# SIMULATED DISTRIBUTIONAL LEARNING IN DEEP BOLTZMANN MACHINES LEADS TO THE EMERGENCE OF DISCRETE CATEGORIES

Paul Boersma

University of Amsterdam
paul.boersma@uva.nl

## ABSTRACT

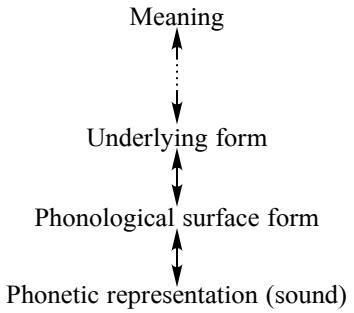
There is a potential close correspondence between multi-level linguistic theories and bidirectional deep artificial neural networks. This paper shows that in a deep Boltzmann machine, simulated distributional learning of spectral content leads to the emergence of appropriate categorical behaviour, both along a one-dimensional continuum (three sibilant places) and along a two-dimensional continuum (five vowels).

**Keywords**: neural networks, deep learning, emergence, distributional learning, categories.

## 1. BIDIRECTIONAL PARALLELLISM

Bidirectional multi-level models of phonology and phonetics [11, 3] are models that contain multiple levels of representation and in which processing works bidirectionally, i.e. the same connections or constraints are used for comprehension and production [cf. 14 for the 1-level case]. Typically, the evaluation of the best path from sound to meaning in comprehension works in parallel across levels [11, 4].

**Figure 1**: Bidirectional phonology and phonetics.



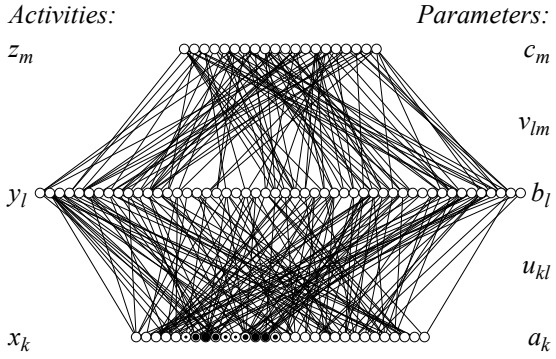

Models like the one in Fig. 1 have been shown to work nicely for many kinds of phenomena that involve existing phonological categories, but are not good at all at informing us of where those categories come from in the first place (for an attempt, see [2]).

## 2. NETWORK STRUCTURE

The present paper shows that the deep Boltzmann machine [13, 12] in Figure 2 is capable of showing how discrete phonological categories can emerge in a simulated first-language learner as a result of auditory-driven distributional learning alone, i.e. without any top-down supervision from a lexicon.

For simplicity, our network has three levels of nodes. As for the network state during processing (in our case only listening), the input nodes have *activities* $x_k$, where $k$ runs from 1 to $K = 30$, the middle level of nodes has activities $y_l$, with $l$ running from 1 to $L = 50$, and the top level has activities $z_m$, with $m$ from 1 to $M = 20$. As for the long-term memory (the *parameters*) of this network, each node $k$ at the input level is connected to each node $l$ at the middle level by a strength (*weight*) $u_{kl}$, and the middle level is fully connected to the top level by weights $v_{lm}$; also, the input nodes have *biases* $a_k$, the middle nodes $b_l$, and the top nodes have biases $c_m$.

**Figure 2**: Our deep Boltzmann network (only 10 percent of the connections is visible).




The lowest level $x_k$ represents any auditory-phonetic continuum. In this paper it reflects the basilar membrane, with low frequencies at the left and high frequencies at the right. In Fig. 2, $x_k$ is activated as for a token of the vowel /o/, with an F1 around node 8 and an F2 around node 13.5 (i.e. between nodes 13 and 14). The higher levels contain hidden nodes that are *binary*, i.e. their activity is either on or off (black or white in Fig. 2); note that none of these levels is to be equated (yet) with any of the levels in Fig. 1.

Just as the model in Fig. 1, the network in Fig. 2 is *bidirectional*, in the sense that information can flow down from $z_m$ to $x_k$ using the exact same connection weights as information flowing up from $x_k$ to $z_m$, and the intermediate biases $b_l$ influence information flowing up from $x_k$ to $y_l$ in the same way as they influence information flowing down from $z_m$ to $y_l$.

## 3. TRAINING PROCEDURE

To train the network, we apply multiple sounds to its input level, and let the network process each sound freely without supervision, i.e., we do not tell the network how many categories it should create or whether it performs "correctly". Each incoming datum is handled in four consecutive phases: initial settling, Hebbian learning, dreaming, and anti-Hebbian learning.

