Modeling effects of incremental memory and prediction pressures on phoneme learning from speech

Cory Shain and Micha Elsner, Ohio State

What learning signals enable infants to discover linguistic patterns from a noisy, information-rich perceptual stream? Some theories of language acquisition invoke *memory pressures* to explain infant learning, arguing that linguistic representations constitute efficient compression codes whose discovery might optimize long-term storage demands [24, 29] and/or working memory demands during real-time speech processing [3]. This view is supported by experimental [18] and modeling [9, 33] evidence, but other work has questioned the efficiency of human mental codes [23] and the utility of memory pressures for language learning [27]. An alternative class of theories invokes *prediction pressures* as a learning signal [31, 15, 1], since knowledge of linguistic regularities might make speech more predictable. Recent work has argued that incremental language models [16, 30] acquire language-like representations from a prediction objective [22] and covary with measures of human processing [13, 36]. This discussion mirrors related discussion about the relative importance of memory and prediction in theories of adult sentence processing [20, 11] and broad neuronal-level learning [2, 26, 5, 17, 34], and, as in those fields [21], memory and prediction pressures may play complementary roles in infant language learning.

In this study, we develop a broad-coverage unsupervised neural network model (Fig 2) to examine possible influences of memory and prediction pressures on infant phoneme learning from speech. Cochlear output is submitted to a hierarchical multiscale recurrent neural network (HM-RNN) [6] speech processor. Each layer of the network processes representations from the layer below, dividing them into discrete segments; at predicted segment boundaries, the layer both (1) emits its segment label (hidden state) to the layer above, and (2) flushes its working memory and refreshes it with top-down guidance from the layer above. In this way, the encoder generates a sparse, hierarchical speech representation over multiple timescales. We apply a novel incremental objective function that at any point in time attempts to reconstruct B segments into the past and predict F segments into the future from the layer below, applied only at incoming segment boundaries. Learning is driven only by these objectives, without supervision for boundary locations or segment labels. Our model implements several independently supported cognitive constraints: incrementality [35]; hierarchically organized [14, 25], feature-rich [7, 28] segmental [32, 19] representations; interactive top-down and bottom-up information flow [38, 10]; modeling of its own sequence of latent representations [12]; and local error signals that are plausibly supported by human working memory [4, 8].

We use the model to study phoneme learning from English and Xitsonga speech in the Zerospeech 2015 dataset [37], (1) experimentally manipulating $B \in \{0, 5, 25, 50\}$ (strength of memory pressure) and $F \in \{0, 1, 5, 10\}$ (strength of prediction pressure), along with number of layers $L \in \{2, 3, 4\}$, and (2) evaluating the impact of these manipulations on three measures of phoneme induction quality: (i) alignment between modeled and human-annotated phoneme boundaries, and (ii) phoneme and (iii) phonological feature (e.g. [$\pm$voice]) classification accuracy from a linear probe of the first layer's hidden state (the most phoneme-like in tuning experiments). Evaluations on a held-out test set (Fig 1) show statistically significant benefits of working memory pressures (better performance when $B > 0$), prediction pressures (better performance when $F > 0$), and depth (better performance when $L > 2$), with a general peak in performance when $B \approx 25$ and $F \approx 5$. The optimality of these values is intriguing because they correspond respectively to 250ms and 50ms intervals, which fall within estimates of the storage duration in humans of the unanalyzed acoustic traces [8] that are needed to compute the objectives. Performance patterns are largely consistent across languages and metrics, supporting a language-general, complementary influence of memory and prediction pressures on overall phoneme learning. We also compare against an architecturally matched untrained baseline (Baseline U) and against an architecturally matched cross-language baseline (Baseline X, i.e. English training and Xitsonga evaluation, or *vice versa*). Baseline U measures architectural inductive bias (how much phoneme knowledge can be derived from processor design, without learning), while Baseline X measures domain inductive bias (how much phoneme knowledge can be derived from knowledge of some form of human speech, without exposure to the target language). Both kinds of biases might plausibly be innately specified, and comparisons indicate that the full model systematically outperforms them only in the presence of both memory and prediction pressures. Memory and prediction pressures thus modulate not only absolute acquisition performance, but also the utility of language experience *vis-a-vis* plausible inductive biases. Our study therefore suggests that both memory-driven and prediction-driven learning signals may be available to infants during early phoneme acquisition.

