# Taxonomic and Whole Object Constraints: A Deep Architecture

**Mattia Cerrato (mattia.cerrato@unito.it)**
Dipartimento di Informatica, University of Torino
Corso Svizzera 185, 10149, Torino, Italy

**Edoardo Arnaudo (edoardo.arnaudo@edu.unito.it)**
Dipartimento di Informatica, University of Torino
Corso Svizzera 185, 10149, Torino, Italy

**Roberto Esposito (roberto.esposito@unito.it)**
Dipartimento di Informatica, University of Torino
Corso Svizzera 185, 10149, Torino, Italy

**Valentina Gliozzi (valentina.gliozzi@unito.it)**
Center for Logic, Language, Cognition & Dipartimento di Informatica, University of Torino
Corso Svizzera 185, 10149, Torino, Italy

**Abstract**

We propose a neural network model that accounts for the emergence of the taxonomic constraint and for the whole object constraint in early word learning. Our proposal is based on Mayor and Plunkett (2010)'s neurocomputational model of the taxonomic constraint and extends it in two directions. Firstly, we deal with realistic visual and acoustic stimuli. Secondly, we model the well-known whole object constraint in the visual component. We show that, despite the augmented input complexity, the proposed model compares favorably with respect to previous systems.

**Keywords:** Neural Networks; Children; Language acquisition.

## Introduction

How do infants learn the referent of words? As Quine (1960) famously pointed out, for every word heard in a given circumstance, there are several possible referents: in order to infer the appropriate one, infants have to rule out several possible alternatives. But how? Markman (1989) proposed that infants rule out inappropriate referents by means of three constraints. By the *taxonomic constraint* children extend words to taxonomically-related objects: when a child hears the word "dog" pronounced by a caregiver while pointing at a specific dog, she generalizes the referent of "dog" to all dogs, not just to the one in front of her. By the *whole object constraint* children assume that novel words refer to objects as a whole, rather than to their parts, substance, color, or the visual context in which it appears. Lastly, by the *mutual exclusivity constraint* children assume that two labels usually do not refer to the same object.

This paper concerns the first two constraints, namely the taxonomic and the whole object constraint.

Our starting point is Mayor and Plunkett (2010)'s neurocomputational model of the taxonomic constraint. Their model provides an account of how the taxonomic constraint may emerge from infant experience, as the result of the interplay between (i) taxonomic organization of visual inputs in visual areas, (ii) phonetic organization of the acoustic inputs in acoustic areas, (iii) Hebbian learning developing connections between the two organizing areas. The model uses self-organizing maps (Kohonen, 2001) and Hebbian learning (Hebb, 1949), which are considered cognitively plausible mechanisms, describing at an abstract level realistic forms of information organization in the brain (Hebb, 1949; Miikkulainen, Bednar, Choe, & Sirosh, 2005). The powerful interplay between these structures allows word-object associations to taxonomically generalize after a single (*one-shot*) joint word object presentation[1].

Here we extend Mayor and Plunkett (2010)'s seminal model in two directions:

1. We intend to investigate whether the taxonomic constraint can emerge from experience if we consider *realistic visual and acoustic* stimuli (photographic images with different size, color, location in the picture, point of view, etc. and audio excerpts embodying spoken words synthesized via software) instead of the very simple, artificially built stimuli examined in the original model. A first effort in this direction was undertaken by (Fenoglio, Esposito, & Gliozzi, 2017), in which, however, only realistic visual stimuli were considered. Here we enrich that proposal by considering visual *and acoustic* realistic stimuli (as well as the whole object constraint, see below). To this purpose, we insert in the model two deep architectures, one convolutional to process visual stimuli and the other recurrent to process realistic acoustic stimuli.

2. We insert the *whole object constraint* in the model. Whether early learned or innate, the capacity of picking up the objects in a scene is present in early infancy (see e.g. Spelke, 1990). However, this primacy of the object concept

[1] For a critical discussion of the breadth of one shot learning and fast mapping see for instance (Yurovsky, Fricker, Yu, & Smith, 2014) or (McMurray, Horst, & Samuelson, 2012).

in visual scene analysis is not present in most recent convolutional neural network (CNN) models, that are the state of the art in vision tasks. In fact, these models usually process visual images as a whole (object and background context together), see e.g., (Zhu, Xie, & Yuille, 2017). Here we overcome this limitation of CNNs by inserting a segmentation module that extracts the object from the visual scene before feeding it to the CNN for feature extraction. In the experimental section we show that the whole object constraint improves the performance of the model.