**3.1. The initial settling phase.** The activity spreads up from the input level, which is "clamped" (held constant) at the sound $x_k$, and down from the highest level, whose nodes $z_m$ start out with activity 0. This determines the middle-level activities $y_l$: for all $l$ from 1 to $L$,

$$y_l \leftarrow \sigma\left(b_l + \sum_{k=1}^{K} x_k u_{kl} + \sum_{m=1}^{M} v_{lm} z_m\right) \quad (1)$$

where $\sigma(\,)$ is a monotonic nonlinearity, here the standard logistic function

$$\sigma(x) \coloneqq 1/(1 + \exp(-x)) \quad (2)$$

After this initial activation of the network, the network is allowed to resonate into a near-final state [12]. First, new top-level $z_m$ are computed ($x_k$ does not change): for all $m$ from 1 to $M$,

$$z_m \leftarrow \sigma\left(c_m + \sum_{l=1}^{L} y_l v_{lm}\right) \quad (3)$$

The sequence of (1) and (3) is then repeated 10 times, taking the network close to an equilibrium state in a deterministic way (*mean-field approximation*).

**3.2. The Hebbian learning phase.** After having thus spread the influence of the input throughout the network, we can perform the first learning step, which we call the *Hebbian phase*: any connection between two active nodes is strengthened, so that these nodes will be even more often simultaneously active in the future [9, 6], and any node that is active receives a higher bias, so that this node will be even more often active in the future:

$$a_k \leftarrow a_k + \eta x_k \quad (4)$$
$$b_l \leftarrow b_l + \eta y_l \quad (5)$$
$$c_m \leftarrow c_m + \eta z_m \quad (6)$$
$$u_{kl} \leftarrow u_{kl} + \eta x_k y_l \quad (7)$$
$$v_{lm} \leftarrow v_{lm} + \eta y_l z_m \quad (8)$$

where $\eta$ is a small learning rate (0.001).

**3.3. The dreaming phase.** In the next phase we have the network dream up its own pattern [8]. From [12] we take the idea that this is randomly generated, and from [7] that this can be based on the actual input, but without clamping. Thus, we now let the input level $x_k$ be influenced by the middle level $y_l$:

$$x_k \leftarrow a_k + \sum_{l=1}^{L} u_{kl} y_l \quad (9)$$

(note that there is no logistic function for $x$, which should continue to resemble a continuous input).

We then compute new values for $z_m$ stochastically (i.e., with random variation), after which new values for $y_l$ are computed, also stochastically:

$$z_m \sim \mathcal{B}\left(\sigma\left(c_m + \sum_{l=1}^{L} y_l v_{lm}\right)\right) \quad (10)$$
$$y_l \sim \mathcal{B}\left(\sigma\left(b_l + \sum_{k=1}^{K} x_k u_{kl} + \sum_{m=1}^{M} v_{lm} z_m\right)\right) \quad (11)$$

where $\mathcal{B}(\,)$ denotes a Bernoulli deviate, i.e. $x \sim \mathcal{B}(p)$ will put the number 1 into $x$ with probability $p$, and put the number 0 into $x$ with probability $1 - p$. The sequence (9)–(11) is performed ten times (*Gibbs sampling*). The stochasticity, together with starting from real inputs, should ensure that, in the long run, the activation patterns found in the network in this phase sample the distribution of possible activation patterns faithfully.

**3.4. The anti-Hebbian learning phase.** After the network settles again, the parameters are again updated in the second, *anti-Hebbian phase* [8 ,7, 12]:

$$a_k \leftarrow a_k - \eta x_k \quad (12)$$
$$b_l \leftarrow b_l - \eta y_l \quad (13)$$
$$c_m \leftarrow c_m - \eta z_m \quad (14)$$
$$u_{kl} \leftarrow u_{kl} - \eta x_k y_l \quad (15)$$
$$v_{lm} \leftarrow v_{lm} - \eta y_l z_m \quad (16)$$

## 4. ONE-DIMENSIONAL CONTINUUM

The input continuum is the cortical representation of a basilar excitation pattern, where node 1 is the lowest basilar frequency and node 30 the highest.

