ITERATED DISTRIBUTIONAL AND LEXICON-DRIVEN LEARNING IN A SYMMETRIC NEURAL NETWORK EXPLAINS THE EMERGENCE OF FEATURES AND DISPERSION

Klaas Seinhorst, Paul Boersma and Silke Hamann

University of Amsterdam
{seinhorst,paul.boersma,silke.hamann}@uva.nl

ABSTRACT

We present a neural network model of phonetic and phonological acquisition that can handle two distinct phenomena: category creation and auditory dispersion. Within a single neural network, learning proceeds in two stages. The first stage is distributional learning, during which the model induces phonological features from an auditory input distribution; in the second stage, the model acquires knowledge about the relation between lexical categories and the auditory input distribution. The model can be used bidirectionally: once perceptual learning is complete, the network can also be asked to speak. In the production direction, effortful, perceptually peripheral tokens are avoided. In a chain of iterated learners, in which the output of one generation serves as the input to the next, sound systems emerge that maintain sufficient contrast at a moderate articulatory cost, regardless of the initial distribution.

Keywords: neural networks; phonological features; auditory dispersion; sound change.

1. INTRODUCTION

The speaker-listener is often assumed to be subject to two competing forces: on one hand, the pressure to be clear; on the other hand, the pressure to be lazy [2, 20, 27]. These pressures oppose each other, since a larger auditory contrast entails increased articulatory effort as well. The two forces can be seen at work both at the level of the individual speaker-listener and at the level of the linguistic system: e.g. speakers reduce contrasts in casual speech [9, 11]; and in sound systems, the auditory correlates of phonological categories are distributed in a way that maintains sufficient perceptual distinctiveness between categories [18, 19].

1.1 Auditory dispersion

The maintenance of sufficient contrast in sound systems is sometimes regarded as a synchronically functionalist process [14, 17, 25]; Boersma & Hamann [7] (hereafter B&H), on the other hand, model it as an emergent phenomenon. In their formalization, the learner first goes through a stage of lexicon-driven perceptual learning: she acquires a grammar expressing knowledge about the relation between auditory cues and phonological categories. Here a prototype effect occurs, i.e. the learner prefers a token in perception that is less confusable and more peripheral than the token that was most frequent in her input. B&H's model is bidirectional, so the learner uses this same knowledge in production as well; however, as a speaker she tries to avoid auditorily peripheral tokens, since these require more articulatory effort. Using simulations of iterated learning chains, B&H show that in a stable inventory, the articulatory and prototype effects cancel each other out; when the contrast in the initial distribution is exaggerated, the articulatory effect will push the categories towards more central values; when the initial contrast is confusing, the prototype effect will push the categories outwards. In all cases, a stable, optimally dispersed inventory emerges within a number of generations, without any goal-oriented elements in the model.

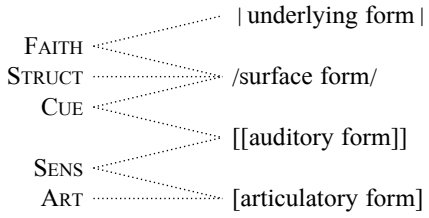
1.2 Category creation

Most mainstream phonological frameworks presuppose the existence of discrete categories. Since we want to model category emergence, we work with neural networks, which have already been used to model feature discovery [1, 6, 10, 29]. A neural network consists of layers of nodes, connected to each other with connections that can be either excitatory or inhibitory. Nodes can be activated, and activity can flow through the network: when a node is active, it activates those nodes that it is connected to with excitatory connections (i.e. connections with positive weights), and inhibits those nodes that it is connected to with inhibitory connections (i.e. connections with negative weights).

2. THE MODEL

We present a neural network model that is capable of category creation, in which optimal auditory dispersion emerges as well. Our formalization is couched in Boersma's bidirectional model of phonology and phonetics (BiPhon) [5], shown below.

Figure 1: The architecture of the BiPhon model.



The BiPhon model assumes at least two phonological and two phonetic levels of representation. The phonological levels are an underlying morphemic level (UF), and a prosodically structured surface form (SF) [28]; in addition, there are an auditory-phonetic representation (AudF), representing auditory cues like formants and plosive release bursts, and an articulatory-phonetic form (ArtF) containing a plan of articulatory gestures. Faithfulness knowledge evaluates the mapping between UF and SF [21]; structural restrictions evaluate the phonotactic wellformedness of SF [28]; cue knowledge expresses the relation between phonological categories and auditory cues [4, 12]; sensorimotor knowledge evaluates the relation between auditory cues and their articulatory implementation [3]; and articulatory constraints evaluate articulatory effort [2, 17].

