Self-production facilitates and adult input interferes in a neural network model of infant vowel imitation

Anne S. Warlaumont[1], Gert Westermann[2], and D. Kimbrough Oller[1]

Abstract. It is well known that greater amounts of adult input facilitate a child's language development. Thus, one might expect that increased amounts of adult input would help an infant learn to accurately imitate the vowels of his/her native language. In addition, an infant's own production of sounds during cooing, babbling, etc. is known to be important to the development of speech abilities. We simulate infant vowel development using a neural network that contains a layer of auditory neurons, a layer of motor neurons, and bidirectional connections linking these perceptual and motor layers. During an initial babbling phase, the system produces random motor activations, hears the acoustic consequences of these motor activations, and adjusts the weights between its auditory and motor layers in a Hebbian fashion. In simulations, passive auditory input from an external "caregiver" is also included during the babbling phase, and is used to update existing auditory-motor connections. In a testing phase, the model is given adult vowels as auditory input and asked to imitate them. Results indicate that self-productions do promote the development of the ability to imitate, but, somewhat counter-intuitively, the more adult input this model receives during babbling, the less accurate its imitations are during test. Explanations and implications of this finding are discussed.[12]

1 INTRODUCTION

Numerous studies have shown that language input from caregivers has a positive effect on language acquisition. For example, a canonical finding is that the number of words a child hears from his/her caregivers predicts later vocabulary size and language test scores [1]. In the phonological domain, research suggests that infants tend to produce sounds that resemble those of the language spoken by their caregivers as opposed to other languages and to produce vocalizations that sound like those they have just recently heard [2-4] (but see [5] for a critical review).

For example, Kuhl & Meltzoff [2] presented 12- to 24-week old infants with recordings of a female adult producing exemplars of a single American-English vowel: /a/, /i/, or /u/. They recorded the cooing vocalizations produced by the infants during this exposure period. The infants' vocalizations were transcribed into broad phonetic categories and it was found that /a/-like vowels tended to correspond to sessions where adult /a/ vowels were played, /i/-like vowels tended to correspond to sessions where /i/ vowels were played, and /u/-like vowels tended to correspond to sessions where /u/ vowels were played. Understanding how this ability to imitate is achieved is important because the ability to imitate is thought to provide an important foundation for language learning in general [6]. It was proposed that two factors drove the observation in [2]: (1) perceptual re-organization based on hearing the auditory input and (2) learning of auditory-motor mappings based on self-production. These two factors were noted to be theoretically separable.

A number of connectionist modeling studies have demonstrated that artificial neural networks are sensitive to external input. Such work has shown how that input can be beneficial from the standpoint of helping the neural network develop language ability, including imitating the sounds of its ambient language. For example, Heintz et al. [7] show that a model consisting of a layer of auditory neurons and a layer of motor neurons, connected to each other by weighted Hebbian connections, can learn to correctly imitate adult vowels. In their model, a training trial consists of jointly presenting acoustic features of an adult vowel such as /i/ with the positions of vocal tract organs, such as the tongue and lips, required for the child to produce that same vowel.

Li, Zhao, and MacWhinney's connectionist word-learning model, DevLex-II [8], also learns the sounds of its language from external input and also contains layers (in their case phonological input, phonological output, and semantic layers) connected by weighted Hebbian connections. During training, the Hebbian weights between the phonological input and the semantic layers are updated in response to simultaneous presentation of phonological and semantic representations and the Hebbian weights between the semantic and the phonological output layers are also updated in response to simultaneous presentation of phonological and semantic representations. In addition to Hebbian weights between layers, each layer also has its phonetic or semantic features updated using a self-organizing map algorithm. Words are presented with frequencies corresponding to those observed in real caregivers' speech. After training, the model is successfully able to comprehend and produce words in its language.

