Neural Networks 38 (2013) 64–75

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Prespeech motor learning in a neural network using reinforcement

Anne S. Warlaumont [a,*], Gert Westermann [b], Eugene H. Buder [c], D. Kimbrough Oller [c]

[a] Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Rd., Merced, CA 95343, USA

[b] Department of Psychology, Lancaster University, Lancaster LA1 4YF, UK

[c] School of Communication Sciences and Disorders, University of Memphis, 807 Jefferson Ave., Memphis, TN 38105, USA

ARTICLE INFO

Article history:
Received 5 January 2012
Revised and accepted 13 November 2012

Keywords:
Infant vocalization
Motor development
Neural network
Reinforcement
Neuromuscular control
Articulatory speech synthesis

ABSTRACT

Vocal motor development in infancy provides a crucial foundation for language development. Some significant early accomplishments include learning to control the process of phonation (the production of sound at the larynx) and learning to produce the sounds of one's language. Previous work has shown that social reinforcement shapes the kinds of vocalizations infants produce. We present a neural network model that provides an account of how vocal learning may be guided by reinforcement. The model consists of a self-organizing map that outputs to muscles of a realistic vocalization synthesizer. Vocalizations are spontaneously produced by the network. If a vocalization meets certain acoustic criteria, it is reinforced, and the weights are updated to make similar muscle activations increasingly likely to recur. We ran simulations of the model under various reinforcement criteria and tested the types of vocalizations it produced after learning in the different conditions. When reinforcement was contingent on the production of phonated (i.e. voiced) sounds, the network's post-learning productions were almost always phonated, whereas when reinforcement was not contingent on phonation, the network's post-learning productions were almost always not phonated. When reinforcement was contingent on both phonation and proximity to English vowels as opposed to Korean vowels, the model's post-learning productions were more likely to resemble the English vowels and vice versa.

1. Introduction

1.1. Human infant vocal development

During the first year of life, human infants make considerable progress in learning to produce speech-like sounds. One of the first achievements in prelinguistic vocal development is acquiring the ability to control phonation, producing voiced sounds at will (Oller, 2000). Basic modal phonation is so readily produced by a healthy adult that its complexities may easily be overlooked. In fact, phonation involves active settings of a number of muscles that contribute to the positions, compressions, and stresses in the tissues of the larynx (Titze, 1994). To further complicate things, it has recently become clear that the larynx and the upper vocal tract interact nonlinearly (Titze, 2008). How infants learn to control this system in order to support phonation is an open question.

Soon a number of other milestones are achieved, such as expansion of the range of pitches, durations, and vocal qualities produced, and the emergence of syllabic consonant–vowel timing (Koopmans-van Beinum & van der Stelt, 1986; Oller, 2000; Oller & Lynch, 1992; Stark, 1980). Toward the end of the first year of life, infant vocalizations have been reported to begin to show adaptation to the phonetic characteristics of the particular language environment as opposed to those of other languages (de Boysson-Bardies, Halle, Sagart, & Durand, 1989; de Boysson-Bardies & Vihman, 1991). For example, a study by de Boysson-Bardies et al. (1989) of 10-month-old infants from monolingual French, English, Cantonese, and Arabic speaking households compared the vowel sounds produced during canonical babbling by each infant to the vowels and their frequencies in adult speech in the household language. The study found that mean first and second formant frequencies of vowels produced by infants were significantly different across language backgrounds and that the patterns of differences matched those estimated for adult speech for the four languages. The results were taken as evidence that a child's language environment influences the range of movements of the infant's articulators, particularly the tongue and lips, supporting the development of the vowel system of the target language.

* Corresponding author. Tel.: +1 6072273726; fax: +1 4847273726.
E-mail addresses: anne.warlaumont@gmail.com, awarlaumont2@ucmerced.edu (A.S. Warlaumont), g.westermann@lancaster.ac.uk (G. Westermann), ehbuder@memphis.edu (E.H. Buder), koller@memphis.edu (D.K. Oller).
URLs: http://www.annewarlaumont.org/ (A.S. Warlaumont), http://www.psych.lancs.ac.uk/people/gert-westermann (G. Westermann), http://profiles.memphis.edu/ehbuder (E.H. Buder), http://profiles.memphis.edu/koller (D.K. Oller).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.11.012

1.2. Reinforcement in early vocal development

The human infant develops within a social environment of interaction with parents and other adults and children. For this reason, the developing social brain has recently become a focus in infancy research (e.g., Blakemore, 2010; Grossmann & Johnson, 2007). Speech production development is one of the many behaviors that develops in the context of and is shaped by social interaction. Caregivers direct vocalizations (such as acknowledgments, imitations, playful vocalizations, and object labels) toward their infants as well as smiling at, looking at, and touching their infants. These caregiver behaviors, particularly the vocal ones, are modulated in response to infants' vocalization behaviors (Gros-Louis, West, Goldstein, & King, 2006; Papoušek & Papoušek, 1989) and they serve as reinforcers to the infant: experimental work has shown that contingency of maternal responses on infant vocalization leads to increased infant rates of vocalizing (Goldstein, King, & West, 2003).

In addition to social sources, reinforcement may also come from internal sources. For example, the high auditory salience of a self-produced sound or its matching to the infant's auditory preferences may function as reinforcers. It is likely that auditory salience and preference are influenced both by innate factors and by exposure to ambient language input. Salience-based reinforcement-learning, though it has not been addressed in research on development of vocalization abilities, has been shown to be feasible in a non-neural-network computational model of eye movements for joint attention (Lewis, Deák, Jasso, & Triesch, 2010). Whether it originates from social sources or from internal preferences, the idea is that positive reinforcement for producing speech-like vocalizations facilitates the development and increased usage of the reinforced vocalizations, consistent with the principles of operant conditioning (Domjan, 2010).

Functionally, positive reinforcement provides an agent with feedback that its vocalization was on the right track, without directly indicating what the motoric target is. It is useful to compare reinforcement-based learning to two other types of learning, unsupervised self-organization (e.g., in learning by Kohonen maps and Hebbian networks) and supervised learning (e.g., utilized by feed-forward and simple recurrent networks that learn via the delta rule and backpropagation). On the one hand learning from reinforcement does, unlike unsupervised learning, rely on the model's receiving feedback about how well it performed. However, this feedback is not as targeted as in supervised learning in that the exact desired modeled behaviors are not assumed to be known by the entity providing the feedback.