We make the network listen to 100,000 pieces of data from a language with three sibilant categories whose population-average centre frequencies lie at nodes 8, 16 and 23. To generate each datum, we choose (unbeknownst to the learner) one of the three categories with equal probability (1/3), then determine a basilar centre frequency by sampling it from a Gaussian distribution with mean $\mu = 8$, 16 or 23 (depending on the category) and a standard deviation of $\sigma = 2.0$:

$$f_c \sim \mathcal{N}(\mu, \sigma) \quad (17)$$

As a result, the total distribution of centre frequencies is as given in Figure 3.

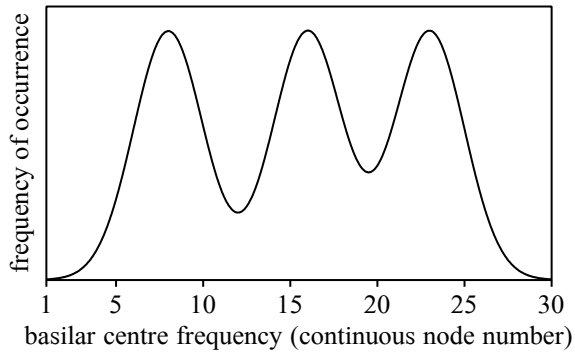
For our sampled $f_c$, the network's input activation is then the basilar excitation pattern (for $k = 1..30$)

$$x_k = 5\, e^{-\frac{1}{2}\left(\frac{k - f_c}{w}\right)^2} - 0.5 \quad (18)$$

where $w = 1.5$ is the half-width of the Gaussian peak that the sound produces on the basilar membrane (this

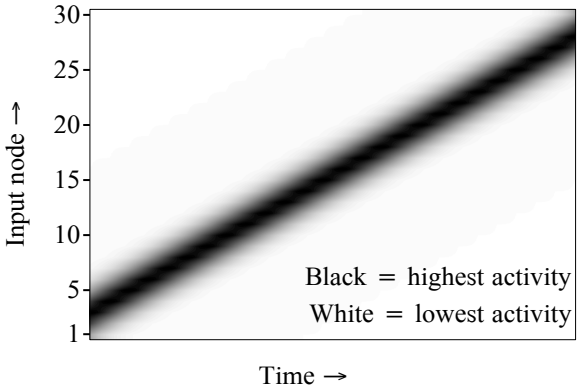
is admittedly unrealistically sharp for sibilants). Then follow phases 3.1 through 3.4, all for each incoming piece of data.

**Figure 3**: Distribution of centres of 1-peak inputs.




After having received, processed and learned from 300 pieces of data, we test (or *measure*) our network for the first time. Measuring is performed by applying the sweep in Figure 4. That is, we apply centre frequencies from 3.0 to 28.0, in 251 steps spaced 0.1 nodes apart, to the input of the network.

**Figure 4**: Applied input sweep.




With each of the 251 centre frequencies we determine the initial state of the network according to 3.1, and then resonate with open input according to (9), (3) and (1) ten times (no learning has to take place). The resulting "reflected" $x_k$ are shown in Figure 5. The network has arrived in a structure where on *any* input sound, the network reflects the same activation pattern, namely an activation pattern that mimics the pooled distribution of Fig. 3, convolved with (18).

After training for 2700 more pieces of data, the network is measured again by the sweep of Figure 4, and the network echoes the activity in Figure 6. This is **categorical behaviour**: although the network can handle a continuous range of possible inputs, the network, on *any* input, has to settle in one of only three possible states, i.e., the organism containing this network will be able to extract only three possible different messages from the incoming sound. This is what categorization is all about!