Yoshikawa et al. [9] use a similar neural network architecture but a different training approach to model the development of vowel imitation ability. An auditory self-organizing map and a motor self-organizing map are linked to each other by Hebbian connections. The model is trained by having it produce a random action of a robotic vocal tract. A human "caregiver" judges whether the sound produced by the robot's vocal tract is similar to a vowel in their repertoire. If so, the human imitates the robot, and the first four formant frequencies of the human caregiver's imitation are fed to the model's auditory layer. The Hebbian connections between the auditory and motor layer are then

[1] School of Communication Sciences and Disorders, Univ. of Memphis, 38105, USA. Email: {awarlmnt, koller}@memphis.edu.

[2] Dept. of Psychology, Oxford Brookes Univ., Oxford OX3 OBP, UK. Email: gwestermann@brookes.ac.uk.

updated to reflect the correspondence between the caregiver's production and the child's.

Westermann and Miranda [10,11] show that a model consisting of an auditory and a motor layer, again linked by weighted Hebbian connections but without self-organization of its perceptual and motor nodes' tunings to the external world, can learn to adapt its auditory percepts of vowels to the language-specific input it has heard (it also adapts those same percepts to reflect the auditory correlates of sounds produced during random babbling training trials). A unique feature of this model is that the correspondence between the sensory and motor pairings for a given speech sound is not assumed beforehand. The present study makes this same conservative assumption regarding what information is available to the child, but rather than focusing on changes in perceptual representations resulting from self-production and caregiver input, we focus on changes in imitation ability as a function of self-production and caregiver input. Given that modification of Hebbian auditory-motor connections based on adult input was sufficient to achieve language-specific perceptual reorganization, one might expect the same kind of mechanism to facilitate imitation.

The present study describes a connectionist model of vowel perception and production development. The model is tested on its ability to imitate adult vowels as in [2]. The approach is similar to some of the other connectionist models described above in that it contains an auditory neuron layer connected via Hebbian weights to a motor neuron layer. However, unlike some of the other models that are tested on the ability to imitate adult input, e.g. [7, 9], it makes the more conservative assumption that activations of the model's motor neurons can only be achieved (1) through the action of the model itself and subsequent perception of self-produced vocalizations or (2) through propagation of adult-generated activation on the auditory input layer via Hebbian connections to the motor layer. In other words, our study is novel because we test how well a model can learn to imitate when it is not given any direct information about which of its own motor articulations correspond to the adult targets. We systematically vary the number of adult-input trials to see how much passive adult stimulation acting through existing auditory-motor connections contributes to the model's development of the ability to imitate an adult. We hypothesized that, as [2] suggests, both self-production trials and passive-adult-input trials would contribute to learning.

2 METHOD

2.1 Auditory and motor neural networks

The model architecture is illustrated schematically in Fig. 1. It has two layers of neurons: an auditory layer and a motor layer. The auditory and motor layers are fully interconnected via modifiable weighted connections.

The auditory layer contains 25 neurons. Each node in the auditory layer has a set of weights to each acoustic input feature (relative first and second formants; see the Vowel Synthesis section below). A neuron's set of weights to input features defines the center of the neuron's receptive field; the closer an input gets to the center of the receptive field, the greater the activation of the neuron. An acoustic input activates the auditory neurons by multiplication (dot product) with these weights.

The motor layer contains 100 neurons. Each node in the motor layer has a receptive field defined by its set of weights to each upper vocal tract muscle (see the Vowel Synthesis section below).

A winner-takes-all function is applied to each layer of neurons before allowing its activation to spread to other layers and before making any Hebbian updates to the weights connecting the two layers. This prevents the auditory and motor representations from being heavily biased toward central regions in the input and output spaces, respectively.

During training, when the auditory and motor networks are simultaneously activated, the connection weights between two networks are updated according to the following Hebbian learning with decay rule:

$$W(t+1) = W(t) + \alpha(a \cdot m' - W) \quad (1)$$

where t is the current learning trial, $t+1$ is the next learning trial, W is a matrix representing the weights from each auditory node to each motor node, a is the vector representing the set of auditory neuron activations, m is the vector representing the set of motor neuron activations, and α is a learning rate parameter that starts at .1 and decreases by a factor of .99 on each learning trial until it reaches a minimum value of .01. Weights are initialized to zero at the start of training.