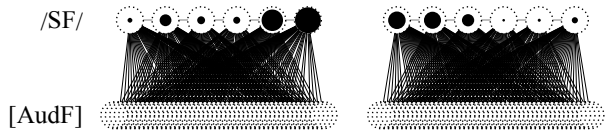
In the neural network version of the BiPhon model, every level of representation is implemented as a layer of nodes [6]. In this paper, we model two independent auditory continua (e.g. centre of gravity in sibilants, and VOT), resulting in two separate AudF layers and two SF layers. We assume perfect sensorimotor knowledge; instead, the AudF and ArtF layers are connected with connections whose weights are more negative at the edges of the continuum, inhibiting activities there more strongly.

The neural network learns in two stages: a distributional learning stage followed by a lexicon-driven learning stage (§3). Once learning is complete, the network will produce an output that may serve as the input to a new network (§4), i.e. we create a chain of iterated learners [16]. We will show that all three types of learning play a role in the emergence of features as well as auditory dispersion.

3. EMERGENT FEATURES

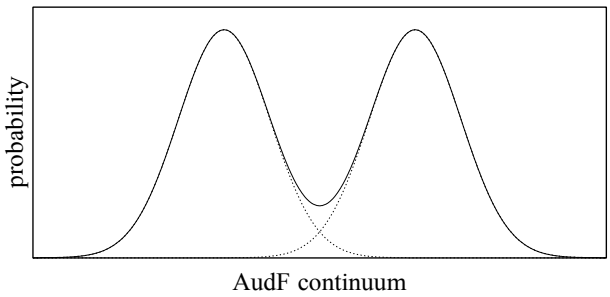
Fig. 2 shows a neural network in its initial state. Per auditory continuum, there are 48 AudF nodes and 6 SF nodes, connected with excitatory cue connections (drawn in black). Within each SF layer, all nodes are connected with inhibitory connections [29] (drawn in grey). Before learning begins, the activities of the SF nodes and the weights of the cue connections are small and random. A larger black circle drawn inside a node indicates stronger activation.

Figure 2: A neural network before learning.



All simulations are run with a script in Praat [8]. The first step in the simulation is to determine the initial language, based on parameters set in the script: the user indicates where the category peaks in the auditory environment of the first generation lie. The probabilities of the input tokens are normally distributed around the category peaks, as in Fig. 3.

Figure 3: The auditory environment of the learner.



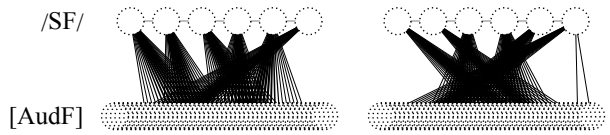
Since the continuum on the horizontal axis is divided into 48 nodes, the probability of each individual node at AudF is computed. In Fig. 3, there are two categories on the continuum, one with a peak at 35% of the continuum, one with a peak at 65%. The dotted curves in the figure show the distributions of both categories, the solid curve shows the sum of these categories. Learning proceeds in two stages.

Stage 1: distributional learning. We assume that there is no lexicon in place at this stage yet: the network only learns from the auditory environment, without having acquired any category labels yet. A learning step in the first stage proceeds as follows. From the cumulative distribution, i.e. the solid curve in Fig. 3, an auditory value is drawn, based on its probability. A bit of transmission noise [7, 24], whose amount is normally distributed around mean 0, is added to the token; a normally distributed activation pattern, centered around the token, is applied to the AudF layer. This last step is motivated by the fact that the biological correlate of the AudF layer is the basilar membrane, and incoming sounds on the basilar membrane activate adjacent hair cells as well [23]. Subsequently, the activities from the AudF layers are spread to the corresponding SF layers in 100 time steps. The inhibitory connections within each SF layer cause a node that becomes activated to simultaneously deactivate the other nodes in the layer; as a result, nodes become specialized in parts of the auditory continuum ("competitive learning").

Once activity spreading is done, the weights of the cue connections are updated according to the 'inoutstar' algorithm [6], a bidirectional algorithm that combines properties of the instar and outstar learning rules [15, 29].