Reinforcement-based learning is suitable for situations where the optimal behavioral or motoric output is unknown, as when a modeler or roboticist wishes to make a realistic synthesizer produce certain types of sounds. Infants also may not have direct access to the correct motor configurations for producing target vocalizations, so reinforcement from caregivers or the infants' own learned or innate auditory preferences may serve as useful guides in the infants' learning to produce vocalizations of a given type.

1.3. Previous vocal development models

Additional mechanisms likely also play important roles in learning to produce speech-like sounds. One proposal is that adaptations of infant vocalizations to the ambient language result from self-organized perceptual and perceptual-motor learning. For example, it has been argued that by monitoring their own vocalizations, infants learn sensorimotor mappings that enable them to reproduce sounds heard from others (Kuhl & Meltzoff, 1996; Vihman, 1993). Most computational neural network modeling work to date has focused on this mechanism and not on reinforcement (note that the two are not mutually exclusive) (Guenther, Ghosh, & Tourville, 2006; Heintz, Beckman, Fosler-Lussier, & Ménard, 2009; Oudeyer, 2005; Warlaumont, Westermann, & Oller, 2011; Westermann & Miranda, 2004; Yoshikawa, Asada, Hosoda, & Koga, 2003).

The DIVA (Directions Into Velocities of Articulators) model (Guenther et al., 2006; Guenther, Hampson, & Johnson, 1998) focuses on self-organizing synaptic mappings between auditory, somatosensory, and motor brain regions. The DIVA model is assumed to have knowledge about which specific vowels and consonants exist in its language and their acoustic properties (for example, the first three formant frequencies). During a "babbling" phase, the model randomly moves its articulators, i.e., its tongue, jaw, and lips. Learning consists of updating the synaptic mappings between the motor and sensory cortices to reflect the associations between articulatory motor commands and their somatosensory and auditory consequences discovered during the babbling experience. When the model's random movements happen to produce a synthesized sound that corresponds acoustically to a sound in its language, the synaptic mappings from a premotor speech sound layer to motor cortex and to sensory cortices are also updated. The effect is that future activation of the speech sound simultaneously activates the appropriate motor commands and inhibits the appropriate auditory and somatosensory expectations. The inhibition of auditory and somatosensory regions enables the model to detect if there is any error in its production of the sound and if so to make appropriate motor corrections.

The DIVA model is the most comprehensive and well-tested model of human speech sound learning to date. It has been compared to adult fMRI data and has been used to model normal adult performance under various experimental manipulations, differences in hearing impairment and stuttering, and robustness in the face of developmental changes across childhood in the size and shape of the vocal tract (Callan, Kent, Guenther, & Vorperian, 2000; Guenther et al., 2006; Max, Guenther, Gracco, Ghosh, & Wallace, 2004; Perkell et al., 2007). However, there are a number of aspects of early vocal learning that it has not yet addressed. For one, it has not yet been used to model self-initiated behavior; instead speech sounds are activated directly by the modeler (Guenther et al., 2006). Relatedly, it does not directly address the role that reinforcement might play in shaping spontaneous vocal behavior. Finally, it does not address phonatory learning, i.e. learning to produce voicing and learning to control the pitch, amplitude, etc. of vocalizations, despite this being a major aspect of early speech development.

Several other models, narrower in scope than the DIVA model, aim to explain how infants might learn to imitate vocalizations produced by others via Hebbian learning of perceptual-motor connections (Heintz et al., 2009; Warlaumont et al., 2011; Yoshikawa et al., 2003). These models each consist of two layers of neurons, one auditory and one motor, with weighted connections between the two layers. As in the DIVA model, learning in these models involves having the model produce random motor outputs and determining vocal tract configurations, which in turn determine the acoustics of synthesized vocalizations. In Yoshikawa et al. (2003), each model production is then imitated by a human adult, and sensorimotor connections are updated in a Hebbian fashion so as to link the acoustics of the adult imitation to the motor outputs of the model. After training, adult vowels can be input and the model produces correct vowel imitations. In Heintz et al. (2009) and Warlaumont et al. (2011), learning from model productions is based on Hebbian associative learning between the acoustics of the model's own vocalization and its motor outputs. In addition to learning based on self-production, these models include passive listening events, in which the model receives external auditory input, as if from a caregiver, and the model self-organizes its perceptual receptive fields and/or its Hebbian perceptual-motor

connections as a result. However, in these models, the utility of such passive learning from adult input for improving imitation accuracy has not been established, although in a similar model by Westermann and Miranda (2004) it has been shown that such adult input does produce ambient language effects on perceptual representations. Presumably even if passive perceptual input does not produce improvements in imitation accuracy, it is possible that were the post-learning spontaneous vocalizations of these models to be explored, ambient language effects of the sort shown in the literature on human infants might be observed. This possibility has not yet been examined.

Kanda, Ogata, Takahashi, Komatani, and Okuno (2009) have also addressed learning to produce the vowels of a given language. Their model is a recurrent neural network with parametric bias (RNNPB). In a first phase of learning, inputs are sequences of adult vowels. The model is trained to predict, on the basis of the acoustics and corresponding motor parameters at the current and previous time steps, the acoustics and corresponding motor parameters that will be input at the next time step. After this first phase of training, the model is able to segment sequences of vowels based on where prediction errors are highest. In a second phase of learning, the model learns to represent segmented vowels as constant values of two "parametric bias" neurons. After this second phase of learning, the parametric bias neurons can be activated by the modeler and the network accurately produces the correct vowels. Although the model performs well on segmentation, recognition, and production tasks, its plausibility is questionable. It is assumed that during training the model knows, for each adult vowel, both its acoustic parameters and the precise articulatory motor parameters that generated the vowel. Such an assumption is consistent with Liberman and Mattingly's motor theory of speech perception (Liberman & Mattingly, 1985), which posits that from birth, infants' perception of speech sounds is innately linked with the articulatory gestures that produce those speech sounds. However, whether infants innately, without any prior learning, possess direct access to the precise motor commands they would need in order to produce a sound that they hear someone else in their environment has produced is a strong assumption, especially given the fact that infants do not at birth or even within the first few months of life produce vocalizations that sound like speech, except perhaps accidentally (Oller, 2000).

