Computer modeling suggests patterns of perceptual availability of phonological structure during infant language acquisition

Cory Shain and Micha Elsner, Ohio State

Distinctive features like [±voice] and [±sonorant] have been a core construct of phonological theory for many decades [44, 25, 7, 8], and psycholinguistic evidence suggests that they are cognitively available to both adults [6] and infants [27, 22, 47]. Nonetheless, distinctive features are not directly observed by humans; they are abstractions that must be inferred from dense perceptual information (sound waves) during language acquisition and comprehension, which raises questions about how they are learned and recognized. Recent work on child language acquisition has stressed the importance of top-down (e.g. lexical and phonotactic) information for acquiring phonemic categories [34, 42, 17, 28, 32, 18, 12]. But to prevent the acquisition process from being circular, the acoustic signal must also provide evidence for phonemic categories. Furthermore, top-down guidance is likely less reliable to young infants, who must therefore rely more heavily on bottom-up perceptual information. To a learner faced with the immense challenge of discovering structure in dense perceptual input, do theory-driven phonological features "stand out" or are they swamped by noise?

We address this question through computational acquisition modeling, which permits fine-grained analysis of the learned representations that is not possible to obtain from human infants. Our acquisition model takes as a starting point cognitive evidence that brains actively model their perceptual world [16, 41, 49], that autoassociation characterizes the behavior of many brain regions [43, 39], that language comprehension and production might be linked through a sensorimotor loop [24, 15, 46, 48, 38, 26, 4], that limited auditory memory requires austere compression of dense acoustic percepts during real-time language comprehension [3, 2, 14], that featural decomposition of phone segments occurs during the acquisition process [27, 22, 47], and that there are broad tendencies toward categorical perception in human cognition [20], including that of infants [13]. The computational learners used in this study have all of these characteristics: they are deep neural autoencoders (percept modeling, autoassociation, sensorimotor loop) that force acoustic information from phone segments through a tight 8-dimensional representational bottleneck (compression) consisting of discrete binary stochastic neurons or *BSNs* (feature decomposition, categorical perception). Our learners thus have a representational capacity of 256 discrete phone categories, decomposable along 8 binary feature dimensions, with which to describe their variegated perceptual world. This setup allows us to evaluate degrees of correspondence between these perceptually-driven unsupervised representations and theory-driven phonological representations.

We deploy these models to answer two questions about the data available to young learners whose training signal must primarily be extracted from bottom-up perceptual information: (1) to what extent can phoneme categories emerge from a drive to model auditory percepts, and (2) how perceptually available are theory-driven phonological features? We apply our models to naturally-occurring acoustic phone segments from two typologically unrelated languages: the Xitsonga [10] and English [35] corpora from the Zerospeech 2015 shared task [45]. Unsupervised phone classification metrics *homogeneity* (H), *completeness* (C), and *V-measure* (V) [40] are given in Tables 1a & 1b. As shown, much phonemic structure is perceptually available from acoustics alone (20-40x clustering improvement over a random baseline). We further analyze the recoverability of theory-driven phonological features from the learners' latent bit patterns, using random forest classifiers [33] to fit propositional logical statements that map from latent bits to binary featurizations of the true segment labels [21, 19]. Precision (P), recall (R), and F-scores (F) are given in Tables 1c & 1d. Patterns of feature availability are remarkably consistent across languages, suggesting that the models are capturing generalized perceptual patterns. Furthermore, there are strong asymmetries in perceptual availability, with good recovery of voicing features and features that distinguish prototypical consonants from prototypical vowels, along with comparatively poorer recovery of e.g. certain place and manner distinctions. These findings align with attested patterns of infant phone discrimination [1, 11, 31, 36, 30, 5, 29, 37, 9, 23].

Our results show (1) that phonemic structures emerge naturally but imperfectly from perceptual reconstruction and (2) that theory-driven features differ in degree of perceptual availability. Together, these findings suggest that reliable cues to phonemic structure are immediately available to infants from bottom-up perceptual characteristics alone, but that these cues must eventually be supplemented by top-down lexical and phonotactic information to achieve adult-like phone discrimination. Our results also suggest fine-grained differences in degree of perceptual availability between features, yielding testable predictions as to which features might depend more or less heavily on top-down cues during child language acquisition.

