2014 Joint IEEE International Conferences on Development and Learning
and Epigenetic Robotics (ICDL-Epirob)
October 13-16, 2014. Palazzo Ducale, Genoa, Italy

# An iterative probabilistic model of speech-related vocalization rate growth due to child-caregiver interaction

Anne S. Warlaumont
Cognitive and Information Sciences, University of California, Merced
www.annewarlaumont.org, awarlaumont2@ucmerced.edu

***Abstract*—Over the course of the first four years of life, the proportion of children's vocalizations that are speech-related increases steadily. The rate of this growth is reduced for children with autism spectrum disorder (ASD) and for children from households with relatively lower socioeconomic status (SES). The present study attempts to model this set of findings, treating adult responses as reinforcers of child behavior. The model starts with a 50% chance of producing a speech-related vocalization and gradually increases this probability by updating its speech-related vocalization and not-speech-related vocalization probabilities each time a response is received. Numbers of vocalizations per day and rates of adult responding to the two vocalization types are drawn from human data to create high SES typically developing (TD), high SES ASD, low SES TD, and low SES ASD versions. The model shows growth in speech-related vocalizations that matches well to that observed for the human children and that matches the differences observed across clinical and SES groups. Some aspects of speech-related vocalization development are not well accounted for by the model; possible explanations and extensions are proposed.**

## I. Introduction

Over the first few years of life, children's vocal utterances become more and more adult-like. This increase in communication abilities is affected by clinical differences that originate within the child. For example, children with or at heightened risk for autism spectrum disorder (ASD) exhibit reduced increases with age in speech-related vocalization rate [1], [2], delayed achievement of babbling milestones [1], [3], [4], delayed development of mature vocalization acoustics [5], delayed acquisition of adult-like prosody [6], and slower development of language more generally [7]. Development of communication abilities is also affected by factors originating outside the child. In particular, higher socioeconomic status (SES) is associated with faster child language development compared to children from families of lower SES [8]–[10].

In a previous study [2], we too found an increase in speech-related vocalization rate as a function of age, with this growth being slower both for children with ASD and for children of lower SES (Fig. 1). The study analyzed the child and adult vocalizations produced during daylong naturalistic audio recordings. The child and adult vocalizations were automatically tagged and child vocalizations were automatically classified by the LENA system [5], [11], [12] as either speech-related (speech, babble, song, etc.) or as not-speech-related (cry, laugh, burp, effort grunt, etc.).

It was proposed that the growth in speech-related vocalization over age as well as the ASD- and SES-differences in this growth could be explained by a social feedback loop involving the micro dynamics of interaction (see also [13]). In the proposed feedback loop, adults respond contingently to children's vocalizations and children's vocalizations are in turn contingent on previous adult responses (Fig. 2). Specifically, adult responses are more likely for speech-related vocalizations than for not-speech-related vocalizations and a child's vocalization is more likely to be speech-related than not-speech-related if the child's previous speech-related vocalization received an adult response. Adult responses were operationalized as any adult vocalization starting within a 1-second window following the offset of a child vocalization. Analysis of the responses to child vocalizations and of children's responses to adult responses provided support for the two contingencies of adult on child and of child on adult (for related experimental findings, see [14]).

Two aspects of the feedback loop were found to differ in ASD vs. TD. First, children with ASD produced fewer vocalizations of any sort. This would be expected to reduce the number of learning opportunities for the ASD group (see also [15] and [16]). Second, adult responses to the children with ASD were less contingent on child vocalization type, i.e., adult response likelihoods were more similar across vocalization types. This would also be expected to negatively impact the children's growth in speech-related sounds, because the feedback they received was less favoring of speech-related sounds. The same two differences were also found for children of lower socioeconomic status (as indicated by maternal education level) compared to children of higher socioeconomic status, consistent with other previous studies that have found SES differences in kinds of parent speech input to their children, as well as in parent-child interactions [8], [9]. Different input patterns could derive from differences across the SES groups in parents' general conversational styles, beliefs about how much their behavior affects their children's development, economic stresses, as well as other factors [17].

We built a simple computational model to verify that such contingencies could in principle lead to the differences in growth of speech-related vocalizations across groups (see the Supplemental Online Material of [2]). Indeed, the model showed slower learning when child vocalization rate and adult