Other work by Oudeyer (2005) is unique in that it does explicitly address ambient language effects on spontaneous vocalizations. The model consists of multiple agents, each with a layer of auditory neurons connected to a layer of motor neurons that in turn connect to three articulatory parameters: lip rounding, tongue height, and tongue position. At each iteration, an agent is randomly chosen and its motor neurons are randomly activated. The agent adjusts, in a self-organizing manner, its neuro-articulator weights as well as the connection weights between the two layers. The topographically closest neighbor hears the first agent's vocalization, has activation propagated from its auditory to its motor layer and then also updates its neuro-articulator weights. In this way, the second agent becomes more likely to spontaneously produce sounds similar to those of the first agent. The model provides an impressive demonstration of how self-organized learning and interaction among agents can affect clustering of the vowel space as well as adaptation of vocal productions to others in the environment. However, by design it does not include modeling of either social or intrinsic reinforcement effects. Also, like the other models, it does not address phonatory learning.

Thus, despite the insights obtained from previous work, many aspects of early vocal motor learning in human infancy remain to be modeled. For one, reinforcement has not been incorporated, despite its important role in the empirical human infancy literature. Second, the focus has been on responses to caregiver vocalizations or production of given sequences of phones and has rarely (an exception being Oudeyer, 2005) addressed spontaneous productions. Third, previous work has focused heavily on learning vowel sounds and has not addressed development of control over phonation, which is also an important aspect of speech production. In the present study, we introduce a neural network architecture that addresses each of these three aspects of early vocal motor learning.

1.4. Our model

Our model consists of a topographically organized layer of neurons that control a physiologically realistic vocalization synthesizer (Boersma, 1998; Boersma & Wennink, 2010) via neuromotor connections. During learning, the model explores its vocalization capabilities. If and only if it produces a vocalization that is reinforced, its neuromotor connections are updated to reflect its current neuronal and muscle activations. This dependence of learning on reinforcement is consistent with neurophysiological findings that learning in motor cortex is modulated by dopamine, a neurotransmitter strongly associated with reinforcement (Molina-Luna et al., 2009). Updating of neuromuscular weights follows the learning procedure for the self-organizing map (Kohonen, 1990), a popular type of neural network consisting of a layer of neurons with topographically-organized receptive fields that adapt to the environment. The topographic organization corresponds to the topographic organization observed throughout the brain.

The combination of self-organizing topographic map learning and reinforcement gating represents a novel neural network modeling approach. Note that the approach has a different emphasis from most computational reinforcement learning work such as that focusing on temporal difference learning and related methods (Sutton & Barto, 1998). For example, we do not consider reinforcement that is delayed. Another difference is that in our model the primary function of reinforcement is to gate the learning of neuromotor connections. While reinforcement learning systems have been developed that use neural networks for processing sensory inputs, only a few attempts have been made to make neural networks that use reinforcement to learn how to produce behavioral outputs (Barto, 1995; Izhikevich, 2007). Those that have have shown promising results, but have not to our knowledge used the self-organizing map network or addressed problems of learning to produce complex motor output patterns such as controlling over a dozen muscles as is done here. The benefit of integrating reinforcement into neural networks is that, if successful, it could extend the application of self-organized neural network learning to problems of motor learning. In perceptual learning, self-organizing processes can take advantage of statistical regularities in the sensory environment in order to learn structured representations. In motor learning that is driven by random exploration, however, purely self-organized learning is of limited value since there are not statistical regularities in the motor productions — they are produced at random. Using reinforcement to gate learning is a simple modification of self-organized learning that allows the self-organized learning process to take advantage of statistical regularities in the motor space with regard to what actions lead to reward compared to what actions do not.

2. Method

2.1. Vocalization synthesis and analysis

All of the simulations in the present study used Boersma's articulatory speech synthesizer, implemented in Praat, a free speech analysis and synthesis software (Boersma, 1998; Boersma & Wennink, 2010). The synthesizer consists of a model of the human vocal tract, including the lungs, trachea, larynx, pharynx, oral cavity, and

nasal cavity. The walls of the vocal tract are modeled as coupled mass-spring systems. The synthesizer includes several options for the number of masses used in modeling the vocal folds; for the present study, we used the default two-mass option. The synthesizer also offers three sizes of vocal tract: adult female, adult male, and child; we used the default adult female version, since our target vowel acoustic measurements came from a study of adult female speakers. Based on the volume of air in the lungs and the activation of laryngeal and upper vocal tract (i.e., pharynx, oral cavity, and nasal cavity) muscles, specified by the user, the synthesizer calculates the positions and mechanical parameters of the vocal tract walls and the air pressures at each section of the vocal tract over time. The fluctuating air pressure at the mouth determines the synthesized sound. An advantage of using this synthesizer over the synthesizers used in most previous models of infant vocal development is that it allows for motor control of the larynx to be modeled, which is necessary for phonatory development to be addressed.

For this study, all synthesized sounds lasted 0.5 s. Similar to the example given by Boersma (1998), the Lungs parameter, which represents the speaker's lung volume, was set to 0.2 at time 0 s and to 0 at time 0.1 s (−0.5 corresponds to maximum exhalation and 1.5 corresponds to maximum inhalation). The activations of twenty muscle parameters, listed in Table 1, varied across vocalization events according to the procedures described below. Within a vocalization event muscle activations were static, i.e. there was no intra-vocalization variation. How each muscle's activation for a given vocalization event was determined is described below in Section 2.3.

