A NEURALLY PLAUSIBLE IMPLEMENTATION OF A VOWEL RECOGNITION SYSTEM

ROELANT OSSEWAARDE[1]

[1] HU University of Applied Science, Computer Science, Utrecht, The Netherlands, e-mail: roelant.ossewaarde@hu.nl

Abstract We present research-in-progress aimed at developing a sensory recognition system with good performance in sparse training contexts. The system is a machine learning model, structured after a proposed neural architecture of early stage object recognition by humans. It stores representations in parallel systems that mimic associative and declarative memory systems. We present mathematical formulations of the underlying system of storage and apply this to the problem of vowel recognition by infants. The stored representation makes use of the distance between the formant frequencies of vowels, as analog magnitudes, rather than their absolute acoustic valuations. Our formulation allows for a learning strategy that is both neurally plausible and computationally tractable.The resulting system can be used in any environment that requires the system itself to recognize invariant properties of objects, visual or acoustic.

Keywords: object recognition, neural system, analog magnitude, vowel detection, machine learning.

University of Maribor Press

DOI https://doi.org/10.18690/978-961-286-362-3.47
ISBN 978-961-286-362-3

1 Introduction

Neural networks (deep learning) represent the state of the art in current artificial intelligence applications. They dominate in the fields of visual and aural perception, even if some of their shortcomings have proven to be pervasive.

One of these shortcomings is that they require extensive training in order to accommodate variations of a stimulus. Perception AI systems typically have great difficulty detecting objects under transformations such as rotation, scaling and color changes, or of objects that are partially hidden or obscured by other objects. Human perception is hardly affected by either affine transformations or partial obscurity of objects.

Two of the architectural traits of the human brain that make it especially competent in fast object detection are the hierarchical architecture of the brain (Hawkins et al., 2019) and the specific division of labor between episodic and associative memory systems(Poggio & Anselmi, 2016). It is assumed that these systems build partial representations that are invariant for spatial transformations or partial obscurity ("invariants"). At different layers of the hierarchical structure, these objects are represented as imprecise approximations, of objects.

The overall goal is to develop a model that describes how infants learn to distinguish the sounds of their mother language(s) from foreign ones. A previous study (Vallabha et al., 2007) uses an Expectation-Maximization based approach (parametric) and a topography based approach (nonparametric) to learn the specific spaces of English and Japanese vowels. At the heart of their approach is the notion that some probability distribution is an adequate description of the amount of acoustic energy that characterizes each individual vowel. As long as the distributions are sufficiently distinct, software can learn to distinguish vowels after learning to decode the speech signal into an acoustic profile.

This study presents work-in-progress to develop the computational architecture for a neural perception system that mimics the way in which human brains can learn such distributional distinctions as invariant representations of sounds and then use transformations, in a similar way as affine transformations in the visual domain, to facilitate recognition.

Systems that employ invariants in general require far less training than conventional neural systems, and are especially robust to detect stimuli with representations that vary to a large degree (Dupoux, 2018).

1.1 The analog properties of auditory stimuli

The perception of sounds involves the determination of aspects of the stimulus. One aspect that is used to discriminate sounds is the distribution of energy over the spectrum of the frequencies that together form the sound. In particular vowel sounds are well distinguishable based on their differences in fundamental frequency (*F0*) and other local spectral peaks (formamts) (*F1* and *F2*). Vowel production shows variation between individual speakers. Vowel recognition is based on the relative difference between *F0*, *F1* and *F2* rather than on their absolute values.

The recognition of vowels requires the comparison between the magnitudes of the spectral energy at different formant positions. The neural system to represent magnitudes is usually modeled as a Analog Magnitude Accumulator (AM, Whalen et al., 1999) a process by which each events are enumerated or represented as an impulse of activation from the nervous system. The representation of a magnitude through this system is an approximation of the total number, with some margins for noise and error. Analog Magnitudes are subject to the so-called *magnitude* and *distance* effects (Dehaene, 1997; Flombaum et al., 2005).