978-1-4799-7540-2/14/$31.00 ©2014 IEEE

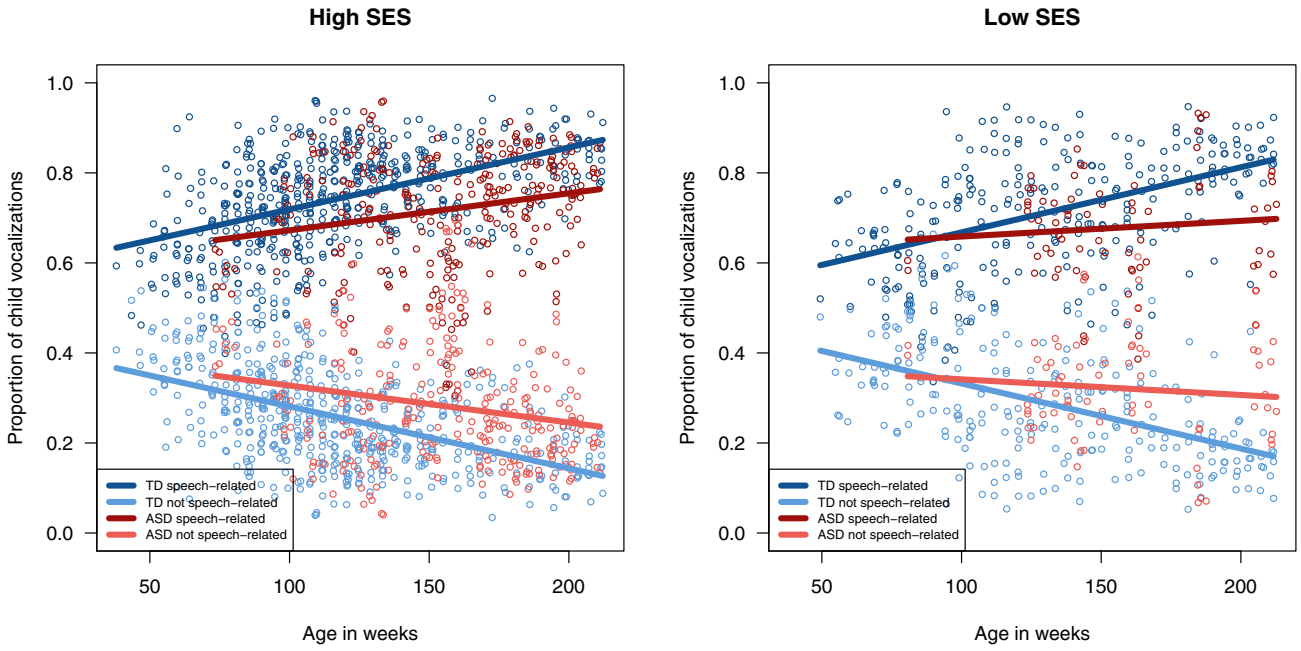


Fig. 1. The proportion of vocalizations produced by human children that are speech-related increases as a function of age. The proportion of vocalizations that are not-speech-related decreases. The speech-related vocalization increase is faster for TD children (blue) than for children with ASD (red) and is faster for children from families of higher socioeconomic status (left) than for children from families of lower socioeconomic status (right). Data for the ASD group start at 16 months of age due to limits on how early the disorder can be identified. Note that the parameters used in the present modeling study and the measures against which the model is compared are based on a subset of this sample that was matched for gender, age, and SES between the TD and ASD groups. This figure is adapted from [2].

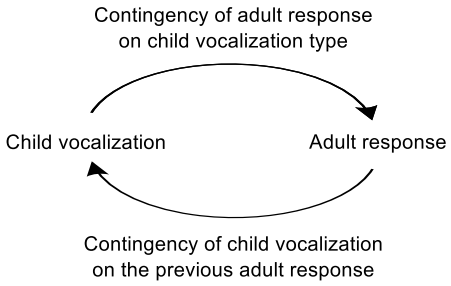


Fig. 2. Feedback loop proposed as a mechanism underlying the patterns of growth in Fig. 1. When a child produces a vocalization and subsequently receives an adult response, this affects the child's future vocalizations. Adult responses are contingent on child vocalization type, such that speech-related child vocalizations are more likely to receive a response than not-speech-related ones. When children's subsequent vocalizations are biased toward those types that previously received responses, over time, speech-related vocalizations will become progressively more prominent. Differences in this feedback loop, such as fewer child vocalizations, reduced adult response rates, reduced contingency of adult responses, or reduced contingency of child behavior (i.e. child learning), due to conditions such as ASD or low SES, should result in differences in child speech development. This figure is adapted from [2].

response probability parameters were matched to those of the ASD group than when matched to those of the TD group. The present study extends that initial modeling effort. The model and results are described in much greater detail, visualizations of the model and its performance are provided, and the model is extended to address SES differences in addition to TD-ASD differences.

The next section will present how the model works, then compare it to two other modeling approaches, then describe how different versions were created and how the model was tested. In the Results section, the model is evaluated in two ways: (1) its production of speech-related versus not-speech-related sounds across development is compared to what was observed for human children and (2) contingency of child vocalization type on most recent adult response is compared to that observed in human data. This is followed by a discussion of the implications of the model for understanding how children learn to adapt their vocal behaviors using contingent adult responses and some suggestions for specific future directions.