Remarkably, our model replicates Mayor and Plunkett (2010)'s performance with realistic visual and acoustic stimuli, albeit requiring *very-few* joint presentations of image and spoken word pairs.

It is worth mentioning here that we do not try to maintain that CNNs or LSTMs are cognitively plausible models of how realistic stimuli are processed in biological brains – more details about this point in the "New Model" Section. We are just validating the hypothesis that the (Mayor & Plunkett, 2010) model could generalize to more complex stimuli and that the whole-object constraint can be helpful in this model.

## Mayor and Plunkett (2010)'s model

Mayor and Plunkett (2010) neurocomputational model of taxonomic constraint (Figure 1) is based on (i) a visual self-organizing map (SOM) that processes visual inputs, (ii) an acoustic SOM that processes acoustic inputs, (iii) Hebbian connections between the two maps. Both self-organizing maps and Hebbian learning are considered cognitively plausible mechanisms (Hebb, 1949; Miikkulainen et al., 2005)

Firstly, the two maps are independently trained (using the standard learning algorithm for self-organizing maps, see Kohonen, 2001) to categorize the visual and the acoustic stimuli. This first learning phase is preliminary to word learning, and unsupervised, proper word learning starting to occur once infants have already started to learn to organize visual and acoustic information in isolation.

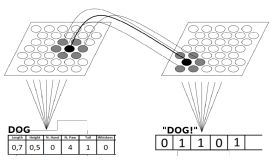



Figure 1: Mayor and Plunkett (2010)'s model

In this way, the two maps learn to represent the stimuli of their training set in a topologically significant way: close units respond (activate) similarly to similar stimuli. The *neural activation* $a_j$ of a neuron $j$ in response to a stimulus $x$ is defined as: $a_j = e^{-\frac{q_j}{\tau}}$, where $q_j$ is the *quantization error* (i.e., the distance between the input vector $x$ and $j'$s weight vector: $q_j = \|\mathbf{x} - \mathbf{w}_j\|$)), and $\tau$ is a parameter that modulates the neural activation. The neuron having the strongest activation is the stimulus' Best Matching Unit (BMU).

Once this first phase of learning is complete, the actual word learning can start. This is the Hebbian Learning phase, in which visual and acoustic stimulus are presented to their respective maps and the synapses between the two maps are strengthened. In particular, for each neuron $v$ on the visual map and neuron $p$ on the acoustic map, the Hebbian connection $u_{v,p}$ is strengthened proportionally to the resulting neural activations $a_v$ and $a_p$, as follows:

$$u'_{v,p} = u_{v,p} + 1 - e^{-\lambda a_v a_p}$$

where $\lambda$ is the Hebbian training learning rate, and $u'_{v,p}$ is the Hebbian connection after the update.

A single Hebbian learning event, combined with the previously acquired categorization capabilities of the visual and acoustic SOMs, allows the model to generalize the association to other stimuli belonging to the same category.

*Comprehension* is assessed by considering what visual category is retrieved when a word is presented to the auditory map and its activation is propagated via Hebbian connections. *Production* is assessed by considering what word is produced by the auditory map when a visual stimulus is presented to the visual map and the resulting activation is propagated through Hebbian connections.

The ability of the model to extend the learned word-object associations to other words and objects belonging to the same category is measured by the *Taxonomic Factor*, which is the percentage of correct word-object associations generated by the model (i.e., the average of the Production and Comprehension statistics). Results show that when the SOMs are adequately trained the Taxonomic Factor reaches 80% after a single joint word-object presentation.