It may be meaningful for a speaker to produce sounds in a particular part of their range, for example to indicate prominence, segmentation boundaries (Ladd, 2008) or information structure (Wennerstrom, 2001). When a speaker raises or lowers the pitch (*F0*), all other formants change as well. It is the distance between formants that determines which vowel is perceived. This distance must be large enough for speakers to be able to tell the formants apart. This is captured by the notion of "just-noticeable difference interval" (JNI), which has been studied for different kinds of stimuli. It is often assumed that the Weber-Fechner's law of psychophysics governs the function that predicts when two stimuli are distinct enough to be judged different. Applied to formants, the Weber-Fechner law expresses that our ability to detect a difference between two formant values depends on the base pitch height itself (Weber's law), and that the relationship between stimulus difference and discriminability is logarithmic (Fechner's law). Interpreted in this domain: the JNI is greater for pitches higher in the spectrum than for pitches lower in the spectrum, and best plotted on a logarithmic scale.

Both human infants and human adults can generate numerosity estimates for up to three AM-sets in parallel (Halberda et al., 2006; Zosh et al., 2007), which at least in theory would enable an infant in the very earliest stage of language learning to be able to compare the quantities involved with vowel discrimination.

2 General architecture

As the general computational architecture to model the analog magnitudes necessary for vowel discrimination, we follow the general proposal of cortical organization detailed in Hawkins et al. (2019), the representational model of invariants by Poggio & Anselmi (2016) and the strategies for storing approximations from Leibo et al. (2015).

Hubel and Wiesel proposed a distinction between *simple* and *complex* cells (henceforth: S-, respectively C-cells) where C-cells pool S-cells in a network. One C-cell with its S-cells is a Hubel-Wiesel module (HW-module). The response of a C-cell that pools S-cells to a stimulus x is denoted as $\mu_k(x)$ for the k-th element of the signature of the concept. An HW-module features as a computational structure in most current theories of neural networks, including convolutional, HMAX and Nearest Neighbors networks. Concepts as stored in the brain are referred to as templates (τ_k) with as actual manifestations a set of k signatures.

2.1 Cortical columns store invariants

An HW-layer consists of one to many HW-modules, and the hierarchical organisation of HW-layers is called an HW-architecture. In the human cortex there are six of such layers, with the bottom one (denoted V1 for modules involved in visual perception) connected to the sensorimotor areas of the brain, and the highest one (IT) assumed to be the most abstract. Any form of cognition involves an interplay between the higher and the lower levels.

An invariant representation can be modeled as a particular activation pattern of different HW-modules. Invariant representations are encoded in the brain as early as in the lowest HW-layers. Because these layers are activated within the first 100ms of the ventral stream's exposure to a stimulus, feedback from the cortex cannot play a role yet in recognition. There simply isn't enough time for synapses to go roundtrip from the lowest to the highest level of the cortex in 100ms. We model in a computational way that early stage of recognition, assuming that the higher levels that do receive feedback from the top levels (or other HW-modules) are organized in a similar way.

Invariant representation for a stimulus at that low level requires the generation of representations of the perceived stimulus under a known set of transformations (in the auditory domain: pitch variations, voice quality, loudness, duration).

An HW-module may be considered a data structure that encodes a signature of τ_k. Under that assumption, it has a set of values and operations to access and update the atoms of data that the particular HW-module store. Learning means: a sequence of inserts or updates in the HW-module, given a particular activation caused by a stimulus. Each of the K HW-modules stores data D that corresponds with the signature of template τ_k.

In the brain, HW-modules are composed of neural cells, interconnected through dendrites that connect to axons. Computationally, if we denote the number of potential connections as n, a dendritic segment is represented as a binary vector $D = [b_0.., b_{n-1}]$ where a non- zero value b_i represents a synaptic connection to presynaptic cell i and $s = D$ indicates the number of synapses on that segment. At a given moment, 20-300 synapses (s) are typically active, over a much larger number of potential connections. A possible representation of the synaptic configuration of HW-modules is by sparse distributed representations (Ahmad & Hawkins, 2016).