**Figure 5**: Reflection after 300 pieces of data: undertrained behaviour.

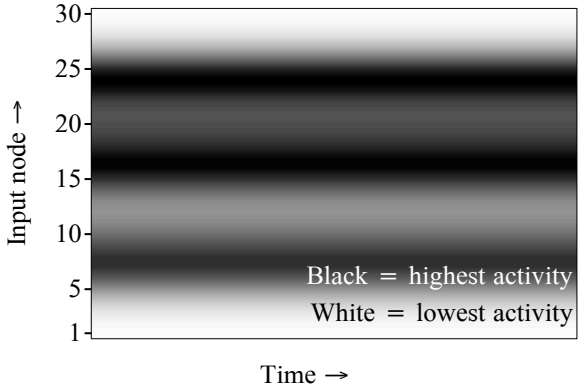



**Figure 6**: Reflection after 3,000 pieces of data: categorical behaviour.

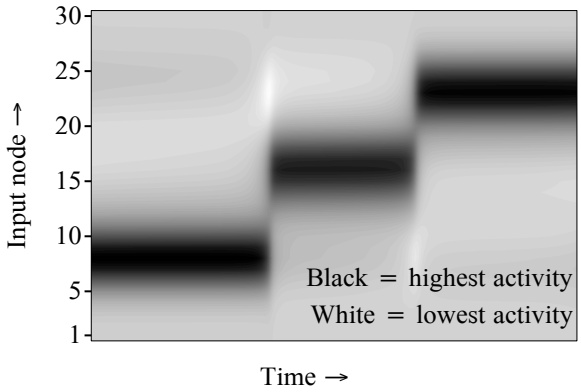
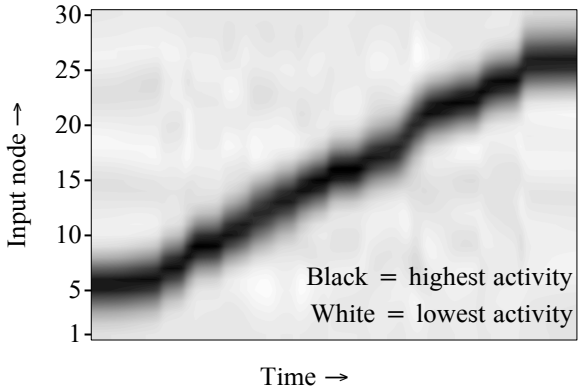



**Figure 7**: Reflection after 30,000 pieces of data: overtrained behaviour.




After training with 27,000 more pieces of data, the network has become good at echoing the input sweep, as we can see in Figure 7. The network's behaviour is no longer categorical.

The final situation is a network that returns its input nearly perfectly. The crucial stage here, however, is that of Figure 6: the change between input and reflection (*perceptual warping*) can be regarded as the perceptual magnet effect observed by Kuhl et al. [10] for human infants, which is a phenomenon widely held to be associated with phonological category creation.

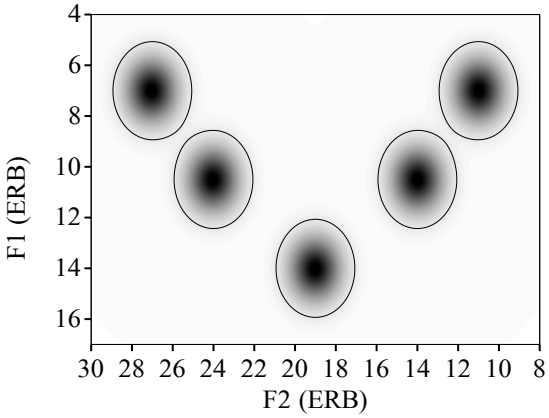
## 5. TWO-DIMENSIONAL CONTINUUM

The next more complex case is where each category is defined by *two* peaks on the basilar membrane. In a language with five vowels, we select one of the vowels /a,e,i,o,u/ (again, the network does not know which we chose, or that there are supposed to be five) randomly with equal probability (1/5), then sample an F1 and F2 value (in ERB) randomly from around the mean F1 and F2 for that vowel (with $\sigma = 0.9$ ERB). The excitation pattern on the input, where node numbers correspond to ERB values, becomes

$$x_k = 5\left(e^{-\frac{1}{2}\left(\frac{k-\mathrm{F1}}{w}\right)^2} + e^{-\frac{1}{2}\left(\frac{k-\mathrm{F2}}{w}\right)^2}\right) - 0.5 \quad (19)$$

The overall distribution of F1 and F2 values that our virtual learner is confronted with, is shown in Fig. 8.

**Figure 8**: Input distribution in a 5-vowel language. The ellipses mark a relative height of 10%. Nodes equal ERB values, e.g., node 16 represents 16 ERB.




After learning from 1,000 pieces of data, the network is measured by applying 200 randomly chosen F1–F2 combinations from the language to the network.

**Figure 9**: Perceptual magnet effect for vowels. Each arrow points from an input vowel token to its reflection.

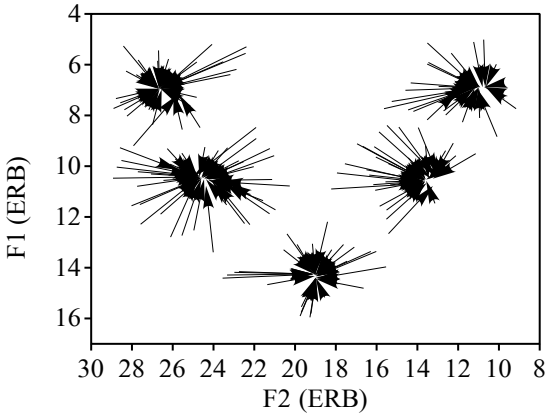



Figure 9 shows how the network reflects these 200 vowel tokens. Just as in the one-dimensional case, the network is seen to go through a stage of categorical behaviour: the arrows tend to point at only five centres of attraction. When training further, we find that the arrows get shorter and shorter, until the reflections become identical to the inputs.