Prior to training, all auditory and motor receptive field weights are set to random uniformly distributed values. For the main model version, these receptive fields remain static throughout the course of training. In alternate model versions, the auditory receptive fields and/or the motor receptive fields are updated with each auditory input or motor production, respectively. This updating is done according using the standard self-organizing map algorithm [12]. The algorithm specifies that neurons in each layer be assigned locations on a square grid. On a given trial, the most activated node as well as its neighbours have their receptive field centers (i.e., their weights to acoustic features or muscle activations) modified to more closely resemble the current acoustic features or muscle activations. Such updates occur before the winner-takes-all function is applied.

2.2 Vowel data

The model simulations rely on a database of 4,022 synthesized vowels and a set of 30 real adult vowels.

The synthesized vowel database was created using the articulatory synthesis and formant and pitch extraction tools available as part of Praat, a free phonetics program [13]. Sounds were generated by randomly varying fourteen upper vocal tract muscle parameters related to the face, mouth, tongue, and pharynx. These were superimposed on a 1-second fixed pattern of lung volume and laryngeal muscle parameters. Praat uses these lung, larynx, and upper vocal tract parameters to define a system of masses and springs that represent the vocal tract boundaries in an adult female. Praat then derives the air pressures in this vocal tract model, which determine the synthesized vocal sound. Fundamental frequency (f_0), first formant frequency (F1), and second formant frequency (F2) traces were estimated for each resulting sound and sounds that did not contain at least 40 consecutive milliseconds where an f_0

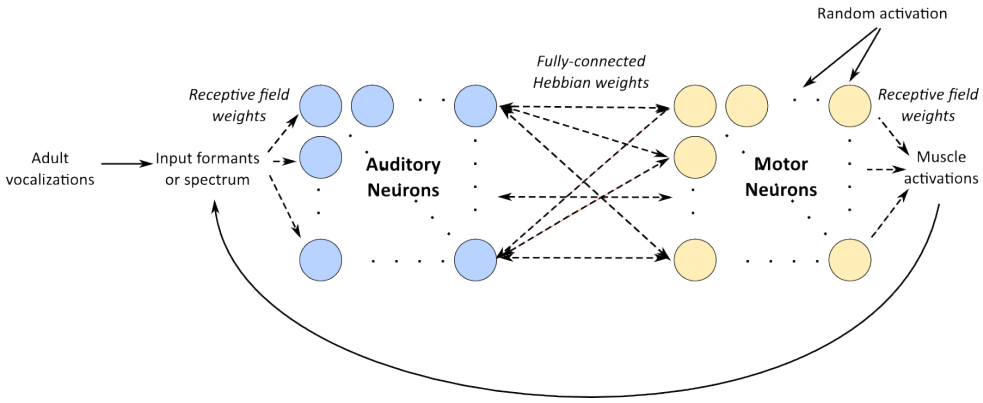

Figure 1. Schematic diagram illustrating the model architecture.

was detectable were discarded. For each remaining sound, we measured the mean F1 minus mean f_0 and mean F2 minus mean F1 over all portions of the sound where there were at least 40 consecutive ms of detectible f_0. Each database entry was thus comprised of a set of 14 muscle activation values and several acoustic measurements on the resulting sound.

Adult sounds consisted of 10 exemplars each of the English /a/ /i/ and /u/ vowels, produced by a female adult American English speaker. F1-f_0 and F2-F1 were obtained for these vowels using the same procedure as for the synthesized sounds. Relative formants were normalized to the combined range observed in the synthesized and human adult data.