The model code and simulation data can be downloaded from http://dx.doi.org/10.5281/zenodo.11622

## II. Methods

### A. Learning algorithm

A schematic diagram of how the model works is provided in Fig. 3. In each simulation, the model's initial probability of producing a speech-related vocalization, $P(sp, 0)$, and initial probability of producing a not-speech-related vocalization, $P(nsp, 0)$, are set to:

$$P(sp, 0) = P(nsp, 0) = .5. \quad (1)$$

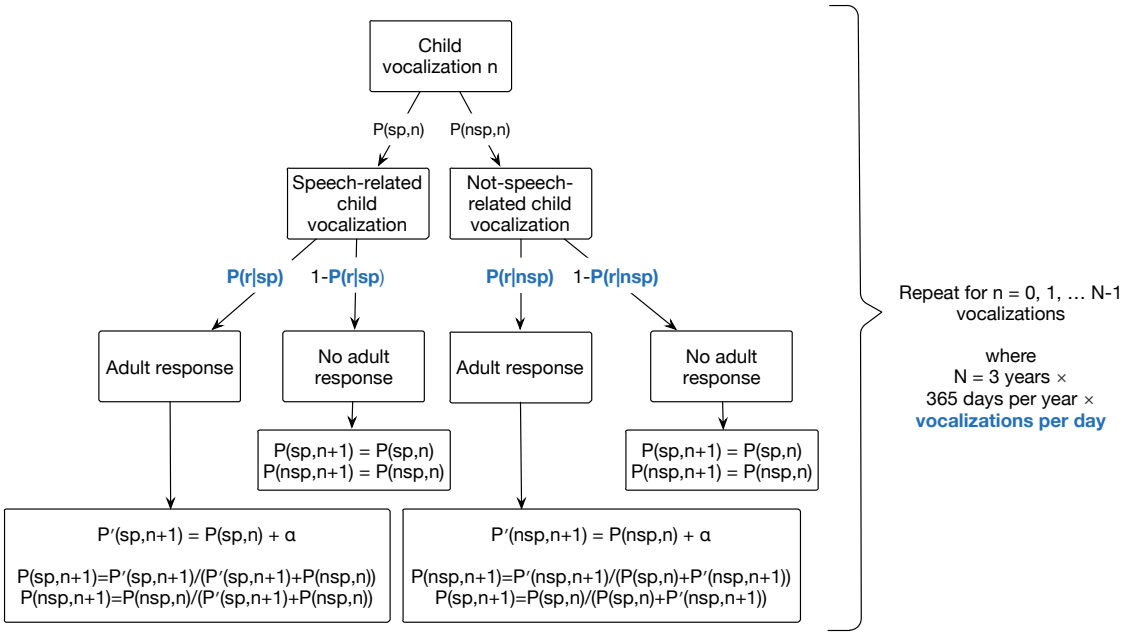


Fig. 3. Schematic diagram of the computational model. Each iteration starts with the child making a vocalization. That vocalization is probabilistically determined to be either speech-related or not-speech-related. Whether or not the vocalization receives an adult response is also probabilistically determined. If a response is received, the probability of producing the sound type that preceded that response is incremented by a small amount, and then the child vocal type probabilities are rescaled. This whole process is repeated many times, corresponding to several years' worth of child vocalizations. The parameters that were taken directly from the human data and that differ across groups are shown in blue. $n$ is the index of the current child vocalization. $P(sp, n)$ is the probability of vocalization $n$ being speech-related. $P(nsp, n)$ is the probability of vocalization $n$ not being speech-related. $P(r|sp)$ is the probability of adult response when the child vocalization is speech-related and is taken from the human data. $P(r|nsp)$ is the probability of adult response when the child vocalization is not-speech-related and is also taken from the human data. $\alpha$ is the one free parameter in the model and determines the learning rate. $N$ is the total number of vocalizations the model produces. The number of vocalizations per day is taken from the human data.

The model then undergoes a number of vocalization episodes, N, where

$$N = 3 \cdot 365 \cdot v \tag{2}$$

and $v$ is the number of vocalizations produced per day. Thus, $N$ is three years' worth of vocalizations. $v$ was taken directly from the human data (see Table I) and was set differently depending on the group being simulated (described in section C below).

At each vocalization episode, $n$, the vocalization is randomly chosen to be speech-related or not-speech-related with probability $P(sp, n)$ and $P(nsp, n)$, respectively. It is then randomly determined whether the vocalization should receive an adult response, with $P(r|sp)$ being the probability of an adult response to a speech-related vocalization if the vocalization was speech-related and $P(r|nsp)$ being the probability of an adult response if the child vocalization was not-speech-related. $P(r|sp)$ and $P(r|nsp)$ are taken directly from the human data and are set differently depending on the group being simulated (see Table I and subsection C).