The synthesized sounds were analyzed automatically, also in Praat, to get estimated measures of fundamental frequency (f_0) and first and second formant frequencies (F1 and F2) at 250 ms after the start of vocalization synthesis. When Praat could not identify an f_0 at this time in the sound, which tends to happen when the synthesized sound is silent or breathy but lacking phonation, then the f_0 was considered undefined. Ellis's RASTAMAT toolbox (Ellis, 2007) was used to convert frequencies from Hz to mel, as the nonlinear mel scale better reflects the frequency scaling of the human auditory system. f_0, F1, and F2 were the quantities that determined whether or not a given vocalization was reinforced, as described in Section 2.5 below.

Table 1
The vocal tract synthesizer muscles controlled by the neural network. Laryngeal muscles are those mainly involved in phonation and articulatory muscles are those mainly involved in controlling the shape of the upper vocal tract.

Muscle number	Name	Grouping
1	Interarytenoid	Laryngeal
2	Cricothyroid	Laryngeal
3	Vocalis	Laryngeal
4	Thyroarytenoid	Laryngeal
5	Posterior Cricoarytenoid	Laryngeal
6	Lateral Cricoarytenoid	Laryngeal
7	Styloglossus	Articulatory
8	Masseter	Articulatory
9	Upper Tongue	Articulatory
10	Lower Tongue	Articulatory
11	Orbicularis Oris	Articulatory
12	Vertical Tongue	Articulatory
13	Transverse Tongue	Articulatory
14	Levator Palatini	Articulatory
15	Risorius	Articulatory
16	Genioglossus	Articulatory
17	Hyoglossus	Articulatory
18	Mylohyoid	Articulatory
19	Lateral Pterygoid	Articulatory
20	Buccinator	Articulatory

2.2. Neural network architecture

The neural network contained 25 neurons arranged on a 5 × 5 grid. Each neuron had a spatial location defined by (x, y) coordinates (see Fig. 1) and each neuron had modifiable connection weights to each of the twenty muscles. The connection weights from each neuron to the set of all muscles determined a specific state of the synthesizer's vocal tract. In turn, each vocal tract state was associated with a synthesized vocalization for which f_0, F1, and F2 traces could be automatically estimated (although they could be undefined at the measured point in time).

2.3. Learning

Prior to learning, the neurons' connection weights to the vocal tract muscles were chosen from a uniform random distribution ranging between 0 and 1. Each simulation had 1000 learning

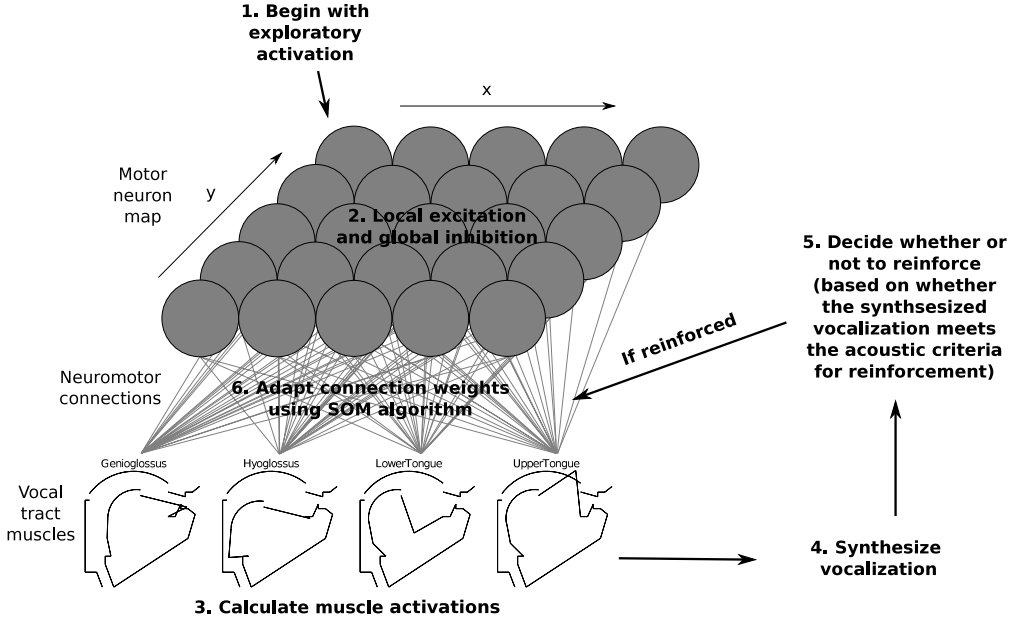

Fig. 1. Schematic diagram of the neural network model.

events, each of which corresponded to a discrete time step. A learning event began by randomly activating the motor neurons in an exploratory fashion. The extent of this random exploration depended on whether the previous vocalization event had been reinforced. If the model had not received reinforcement on the previous event, its activation was drawn from a uniform random distribution ranging from zero to one. Alternatively, if the model had indeed received reinforcement for its previous vocalization, instead of resetting the neuronal activations, a small amount of noise, ranging from −0.25 to 0.25 was added to the previous learning event's neuron activations, subject to the constraint that the resulting activations had to remain between 0 and 1. Thus, if the previous vocalization had been reinforced, exploration was more precisely targeted. At this point, the most active node (i.e. the node with the highest activation value as determined by the procedure just described) and its closest neighbors were identified and local excitation and lateral inhibition was effected as follows: The most active neuron had 2, 3, or 4 closest neighbors depending on whether it was located on a corner, on an edge, or on the interior of the motor neuron grid, respectively. These neurons remained excited. All other neurons besides that most active neuron and its 2–4 neighbors were inhibited by setting their activations to zero.

Activation was then propagated from the neurons to the muscles. Muscle activations were given by

$$\overline{m} = \frac{\overline{a}}{\sum \overline{a}} W + \overline{n}$$

where $\overline{m}$ is a row vector representing the activation level of each vocal tract muscle, $\overline{a}$ is a column vector representing the activation of each neuron, and W is a matrix giving the connection weights from each neuron (in rows) to each muscle (in columns). Thus, muscle activations were a function of the normalized neuron activations propagated through the weighted neuromuscular connections. The $\overline{n}$ is Gaussian noise added at the muscular level, intended to model incidental variation in the shape of the vocal tract. Such variation would correspond to changes in infants' vocal tract positioning due to feeding, mouthing of objects (Fagan & Iverson, 2007), or postural stabilization. Vocal-tract-level "noise" facilitated broad exploration by the model of its full range of vocal capabilities. As with the exploratory activation at the neuron level, noise at the muscular level was dependent on whether the previous vocalization had been reinforced. Muscular noise was more restricted if the model had previously been reinforced, having a standard deviation of 1 if the previous vocalization had not been rewarded and a standard deviation of 0.25 if it had. After the muscle activations for the current event were determined, a vocalization corresponding to those activations was synthesized, the vocalization's f_0, F1, and F2 were estimated, and based on these values it was determined whether the network would receive reinforcement for the vocalization event. The specific acoustic criteria for reinforcement are described in more detail in Section 2.5.

