Bimodal Incremental Self-Organizing Network (BiSON) with Application to Learning Chinese Characters

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

[1] Monash University, Australia
Andrew.Paplinski@monash.edu
[2] University of New South Wales, Australia
w.mount@adfa.edu.au

Abstract. We present a recurrent learning system that can incrementally integrate stimuli in two modalities, visual and auditory. The system consists of five self-organizing modules, each mapping input stimuli into respective latent spaces. Two sensory modules convert the input stimuli into an internal 3-D "neuronal code". The central module integrates the bimodal information, and through modulatory top-down feedback influences the organization of data in two unimodal association units. Two feedback gains control the strength of the feedback connection. As an example we selected a set of Chinese characters and related spoken words. It is shown that the learning system can build a stable neuronal structure for incrementally applied visual and auditory stimuli.

Keywords: Multimodal Learning, Visual and Auditory stimuli, Recurrent networks, Self-organization, Chinese characters.

1 Introduction

It is well acknowledged that human languages are inherently cross-modal, requiring both written and spoken components to realize their full potential. Interesting accounts of the origins of human written and spoken language can be found in [7] and many others.

Due to the redundancy between the visual-signing (gestural, drawing or writing) and auditory-speech systems, spoken cross-modal references to symbolic names, as well as written representations of spoken signals, allowed for an increasingly rich repertoire of utterances, words and characters. These could be combined to describe the physical and mental world in more abstract terms and the argument goes that as languages became more sophisticated, they became increasingly embedded in the complex culture within which they co-evolved [3].

Some important differences in the way the human brain processes pictographic languages in general and Chinese in particular is described in [2]. As processing of the radicals and oriented brush strokes comprising the 50,000 or so known Chinese characters is very different from that of phonetically based languages such

M. Lee et al. (Eds.): ICONIP 2013, Part I, LNCS 8226, pp. 121–128, 2013.

as English, a different set of phonological and orthographic skills are required in Chinese language acquisition [16].

Language processing is fundamental to human cognitive ability and involves multiple cortical networks and pathways across visual, auditory and other modalities. Large brain networks used in reading are discussed in [5]. See also [14] and [15] for some of our own efforts to develop simplified models of such networks.

While language comprehension and production requires the function of multiple cortical areas acting in concert, a region on the left hemisphere, the left superior temporal sulcus (STS) has been advanced as the main site for integration of visual and auditory speech information [3]. Recent fMRI studies support the key role of this region in the fusion of letters and speech sounds in the human brain[1].

Previous results of applying our models to the problem of integration of phonemes and letters in Chinese and Swedish are reported in [4] and [8] respectively. A related modelling framework is used in this case, however later enhancements for sequential feed-forward and recurrent learning [15] and incremental learning [11] provides an opportunity to revisit the problem of learning Chinese characters and associated sounds.

By incrementally building up sensory, unimodal associative and fused bimodal representations within our simplified five module network, a consistent way in which a human child or a computational agent may learn essential features of Chinese or any other spoken and written language is suggested.

2 The Structure of the Learning System

The structure of our bimodal incremental learning system is presented in Fig. 1. The function of the system is to receive sensory information across two modalities, visual and auditory, and integrate these representations. As an example, we use Chinese characters and their utterances as inputs to our system. We have experimented previously with Chinese characters in [4] and more recently [13].

The main part of the learning systems consists of five interconnected self-organizing modules. Two **sensory** level modules, **Vis** and **Aud**, process visual and auditory stimuli, respectively, converting coded sensory information, $\mathbf{x}_V$ and $\mathbf{x}_A$ into the standard internal representation of signals $\mathbf{y}_V$ and $\mathbf{y}_A$. In the next hierarchical level, two **unimodal association** modules, **UV** and **UA**, combine the signals from the sensory level, $\mathbf{y}_V$ and $\mathbf{y}_A$, with the modulating top-down feedback signals, $\mathbf{y}_{VA}$, produced by the top level **bimodal association** module, **V+A**. The strength of the top-down modulatory feedback is controlled by two gain parameters, g_{UV} and g_{UA}, at the input to the respective unimodal association modules.

The bimodal association module is presented here as a central part of the learning system. We can hypothesize that this module may also be activated by endogenous thoughts and can be used to drive modal effector systems, one for writing and one for articulation.