If the model produces a speech-related vocalization and is reinforced, then $P(sp, n+1)$ and $P(nsp, n+1)$ are updated according to the following difference equations:

$$P'(sp, n+1) = P(sp, n) + \alpha, \tag{3}$$

which increments the probability of producing the just-reinforced behavior, followed by

$$P(sp, n+1) = \frac{P'(sp, n+1)}{P'(sp, n+1) + P(nsp, n)} \tag{4}$$

and

$$P(nsp, n+1) = \frac{P(nsp, n}{P'(sp, n+1) + P(nsp, n)}, \tag{5}$$

which normalize the probabilities of producing each vocalization type. $\alpha$ is the one free parameter in the model and was set to 0.000008, which in pilot work was found to yield final $P(sp, N)$ and $P(nsp, N)$ values that were roughly similar in magnitude to the $P(sp)$ and $P(nsp)$ values observed for the human children's recordings from ages 3;0–3;11 (see Fig. 1 and the rightmost column of Table I). Similarly, if the model produces a not-speech-related vocalization and is reinforced, then the update equations are:

$$P'(nsp, n+1) = P(nsp, n) + \alpha \tag{6}$$

$$P(nsp, n+1) = \frac{P'(nsp, n+1)}{P(sp, n) + P'(nsp, n+1)} \tag{7}$$

$$P(sp, n+1) = \frac{P(sp, n)}{P(sp, n) + P'(nsp, n+1)}. \tag{8}$$

If the model is not reinforced, then $P(sp, n+1) = P(sp, n)$ and $P(nsp, n+1) = P(nsp, n)$. In other words, no learning takes place.

TABLE I
AVERAGE HUMAN AND MODEL VOCALIZATION RATES, ADULT RESPONSE PROBABILITIES, AND CHILD VOCALIZATION TYPE PROBABILITIES. SQUARE BRACKETS GIVE 95% CONFIDENCE INTERVALS.

| Group | Vocalizations per day ($v$) | $P(r\|sp)$ | $P(r\|nsp)$ | Initial model $P(sp,0)$ | Final model $P(sp,N)$ | 3-year-old human $P(sp)$ |
|---|---|---|---|---|---|---|
| High SES TD | 2,796 | .214 | .149 | .5 | .830 [.829,.832] | .804 |
| High SES ASD | 2,149 | .205 | .154 | .5 | .721 [.720,.723] | .748 |
| Low SES TD | 1,994 | .192 | .130 | .5 | .748 [.746,.750] | .766 |
| Low SES ASD | 2,044 | .173 | .133 | .5 | .673 [.671,.675] | .674 |

The whole sequence of steps is repeated for three years' worth vocalization episodes, i.e. for $n = 0, ..., N-1$.

### B. Comparison to other approaches

Prior to developing this particular model, adaptations of commonly used reinforcement-based models were tried. As described in the following paragraphs, these approaches did not match the human trajectory of increasing speech-related vocalization very well. The first approach that was tried was a Rescorla-Wagner (RW) model adapted for operant conditioning. The second approach was based on Q-learning from the reinforcement learning literature. In all cases, initial $P(sp)$ and $P(nsp)$ were set to .5.

The RW model is a model of classical conditioning [18]. It has recently been successfully applied to explain findings from human child and adult word learning experiments [19], [20]. In order to adapt it for our problem, which is an operant learning problem rather than a classical conditioning one, the model was set to always have two possible contexts, speech-related vocalization and not-speech-related vocalization. The model entered each context with probability $P(sp)$ and $P(nsp)$, respectively, and then observed whether or not a response was given. From this, the RW model learned to associate each context with a reward probability. These associations were stored as updates to $P(sp)$ and $P(nsp)$. In the end, this model ended up matching its probability of producing a speech-related vocalization, $P(sp)$, to the relative balance of $P(r|sp)$ to $P(r|nsp)$. This resulted in a much lower final $P(sp)$ than was observed for the human data, and also resulted in quite suboptimal rates of receiving reward.

In adapting Q-learning [21] to the present problem, it seemed most sensible to start off assuming there was only one state. The value of $Q$ for each of two policies, (1) producing a speech-related vocalization and (2) producing a not-speech-related vocalization, was the thing to be learned. The model chose at each vocalization episode to follow the policy with the highest $Q$. Perhaps unsurprisingly in retrospect, this model converged very quickly, within a few iterations, to a $P(sp)$ of 1 and a $P(nsp)$ of 0. This very fast learning did not match the much more gradual learning exhibited by the human children.

### C. Model versions and evaluation

Four versions of the model were created, a high SES typically-developing (TD) version, a high SES autism spectrum disorder (ASD) version, a low SES TD version, and a low SES ASD version. In a subsample of the human data that had been matched across TD and ASD groups for gender, SES, and age [2], each of the four groups had a different mean number of child vocalizations per day, a different mean probability of adult response within 1 s for child speech-related vocalizations, and a different mean probability of adult response within 1 s for not-speech-related vocalizations. These average values for the human data were used to create each model version's $v$, $P(r|sp)$, and $P(r|nsp)$ (see Table I).

100 simulations of each model version were run. Across all the simulations for a given group, the average final probability of producing a speech-related vocalization, $P(sp,N)$, was obtained, as was its 95% confidence interval. These final probabilities were then compared to the same probabilities for the older, 3-year-old (3;0–3;11) child recordings from the matched human sample.

