Improving Spiking Neural Network Model of Canonical Babbling by Implementing an Automated Babbling Evaluation Algorithm

During the first year of life, human infants acquire many communicative skills. Two of these skills are 1) they start to produce speech-like syllables with consonants and vowels (also known as canonical babbling; Oller, 1980; Stark, 1980), and 2) they learn to take smooth turns with caregivers with minimum gaps or interruptions (Hilbrink, Gattis & Levinson, 2015; Snow, 1977). To understand how infants learn to produce speech-like vocalizations (the first kind of communicative skill mentioned above) enabled by the nervous system, Warlaumont and Finnegan created a biologically motivated model to simulate the acquisition of canonical babbling.

This computational model consists of a recurrent neural network, dopamine-modulated spike-timing-dependent plasticity (Izhikevich, 2007), and a vocal tract. When the vocal tract produces a babble, the model estimates the saliency of the sound and sends a reward when its saliency is high. They found that over time the model produced more vocalizations that were salient and speech-like, while the randomly rewarded yoked control did not. Their paper is a first attempt to present a neurophysiologically informed model of the acquisition of babbling. While it achieves good results and provides valuable predictions about infant vocal development, the model simplifies the evaluation of babbling. It uses a model, trained on sung melodies without words as well as sung unaccompanied folk songs (Coath et al., 2007, 2009) to estimate the auditory saliency of babbling, and it gives rewards based solely on auditory saliency. However, in addition to auditory saliency, caregivers in reality may also respond to infant vocalizations according to their frequency, amplitude, duration, speech-likeness (Warlaumont, Richards, Gilkerson & Oller, 2014), affect, contexts, and etc. There are many features and dimensions of a sound that may lead adults to respond to infants in different ways. In this study, I will improve the model presented by Warlaumont and Finnegan (2014), by replacing the auditory saliency model with a babbling-evaluation algorithm that approximates human judgments.

The proposed babbling-evaluation algorithm will be trained in a supervised way, on a dataset of adults’ naturalistic responses to infant babblings. It will automatically extract features from infant babbling that are relevant in predicting adults’ response. I compiled a set of infant vocalizations of different maturity levels from previous experimental recordings. Next, I recruited naive adult participants (n = 50 Cornell undergraduates) and let each of them listen to 100 babblings, and let them choose whether they will respond to each babbling or not. Then I trained the babbling-evaluation algorithm on 60% of the human adults’ decision data, tune hyperparameters of the algorithm using 20% of the data, and tested the algorithm with the remaining 20% of the data. The correlation between human adults’ and algorithm’s decisions will inform us how well the algorithm approximates real human judgments of the different babbling sounds. I will also get the correlation between human adults’ and the Warlaumont & Finnegan model’s decision; I will compare whether our babbling-evaluation algorithm or their auditory saliency model approximates human judgment better. Lastly, I will graph the saliency and speech-likeness of babbling produced by the model overtime, to see how the model acquires canonical babbling based on the algorithm’s feedback.

If the babbling-evaluation algorithm has good performance that mimics human evaluation of infant vocalizations, I will be one step closer to comparing model learning progress side-by-side with real infants’ vocal learning progress. By controlling “social feedback” from caregivers or the algorithm (if they correlate significantly), I can implement neurons and tweak parameters in the neural network model so that its learning progress approximates human infants’. This side-by-side comparison may provide further insights of how infants’ nervous systems support their vocal acquisition.

To train the classifier, I started with a 1-dimensional convolutional neural network that was pretrained on the SpeechCommands dataset. The model uses the M5 architecture from (Dai et al. 2016), which contains four convolutional layers (each containing batch normalization and max pooling) followed by a fully-connected layer. Since our dataset only has 2 classes while the original model was trained with 35 classes, I replaced the final layer with a randomly-initialized linear layer with 2 outputs, and kept the pretrained weights for the other layers. I then fine-tuned the entire network for 50 epochs on our dataset, using the negative log-likelihood loss and the Adam optimizer with learning rate of 1e-3 (which was reduced by a factor of 10 every 20 epochs) and weight decay of 1e-4. On a held-out validation dataset, I achieved an accuracy of 72.2%, precision of 78.5%, recall of 71.1%, and F1 score of 74.6%. Note that a model that always predicts the positive class would achieve 57.4% accuracy on this dataset, indicating that this model performs substantially better than chance but there is still room for further improvement.

References

Coath, M., Denham, S. L., Smith, L., Honing, H., Hazan, A., Holonowicz, P., & Purwins, H.

(2007). An auditory model for the detection of perceptual onsets and beat tracking in

singing. In *Neural information processing systems, workshop on music processing in the*

*brain*.

Coath, M., Denham, S. L., Smith, L. M., Honing, H., Hazan, A., Holonowicz, P., & Purwins, H.

(2009). Model cortical responses for the detection of perceptual onsets and beat tracking

in singing. *Connection Science*, *21*(2-3), 193-205.

Hilbrink, E. E., Gattis, M., & Levinson, S. C. (2015). Early developmental changes in the timing

of turn-taking: a longitudinal study of mother–infant interaction. *Frontiers in psychology*,

*6*, 1492.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and

dopamine signaling. *Cerebral cortex*, *17*(10), 2443-2452.

Oller DK. The emergence of the sounds of speech in infancy. In: Yeni-Komshian GH, Kavanagh

JF, Ferguson CA, editors. Child phonology, vol. 1: Production. New York: Academic

Press; 1980. p. 93–112.

Stark RE. Stages of speech development in the first year of life. In: Yeni-Komshian GH,

Kavanagh JF, Ferguson CA, editors. Child phonology, vol. 1: Production. New York:

Academic Press; 1980. p. 73–92.

Snow, C. E. (1977). Mothers’ speech research: From input to interaction. *Talking to children:*

*Language input and acquisition*, *3149*.

Warlaumont, A. S., & Finnegan, M. K. (2016). Learning to produce syllabic speech sounds via

reward-modulated neural plasticity. *PloS one*, *11*(1), e0145096.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop

for speech development and its reduction in autism. *Psychological science*, *25*(7),

1314-1324.