## 6. DISCUSSION

The perceptual magnet effect [10] was modelled before with unidirectional neural maps by [5] and with bidirectional competitive neural networks by [1]. The present paper models it with a method borrowed from the field of machine learning, namely bidirectional deep (i.e. multi-level) restricted Boltzmann machines [12]. Both bidirectional implementations allow a connection to bidirectional multi-level theories of phonology and phonetics, but the present model does so in a computationally efficient way, because the activity of a level can be computed efficiently as there are no connections *within* levels (this is what the term *restricted* means); in this respect the model differs from the competitive network implementation [1], which requires inhibitory connections within levels. The latter model was also deemed "brittle" [1], whereas the present model may be more robust, in the sense that it works similarly when there are only two levels (i.e. it is a single restricted Boltzmann machine), or when there are more than three levels, in which case the settling scheme of repeating (1)–(3) or (9)–(11) must generalize to simultaneous activation spreading from all odd-numbered to all even-numbered levels, followed by simultaneous activation spreading from the even- to the odd-numbered levels, and this repeated several times [12].

The states of categorical behaviour that we found were only transient, with subsequent learning wiping the categories out. Whether this is a disadvantage remains to be seen: what for machine learning purposes might be a disadvantage, does not have to be so for purposes of cognitive modelling. One wonders, for instance, whether the emergence of even higher levels of representation, such as the lexicon, will be able to maintain the network's categoricality.

## 7. CONCLUSION

We hope that future computer simulations with this type of networks can not only account for category emergence but also for the phenomena handled by earlier models with fixed categories such as [11, 3, 4]. If that succeeds, we will have taken crucial steps toward providing a comprehensive theory of language comprehension, production, acquisition and change. We are looking forward to including multiple segments in time and to seeing multiple phonological features emerge both bottom-up from the phonetics and top-down from alternations and the lexicon.

## 8. REFERENCES


[1] Boersma, P., Benders, T., Seinhorst, K. 2018. Neural networks for phonology and phonetics. Manuscript, University of Amsterdam.

[2] Boersma, P., Escudero, P., Hayes, R. 2003. Learning abstract phonological from auditory phonetic categories: an integrated model for the acquisition of language-specific sound categories. *Proc. 15th ICPhS* Barcelona, 1013–1056.

[3] Boersma, P., Hamann, S. 2009. Loanword adaptation as first-language phonological perception. In: Calabrese, A., Wetzels, W.L. (eds.), *Loanword Phonology*. Amsterdam: John Benjamins, 11–58.

[4] Boersma, P., Van Leussen, J.-W. 2017. Efficient evaluation and learning in multi-level parallel constraint grammars. *Linguistic Inquiry* 48, 349–388.

[5] Guenther, F.H., Gjaja, M.N. 1996. The perceptual magnet effect as an emergent property of neural map formation. *J. Acoust. Soc. Am.* 100. 1111–1121.

[6] Hebb, D.O. 1949. *The Organization of Behavior*. New York: Wiley.

[7] Hinton, G. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation* 14, 1711–1800.

[8] Hinton, G., Sejnowski, T.J. 1983. Optimal perceptual inference. *Proc. IEEE Conference on Computer Vision and Pattern Recognition* Washington, 448–453.

[9] James, W. 1890. *The Principles of Psychology.* Dover Publications.

[10] Kuhl, P.K. 1991. Human adults and human infants show a "perceptual magnetic effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50, 93–107.

[11] McClelland, J.L., Elman, J.L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18, 1–86.

[12] Salakhutdinov, R.R., Hinton, G.E. 2009. Deep Boltzmann machines. *Proc. International Conference on Artificial Intelligence and Statistics*.

[13] Smolensky, P. 1986. Information processing in dynamical systems: Foundations of harmony theory. In: Rumelhart, D.E., McClelland, J.L. (eds.), *Parallel Distributed Processing*, volume 1. Cambridge MA: MIT Press, 194–281.

[14] Smolensky, P. 1996. On the comprehension/ production dilemma in child language. *Linguistic Inquiry* 27, 720–731.