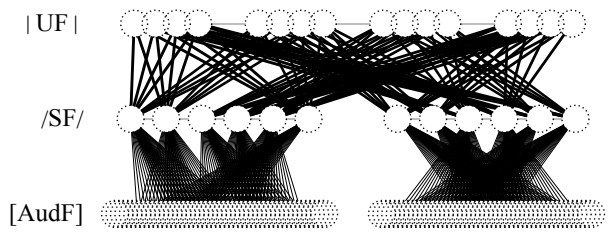
Fig. 4 shows the network from Fig. 2 after 8,000 learning steps. On the leftmost continuum, SF nodes 1, 2 and 6 have become specialized in the left side of the continuum; nodes 3, 4 and 5 are specialized in the right side. On the rightmost continuum, nodes 3, 5 and 6 have become connected to the left side of the continuum; and nodes 1, 2 and 4 to the right side. This means that when we spread activity from an AudF node to SF on the left continuum, two patterns are possible: either nodes 1, 2, and 6 are on while 3, 4 and 5 are off, or 3, 4 and 5 are on while 1, 2 and 6 are off. In other words: binary features have emerged at SF. The same is true for the other continuum. This categorical behaviour is directly observable in our model, while it needs to be inferred in other computational models of distributional learning [13, 22, 26].

Figure 4: The network after 8,000 learning steps.



Stage 2: lexicon-driven learning. After the 8,000th learning step, lexicon-driven learning begins. The network is extended with an UF layer, which has four nodes per lexical category and is connected to both SF layers with excitatory connections. In this second stage of learning, the input to the network consists of AudF-UF pairs: the auditory tokens are drawn from the dotted distributions in Fig. 3, i.e. the distributions with lexical category labels, and the corresponding category nodes at UF are switched on as well. Since the activity between AudF and UF necessarily flows through SF, the newly emerged categories mediate this process. Once the activity spreading is done, the weights of the faithfulness and cue connections are updated with the same algorithm from the previous learning stage. Fig. 5 shows the network from Figs. 2 and 4 after 16,000 tokens.

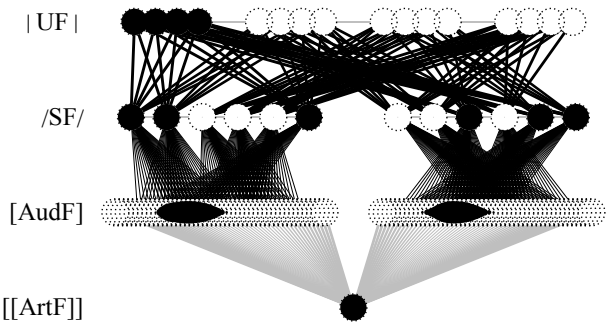
Figure 5: The network after 16,000 learning steps.



4. EMERGENT DISPERSION

The bidirectionality of the BiPhon model is ensured because the neural network is symmetric, i.e. the connection from node A to node B has the same weight as that from node B to node A. Therefore, the network can be used in the production direction as well. In the production direction, we add an ArtF node, which is connected to both AudF continua with inhibitory connections, whose weights are more negative at the edges of the continua. The production of a lexical category entails the activation of its UF nodes, then spreading activity down through SF to AudF. ArtF needs to be activated as well, spreading activity up to AudF, inhibiting the edges of both continua more strongly, cf. Fig. 6. The resulting activities at the AudF layer can be interpreted as probabilities. We can sample this probability distribution to get inputs for training a new network, then use the production of that network to train a third, and so on: that is, we can create a chain of iterated learners, and track the evolution of an inventory across multiple generations.

Figure 6: The network producing a lexical category.



We explore the evolution of two types of initial distribution, both with four categories: a "standard" inventory, whose peaks lie at 35% and 65% of both continua, and a skewed, exaggerated inventory.

Fig. 7 is a 48×48 grid, showing the initial probability distribution of all combinations of AudF nodes on both continua in the standard inventory.

Figure 7: A standard contrast.

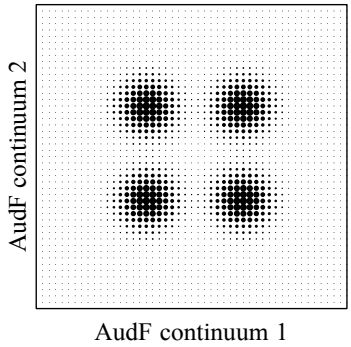


Fig. 8 shows the evolution of this inventory over 10 generations, averaged over 5 runs. Black curves indicate the average input nodes; grey curves indicate the average ± 1 sd. Fig. 8 shows that the standard inventory remains stable over the generations.

Figure 8: The evolution of a standard contrast.