## New Model

We have enriched the original Mayor & Plunkett model (2010) (and Fenoglio et al. (2017)) so that (i) it can deal with realistic visual *and acoustic stimuli*, and (ii) it captures *the whole object constraint*. A graphical representation of the overall model is contained in Figure 2. For the visual and the acoustic stimuli we propose to use Deep Neural Networks to act as powerful feature extractors: we use two deep convolutional neural networks (Mask R-CNN and Inception V3) to process visual information and a deep recurrent neural network (Deep Speech) to process acoustic information. These models have been widely adopted for this purpose by the Machine Learning community, as they are able to output highly discriminative features (Razavian, Azizpour, Sullivan, & Carlsson, 2014; Graves, Mohamed, & Hinton, 2013; Hannun et al., 2014). Even if it is not the main focus of this paper, it is worth mentioning that these models have also been proposed as realistic models of visual and acoustic processing. Several studies establish a parallel between the representations of the visual input created by the different levels of CNNs and the way in which visual stimuli are processed by the visual cortex (Serre, 2016; Kriegeskorte, 2015; Khaligh-Razavi & Kriegeskorte, 2014). Furthermore, comparisons have been drawn between Recurrent Networks units (specifically the LSTM cell (Hochreiter & Schmidhuber, 1997)) and biologically plausible models of working memory such as the

PBWM model (prefrontal cortex and basal ganglia) (O'Reilly & Frank, 2006).

### Visual Component

The visual stimuli that we consider are images taken from the Common Objects in COntext (COCO) dataset (Lin et al., 2014). In this dataset images are labelled pixel-wise, meaning that it is possible to extract the foreground objects from the background scene (i.e. performing image *segmentation*). As a first component of the visual module, we included a Mask R-CNN segmentation model (He, Gkioxari, Dollár, & Girshick, 2017), which separates foreground objects from the background content. Then the foreground object is cut from the background, the background erased and the new image so obtained is fed into an InceptionV3 Deep Convolutional Network (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), an architecture which displays human-like performance on Object Recognition tasks. The deep network processes the object and builds a representation for it. We extract the representations (i.e. the neural activations) that are found in the deepest layer before the fully connected neural classifier (as they contain the most abstract features which have the best chance to depict the abstract concept the object instance refers to), and feed these representations to the visual self-organizing map.

To summarise, we employ a stack of two Deep Neural Networks in our visual module: the first one segmenting the object from the context; the second one analysing that output by means of a standard convolutional deep network. This architecture allows the model to overcome one limitation of standard deep convolutional models that, differently from humans (Spelke, 1990), do not use the notion of object when processing an image, and, on the contrary, rely very much on background information in object recognition tasks (Zhu et al., 2017).

### Acoustic Component

We process spoken words using a Deep Speech Recurrent Network (Hannun et al., 2014) which is close to the state of the art in the Speech Recognition (i.e. parsing speech into text) task. This network is able to extract highly discriminative representations from our input stimuli, which have been generated using a realistic voice synthesizer that can be set up to use both male and female voices as well as different regional English accents. In our experiments, we had the generator pronounce labels from the COCO dataset. Similarly to the visual module, we extract features by concatenating the hidden state of the recurrent units after each time step. The resulting vector representations are then truncated to the same length and reduced in dimensionality by means of principal component analysis.

### Overall Model

The upper component of the model, comprising the visual and acoustic self-organizing maps and their Hebbian connections, is trained as in the original (Mayor & Plunkett, 2010) model:

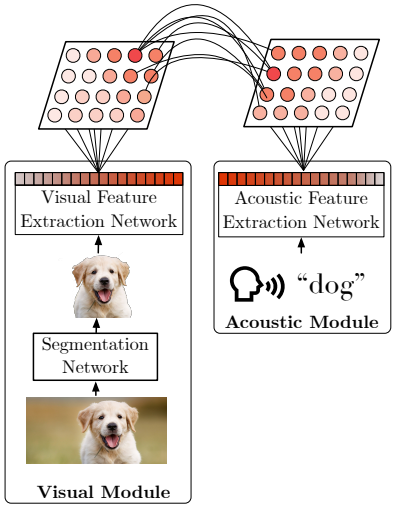



Figure 2: Our model

at first the visual and acoustic self-organizing maps are separately trained to organize their stimuli, then Hebbian learning starts.

