2 for shape processing. Neural Computation **21**(3) (2009) 762–785
12. Miikkulainen, R., Kiran, S.: Modeling the bilingual lexicon of an individual subject. In: Lect. Notes in Comp. Sci. Volume 5629., Springer (2009) 191–199
13. Monner, D., Reggia, J.A.: An unsupervised learning method for representing simple sentences. In: Proc. Int. Joint Conf. Neural Net., Atlanta, USA (2009) 2133–2140
14. Chou, S., Papliński, A.P., Gustafsson, L.: Speaker-dependent bimodal integration of Chinese phonemes and letters using multimodal self-organizing networks. In: Proc. Int. Joint Conf. Neural Networks, Orlando, Florida (2007)
15. Gustafsson, L., Jantvik, T., Papliński, A.P.: A multimodal self-organizing network for sensory integration of letters and phonemes. In: Proc. IASTED Int. Conf. Artif. Intell. Soft Comp., Palma De Mallorca, Spain (2007)
16. Gliozzi, V., Mayor, J., Hu, J.F., Plunkett, K.: Labels as features (not names) for infant categorisation: A neuro-computational approach. Cog. Sci. **33**(3) (2009) 709–738
17. : Merriam-Webster online dictionary. http://www.merriam-webster.com (2010)
18. Brooks, M.: Voicebox: Speech processing toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (2010)
19. TheMathWorks: MATLAB. The language of technical computing. http://www. mathworks.com (2010)