Following our previous works [14,15,8], the building block of our system is a self-organizing module (map) with the following characteristics (see Fig. 2):

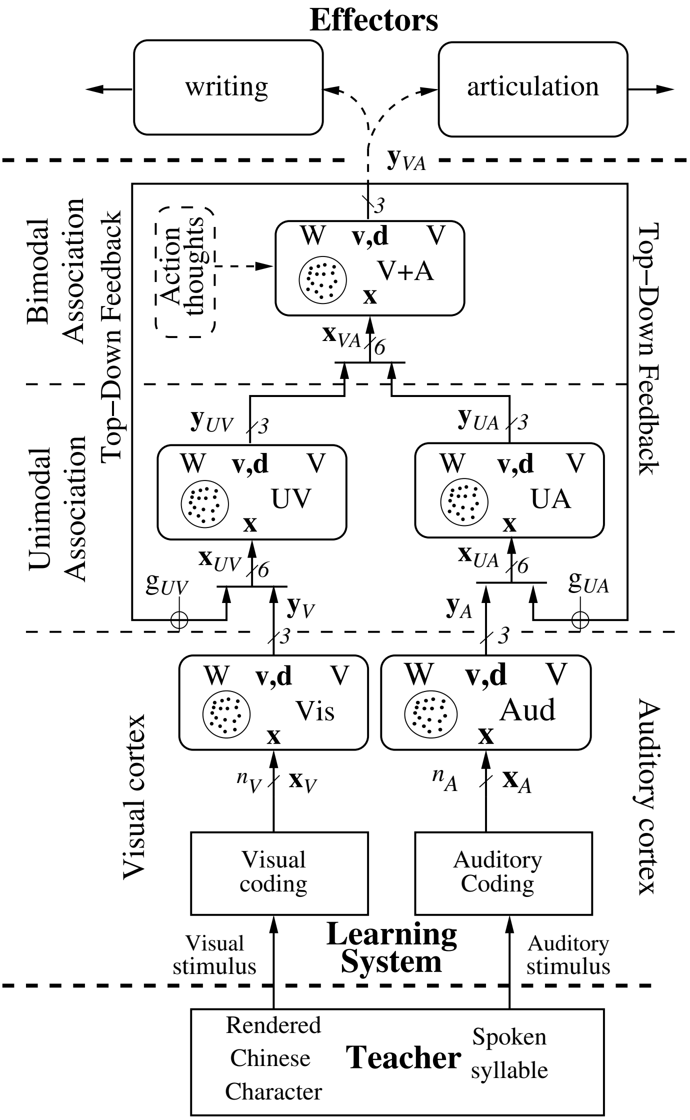


Fig. 1. The structure of the incremental learning system. The main part of the system consists of five self-organizing modules: Vis, UV, Aud, UA and V+A.

- The neuronal units (shown as yellow dots in Figures 2 and 4) are randomly distributed inside a unit circle, rather than on a uniformed rectangular grid.
- A constant number (stochastically) of neuronal units per stimulus, ϵ, is maintained to simulate the redundancy observed in biological systems.
- All stimuli vectors are projected on a unity hypersphere. Therefore, a simplified "dot-product" version of the Kohonen learning law [9] may be used.

Central to our processing architecture are the outputs from self-organizing modules, for example, $\mathbf{y}_V$ in Fig. 1. Note that dimensionality of all output vectors is 3. Such an output vector is a concatenation of the 2-D position vector of the winning neuron and its postsynaptic activity, namely,

$$\mathbf{y} = [\mathbf{v}_w \; d_w] = \mathcal{K}(\mathbf{d}), \quad \mathbf{d} = W \cdot \mathbf{x} \tag{1}$$

where W is an $M \times D$ matrix of parameters, the weight matrix, M being the number of nodes (neurons), and $\mathcal{K}$ is the Winner-Takes-All function identifying the position of the neuronal node $\mathbf{v}_w$ for which the post-synaptic activity $\mathbf{d} = W \cdot \mathbf{x}$ attains the maximum.

These output signals implement a ubiquitous "neuronal code", providing a unified way for information labels to be exchanged between modules of the network. It should be emphasized that the positions of neurons are considered in a latent space, not the physical one. This implies that during incremental learning the physical position of participating neurons is not affected.

3 The Incremental Learning Process

Our incremental learning process for a single iSOM has been introduced in [11]. We refer to this paper for detailed comparisons with other structures that may appear similar, in particular, a variety of growing SOMs. One fundamental difference is that during the learning process, we maintain a stochastically constant ratio between the number of neuronal units and the number of current stimuli.