The contingency of the child vocalization type on whether or not the previous speech-related vocalization received a response was also determined for each simulation, using the same method as the previous study with human children [2]. Specifically, child contingency, $C_c$ was defined as

$$C_c = \frac{\sum_{n=0}^{N-1}[m(n) \in \{sp|(r|sp)\}]}{0.01 + \sum_{n=0}^{N-1}[m(n) \in \{sp|(r|sp), nsp|(r|sp)\}]} - \frac{\sum_{n=0}^{N-1}[m(n) \in \{sp|(\neg r|sp)\}]}{0.01 + \sum_{n=0}^{N-1}[m(n) \in \{sp|(\neg r|sp), nsp|(\neg r|sp)\}]}, \quad (9)$$

where $n$ is the vocalization episode number, N is the total number of vocalizations produced across the entire simulation, and $m(n)$ is the combination of information about the current vocalization type and whether its most recent speech-related vocalization received a response. $sp|(r|sp)$ is a speech-related vocalization for which the most recent speech-related vocalization received a response, $nsp|(r|sp)$ is a not-speech-related vocalization for which the most recent speech-related vocalization received a response, $sp|(\neg r|sp)$ is a speech-related vocalization for which the most recent speech-related vocalization did not receive a response, and $nsp|(\neg r|sp)$ is a not-speech-related vocalization for which the most recent speech-related vocalization did not receive a response. Square brackets indicate that the value within the summation is 1 if the expression within the brackets is true and 0 if it is false. The addition of 0.01 to each denominator prevents division

by zero. A positive $C_c$ indicates that child vocalizations were more likely to be speech-related when the most recent child speech-related vocalization received a response than when it did not receive a response.

## III. Results

Fig. 4 presents the average change in each version of the model's $P(sp, n)$ as a function of age. The TD simulations exhibited faster growth in speech-related vocalization proportion compared to the ASD simulations. In addition, the high SES simulations exhibited faster growth than the low SES simulations. At the end of training, the high SES TD simulations had the highest $P(sp)$, averaging .830 (compare to .804 for the human data), and the low SES ASD simulations had the lowest $P(sp)$, averaging .673 (compare to .674 for the human data); see Table I. The high SES ASD and low SES TD had intermediate final $P(sp)$, at .721 and .748, respectively (compare to .748 and .766 for the human data). These results match the human data quite well, having the same ordering of $P(sp)$ across the four groups and a decent quantitative fit as well. Note that the human study used linear regressions to characterize the change in $P(sp)$ with age, whereas the model's growth in $P(sp)$, while it could be reasonably well approximated by a linear function within the observed time range, is in fact nonlinear, decreasing in slope over time.

With regard to the contingency of child vocalization type on immediately previous adult response, $C_c$, the model did not provide as good a fit to the human data. The $C_c$ values reported for human children in [2] were the same for both ASD and TD groups, 0.042 (this was calculated over the whole dataset, not the matched subsample, and values were not broken down by SES). For both groups this $C_c$ was highly statistically significantly greater than zero. No statistically significant relationship between $C_c$ and ASD or SES was detected in the human data. In contrast, the model had $C_c$ values of 0.000 for all four groups, indicating a lack of a measurable local child contingency on previous response. One possible reason for this is that the model's learning rate, $\alpha$, was very low, so that learning from any single vocalization episode had only a tiny, hard to detect effect on $P(sp)$. In support of this explanation, when the learning rate was increased to 0.2 and the model was run for only one day's worth of vocalizations (which was all that was needed to reach final $P(sp)$ values similar to the 3-year-old human children), the model did exhibit a positive $C_c$. The mean $C_c$ values [and their 95% confidence intervals] for this fast-learning version were 0.041 [0.015,0.066] for the high SES TD group, 0.057 [0.028,0.086] for the high SES ASD group, 0.034 [0.010,0.058] for the low SES TD group, and 0.043 [0.020,0.066] for the low SES ASD group. Thus, when the learning rate was sped up (and number of vocalization episodes correspondingly decreased), the model did exhibit the child contingencies observed in the human data.

## IV. Discussion

Considering the simplicity of the model presented here, it accounts for the change in human children's speech-related vocalization probability quite well. The model increments the probability of producing a given vocalization type each time that vocalization type is responded to. It makes no changes to its behavior when a vocalization does not receive a response. When adult response probabilities for speech-related sounds are higher compared to response probabilities for not-speech-related sounds, as has been observed for human adults interacting with young children, this leads to a gradual increase over time in the child's production of speech-related vocalizations compared to not-speech-related vocalizations. The model's adult response probabilities and number of vocalization episodes were set to the exact same values as were observed in a sample of daylong audio recordings of human children ages 1;6 to 4;0. Four groups of children were modeled in this way: high SES TD, high SES ASD, low SES TD, and low SES ASD. The model's final speech-related vocalization probabilities for each group matched quite well to those for the 3-year old human child recordings. Small differences in vocalization rate and in adult response probabilities led to diverging behaviors of the four groups over the developmental period studied here [2], [15], [16], [22], [23], with both ASD and low SES reducing the speech development rate.