2.3 Learning and test trials

Two types of learning experiences are modeled. The first type of learning trial is the infant production trial, which models the infant's experience of exploring his/her own motor capabilities and hearing the resulting sound. An infant production trial begins with a random activation of the model's motor neurons. This specifies a set of upper vocal tract muscle activations. The item in the synthesized vowel database that has muscle activations most similar to those specified by the winning motor neuron's receptive field is identified. The acoustic features associated with that vowel are then presented to the network, where they cause activation of the auditory layer. The auditory neurons are at the same time stimulated by activation propagating from the motor layer through the auditory-motor connection weights. At this point, both the auditory and motor layers of neurons are active, so the connection weights between them are updated according to the Hebbian learning rule described above. This concludes the infant production trial.

The second type of learning trial is the adult input trial, which models the infant's experience of hearing his/her caregiver vocalize. An adult input trial begins by choosing an item at random from the set of adult vowels. The acoustic features of that item are then presented to the model, which causes its auditory neurons to become active. This in turn causes activation to spread through the auditory-motor connection weights to the motor layer. At this point, both layers of neurons are active and their Hebbian connection weights are updated, concluding the adult input trial.

In the present study, different versions of the model were run, each with with differing amounts of adult input. In no-adult-input simulations, there were 500 infant-production training trials. In adult-input simulations there were either 600, 700, or 800 training trials; at each learning trial the probability of that trial being an adult input trial was proportional to the total number of training trials minus 500.

The model is tested on an imitation task. An imitation trial is initiated by presenting the model with acoustic features of an adult vowel. This activates the model's auditory neurons, which, via the auditory-motor connections, activate the model's motor neurons. The synthesized vowel that best matches the pattern of activation at the motor neuron level is then taken as the model's imitation. The Euclidean distance between the acoustic features of the imitated sound and those of the adult sound are then compared. Smaller distances indicate better performance. The model is tested on its imitation of each of the 30 adult vowels.

3 RESULTS

We ran a large number of simulations, systematically varying model parameters, specifically the number of adult input trials given in addition to the infant production trials and whether or not the auditory and motor layers had self-organizing receptive fields.

Prior to any training, it was common for all inputs to result in the same imitation sound, since the weights between the auditory and motor layers are initialized to zero. Across training, the model's ability to accurately imitate adult input improves as evidenced by the imitations' acoustic features becoming more similar to the input vowels' acoustic features. Figure 2 illustrates this change for one of the simulations. Measurement of the mean distance between the target input and the model's imitation in relative formant space corroborates this observation that performance improves with training (see the leftmost column of Fig. 3).

In contrast, increased amounts of adult input had a negative effect on performance. Figure 3 shows this detrimental effect of adult input for model versions in which receptive fields are static throughout training. This effect can be quantified statistically by regressing the change in mean imitation accuracy across training on the number of adult input trials, yielding r = -.268, $t(148)$ = -3.385, $p < .001$. This effect also held when self-organization of auditory and/or motor layers was turned on and when using different acoustic input features, such as spectra.

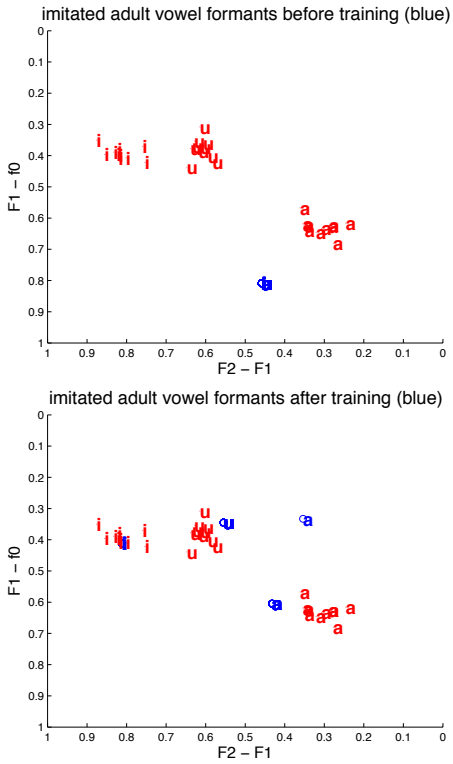

Figure 2. Imitated vowels' normalized formants for one of the model simulations before (above) and after (below) learning. Adult inputs are shown in red and the model's imitations are shown in blue. Letters indicate the adult vowel phone targets.