If no reinforcement was given, the event concluded without any changes to the network weights. However, if reinforcement was in fact given, the weights from the motor neurons to the vocal tract muscles were modified according to a self-organizing map algorithm (Kohonen, 1990),

$$W_{p,t+1} = \begin{cases} W_{p,t} + \alpha(\overline{m} - W_{p,t}) & \text{if } \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} \leq \theta \\ W_{p,t} & \text{otherwise,} \end{cases}$$

where $W_{p,t}$ gives the connection weights from neuron p to the vocal tract muscles at the time of the current event, (x, y) are the coordinates on the motor neuron map for a given neuron, and q is the most active motor neuron. α is the learning rate and was set to 0.8 for the simulations presented here, based on pilot work indicating that compared to smaller learning rates (e.g. $\alpha = 0.2$) there was no substantial difference in performance other than slower learning with the latter. θ is the size of the learning neighborhood and was set to 1. In other words, the neuromotor connection weights were adjusted so that muscle activations similar to those just produced would be more likely to be produced on subsequent events. At this point, the learning event was complete.

2.4. Performance evaluation

At the beginning and end of each simulation, we tested the network to see what kinds of vocalizations it would spontaneously produce. Each simulated network was made to vocalize 25 times in the same manner as in training except that no reinforcement was ever provided and there was no noise added at the muscular level. The muscular-level noise was left out in order to provide a clear view on what the network learned at the neural level.

2.5. Reinforcement criteria

Seven different reinforcement conditions were evaluated, with the goal being to compare the sounds produced and the neural representations developed across the different conditions. We ran 50 simulations for each reinforcement condition.

In the first condition, reinforcement was always given, no matter what the network produced. In the second condition, reinforcement was given if the sound produced by the model had a defined f_0 at time 0.25 s which had the effect of reinforcing voiced (i.e. phonated) but not unvoiced (e.g., silent or breath-only) sounds. Although the reinforcement criterion is quite simple, the act of phonation involves coordination of a number of muscles (see Table 1) in order to cause vibration in a nonlinear system of laryngeal tissues (Buder, Chorna, Oller, & Robinson, 2008; Titze, 2008).

In the third condition, in order to be reinforced the model's vocalization had to not only be phonated (operationalized as having defined f_0) but also had to be similar to one of thirteen American English vowels. Similarity to a vowel was operationalized as Euclidean distance in the two-dimensional space defined by F1–f_0 and F2–F1. Most previous efforts to characterize vowels quantitatively have focused on fundamental, first, second, and sometimes third formant frequencies, with this method of differencing bark-scaled or log values (the mel scale is similar to the bark scale) having precedent in studies of both human vowels and vowels produced by articulatory synthesizers (Heintz et al., 2009; Johnson, 2005). The model had to become increasingly similar to one of the vowels, or else fall within a threshold degree of similarity, in order to be reinforced. The threshold degree of similarity was 3 mels (in other words, the target region around a vowel was a circle with a 3 mel radius). Throughout training, a record was kept, for each American English vowel, of the top ten model vocalizations that were closest to that American English vowel. The increasingly similar criterion for reinforcement was defined such that on a given trial, the model's production, if it did not fall within the 3-mel radius of an American English vowel, had to at least be closer to one of the American English vowels than one of the top ten previous model vocalizations.

The fourth condition was the same as the third except that ten Korean (instead of English) vowel targets were used. The American English and Korean vowel targets were taken from a prior study of vowels produced by adult female native speakers of the two languages (Yang, 1996).

The fifth reinforcement condition was the same as the third except that instead of all American English vowels being targeted, only the vowel /a/ was reinforced. The sixth and seventh conditions were the same as the fourth except that the individual target vowels were /e/ and /u/, respectively. Focusing on single vowel targets

allowed us to see how well the model can learn to produce specific vowels, and to clearly visualize the effects of reinforcement.

The reinforcement in any of these conditions could potentially model both extrinsic and intrinsic reinforcement. An example of extrinsic reinforcement in this domain would be a parent preferentially responding to voiced sounds as opposed to very quiet sounds or silence (parents almost certainly do respond contingently in this way, as discussed in Section 1.2). Similarly, parents may respond contingently to vowel sounds that sound like those in their own language(s), especially when they interpret the child's sound as a word. An example of intrinsic reinforcement would be when a child is made happy, engaged, curious, or some other positive emotion when they produce a sound as opposed to silence — this seems highly likely since the production of voiced sounds at the larynx will stimulate both the auditory system and the somatosensory system. With regard to vowels, intrinsic reinforcement is also possible. In our model, reinforcement for production of vowels from a specific language could correspond to the satisfaction or interest generated in a child when they produce a sound that they recognize as corresponding to sounds they have often heard others, such as their caregivers and siblings, produce. Regardless of the source of reinforcement, the same mechanism of reinforcement-gated learned utilized by our model could be at play.

We tested both phonatory and articulatory performance after learning. Greater post-learning tendency both to produce voiced sounds and to produce sounds resembling the target language would indicate generalizability of the reinforcement-gated self-organized learning approach. In particular, it would show that the model can, using a single learning mechanism, simultaneously learn two foundational speech skills, phonation and vowel articulation.