If an HW-module is seen as a data storage unit, its insert and access operations can be given as in Eqs. 1 and 2, (cf. Leibo et al., 2015). Object categorization is then done by finding the μ_k that minimizes the loss function.

$$\text{INSERT}(\text{D}_\text{k}, \text{t}) : D_k \leftarrow D_k \cup \{t\}. \quad (1)$$
$$\text{QUERY}(\text{D}_\text{k}, \text{t}) : \mu_\text{k}(\text{x}) \leftarrow \max \langle x, t\rangle \quad (2)$$

Biologically, the INSERT-operation of Eq. 1 is implausible. Brains do not directly store information as a database does, but rather, we argue, as an approximation of the strength of an aspect of the stimulus. In this view, the brain stores an approximation of properties such as size, length, temporal structure etc. as the components of an invariant representation of objects in the real world. Rather than storing a discrete number, an analog magnitude is stored, in the form of a subpart of the joint activation of a subpart of a cortical column.

In our computational model, the formulation of the INSERT operation as proposed by Leibo et al. (2015) is used, which formulates two separate operations.

In the first INSERT strategy, the best rank-*r* approximation of a matrix is computed. The set of templates (the invariant properties of vowel sounds) is first expressed as a matrix (T_k), which represents the specific activation patterns of HW-modules. The matrix is then reduced using Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). The INSERT operation is then defined as in Eq. 3, the concatenation of T_k and new information (*t*) resulting from a stimulus.

The second INSERT strategy uses random projections (Bingham & Mannila, 2001) (cf. Eq.4). A random projection is a projection of matrix X (with dimensions $n \times m$) to X' (with dimensions $n \times o$, where

$o < m$ per the Johnson-Lindenstrauss lemma) by the transform over a matrix with random values. Dimensionality reduction using random projections is computationally less intensive but does not result in an outcome with correlated candidates. Hence, storage is quicker but the relations between templates necessary for invariance are lost.

$$\text{INSERT}(Dk, t)\text{: } D\text{k} \leftarrow [\text{T}k \mid t]\ V' \text{ with } U'\Sigma'V' = [\text{T}k \mid t] \quad (3)$$

$$\text{INSERT}(Dk, t)\text{: } D\text{k} \leftarrow [\text{T}k \mid t]\ [R \mid r] \text{ with } U'\Sigma'V' = [\text{T}k \mid t]$$
$$\text{where } \mathbf{r} = \text{random vector s. th. } [\mathbf{R}|\mathbf{r}] \text{ is orthogonal} \quad (4)$$

The two INSERT operations can account for two different ways of storing representations. The first INSERT (dimensionality reduction through SVD/PCA) is slower for insertions but does retain any correlations to other templates and is thus a good candidate for the type of associative memory that is required for invariant representation in the cortical areas. The second INSERT works without reference to any other template, and is computationally simpler - the biological counterpart could be the hippocampus which stores episodic memory. Because this project is limited to the earliest stages of cortical processing, before any synaptic pathway can be excited or inhibited for feedback, the first INSERT (Eq. 3) is the only considered for implementation.

2.2 Acquisition of priors

In a machine learning system, invariance to auditory or visual translations (such as scale, du- ration, pitch height) can be built up by simply memorizing examples. The core of the computational model is to construct a system that does not rely on such extensive memorization. The algorithm we use has been adapted from that used for visual processing as proposed by Poggio and Anselmi (2016):

1. Developmental stage
    a. For each of K isolated templates (cf. τk), memorize a sequence of Λ of |G| frames corresponding to the sound pattern transformations (g_i = 1, ...,|G|). This may include the absence of change which characterizes prolonged vowel duration. The sequence of frames is observed over some time interval.
    b. Repeat for each of K templates.