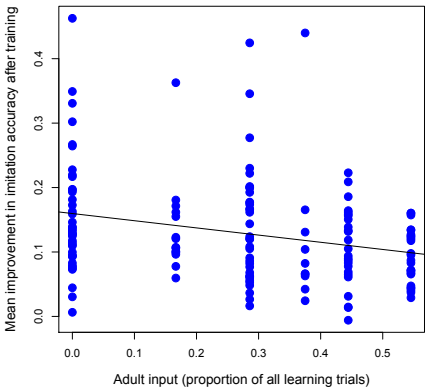

Figure 3. Model performance as a function of amount of adult input during training. Positive values on the y-axis indicate improvement from before training to after training.

4 DISCUSSION

The present study tested the hypothesis that modification of auditory-motor connections based on both self-production and passive adult input would improve performance of a neural network model on a vowel imitation task.

Results indicate that learning from self-productions is important to the model's development of imitation ability. This implies that random motor exploration and perception of the auditory correlates of that motor exploration can be a powerful driver of learning. An implication is that findings of infant vowel imitation in early infancy [2] may be explainable in large part on the basis of mappings achieved during self-production.

On the other hand, we found that modification to auditory-motor connections based on external inputs where the exact motor correspondence is unknown interferes with imitation performance. Given the numerous previous studies such as those reviewed in the Introduction finding that adult input plays a facilitative role in bringing children's language closer to that of their native language, our finding that adult input is associated with worse imitation accuracy is surprising.

One possible explanation is rooted in the fact that imitation in our model is a reinterpretation of the input stimulus within the developed system's own learned sensorimotor mappings. Every infant production trial provides by its nature the completely veridical mapping from motor representation to acoustic representation. In contrast, since adult input in this model does not accompany a known motor representation, adult input may amplify any errors in the model's current mappings. Thus, the present results show that the assumption made by other models [7-9] that the child knows the motor origins of the behavior it observes from a caregiver is nontrivial. Such an assumption makes a difference to performance, so its biological plausibility should be considered.

Since adult input is known to facilitate language learning but does not show such an effect in our model, what mechanisms could underlie its role in real children's language development? One possibility is that passive exposure to adult input affects learning not through modification of the auditory-motor connections but through reorganization of the perceptual system alone, e.g., through adjustment of receptive fields in the auditory system as shown by [14] and modeled in [15].

That being said, adult input effects on perception are not as strong for pre-recorded stimuli [14] and passive TV viewing is associated with reduced rates of language acquisition [16]. Since the TV does not respond differentially to child productions compared to caregiver inputs, which adapt dynamically to the state and abilities of the infant [17,18], the experience of an infant hearing speech on TV might be more like our model's experience hearing adult input. Thus, the finding here that adult input is not associated with increased language performance might not reflect merely a problem with the model but could potentially reflect how an infant might be expected to be affected by passive, non-contingent/non-adaptive input such as that from a TV or radio, especially when such exposure reduces the frequency of the infant's own vocal productions.

Another possibility is that the value of adult input is in actively reinforcing the infant and/or directing the infant's future motor exploration. Reinforcement may help an infant determine when to update neuronal connections, perhaps only updating connections that produce accurate imitations of an adult or updating connections when an adult has imitated the infant and so perceptual activation reflects both the self-vocalization and the caregiver's vocalization, as in Yoshikawa et al.'s model [9]. With regard to shaping exploration, in the model presented here as well as in [7, 9-11], motor activations are drawn completely at random and the entire range of possible motor activations is

covered. The real infant, however, likely starts with a limited repertoire of vocal productions and expands on this. The direction of expansion could presumably be driven by auditory priming from adult input as well as by feedback in the form of perceptual, social, or other rewards [17-19].

