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

Cory Shain and Micha Elsner

shain.3@osu.edu

Ohio State

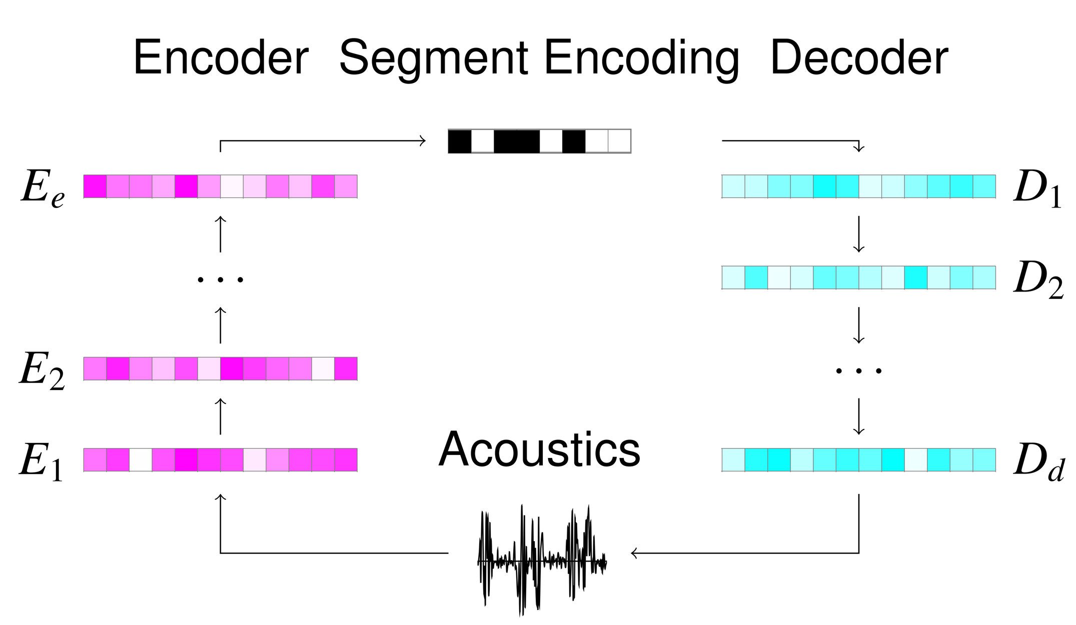
Question

How do young infants learn the phoneme categories and phonological features of their target language? In particular:

- **Q1:** To what extent can phoneme categories emerge from a drive to memorize auditory percepts?
- **Q2:** How perceptually available are theory-driven phonological features?

Background

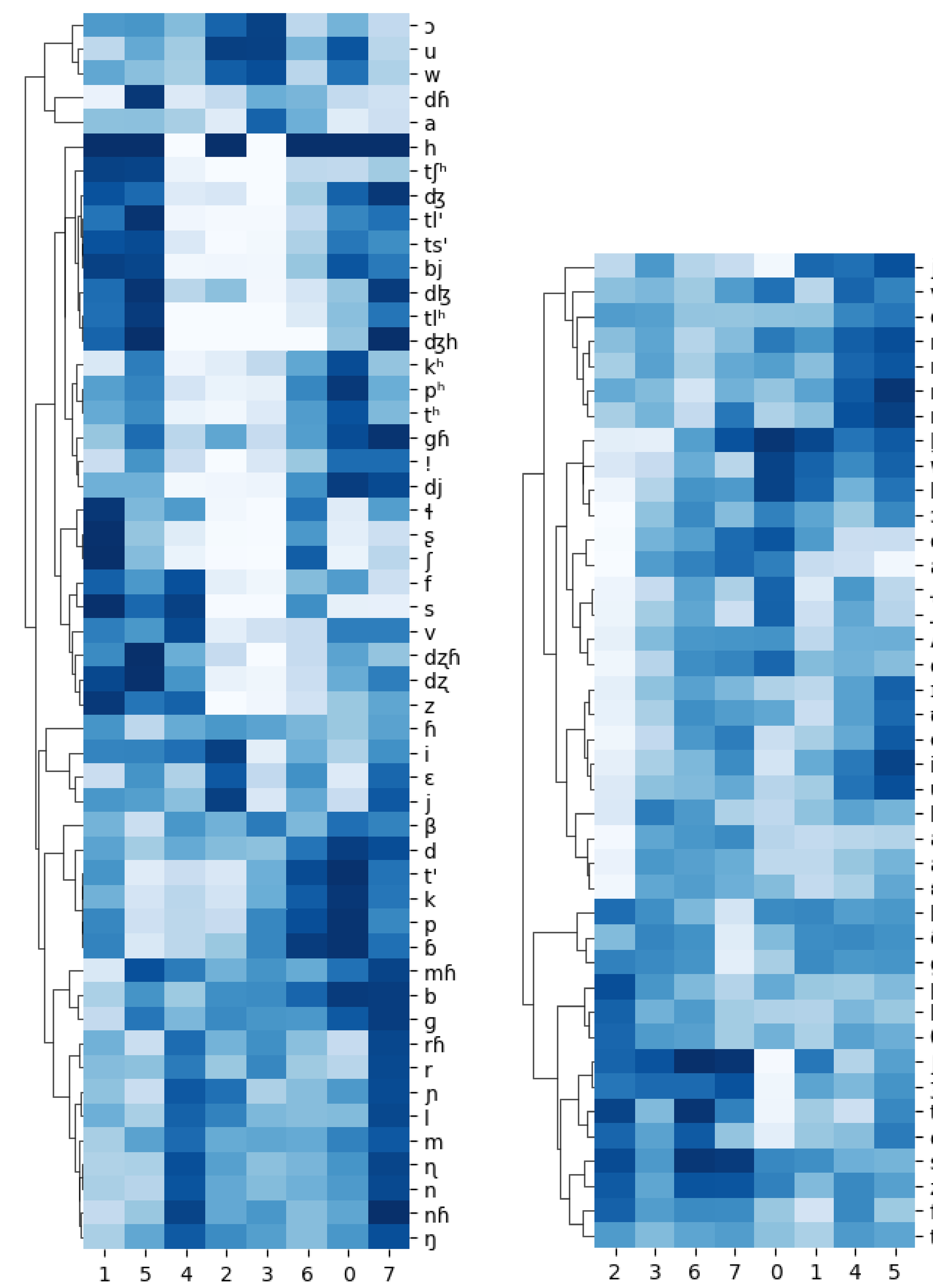
- Acoustics must contain evidence for phoneme categories
- Language comprehension and production might be linked through a sensorimotor loop [6]
- Limited auditory memory requires compression, which may guide learning [1]
- Featural decomposition occurs in acquisition [7]
- Phone perception tends to be categorical, even in infants [4]
- We implement these biases in a computational model and evaluate its acquired representations