One aspect of the human data that the model had difficulty fitting is the local contingency of child vocalization type on whether the most recent child speech-related vocalization received a response ($C_c$). The model could match the human data with regard to this contingency, but only when the learning rate was increased so much that growth in $P(sp)$ was unrealistically fast. One possible explanation for how it is possible for human children to have a gradual increase in speech-related vocalization likelihood over the course of years yet also show a robust local contingency on recent adult responses is that children are learning at a relatively fast rate in the short term, so that local child contingencies on adult response are apparent, and a forgetting or decay process is also at play, causing the overall learning rate to be slower. This idea could be tested by altering the model to give it a larger learning rate ($\alpha$) and adding in a decay of $P(sp)$ and $P(nsp)$ back toward .5. An alternative possibility is that the human $C_c$ values are reflecting some set of processes other than rapid short-term learning.

The decision to set the probabilities of the two vocalization types equally, and to set them to the same values across groups, was the observation from Fig. 1 that at 50 weeks, P(sp) in the human data is about .6 and P(nsp) is about .4. Extrapolating from this data makes it appear plausible that at birth the two probablilities are approximately equal. Testing the validity of this assumption would require a study of the proportions of each type of vocalization in human neonates. Inputting initial P(sp) and P(nsp) values that are drawn directly from human data should affect the simulation outcomes. Seeing how this affects the fit to human speech-related vocalization growth is an important future direction.

It should be pointed out that across the four versions of the model, there were no differences in child learning rate. This might be surprising to some readers, since children with

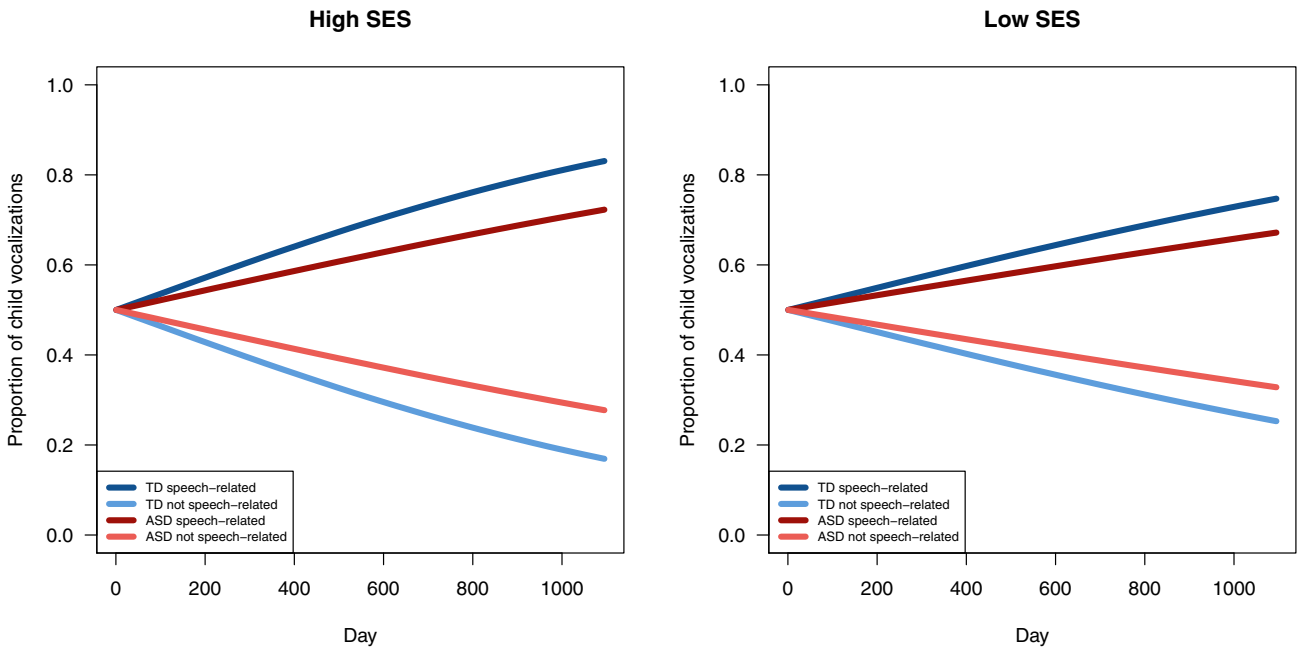


Fig. 4. Change in probability of speech-related vocalization, $P(sp)$, and probability of not-speech-related vocalization, $P(nsp)$, as a function of age in the four versions of the model. The curves represent means over 100 simulations of each model version. TD versions end up with higher speech-related vocalization probabilities than ASD versions, and high SES versions end up with higher speech-related vocalization probabilities than low SES versions. Compare these simulation results to Fig. 1, which shows corresponding human data.