Similarly to the original model, in order to assess the quality of the word/object association (the Comprehension and Production ability), we proceed as follows. For Production, we present to the visual map a visual stimulus, individuate a BMU, propagate its activity through Hebbian connections, and then evaluate the induced activation on the acoustic map with a cascading mechanism where neurons are interrogated in order of activation intensity. The first neuron corresponding to a single spoken word (i.e. a neuron that is the BMU for a single acoustic stimulus category) indicates which word is produced. We say that a Production task is successful if the category of the word matches the category of the visual stimulus. We proceed in a similar way for Comprehension.

## Experiments

In our experimental phase, we set out to answer the following two questions:

1. Does our extension to the original word learning model by (Mayor & Plunkett, 2010) still account for the taxonomic constraint? In other words, is it possible to use realistic auditory and visual stimuli and achieve good word learning performance?

2. Is the whole object constraint beneficial to the word learning process?

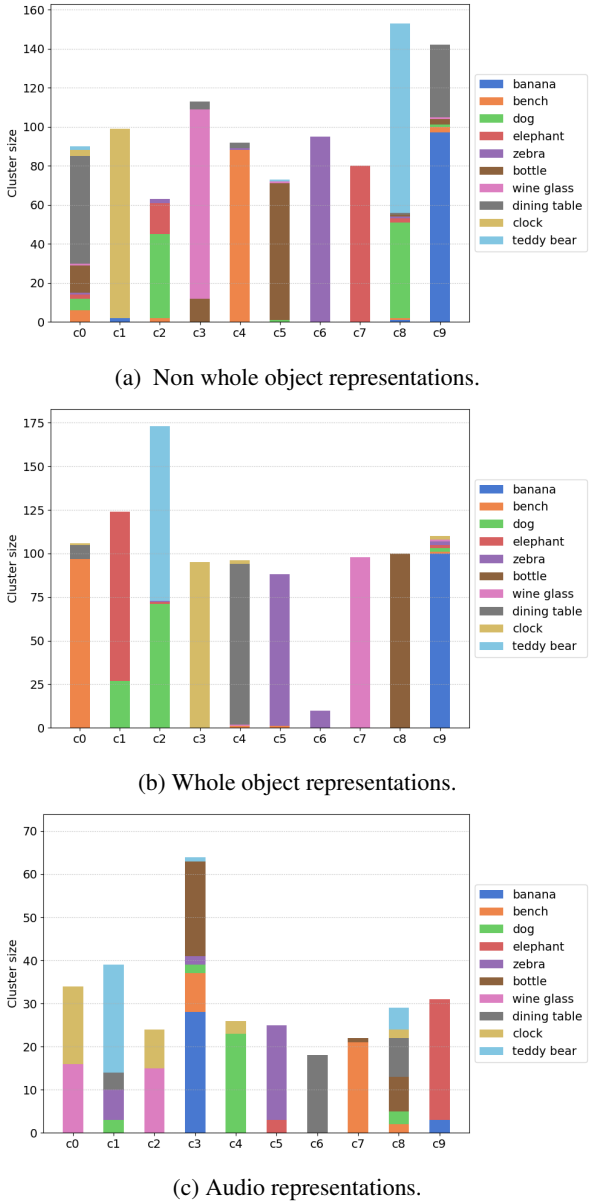
In order to extract whole object and non whole object representations (the inputs of the visual SOM), we trained two separate InceptionV3 networks for the same amount of time (i.e. epochs, full passes of the COCO dataset). However, one network was trained on images where the main object was cut out using the Mask R-CNN model, while the other one em-

ployed images that include a portion of the full visual scene[2]. We refer to these models respectively as the "whole object" and "non whole object" networks. Therefore, we explored the impact of using one network or the other in our visual module as a way to quantify the impact of including the whole object constraint in the overall model. An early evidence of the importance of the whole object constraint is provided by the performances (in terms of Object Recognition accuracy) of the two networks: the whole object network reaches a higher accuracy (93%) than the non-whole object network (77%) after the same amount of training time and similar learning rate schedules[3]. For the experiments that follow, however, we decided to only use as visual stimuli those images that have been correctly classified by both networks; thus, both convolutional models have perfect accuracy on the final visual dataset and can be compared on a fair ground. As far as the acoustic stimuli are concerned, we used a voice synthesizer to generate realistic voice recordings of both male and female voices pronouncing the object categories which appear in the visual dataset. To augment the size of the auditory dataset, we also varied the synthesizer's pronunciation speed. The representations were then extracted using a pre-trained Deep Speech network[4]. We truncated the representations to a length of 25 and kept the 20 most informative factors of variation using PCA. In the following sections, we report representation quality, SOM quality and taxonomic factor measurements for a dataset composed by 1000 visual stimuli and 390 acoustic stimuli belonging to 10 different word-object categories.