BSN autoencoder architecture with encoder layers $E_{1,...,e}$ and decoder layers $D_{1,...,d}$.

Experimental Design

- **Data:**
  - Zerospeech 2015 challenge datasets [12]
    - English [11]
    - Xitsonga [2]
- **Model:**
  - Deep neural autoencoder (percept modeling, autoassociation, sensorimotor loop)
  - 8-dimensional bottleneck (compression)
  - Discrete binary stochastic neurons (BSNs) (feature decomposition, categorical perception)
  - **Inputs/outputs:** MFCC acoustic features from pre-segmented phonemes



Mean activation pattern by gold segment label

Highlights

- Relative clustering improvement over random shows percept modeling provides learning signal
- Asymmetries in feature recovery show some features are more reliably encoded than others
- Similar feature recovery patterns between languages suggests that results reflect general perceptual availability
- Differences between languages may reflect perceptual crowding:
  - Different relative performance on a cluster of vowel and consonant features:
  - **Xitsonga:** vowel > consonant; relatively fewer vowel categories
  - **English:** consonant > vowel; relatively fewer consonant categories
- Aligns with infant phone discrimination patterns:
  - Best recovery of voicing [8] and features that distinguish vowel-like from consonant-like segments, distinctions made early by infants [3]
  - Poorer recovery of e.g. nasal and fricative place distinctions (see clustermaps), as has been shown for infants [10, 9]

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

Xitsonga clustering (2118% relative V-measure improvement)

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

English clustering (4500% relative V-measure improvement)

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

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

English feature recovery

Conclusion

Much phonological structure is perceptually available, but some may not be.

- Based solely on a perceptual modeling objective, our learner partially acquires phoneme categories (**Q1**) and theory-driven features (**Q2**), unsupervised. Evidence for such features is therefore perceptually available.
- Error patterns mimic those of human infants
- Top-down constraints are likely needed in order to refine representations, as has been argued for humans [5]

References

[1] Baddeley, A., Gathercole, S., and Papagno, C. *Psychological Review*, 1 1998.
[2] De Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., De Wet, F., Barnard, E., and De Waal, A. *Speech communication*, 2014.
[3] Dehaene-Lambertz, G. and Dehaene, S. *Nature*, 1994.
[4] Eimas, P. D., Miller, J. L., and Jusczyk, P. W. On infant speech perception and the acquisition of language. 1987.
[5] Feldman, N., Griffiths, T., and Morgan, J. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2009.
[6] Houde, J. F. and Jordan, M. I. *Science*, 1998.
[7] Kuhl, P. K. *Child phonology*, 1980.
[8] Lasky, R. E., Syrdal-Lasky, A., and Klein, R. E. *Journal of Experimental Child Psychology*, 1975.
[9] Narayan, C. R., Werker, J. F., and Beddor, P. S. *Developmental Science*, 2010.
[10] Nittrouer, S. *The Journal of the Acoustical Society of America*, 2001.
[11] Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. *Speech Communication*, 2005.
[12] Versteegh, M., Thiollière, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., and Dupoux, E. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

To appear as: Shain, C. and Elsner, M. (2019). Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In *NAACL 2019*.