ASD might be expected to learn more slowly from social responses than TD children, given their differences in social orienting [15], [22], [24]–[26]. The decision was motivated by the fact that in the previous study with human children, no differences in the local contingency of child behavior on previous adult responding were found, either across groups or as child age changed. That null result could either have been due to a lack of statistical power (second-order contingencies, and differences in them across groups, are harder to detect than first-order contingencies) or due to a true similarity across children in their learning from contingent adult responses. Conservatively we can say that differences in adult response rates and in child vocalization rates are sufficient to generate the observed differences in growth of child speech-related vocalization probability. Future work should experiment with different child learning rates, possibly changing over time, to see how these affect the model's fit to human data. Relatedly, in the human study, adult response rates were found to change over time, and this should eventually also be incorporated into the model.

In our pilot work before setting up the present version of the model, RW and Q-learning models were created and tested. The RW model ended up converging to $P(sp)$ levels that were far from optimal in terms of obtaining many responses, and instead matched the relative adult response rates. The Q-learning model on the other hand converged too quickly to optimal performance. This led to the development of the present model, which turned out to be quite a better fit to the human data. Nevertheless, the particular setups we chose for the RW and Q-learning approaches were not the only conceivable ones. For instance, we could have decided to include more than one state in the Q-learning model, where the state would include information about the model's previous behaviors and perhaps about recent adult behaviors. This might have resulted in a model that fit the human data better. If so, it would be interesting to determine if such a model has anything in common with the present one. Further exploration of various reinforcement-based approaches might yield novel insights about the relationship between these various reinforcement-focused modeling approaches and between the human speech-related vocalization development phenomena studied here and other animal learning phenomena.

Finally, the model does not consider any of the physiological details of children's nervous systems, vocal tracts, or other body parts. Many other computational models have been developed to further our understanding of how children learn to control their vocal tracts and how they learn to combine speech sounds to construct words and sentences (e.g., [27]–[37]). In many cases, these other models explicitly aim to address processing at a neural level. The main strengths of the present work are that it relates very closely to data from real children and it presents a general iterative learning method that is suitable for learning from contingent responses in a manner that is consistent with the rates at which children learn. It might therefore serve as a guide for future work on more detailed models in order to include realistic social response

rates and learning rates while at the same time including more specific neural and physiological bases for the child behaviors and for the learning mechanisms. For example, would existing neural network models of learning to produce speech sounds via reinforcement, such as [35] or [36], if given realistic numbers of trials and adult reinforcement rates, also yield good fits to human data? Relatedly, future extensions of this work both on the human side and on the modeling side should take into account more detailed information about the acoustics and semantics [38] of child and adult vocalizations.

## REFERENCES


[1] R. Paul, Y. Fuerst, G. Ramsay, K. Chawarsk, and A. Klin, "Out of the mouths of babes: vocal production in infant siblings of children with ASD," *Journal of Child Psychology and Psychiatry*, vol. 52, pp. 588–598, May 2011.

[2] A. S. Warlaumont, J. A. Richards, J. Gilkerson, and D. K. Oller, "A social feedback loop for speech development and its reduction in autism," *Psychological Science*, vol. 25, pp. 1314–1324, Jul. 2014.

[3] A. M. Wetherby, D. G. Yonclas, and A. A. Bryan, "Communicative profiles of preschool children with handicaps: implications for early identification," *Journal of Speech and Hearing Disorders*, vol. 54, pp. 148–158, May 1989.

[4] E. Patten, K. Belardi, G. Baranek, L. R. Watson, J. D. Labban, and D. K. Oller, "Vocal patterns in infants with autism spectrum disorder: canonical babbling status and vocalization frequency," *Journal of Autism and Developmental Disorders*, 2014. [Online]. Available: http://link.springer.com/article/10.1007/s10803-014-2047-4

[5] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and atypical development," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 13 354–13 359, Jul. 2010.

[6] S. Peppé, J. McCann, F. Gibbon, A. O'Hare, and M. Rutherford, "Receptive and expressive prosodic ability in children with high-functioning autism," *Journal of Speech, Language, and Hearing Research*, pp. 1015–1028, Aug. 2007.

[7] D. K. Anderson, C. Lord, S. Risi, P. S. DiLavore, C. Shulman, A. Thurm, K. Welch, and A. Pickles, "Patterns of growth in verbal abilities among children with autism spectrum disorder," *Journal of Consulting and Clinical Psychology*, vol. 75, pp. 574–604, Aug. 2007.

[8] B. Hart and T. R. Risley, *Meaningful differences in the everyday experience of young American children*. Brookes, 1995.

[9] E. Hoff, "The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech," *Child Development*, vol. 74, pp. 1368–1378, Oct. 2003.

[10] A. Fernald, V. A. Marchman, and A. Weisleder, "SES differences in language processing skill and vocabulary are evident at 18 months," *Developmental Science*, vol. 16, pp. 234–248, Mar. 2013.

[11] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen, "Signal processing for young child speech language development," in *Proceedings of the 1st Worksho on Child, Computer, and Interaction*, 2008.

[12] D. Xu, U. Yapanel, and S. Gray, "Reliability of the LENA$^{TM}$ language environment analysis system in young children's natural home environmnet," LENA Foundation, Tech. Rep. LTF-05-02, 2009.