3. Results

3.1. Phonation before and after learning

As shown in Fig. 2, before any learning, across simulations the mean number of vocalizations that had identifiable f_0 was approximately 5 (out of a possible 25). When the model was reinforced at every trial, regardless of phonation, the mean number of vocalizations with identifiable f_0 after learning was only 2.28. For the various reinforcement conditions where reinforcement was contingent on phonation, the mean number of vocalizations with identifiable f_0 after learning ranged between 20.2 and 24.5. For each reinforcement condition, the difference between the number of sounds with f_0 before versus after training was highly significant, $p < 0.001$. This indicates that when reinforcement was contingent on voicing (i.e. phonation), the model learned to reliably produce sounds that were clearly voiced (not silent or purely breathy). When reinforcement was given all the time, without regard to voicing, the model's production of sounds that were voiced actually decreased after learning.

Fig. 3 illustrates, for one of the networks that was reinforced for any sound with identifiable f_0, the sounds that were produced before and after learning when each of the 25 neurons was activated in isolation. As can be seen in Fig. 3, most of the spectrograms of sounds produced by the neurons before training showed little acoustic energy and were essentially silent. In contrast, after learning, almost all of the neurons produced sounds with high acoustic energy, indicating that the network learned to produce audibly voiced sounds, that is, to phonate. It can also be seen that there was a range of durations, spectral qualities, and amplitudes: apparently, the simple requirement of defined f_0 at time 0.25 left opportunity for the neural network to develop representations for

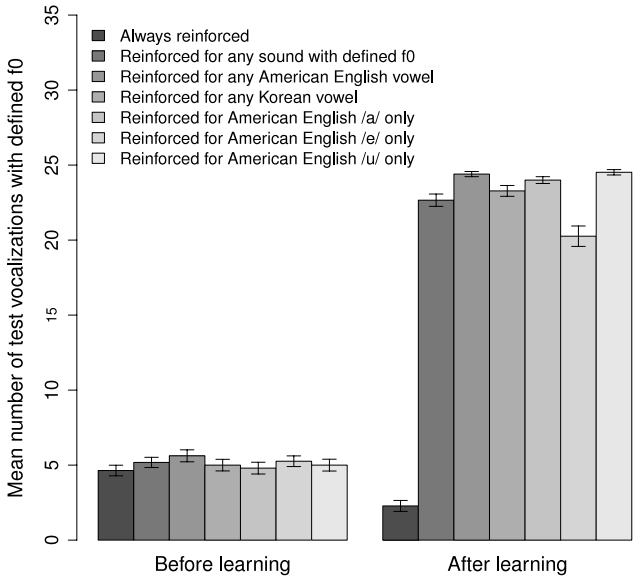

Fig. 2. Mean numbers of vocalizations with identifiable fundamental frequency before and after learning. Means are over the 50 simulations within a given reinforcement condition. Error bars indicate standard errors.

motor control of sounds with a variety of different phonatory characteristics. Finally, note that after learning had taken place, the network exhibited topographic organization — neurons located near each other tended to produce sounds with similar-looking spectrograms.

Fig. 4 shows the laryngeal muscle activations responsible for producing the vocalizations spectrograms in Fig. 3. The figure shows consistencies with what is known about roles of the various laryngeal muscles in phonation. In particular, muscle number 4, the thyroarytenoid, a muscle that courses beside each vocal fold and promotes phonation by adducting the vocal folds (it also relaxes and shortens them), is highly activated, as would be expected. Additionally, muscle number 6, the lateral cricoarytenoid, shows greater activation than muscle number 5, the posterior cricoarytenoid; this corresponds to the fact that the lateral cricoarytenoid is a vocal fold adductor and therefore promotes phonation whereas the posterior cricoarytenoid is a vocal fold abductor, inhibiting phonation.

3.2. Vowel types produced before and after learning

To investigate the types of vowels produced under the various vowel reinforcement conditions, we used the same set of test vocalizations as was used for the phonation evaluations. We compared the simulations in which any sound with identifiable f_0 was reinforced, in which vocalizations resembling any of the American English vowels were reinforced, and in which vocalizations resembling any of the Korean vowels were reinforced. The dependent variables were the number of sounds falling within 3 mel of the American English vowels and the number of sounds falling within 3 mel of the Korean vowels.

As can be seen in Fig. 5, all of the networks produced fewer vowels resembling the vowel targets before learning than after learning. This pattern may be driven in part by the fact that before learning all networks produced fewer sounds with defined f_0, and if a sound did not have defined f_0, it was automatically considered not similar to any of the target vowels.

After learning, the American-English-reinforced model produced the most sounds falling within the 3 mel target range of the American English vowels. A mixed-model regression with vowel