## Representation Quality

First off, we set out to understand whether the representations extracted from the realistic stimuli are well-behaved. To this end, we performed an experiment in which the representations are used as input for the k-Means clustering algorithm with $k$, the number of clusters, set to 10. After fitting the clustering model, we visualize the resulting clusters (see Figure 3) using a histogram plot.

We also assess the trained SOMs' topological organization by visualizing them. In Figure 4, we see that representations belonging to the same category are mapped on neurons that are topologically close. Moreover, we evaluate the organization quality by using the *class compactness* measure; this is computed by averaging the Euclidean distances between neurons that are BMUs for stimuli belonging to the same class and dividing by the average distance between BMUs for any stimulus. Lower values indicate better topological structure.




(a) Non whole object representations.

(b) Whole object representations.

(c) Audio representations.

Figure 3: K-means clusters. Colors show how representations of each category contribute to the clusters.

[2]More specifically, we used the bounding box information for each object in COCO and expanded it by 40% so to preserve a significant amount of visual context.

[3]We trained the whole object network for 60 epochs. We used a learning rate of $10^{-3}$, decreasing it to $10^{-4}$ after 40 epochs. This schedule, however, appeared to be very sub-optimal when training the non whole learning network, as the object recognition accuracy progressed very slowly. Therefore, the second network was trained with a learning rate of $10^{-2}$ and decreased it to $10^{-3}$ after 40 epochs.

[4]https://github.com/mozilla/DeepSpeech/

Averaging this formula over the categories in the dataset results in the overall *SOM compactness* value. We report (Table 1) lower compactness values for the SOM we trained on the whole object representations, and robust compactness for the non whole object and acoustic SOMs.

## Word Learning

As an experimental evaluation of the overall model, we compare the word learning capabilities of our model with and without the whole object constraint. After training with a number of joint word-object presentations, the model has to be able to produce an appropriate acoustic stimulus when presented with an image (*understanding*) and viceversa (*comprehension*). The algorithm used to obtain the final word-object association is described in the "New Model" Section.

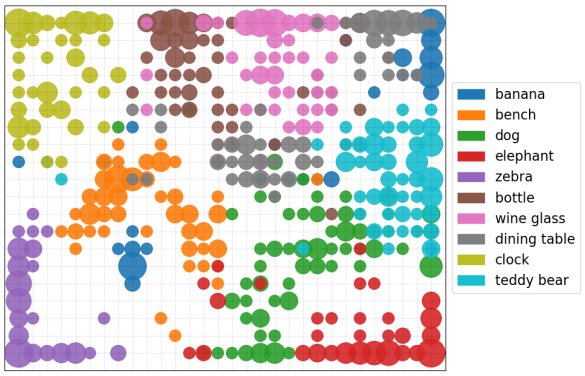
In Figure 5 we report the understanding and comprehension performances alongside the Taxonomic Factor (their average). A set of stimuli (20% of all the visual and acoustic representations) was reserved for testing and was excluded from the training sets.