[13] K. J. Rohlfing and G. O. Deák, "Microdynamics of interaction: capturing and modeling infants' social learning," *IEEE Transactions on Autonomous Mental Development*, vol. 5, pp. 189–191, Sep. 2013.

[14] M. H. Goldstein, A. P. King, and M. J. West, "Social interaction shapes babbling: testing parallels between birdsong and speech," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 8030–8035, Jun. 2003.

[15] P. Mundy and M. Crowson, "Joint attention and early social communication: implications for research on intervention with autism," *Journal of Autism and Developmental Disorders*, vol. 27, pp. 653–676, Dec. 1997.

[16] N. Leezenbaum, S. B. Campbell, D. Butler, and J. M. Iverson, "Maternal verbal responses to communication of infants at low and heightened risk of autism," *Autism*, vol. 18, pp. 694–703, Oct. 2013.

[17] E. Hoff, B. Laursen, and T. Tardif, "Socioeconomic status and parenting," in *Handbook of parenting, vol. 2: Biology and ecology of parenting*, M. H. Bornstein, Ed. Lawrence Erlbaum Associates, 2002, pp. 231–252.

[18] R. A. Rescorla and A. R. Wagner, "A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and non reinforcement," in *Classical conditioning II: current research and theory*, A. H. Black and W. F. Prokasy, Eds. Appleton-Century-Crofts, 1972, pp. 64–99.

[19] M. Ramscar, D. Yarlett, M. Dye, K. Denny, and K. Thorpe, "The effects of feature-label-order and their implications for symbolic learning," *Cognitive Science*, vol. 34, pp. 909–957, Jan. 2010.

[20] M. Ramscar, M. Dye, and J. Klein, "Children value informativity over logic in word learning," *Psychological Science*, vol. 24, pp. 1017–1023, Jun. 2013.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*. MIT Press, 1998.

[22] A. Karmiloff-Smith, "Development itself is the key to understanding developmental disorders," *Trends in Cognitive Sciences*, vol. 2, pp. 389–398, Oct. 1998.

[23] A. Weisleder and A. Fernald, "Talking to children matters: early language experience strengthens processing and builds vocabulary," *Psychological Science*, vol. 24, pp. 2143–2152, Nov. 2013.

[24] A. Klin, "Young autistic children's listening preferences in regard to speech: a possible characterization of the symptom of social withdrawal," *Journal of Autism and Developmental Disorders*, vol. 21, pp. 29–42, Mar. 1991.

[25] P. Mundy and A. R. Neal, "Neural plasticity, joint attention, and a transactional social-orienting model of autism," in *International review of research in mental retardation: autism*, L. M. Glidden, Ed. Academic Press, 2000, vol. 23, pp. 139–168.

[26] G. Dawson, K. Toth, R. Abbott, J. Osterling, J. Munson, A. Estes, and J. Liaw, "Early social attention impairments in autism: social orienting, joint attention, and attention to distress," *Developmental Psychology*, vol. 40, pp. 271–283, Mar. 2004.

[27] G. Westermann and E. R. Miranda, "A new model of sensorimotor coupling in the development of speech," *Brain and Language*, vol. 89, pp. 393–400, May. 2004.

[28] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language*, vol. 96, pp. 280–301, Mar. 2006.

[29] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, pp. 793–809, Sep. 2009.

[30] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, pp. 85–117, Jan. 2011.

[31] H. Kanda, T. Ogata, T. Takahashi, K. Komatani, and H. G. Okuno, "Continuous vocal imitation with self-organized vowel spaces in recurrent neural network," in *IEEE International Conference on Robotics and Automation*, 2009, pp. 4438–4443.

[32] P. Li, I. Farkas, and B. MacWhinney, "Early lexical development in a self-organizing neural network," *Neural Networks*, vol. 17, pp. 1345–1362, Oct.-Nov. 2004.

[33] K. Miura, Y. Yoshikwa, and M. Asada, "Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver," *Advanced Robotics*, vol. 26, Apr. 2012.

[34] C. Lyon, C. L. Nehaniv, and J. Saunders, "Interactive language learning by robots: the transition from babbling to word forms," *PLOS ONE*, vol. 7, p. e38236, Jun. 2012.

[35] A. S. Warlaumont, G. Westermann, E. H. Buder, and D. K. Oller, "Prespeech motor learning in a neural network using reinforcement," *Neural Networks*, vol. 28, pp. 64–75, Feb. 2013.

[36] A. S. Warlaumont, "Salience-based reinforcement of a spiking neural network leads to increased syllable production," in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 2013.

[37] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer, "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation," *Frontiers in Cognitive Science*, p. 1006, Jan. 2014.

[38] M. L. McGillion, J. S. Herbert, J. M. Pine, T. Keren-Portnoy, M. M. Vihman, and D. E. Matthews, "Supporting early vocabulary development: what sort of responsiveness matters," *IEEE Transactions on Autonomous Mental Development*, vol. 5, pp. 240–248, Sep. 2013.