In our case this expected ratio is always greater than one, implying that more than one neuron is used to represent a percept. This can be contrasted with other applications of SOMs where the number of of neurons is typically less than the number of data points. A study into the increased persistence and stability of percepts provided by such neural representations is presented in [6].

The incremental learning process starts with a small number, say $n = 3$, of initial stimuli and consists of two main steps:

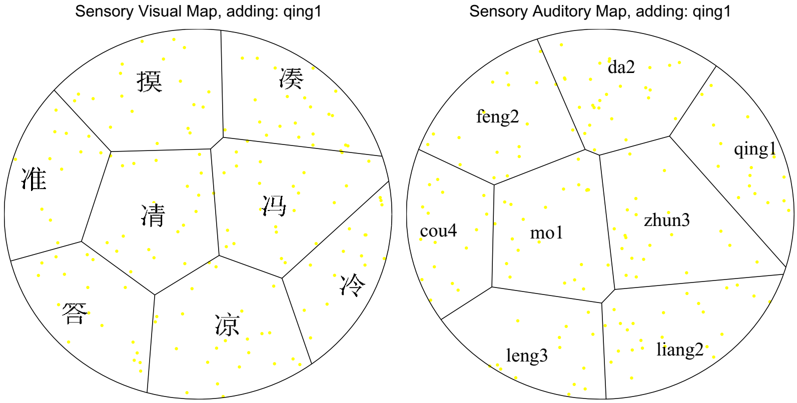
Feedforward learning: We start with setting two feedback gains g_{UV} and g_{UA} to zero, thus opening the feedback loops and

- generate the number of neuronal units proportional to the number of stimuli, $m = n\epsilon$, say $3 \times 16 = 48$
- generate initial weights to be located around the north pole of the unity hypersphere
- perform the "dot-product" learning law for all maps, for all initial stimuli, for a set number of epochs, say 100.

Recurrent learning: We set the feedback gains to required values, e.g., $g_{UV} = g_{UA} = 0.5$ and repeat the learning process with one basic modification: after completing learning **for each stimulus**, we re-evaluate outputs from all 3 interconnected modules, namely, $\mathbf{y}_{UV}$, $\mathbf{y}_{UA}$ and $\mathbf{y}_{VA}$, until the values of the outputs settle. This typically happens after just two steps (see [14,15] for more details)

Adding more stimuli: Now we add new stimuli, one at a time, and repeat the two learning steps above with the following modification:

- generate the additional number of neuronal units proportional to the increment in the number of stimuli, $m_i = n_i\epsilon$, say $1 \times 16 = 16$
- initialize weights of the new m_i units to be equal to the weights of the closest neighbours.
- continue learning in the open and closed loop as above.


Sensory Visual Map, adding: qing1
摸
凑
准
凊
冯
冷
答
凉
Sensory Auditory Map, adding: qing1
da2
feng2
qing1
cou4
mo1
zhun3
leng3
liang2

巴	ba1	模	mo2	打	da3	悌	ti4	冲	chong1	冯	feng2	决	jue2	冺	min3
拔	ba2	抹	mo3	大	da4	苛	ke1	凑	cou4	凎	gan4	冷	leng3	凝	ning3
把	ba3	漠	mo4	梯	ti1	咳	ke2	凗	cui1	凅	gu4	凉	liang2	凄	qi1
爸	ba4	搭	da1	綈	ti2	可	ke3	凋	diao1	减	jian3	冽	lie4	凊	qing1
摸	mo1	答	da2	體	ti3	客	ke4	凍	dong4	凈	jing4	凌	ling2	准	zhun3

Fig. 2. The sensory maps after 8 stimuli. The set of characters and their pinyin names are included.

This incremental process elegantly solves the problem of initialization of weights. Less effective random initialization is performed only for a small number of initial neurons. The result of learning after application of 8 stimuli is shown in Fig. 2.

Topological ordering of the stimuli needs to be considered in the context of the feature vectors used. For the visual channel we used an angular integral of Radon Transform (aniRT) discussed in [13] for the 20,000 Chinese characters and in [12,10] for other types of visual objects.

In the table of Fig. 2, the first 20 characters are grouped according to similarities in pronunciation, while the second set of 20 have a similar structure in terms of the aniRT coefficients. Each rendered character, is converted into a 91 component vector (91 being the size of the diagonal of the image). As described in [13] very few components are required to differentiate between characters, although some more are needed to capture details of the visual object.