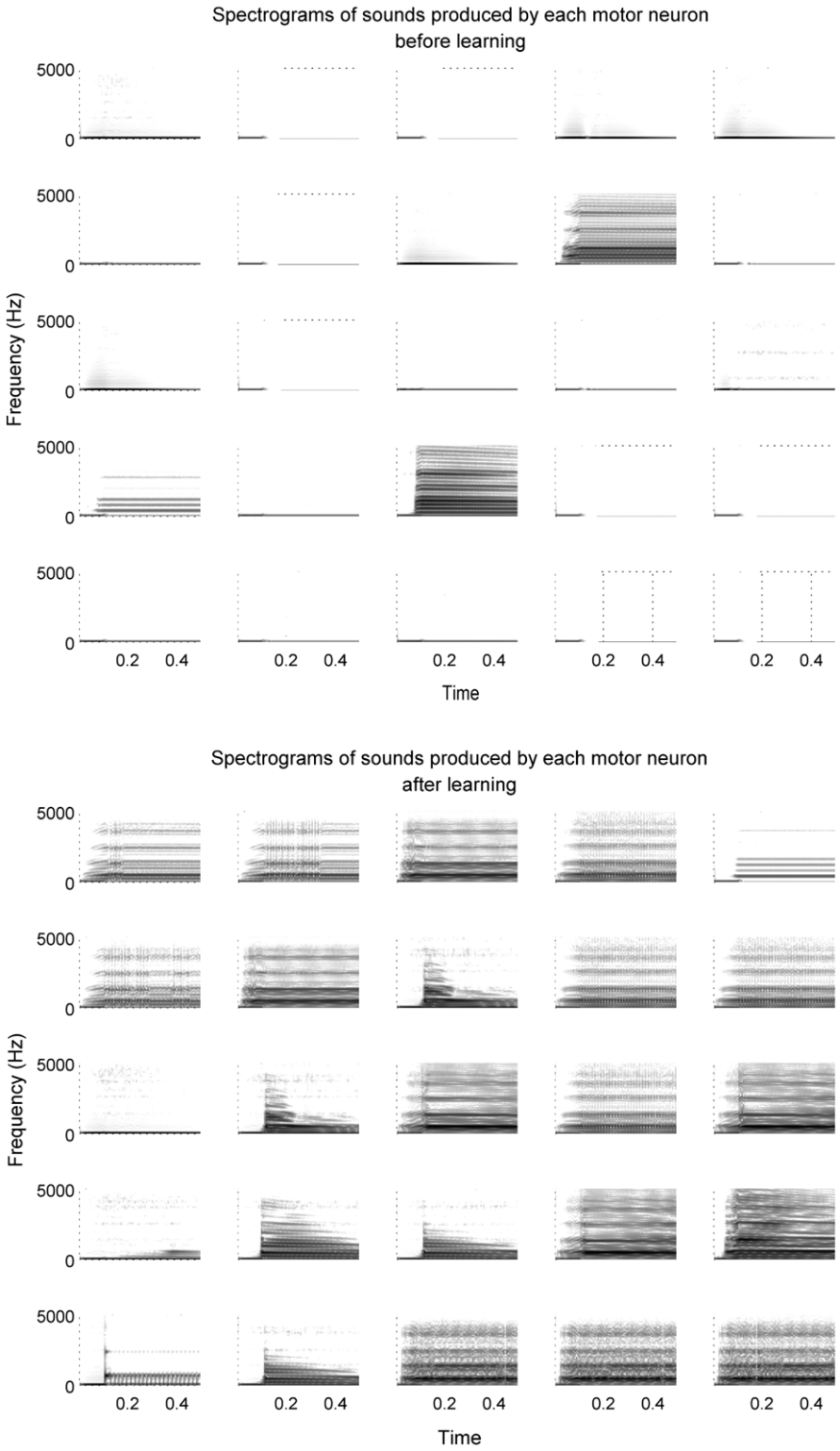

Fig. 3. Spectrograms of sounds produced by individually activating each neuron in one of the simulations from the second reinforcement condition, in which any sound with identifiable f_0 was reinforced. Before learning (pictured at top), three neurons' productions were judged as being voiced: these are located at row 2, column 4; row 4, column 1; and row 4, column 3. After learning (pictured at bottom), all neurons' productions were judged as being voiced except for one, located at row 3, column 1.

and simulation as random effects, reinforcement for American English versus Korean as a fixed effect, and number of vowels resembling American English targets as the dependent variable showed that the difference between reinforcement conditions in the number of American-English-like productions after learning was statistically significant, $\beta = 0.23, p < 0.001$. While the mean number of vowels falling within 3 mel of the Korean targets was overall lower for all reinforcement conditions (some possible explanations for this bias will be given in the Discussion), the Korean-reinforced model was the best-performing. A mixed model regression with number or vowels resembling the Korean targets as the dependent variable revealed the effect of the reinforced language to again be statistically significant, $\beta = 0.28, p < 0.001$. Note that β, the standardized regression coefficient, is comparable