Future computational modeling studies should expand on the foundations supplied by this and the handful of other neural network models of infant vocal imitation, to further explore various mechanisms by which external (i.e., adult) input might shape infant vocal development. For example, perhaps by modifying the model's perceptual representations of speech sounds but not modifying its auditory-motor connections, passive external input could improve performance. In another scenario, perhaps differential reinforcement of the model's productions might be used to adjust the amount of sensorimotor learning on a given trial or to influence where the model concentrates its motor exploration.

ACKNOWLEDGEMENTS

This study was supported by a U.S. Dept. of Energy Computational Science Graduate Fellowship (DE-FG02-97ER25308) and by the Plough Foundation. Thanks to Rick Dale, Eugene Buder, and three reviewers for helpful discussion and feedback.

REFERENCES

[1] B. Hart and T. R. Risley, *Meaningful differences in the everyday experience of young American children*. Baltimore: Brooks (1995).

[2] P. K. Kuhl and A. N. Meltzoff. Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, 100:2425-2438 (1996).

[3] B. Mampe, A. Friederici, A. Christophe, and K. Wermke. Newborns' cry melody is shaped by their native language. *Current biology*. 19:1994-1997 (2009).

[4] Bénédicte de Boysson Bardies and M. Vihman. Adaptation to Language: Evidence from Babbling and First Words in Four Languages. *Language*, 67:297-319 (1991).

[5] D. K. Oller and R. E. Eilers. Interpretive and methodological difficulties in evaluating babbling drift. *Parole*, 7/8:147-164 (1998).

[6] G. E. Speidel and K. E. Nelson, Eds. *The Many faces of imitation in language learning*. New York: Springer-Verlag (1989).

[7] I. Heintz, M. Beckman, E. Fosler-Lussier, and L. Ménard. Evaluating parameters for mapping adult vowels to imitative babbling. In: *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, Brighton, UK (2009).

[8] P. Li, X. Zhao, and B. Mac Whinney. Dynamic Self-Organization and Early Lexical Development in Children. *Cognitive Science*, 31:581-612 (2007).

[9] Y. Yoshikawa, M. Asada, K. Hosoda, and J. Koga. A constructivist approach to infants vowel acquisition through mother-infant interaction. *Connection Science*, 15:245-258 (2003).

[10] G. Westermann and E. R. Miranda. Modelling the Development of Mirror Neurons for Auditory-Motor Integration. *Journal of New Music Research*, 31:367-375 (2002).

[11] G. Westermann and E. R. Miranda. A new model of sensorimotor coupling in the development of speech. *Brain and Language*, 89: 393-400 (2004).

[12] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78:1464-1480 (1990).

[13] P. Boersma and D. Wennink, *Praat: doing phonetics by computer*. Available: http://www.fon.hum.uva.nl/praat/ [Accessed Apr. 4, 2010].

[14] P. K. Kuhl, F. Tsao, and H. Liu. Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America*, 100:9096-9101 (2003).

[15] F. Guenther and M. Gjaja. The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*. 100:1111-1121 (1996).

[16] F. J. Zimmerman, J. Gilkerson, J. A. Richards, D. A. Christakis, D. Xu, S. Gray, and U. Yapanel. Teaching by listening: the importance of adult-child conversations to language development. *Pediatrics*, 124:342-349 (2009).

[17] J. Gros-Louis, M. J. West, M. H. Goldstein, and A. P. King. Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, 30:509-516 (2006).

[18] G. Moran, A. Krupka, A. Tutton, and D. Symons. Patterns of maternal and infant imitation during play. *Infant Behavior and Development*, 10:477-491 (1987).

[19] M. Goldstein, A. King, and M. West. Social interaction shapes babbling: testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America,* 100:8030-8035 (2003).