For the auditory channel, we follow our previous work [14,8,4] where melcepstral coefficients are used to represent frequency of the speech sounds. We use

12 coefficients per frame, with 3 frames overlapping by 50%. We also add the duration of the utterance, so that we have 38-D feature vectors after projecting up on the unity hypersphere. An example of such coding is given in Fig. 3.

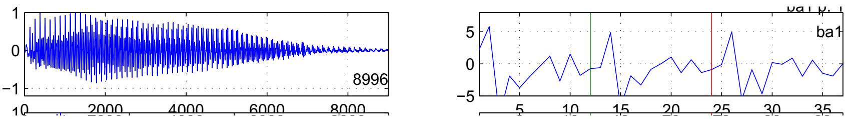


Fig. 3. Representing the 'ba1' sound in the melcepstral domain: 8996 speech samples are coded by 36 melcepstral coefficients. Normalized number of samples is also included as the first coefficient in the right-hand side plot.

If we continue the process of learning adding incrementally more and more stimuli, after 40 stimuli we obtain five maps as presented in Fig. 4. Again, at the sensory level it is easy to spot the topological ordering in both modalities. At the unimodal association level the topological arrangement of the stimuli is influenced by the top down feedback. Finally, the bimodal map presents the fusion of information from two modalities. In this paper we concentrate on the issue of the incremental learning which is performed for the congruent stimuli presented on the visual and auditory channels. The reader is referred to our previous works for considerations related to noisy and incongruent stimuli.

4 Discussion

While the sensory maps develop independently in a feedforward learning mode, the influence of top-down feedback during the recurrent learning phase ensures that cross-modal relationships are encoded in the unimodal and bimodal maps. Significantly, even though crossmodal information is not explicitly contained in either the visual or auditory information presented alone, the BiSON model ensures that the inherently bimodal structure of the words or characters comprising the natural language (in this case Chinese) is effectively encoded and learned.

A further enhancement would be to explore interactive learning and communication through the addition of character articulation and writing effector modules. This could introduce a third sensori-motor modality to our multimodal language framework. Finally, by extending this architecture to include simplified modules for central perceptual, evaluative and task-orientation functions, we hope to develop a sophisticated multilayered learning model where the symbolic elements or tokens of a spoken and written language represent meaningful mental objects and concepts within an interactive setting.

5 Conclusion

We present a recurrent learning system vaguely mimicking some basic cortical areas related to integration of visual and auditory information. In the example,