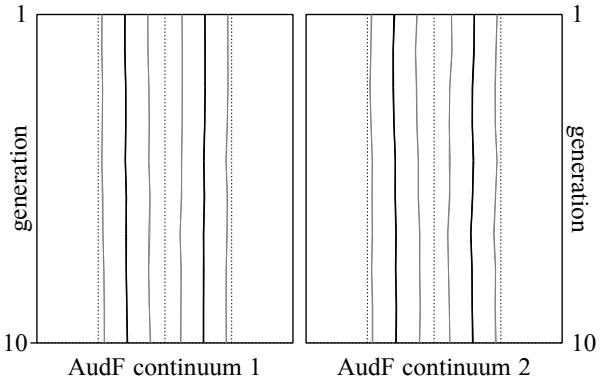


Fig. 9 shows the initial probability distributions of a skewed inventory in a 48×48 grid; the evolution of this inventory over 40 generations, averaged over 5 runs, can be seen in Fig. 10.

Figure 9: A skewed, exaggerated contrast.

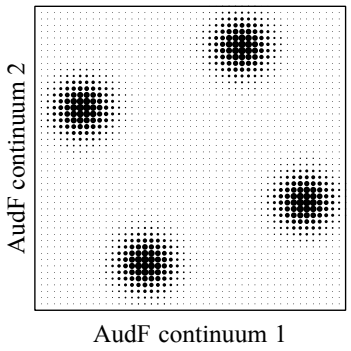
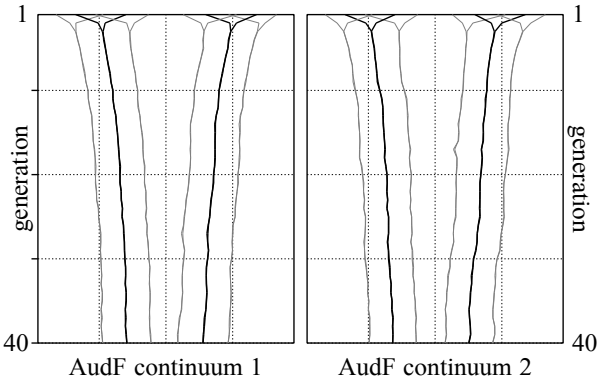


Figure 10: The evolution of a skewed, exaggerated contrast.



Two phenomena can be seen in Fig. 10: **merger** and **dispersion**. The merger proceeds in two steps. Firstly, even though none of the categories shared a peak in its distributions with any other categories initially, the peaks on both continua are sufficiently close to one another that the network induces identical representations at SF during distributional learning, already within the first generation. Subsequently, the resulting bimodal distributions become monomodal (seen in Fig. 10 as the convergence of the black curves), as the inhibitory articulatory connections reduce the size of the most peripheral peak and push it towards the center. Dispersion emerges because the articulatory effect pushes the auditory distributions towards the centre of both continua, after which we end up with the optimally dispersed standard inventory familiar from Fig. 8.

All three types of learning play a role in the evolution of this inventory. The induction of a single feature value from two non-identical auditory distributions can only happen if the lexicon is not yet involved, i.e. during the distributional learning stage. Once the two auditory distributions share the same feature value, both corresponding lexical categories at UF will become connected to SF in the exact same way during the lexicon-driven learning stage, which entails that these categories will also have identical output probability distributions at AudF in production. The iterated learning process is crucial because the initial contrast is too large to be resolved in a single generation; every new generation adds a small articulatory effect.

5. CONCLUSION

We presented a neural network model that involves three types of learning: firstly, a neural network induces features from an auditory input distribution (distributional learning); secondly, it acquires knowledge of the relation between lexical categories and the auditory input distribution (lexicon-driven learning); and thirdly, its speech output is used as the input to a new network (iterated learning). The first two types of learning happen within a single neural network and use the same learning algorithm. The three types of learning account for two distinct phenomena: the emergence of phonological features within every generation, and the emergence of auditory dispersion over multiple generations.

In other computational models of distributional learning [13, 22, 26], categorical behaviour emerged as well, sometimes aided by lexical information [13], but we are not aware of other models that handle both perception and production. Also, to our knowledge, the two different timescales on which features and dispersion emerge on multiple continua have not been unified in a single model before. Future extensions of the model will involve, among other things, the roles of morphological alternations [10] and phonetic context.

6. REFERENCES

[1] Benders, T. 2013. *Nature's distributional-learning experiment*. PhD dissertation, University of Amsterdam.