2. Online determination of invariant signature for a single stimulus from a new object.
    a. For each t_k compute the dot product of the stimulus with each of the |G| transformations in Λ_g.
    b. For each k compute a probability distribution of the resulting values.
    c. The signature is the set of *K* cumulative distributions. It is stored using the INSERT operations defined in Section 2.1.

A template stores the approximate distance between formants and their values. When a stimulus has duration, such as a vowel sound, the transformations learned in the develop- mental stage may be the absence of change. This approach captures both the invariance (by applying known transformations in a generative way) and the analog nature (by using dimensionality reduction) of approximation of the transformation results. In terms of a vowel sound, JNI distances between *F0*, F1, etc. are computed in the generative step, but acceptable deviations from these are accounted for by the proposed approximated storage. A logical extension of the generative part is by positing a two-stage approach, in which the first stage ensures that a final signature is composed by HW-modules by pooling over invariant signatures, and the second stage permits specialized transformations that are object specific.

3 Results

The presented approach is implemented in R. Neural activations are inserted after PCA, modeled as a binary sparse distributed representation (SDR). The advantage of such a distributed representation is that synaptic pathways across HW-modules can be encoded without positing a separate neural level for each different HW-layer. The SVD/PCA models magnitudes and their analog nature.

Because the implementation follows the HW-architecture as discussed in Section 2, it mirrors a biologically plausible neural architecture for perception learning. We are currently in the process of extensive experimentation with real data.

The INSERT-operations using matrices and dimension reduction techniques allow for a plausible way to model the approximateness of properties of concepts. Learning at INSERT takes place by application of Oja's rule, which is a generalized form of PCA using Hebbian learning (Oja, 1982).

Although auditory perception is a relatively straightforward task, the implementation can easily be adapted to other problems that are typically a challenge for training-heavy Machine Learning. Such tasks are for example the fast recognition of previously unseen objects or of objects that may be presented after spatial transformations such as scaling, rotation.

References

Ahmad, S., & Hawkins, J. (2016). How do neurons operate on sparse distributed representations? A mathematical theory of sparsity, neurons and active dendrites. ArXiv:1601.00720 [Cs, q-Bio]. http://arxiv.org/abs/1601.00720

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 245–250.

Dehaene, S. (1997). The number sense. Oxford University Press.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. Cognition, 173, 43–59.

Flombaum, J. I., Junge, J. A., & Hauser, M. D. (2005). Rhesus monkeys (Macaca mulatta) spontaneously compute addition operations over large numbers. Cognition, 97(3), 315–325.

Halberda, J. P., Sires, S. F., & Feigenson, L. (2006). Multiple spatially-overlapping sets can be enumerated in parallel. Psychological Science, 17(7), 572–576.

Hawkins, J., Lewis, M., Klukas, M., Purdy, S., & Ahmad, S. (2019). A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex. Frontiers in Neural Circuits, 12. https://doi.org/10.3389/fncir.2018.00121

Ladd, D. R. (2008). Intonational phonology. Cambridge University Press.

Leibo, J. Z., Cornebise, J., Gómez, S., & Hassabis, D. (2015). Approximate Hubel-Wiesel Modules and the Data Structures of Neural Computation. ArXiv:1512.08457 [Cs, q-Bio]. http://arxiv.org/abs/1512.08457

Oja, E. (1982). Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology, 15, 267–273.
https://doi.org/10.1007/BF00275687

Poggio, T. A., & Anselmi, F. (2016). Visual Cortex and Deep Networks: Learning Invariant Representations. The MIT Press.
http://dl.acm.org/citation.cfm?id=3099710

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. Proc Natl Acad Sci U S A, 104(33), 13273–13278.
https://doi.org/10.1073/pnas.0705369104

Wennerstrom, A. (2001). Intonation and evaluation in oral narratives. Journal of Pragmatics, 33(8), 1183--1206.

Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. Psychological Science, 10, 130–137.

Zosh, J., Feigenson, L., & Halberda, J. (2007). Infants' ability to enumerate multiple spatially-overlapping sets in parallel. Journal of Vision, 7(9), 220–220.