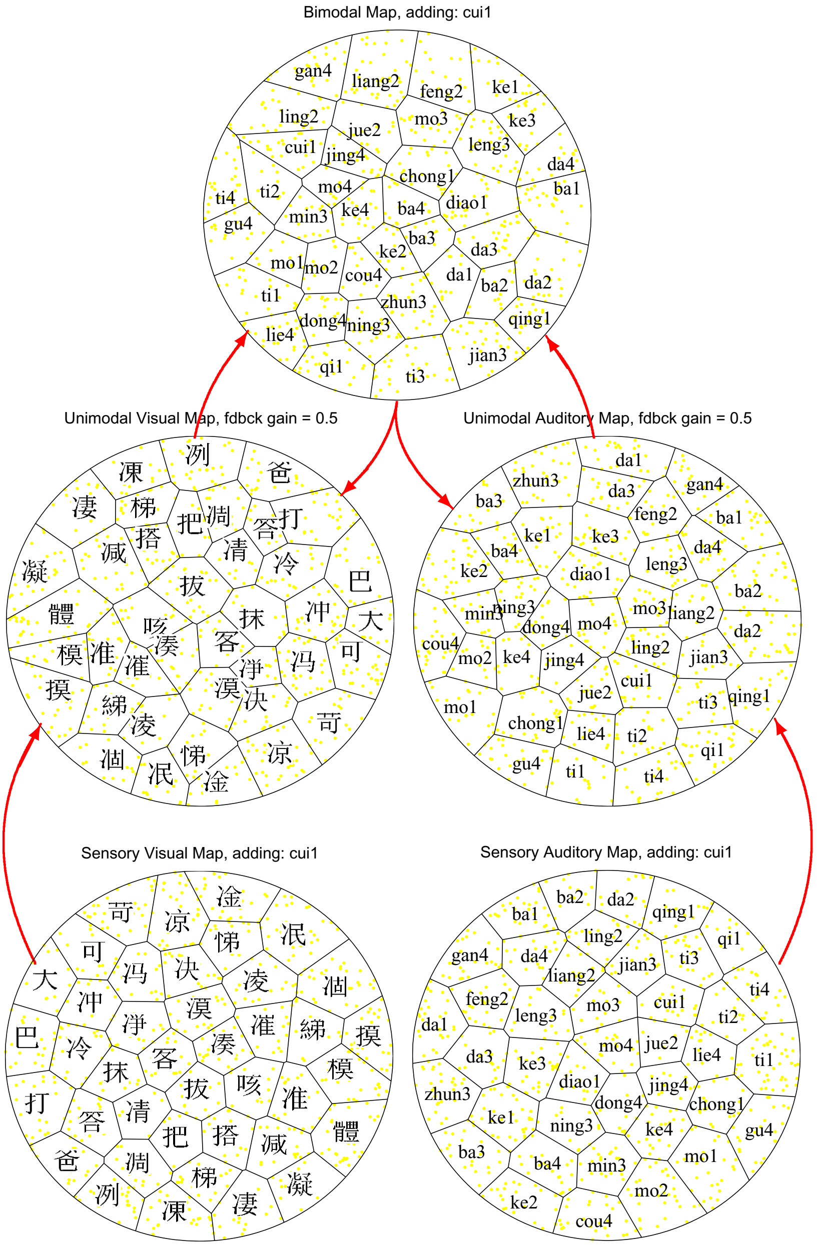


Fig. 4. The five interconnected maps developed after incremental application of 40 stimuli

the system 'reads' Chinese characters and simultaneously 'listens' to their pronunciation. At each stage, we add one more visual-auditory stimulus and the learning system incorporates it into its 5-map structure. Despite the recurrent nature of the system, it converges to a fixed point after a only small number of recurrent iterations. The bimodal module plays the central part of the system for fusion of the bimodal percepts and from which effectors for writing and speaking can be driven.

The software used in this paper is written in MATLAB and is available upon request.

References

1. Atteveldt, N.M.: Speech Meets Script: FMRI Studies on the Integration of Letters and Speech Sounds. University of Maastricht (2006)
2. Bookheimer, S.: How the brain reads Chinese characters. Neuroreport 12(1) (January 2001)
3. Calvert, G.A., Brammer, M.J., Iversen, S.D.: Crossmodal identification. Current Biology 2, 247–253 (1998)
4. 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, pp. 248–253 (August 2007)
5. Dehaene, S.: Reading in the Brain. Viking (2009), `http://pagesperso-orange.fr/readinginthebrain/figures.htm`
6. Druckmann, S., Chklovskii, D.B.: Over-complete representations on recurrent neural networks can support persistent percepts. Advances in Neural Information Processing Systems (2010)
7. Fitch, W.T.: The Evolution of Language. Cambridge University Press (2010)
8. Jantvik, T., Gustafsson, L., Papliński, A.P.: A self-organized artificial neural network architecture for sensory integration with applications to letter–phoneme integration. Neural Computation 23, 2101–2139 (2011)
9. Kohonen, T.: Self-Organising Maps, 3rd edn. Springer, Berlin (2001)
10. Papliński, A.P.: Rotation invariant categorization of visual objects using Radon transform and self-organizing modules. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (eds.) ICONIP 2010, Part II. LNCS, vol. 6444, pp. 360–366. Springer, Heidelberg (2010)
11. Papliński, A.P.: Incremental self-organizing map (iSOM) in categorization of visual objects. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP 2012, Part II. LNCS, vol. 7664, pp. 125–132. Springer, Heidelberg (2012)
12. Papliński, A.P.: Rotation invariant categorization of colour images using Radon transform. In: Proc. WCCI–IJCNN, pp. 1408–1413. IEEE (2012)
13. Papliński, A.P.: The angular integral of the radon transform (aniRT) as a feature vector in categorization of visual objects. In: Guo, C., Hou, Z.-G., Zeng, Z. (eds.) ISNN 2013, Part I. LNCS, vol. 7951, pp. 523–531. Springer, Heidelberg (2013)
14. Papliński, A.P., Gustafsson, L., Mount, W.M.: A model of binding concepts to spoken names. Aust. Journal of Intelligent Information Processing Systems 11(2), 1–5 (2010)
15. Papliński, A.P., Gustafsson, L., Mount, W.M.: A recurrent multimodal network for binding written words and sensory-based semantics into concepts. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part I. LNCS, vol. 7062, pp. 413–422. Springer, Heidelberg (2011)
16. Siok, W., Fletcher, P.: The role of phonological awareness and visual orthographic skills in Chinese reading acquisition. Developmental Psychology 37(6), 886–889 (2001)