## Discussion

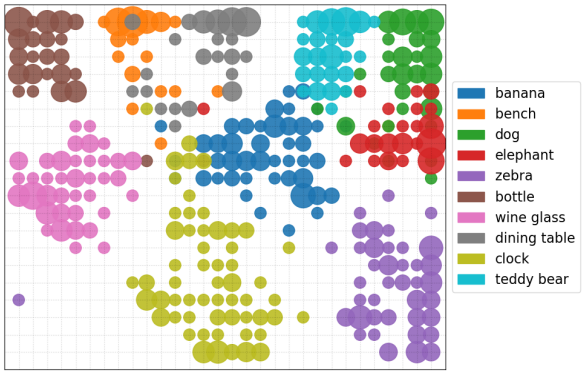
Coming back to the questions we set out to answer at the start of the section, Figure 5 makes apparent that our model manages to perform word learning with appropriate performance (set at a Taxonomic Factor of over 80% in (Mayor & Plunkett, 2010)) after very few word-object presentations. The performance is also in line with previous work on this model (Fenoglio et al., 2017), in which, however, very simplified acoustic stimuli were considered. Therefore, this computational model can still account for the taxonomic constraint even in the face of realistic visual and acoustic stimuli. As for the contribution of the whole object constraint to the model, we first observe that the comparison of the clusters is favorable to the whole object model (Figure 3); furthermore, the self-organizing maps that were trained using the aforementioned representations display good topological organization (Figure 4) and solid compactness values. In addition, as implied by Table 1, the SOM trained with the whole object representations displays stronger topological organization. As for the word learning performance, we obtain a significantly higher Taxonomic Factor when using the whole object representations and conclude that including the whole object constraint in this model is highly beneficial.

# Related Work

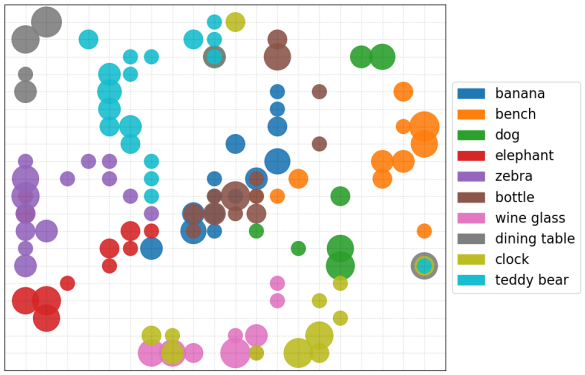
Our model bears some family resemblance to recent models of image-speech association learning (Synnaeve, Versteegh, & Dupoux, 2014; Harwath & Glass, 2015; Harwath, Torralba, & Glass, 2016; Chrupala, Gelderloos, & Alishahi, 2017), which, at least in part, have been proposed as cognitive models of spoken words referent acquisition. Similarly to Synnaeve et al. (2014), here we consider associations between images and single spoken words, whereas Harwath and Glass (2015); Harwath et al. (2016); Chrupala et al. (2017) consider associations between images and




(a) SOM over non whole object representations.




(b) SOM over whole object representations.




(c) SOM over audio representations.

Figure 4: SOM representations. Colors represent different categories, larger circles are for neurons that activate more often.

Table 1: Compactness values for the three SOMs.

| Visual Whole Object | Visual Non Whole Object | Acoustic |
|---|---|---|
| 0.228 | 0.372 | 0.429 |

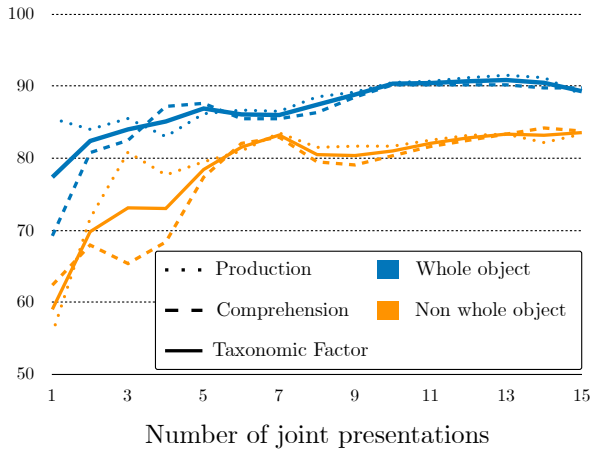



Figure 5: *Taxonomic factor* of the model, using an increasing number of pairs of stimuli per class during the training of the Hebbian Connections (on the x-axis).

more complex acoustic stimuli, as whole spoken sentences (Harwath & Glass, 2015; Harwath et al., 2016; Chrupala et al., 2017). With respect to all these models, the specificity of our model is that it learns to generalize image-speech associations to whole visual categories and all phonetic variants of a corresponding word, out of few *positive* joint image-speech presentations, without any need of explicit counterexamples. This parallels the training schedule by which humans usually learn to associate words (or sentences) to visual stimuli.