Model	H	C	V
Random Baseline	0.023	0.013	0.016
BSN Autoencoder	**0.462**	**0.268**	**0.33**

(a) Xitsonga clustering

Model	H	C	V
Random Baseline	0.006	0.004	0.005
BSN Autoencoder	**0.270**	**0.180**	**0.216**

(b) English clustering

Feature	P	R	F
voice	0.9767	0.9033	0.9386
sonorant	0.9249	0.9085	0.9166
continuant	0.9492	0.7936	0.8645
consonantal	0.8314	0.8915	0.8604
approximant	0.8998	0.8192	0.8576
syllabic	0.8278	0.8523	0.8398
dorsal	0.8935	0.7703	0.8273
strident	0.6991	0.9594	0.8089
low	0.7175	0.8978	0.7976
front	0.6590	0.8101	0.7268
high	0.5875	0.7882	0.6732
back	0.5352	0.8527	0.6577
round	0.5332	0.8551	0.6568
labial	0.5669	0.7725	0.6539
coronal	0.5382	0.8301	0.6530
tense	0.5208	0.8115	0.6344
delayed release	0.5468	0.7226	0.6225
anterior	0.4078	0.8355	0.5481
nasal	0.3635	0.8796	0.5144
distributed	0.2459	0.8537	0.3819
constricted glottis	0.1762	0.9007	0.2948
lateral	0.1536	0.8062	0.2581
labiodental	0.0934	0.7980	0.1672
trill	0.0809	0.7401	0.1458
spread glottis	0.0671	0.5856	0.1204
implosive	0.0041	0.4041	0.0081

(c) Xitsonga feature recovery

Feature	P	R	F
voice	0.9244	0.8567	0.8893
sonorant	0.8544	0.8862	0.8700
approximant	0.8005	0.8370	0.8183
continuant	0.8577	0.7669	0.8098
consonantal	0.8249	0.7357	0.7777
syllabic	0.6624	0.8426	0.7417
dorsal	0.7046	0.7114	0.7080
strident	0.5505	0.9027	0.6839
coronal	0.5758	0.7066	0.6345
anterior	0.5251	0.7280	0.6101
delayed release	0.4413	0.7374	0.5521
front	0.4322	0.7407	0.5459
high	0.3841	0.6931	0.4943
tense	0.3275	0.7101	0.4483
back	0.3128	0.7504	0.4416
nasal	0.2796	0.7544	0.4080
labial	0.2541	0.7077	0.3739
low	0.2410	0.7787	0.3680
distributed	0.2203	0.6881	0.3337
stress	0.2052	0.8027	0.3269
diphthong	0.2039	0.8051	0.3254
round	0.1665	0.7012	0.2692
lateral	0.1484	0.8333	0.2519
labiodental	0.0787	0.6756	0.1410
spread glottis	0.0377	0.6683	0.0714