[2] Boersma, P. 1998. *Functional Phonology: formalizing the interactions of articulatory and perceptual drives*. PhD dissertation, University of Amsterdam.

[3] Boersma, P. 2006. Prototypicality judgments as inverted perception. In: Fanselow, G., Féry, C., Vogel, R., Schlesewsky, M. (eds.), *Gradience in grammar: generative perspectives*. Oxford: Oxford University Press, 167–184.

[4] Boersma, P. 2009. Cue constraints and their interactions in phonological perception and production. In: Boersma, P., Hamann, S. (eds.), *Phonology in perception*. Berlin: Mouton de Gruyter, 55–110.

[5] Boersma, P. 2011. A programme for bidirectional phonology and phonetics and their acquisition and evolution. In: Benz, A., Mattausch, J. (eds.), *Bidirectional Optimality Theory*. Amsterdam: John Benjamins, 33–72.

[6] Boersma, P., Benders, T., Seinhorst, K. 2018. *Neural network models for phonology and phonetics*. Unpublished manuscript, University of Amsterdam.

[7] Boersma, P., Hamann, S. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25, 217–270.

[8] Boersma, P., Weenink, D. *Praat*. Computer program, version 6.0.43. Downloaded from www.praat.org on November 1st, 2018.

[9] Bolinger, D. 1963. Length, vowel, juncture. *Bilingual Review* 3 (1), 43–61.

[10] Chládková, K. 2014. *Finding phonological features in perception*. PhD dissertation, University of Amsterdam.

[11] Ernestus, M. 2000. *Voice assimilation and segment reduction in casual Dutch: a corpus-based study of the phonology-phonetics interface*. PhD dissertation, Free University Amsterdam.

[12] Escudero, P., Boersma, P. 2004. Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition* 26, 551–585.

[13] Feldman, N., Griffiths, T., Morgan, J. 2009. Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2208–2213.

[14] Flemming, E. 1995/2002. *Auditory representations in phonology*. PhD dissertation, University of California, Los Angeles. Published by Routledge (London & New York).

[15] Grossberg, S. 1969. Embedding fields: a theory of learning with physiological implications. *Journal of Mathematical Psychology* 6, 209–239.

[16] Kirby, S. 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation* 5, 102–110.

[17] Kirchner, R. 1998/2001. *An effort-based approach to consonant lenition*. PhD dissertation, University of California, Los Angeles. Rutgers Optimality Archive 276. Published by Routledge (London & New York).

[18] Liljencrants, J., Lindblom, B. 1972. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* 48, 839–862.

[19] Lindblom, B. 1986. Phonetic universals in vowel systems. In: Ohala, J., Jaeger, J. (eds.), *Experimental Phonology*. Orlando: Academic Press, 13–44.

[20] Martinet, A. 1955. *Économie des changements phonétiques: traité de phonologie diachronique*. Berne: Francke.

[21] McCarthy, J. Prince, A. 1995. Faithfulness and reduplicative identity. In: Beckman, J., Walsh, L. Dickey, Urbanczyk, S. (eds.), *Papers in Optimality Theory*. University of Massachusetts Occasional Papers 18. Amherst, MA: Graduate Linguistic Student Association, 249–384.

[22] McMurray, B., Aslin, R., Toscano J. 2009. Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science* 12, 369–378.

[23] Moore, B., Glasberg B. 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am*. 74 (3), 750–753.

[24] Ohala, J. 1981. The listener as a source of sound change. In: Masek, C., Hendrick, R., Miller, M.F. (eds.), *Papers from the parasession on language and behavior*. Chicago: Chicago Linguistic Society, 178–203.

[25] Padgett, J. 2003. Contrast and post-velar fronting in Russian. *Natural Language & Linguistic Theory* 21, 39–87.

[26] Pajak, B., Bicknell, K., Levy, R. 2013. A model of generalization in distributional learning of phonetic categories. In: Demberg, V., Levy R. (eds.), *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics*. Sofia, Bulgaria: Association for Computational Linguistics, 11–20.

[27] Passy, P. 1890. *Étude sur les changements phonétiques et leur caractères généraux*. Paris: Firmin-Didot.

[28] Prince, A., Smolensky, P. 1993/2004. *Optimality Theory: constraint interaction in generative grammar*. Ms, Rutgers University & University of Colorado, Boulder. Published by Blackwell (Malden, MA & Oxford).

[29] Rumelhart, D., Zipser, D. 1985. Feature discovery by competitive learning. *Cognitive Science* 9, 75–112.