Unifying digital phonology with an analogue brain

Joe Collins (UiT)

One frequently encountered criticism of phonological formalisms is that their rigidly discrete nature is fundamentally incompatible with the standard view of the brain as an analogue system, as well as the gradience observed in the phonetic realisation of phonological processes (Port and Leary 2005). This talk will demonstrate that this perceived incompatibility is incorrect. Discrete phonological formalisms are compatible with certain classes of gradient systems which exhibit a high degree of non-ergodicity. The information encoded in such a system will always be highly unevenly distributed, meaning that the system can be characterized at a macro-level abstraction by discrete symbols. An argument that the brain is such a system is presented in the form of a neural network *attractor* model, which demonstrates the emergence of discrete categories from an underlyingly gradient system. It will also be shown how this model can account for incomplete devoicing, which has been argued to be an example of phonetic gradience that discrete phonological models cannot explain (c.f. van Oostendorp 2008). Finally, it will be shown that the formal phonological analysis of devoicing has a higher *Effective Information* (EI), in the sense of Hoel (2017), than the neural model. Thereby demonstrating that not only are phonological formalisms compatible with a gradient view of the brain, but that they are *causally emergent* (*ibid.*), and therefore necessary if we wish to have a complete explanation of natural language grammar.

An attractor network can be defined as a type of dynamical system whose behaviour will always tend asymptotically towards one of a smaller set of networks states, which are referred to as attractors (Hopfield 1982). Their basic structure is typical of most neural networks: a number of identical units connected by "synapses" of varying efficacy. For this reason, they can be employed as an effective model of neural dynamics. One relevant property of attractor networks is that they can function as a way of storing discrete, content addressable memories, which makes them capable of representing the kinds of discrete symbols employed in formal phonological theories.

This fact is demonstrated by the implementation of a toy phonological grammar consisting of 6 possible phones – 3 places of articulation, each with a voiced and voiceless variant – and the capacity to distinguish coda and non-coda positions. The 6 phones are encoded as attractor states in the network, without any process of supervised learning, while information about syllable structure is supplied to the network as a simple inhibitory signal to the network – intended to approximate slow speed neural oscillations (c.f. Ding et al. 2016). What will be demonstrated is that the network can self-organize to voiceless memory, in the case when it is supplied with a voiced input in the context of the inhibitory coda signal. Crucially, those voiceless outputs which are derived from a voiced input, can vary fractionally from those voiceless outputs which are underlyingly voiceless. This small variation is easily interpretable as a small, but consistent, difference in the VOT of the phone during realization. In this way, this simple model is a proof of concept for how a discrete phonological system, when implemented in an underlyingly continuous system, can exhibit the sorts of gradience observed in phenomena such as incomplete devoicing.

Moreover, what can be demonstrated using Hoel's measure of EI, is that the formal, macro-level characterization of the system is more informative regarding causal structure, than the micro-level characterization of the attractor network.

At the macro-level, the toy grammar can be understood as a system having *n=12* possible states *S*={[b]#, [d]#, [g]#, [b], [d], [g], [p]#, [t]#, [k]#, [p], [t], [k]}. The dynamics of the system can be understood as an intervention over each state s_i, at time=*t,* and a resulting effect at time=*t+1.* We can then determine two probability distributions, *Intervension Distribution* (I_D) and *Effect Distribution* (E_D), which can then be used to calculate the *effectiveness* of the system (normalized EI). Following Hoel, the I_D is considered in the maximum entropy case, where $I_D(i)=n^{-1}$. and the E_D is calculated by observing the effects of the interventions at time=*t+1* (see table). These values can then be used to determine the *degeneracy* of the system: $degeneracy = \frac{D_{KL}(E_D|I_D)}{log_2(n)} = log_n(2)\sum_i E_D(i) log_2 \frac{E_D(i)}{I_D(i)}$

I_D *at time=t*	*t+1*	E_D
<do(b#)>=$\frac{1}{12}$	[p]#	<b#>=0
<do(d#)>=$\frac{1}{12}$	[t]#	<d#>=0
<do(g#)>=$\frac{1}{12}$	[k]#	<g#>=0
<do(p#)>=$\frac{1}{12}$	[p]#	<p#>=$\frac{2}{12}$
<do(t#)>=$\frac{1}{12}$	[t]#	<t#>=$\frac{2}{12}$
<do(k#)>=$\frac{1}{12}$	[k]#	<k#>=$\frac{2}{12}$
<do(b)>=$\frac{1}{12}$	[b]	<b>=$\frac{1}{12}$
<do(d)>=$\frac{1}{12}$	[d]	<d>=$\frac{1}{12}$
<do(g)>=$\frac{1}{12}$	[g]	<g>=$\frac{1}{12}$
<do(p)>=$\frac{1}{12}$	[p]	<p>=$\frac{1}{12}$
<do(t)>=$\frac{1}{12}$	[t]	<t>=$\frac{1}{12}$
<do(k)>=$\frac{1}{12}$	[k]	<k>=$\frac{1}{12}$

Which will then allow us to calculate the $effectiveness = [determinism] - degeneracy$. Since our toy grammar is strictly deterministic, the *determinism* is equal to 1. Crunching the numbers gives our toy grammar $eff = \sim 0.93$

Now we can turn to the case of our attractor model. Because the storage capacity of an attractor network can be well defined, relative to the size of the network (i.e. the number units), we can quantify exactly the total number of states in the system. This allows us, in principle, to calculate the lower bound for *degeneracy* in an attractor implementation of our toy grammar. In practice, the state-space of our attractor is so large that it precludes a brute force computation over each and every state. A trinary node implementation of 12 attractors would require 50 units (Ziong and Zhao 2010), giving a system size of $n=3^{50}$. However, demonstrating the high *degeneracy* can be accomplished by averaging over the system. Since each state which is not an attractor will ultimately lead to an attractor, each of the 12 attractor states can be reached by an average of $\frac{3^{50}}{12}$ states, the average value of $\frac{E_D(i)}{I_D(i)} \cong \frac{12^{-1}}{3^{-50}}$, which will cause the *degeneracy* to tend heavily towards 1, resulting in a very low *effectiveness*. Therefore, even when our discrete phonological representations are taken as emergent phonemena from an underlyingly gradient system, such as an attractor network, it is in fact the phonological model which has the highest *effectiveness*, rather than the neurological model.

Finally, it can also be demonstrated that, if the micro-level characterization is a discrete system of the sort repudiated by Port & Leary, then *causal emergence* does not occur. This suggests that formal phonological models are, counter intuitively, **more** valuable in the case where discrete symbols are emergent rather than primitive (pace Port & Leary).

Ding, R., Melloni, L., Zhang, H., Tian, X., Poeppel, D. (2016) Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience* 19, 158–164

Hoel, E. P. (2017) When the Map Is Better Than the Territory. *Entropy* 19:188.

van Oostendorp, M. (2008) Incomplete devoicing in formal phonology, In Lingua, Volume 118, Issue 9, Pages 1362-1374

Port and Leary (2005) Against formal phonology. *Language*, 81 pp. 927-964

Xiong, D., Zhao, H. (2010) Estimates of storage capacity in the q-state Potts-glass neural network. *J. Phys. A: Math. Theor*. 43 445001