(d) English feature recovery

References

[1] Richard N Aslin, David B Pisoni, Beth L Hennessy, and Alan J Perey. *Child development*, 1981.
[2] Alan Baddeley, Susan Gathercole, and Costanza Papagno. *Psychological Review*, 1 1998.
[3] Alan D Baddeley and Graham Hitch. *Working Memory*, 1974.
[4] Johan J Bolhuis, Kazuo Okanoya, and Constance Scharff. *Nature Reviews Neuroscience*, 2010.
[5] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, et al. In *Proc. WMT*, 2007.
[6] Katerina Chládková, Paul Boersma, Titia Benders, and others. In *ICPhS*, 2015.
[7] Noam Chomsky and Morris Halle. *The Sound Pattern of English*, 1968.
[8] George N Clements. *Phonology*, 1985.
[9] Alejandrina Cristià, Grant L McGuire, Amanda Seidl, and Alexander L Francis. *Journal of phonetics*, 2011.
[10] Nic J De Vries, Marelie H Davel, Jaco Badenhorst, et al. *Speech communication*, 2014.
[11] Ghislaine Dehaene-Lambertz and Stanislas Dehaene. *Nature*, 1994.
[12] Gabriel Doyle and Roger Levy. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016.
[13] Peter D. Eimas, Joanne L. Miller, and Peter W. Jusczyk. On infant speech perception and the acquisition of language. 1987.
[14] Micha Elsner and Cory Shain. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.
[15] Luciano Fadiga, Laila Craighero, Giovanni Buccino, and Giacomo Rizzolatti. *European journal of Neuroscience*, 2002.
[16] Jacob Feldman. *Cognition*, 2012.
[17] Naomi H Feldman, Thomas L Griffiths, Sharon Goldwater, and James L Morgan. *Psychological review*, 2013.
[18] Stella Frank, Naomi H Feldman, and Sharon Goldwater. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014.
[19] Kathleen Currie Hall, Blake Allen, Michael Fry, et al. In *14th Conference for Laboratory Phonology*, 2016.
[20] Stevan Harnad. Categorical Perception. 2003.
[21] Bruce Hayes. *Introductory phonology*, 2011.
[22] James Hillenbrand. *Journal of Speech & Hearing Research*, 1985.
[23] Jean-Rémy Hochmann, Silvia Benavides-Varela, Marina Nespor, and Jacques Mehler. *Developmental science*, 2011.
[24] John F Houde and Michael I Jordan. *Science*, 1998.
[25] Roman Jakobson, C Gunnar Fant, and Morris Halle. *Preliminaries to speech analysis: The distinctive features and their correlates*, 1951.
[26] Bernd J Kröger, Jim Kannampuzha, and Christiane Neuschaefer-Rube. *Speech Communication*, 2009.
[27] Patricia K Kuhl. *Child phonology*, 1980.
[28] Andrew Martin, Sharon Peperkamp, and Emmanuel Dupoux. *Cognitive Science*, 2013.
[29] Chandan R Narayan, Janet F Werker, and Patrice Speeter Beddor. *Developmental Science*, 2010.
[30] Thierry Nazzi. *Cognition*, 2005.
[31] Susan Nittrouer. *The Journal of the Acoustical Society of America*, 2001.
[32] Joe Pater and Elliott Moreton. *The EFL Journal*, 2014.
[33] Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, et al. *Journal of machine learning research*, 2011.
[34] Sharon Peperkamp, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. *Cognition*, 2006.
[35] Mark A Pitt, Keith Johnson, Elizabeth Hume, et al. *Speech Communication*, 2005.
[36] Linda Polka, Connie Colantonio, and Megha Sundara. *The Journal of the Acoustical Society of America*, 2001.
[37] Ferran Pons and Juan M Toro. *Cognition*, 2010.
[38] Friedemann Pulvermüller, Martina Huss, Ferath Kherif, et al. *Proceedings of the National Academy of Sciences*, 2006.
[39] Edmund T Rolls and Alessandro Treves. *Neural networks and brain function*, 1998.
[40] Andrew Rosenberg and Julia Hirschberg. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007.
[41] Yosef Singer, Yayoi Teramoto, Ben D B Willmore, et al. *eLife*, 2018.
[42] Daniel Swingley. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 2009.
[43] Alessandro Treves and Edmund T Rolls. *Network: Computation in Neural Systems*, 1991.
[44] Nikolaï Sergeyevich Trubetskoy. Grundzüge der phonologie. 1939.
[45] Maarten Versteegh, Roland Thiollière, Thomas Schatz, et al. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[46] Kate E Watkins, Antonio P Strafella, and Tomáš Paus. *Neuropsychologia*, 2003.
[47] Katherine S White and James L Morgan. *Journal of Memory and Language*, 2008.
[48] Stephen M Wilson, Ayşe Pinar Saygin, Martin I Sereno, and Marco Iacoboni. *Nature Neuroscience*, 2004.
[49] Shaorong Yan, Francis Mollica, and Michael K Tanenhaus. In *Proceedings of the 40th Annual Cognitive Science Society Meeting*, 2018.