Vinyals, Blundell, Lillicrap, Kavukcuoglu, and Wierstra (2016) address the problem of *One Shot Learning*: how to build models that reproduce the crucial ability of humans, infants and adults, of learning out of *few examples*, as opposed to the massive training currently used for many neural network models? The proposed model is trained to integrate in one-shot new observations into pre-existent knowledge, represented by a support set. Similarly to our work, representations extracted from pre-trained neural networks are also employed. The authors test their model on classification tasks in which the training dataset is composed by 1 or 5 examples for each category; while a direct comparison would not be proper, as the experimental setups and datasets are fundamentally different, it is worth mentioning that word learning in the present approach does not rely on the supervised, gradient-based optimization of a training objective (i.e. a loss function). On the contrary, in our model word learning emerges after the unsupervised training of the SOMs and a few joint, positive presentations of word-object pairs.

## Conclusions

In this paper we expand on the the model originally introduced by (Mayor & Plunkett, 2010) and extended by (Fenoglio et al., 2017). Our work focused on two objectives: allowing the model to process realistic acoustic stimuli, and injecting the whole object constraint into it. We also introduce experiments allowing one to assess the effects of these two changes to the model.

In summary, the empirical evidence shows that the realistic stimuli are not hindering the ability of the model to learn the association between objects and word. In fact, even though the greater complexity of the stimuli representation makes the task harder, the system only requires a few joint presentations to reach the 80% taxonomic accuracy performance shown in the original work by Mayor and Plunkett (2010).

For what concerns the whole object constraint, the evidence demonstrates the remarkable impact of this constraint on the performances of the system. In practice the whole object constraint allows for better performances with respect to the model by Fenoglio et al. (2017) even considering that the latter is dealing with simpler acoustic stimuli. It is worth debating whether the whole-object representations extracted by the visual module contain all the parts of the original objects. Indeed, given the discriminative nature of the CNN training process, the representations may only contain few, very specialized features which suffice for the classification task. As a future work, one may investigate this problem by designing experiments in which one studies whether the activation of the visual SOM, elicited by an acoustic stimulus (a word), allows one to reconstruct a prototypical version of the object referenced by the word. Furthermore, we intend to investigate how to cope with the uttering of whole sentences instead of single words.

## References

Chrupala, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. *ACL 2017*.

Fenoglio, G., Esposito, R., & Gliozzi, V. (2017). A neural network model for taxonomic responding with realistic visual inputs. *COGSCI 2017*, 1–6.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *ICASSP 2013*, 6645–6649.

Hannun, A. Y., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *CoRR*, *abs/1412.5567*.

Harwath, D., & Glass, J. R. (2015). Deep multimodal semantic embeddings for speech and images. *ASRU 2015*.

Harwath, D., Torralba, A., & Glass, J. R. (2016). Unsupervised learning of spoken language with visual context. *NIPS 2016*.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. (2017). Mask R-CNN. *CoRR*, *abs/1703.06870*.

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*. Weley & Sons.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, 6). Deep Supervised, but Not Unsupervised, Models May Explain

IT Cortical Representation. *PLOS Computational Biology*, *10*(11).

Kohonen, T. (2001). *Self-organizing maps*. Springer Berlin.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. *ECCV 2014*, 740–755.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological review*, *117 1*, 1–31.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*, 831–877.

Miikkulainen, R., Bednar, J., Choe, Y., & Sirosh, J. (2005). *Computational maps in the visual cortex*. Springer.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, *18*(2), 283–328.

Quine, W. V. O. (1960). Word and object. *MIT Press*.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. *CVPRW 2014*, 806–813.

Serre, T. (2016). Models of visual categorization. *Cognitive Science*, *7*(3), 197–213.

Spelke, E. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29-56.

Synnaeve, G., Versteegh, M., & Dupoux, E. (2014). Learning words from images and speech. *NIPS Workshop on Learning Semantics 2014*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *CVPRW 2016*, 2818–2826.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. *NIPS 2016*, 3630–3638.

Yurovsky, D., Fricker, D., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bullettin Review*, *21(1)*, 1–22.

Zhu, Z., Xie, L., & Yuille, A. L. (2017). Object recognition with and without objects. *IJCAI 2017*, 3609–3615.