Selected References

[1] Apfelbaum, K. S. and McMurray, B. *Cognitive science*, 2017.
[2] Attneave, F. *Psychological review*, 1954.
[3] Baddeley, A., Gathercole, S., and Papagno, C. *Psychological Review*, 1 1998.
[4] Baddeley, A. D., Thomson, N., and Buchanan, M. *Journal of Verbal Learning and Verbal Behavior*, 1975.
[5] Bialek, W., Nemenman, I., and Tishby, N. *Neural computation*, 2001.
[6] Chung, J., Ahn, S., and Bengio, Y. In *International Conference on Learning Representations 2017*, 2017.
[7] Clements, G. N. *Phonology*, 1985.
[8] Cowan, N. *Psychological bulletin*, 1984.
[9] Elman, J. L. *Cognition*, 1993.
[10] Feldman, N., Griffiths, T., and Morgan, J. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2009.
[11] Ferreira, F. and Chantavarin, S. *Current directions in psychological science*, 2018.
[12] Friston, K. *Nature reviews neuroscience*, 2010.
[13] Goodkind, A. and Bicknell, K. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 2018.
[14] Hasson, U., Chen, J., and Honey, C. J. *Trends in cognitive sciences*, 2015.
[15] Johnson, M. A., Turk-Browne, N. B., and Goldberg, A. E. *Behavioral and Brain Sciences*, 2013.
[16] Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. *arXiv preprint arXiv:1602.02410*, 2016.
[17] Keller, G. B. and Mrsic-Flogel, T. D. *Neuron*, 2018.
[18] Kersten, A. W. and Earles, J. L. *Journal of Memory and Language*, 2001.
[19] Kooijman, V., Junge, C., Johnson, E. K., Hagoort, P., and Cutler, A. *Frontiers in psychology*, 2013.
[20] Levy, R. *Cognition*, 2008.
[21] Levy, R., Fedorenko, E., and Gibson, E. *Journal of Memory and Language*, 2013.
[22] Linzen, T., Dupoux, E., and Goldberg, Y. *Transactions of the Association for Computational Linguistics*, 2016.
[23] McMurray, B., Tanenhaus, M. K., and Aslin, R. N. *Cognition*, 2002.
[24] Newport, E. *Cognitive Science*, 1990.
[25] Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E., Feldstein, N. A., McKhann, G. M. V., Schevon, C., Flinker, A., and Mesgarani, N. *bioRxiv*, 2020.
[26] Olshausen, B. A. and Field, D. J. *Nature*, 1996.
[27] Perfors, A. *Journal of Memory and Language*, 2012.
[28] Pierrehumbert, J. B. *Frequency and the emergence of linguistic structure*, 2001.
[29] Pinker, S. *Science*, 1991.
[30] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. *OpenAI Blog*, 2019.
[31] Rohde, D. L. T. and Plaut, D. C. *Cognition*, 1999.
[32] Sanders, L. D. and Neville, H. J. *Cognitive Brain Research*, 2003.
[33] Shain, C. and Elsner, M. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.
[34] Singer, Y., Teramoto, Y., Willmore, B. D. B., Schnupp, J. W. H., King, A. J., and Harper, N. S. *eLife*, 2018.
[35] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. E. *Science*, 1995.
[36] van Schijndel, M. and Linzen, T. In *EMNLP 2018*, 2018.
[37] 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.
[38] Warren, R. M. *Science*, 1970.

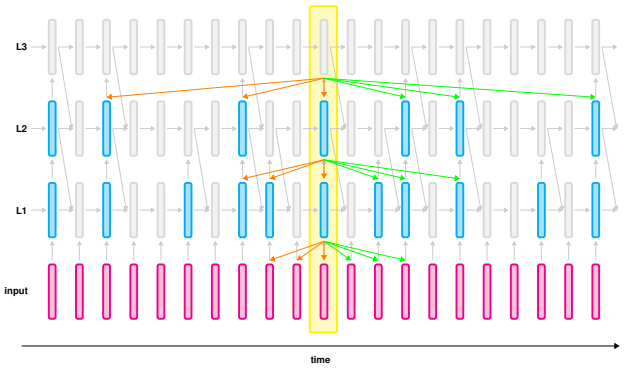

Figure 2: *Model.* Cyan indicates boundaries, grey arrows show encoder information flow, and orange and green arrows respectively show backward and forward decoder information flow.

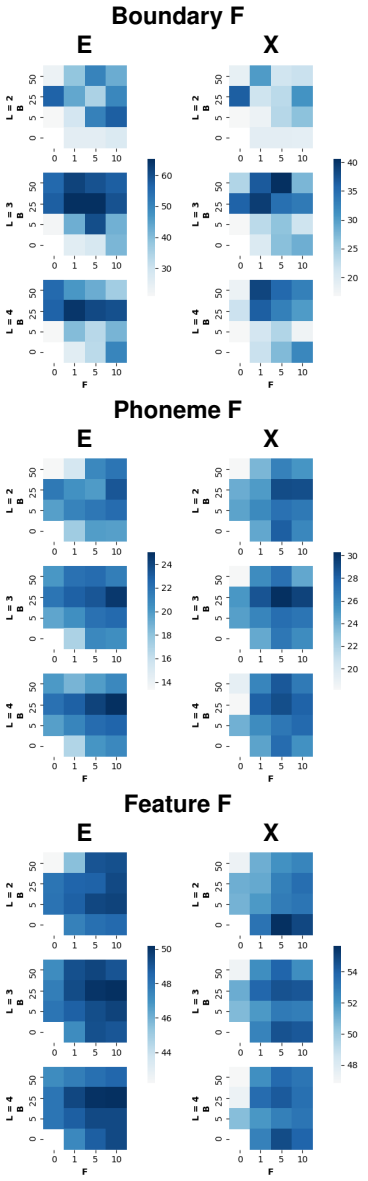

Figure 1: *Acquisition performance.* Phoneme boundary, label, and feature learning scores respectively (darker is better) in English (E) and Xitsonga (X) as modulated by memory pressure (y-axes), prediction pressure (x-axes), and depth (top to bottom by column, shallowest to deepest). All three variables improve phoneme acquisition.