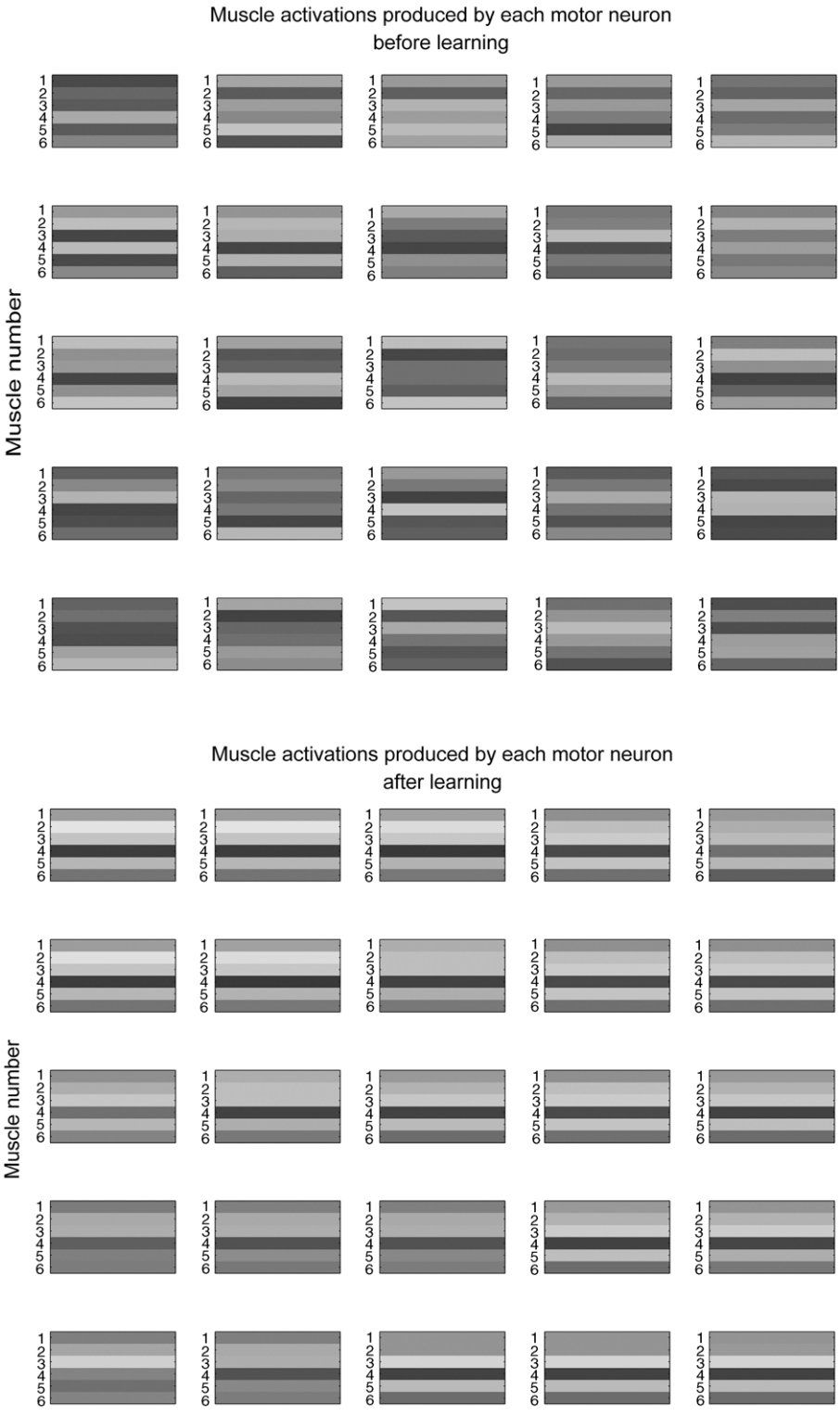

Fig. 4. Connection weights from each neuron in one of the simulations from the second reinforcement condition (the same simulation as in Fig. 3) to each of the laryngeal vocal tract muscles (see Table 1). Darker colors indicate higher weights and thus higher muscle activations.

across the two target languages, indicating that the effect of reinforced language was similar in magnitude for both. Fig. 6 shows the relative formants of the productions from each version of the model after training. In sum, the model learned to produce more of the vowels from the language for which it was reinforced.

For a closer look at the model's learning of specific vowels, we compared the simulations in which only the American English /a/ was reinforced, in which only the American English /e/ was reinforced, and in which only the American English /u/ was reinforced. The dependent variables were the number of sounds falling within 3 mel of /a/, /e/, and /u/. As can be seen in Fig. 7, it was the simulations that were reinforced for /a/ that produced the most vowels resembling /a/ after learning. The differences between /a/-reinforcement and /e/-reinforcement and between /a/-reinforcement and /u/-reinforcement were both statistically significant with $p < 0.001$ in both cases. Similarly, the simulations

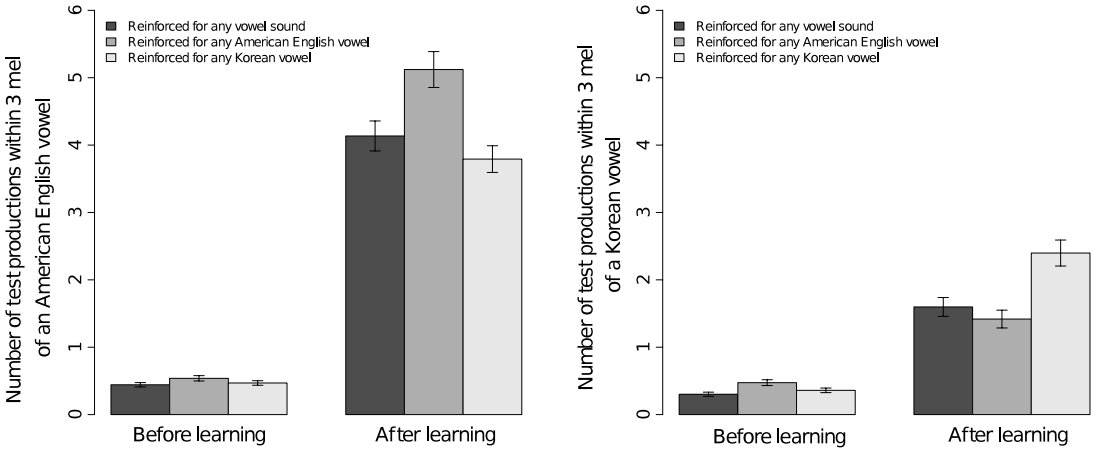

Fig. 5. Mean numbers of vowels within 3 mel of American English (left) and Korean (right) vowels for models in three different reinforcement conditions before and after learning. Means are over the 50 simulations within a given reinforcement condition. Error bars indicate standard errors at the item level.

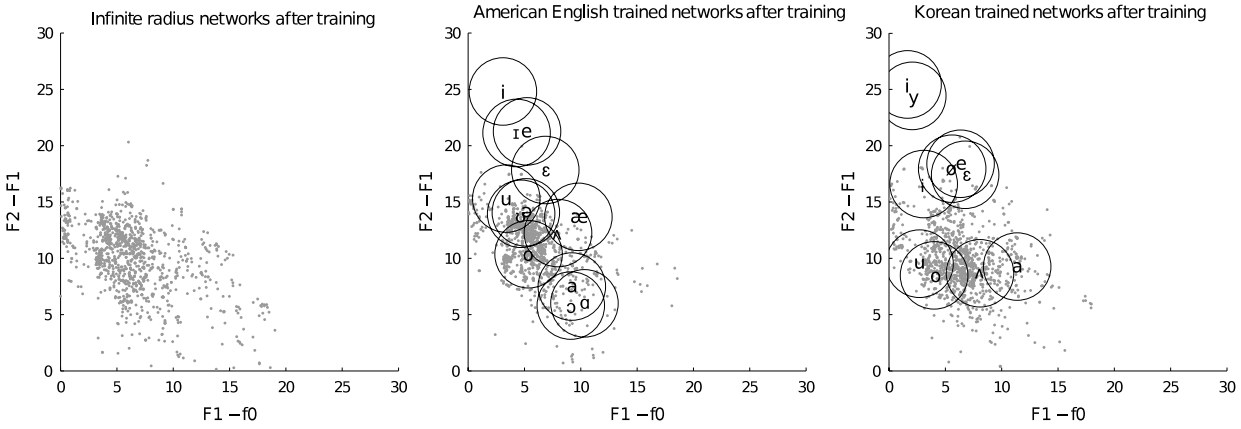

Fig. 6. Relative vowel formants of the vocalizations produced by individually activating each motor neuron from all the simulations in each of three different reinforcement conditions. Left: reinforced for any sound with defined f_0. Middle: reinforced for any American English vowel. Right: reinforced for any Korean vowel. Each gray dot represents one neuron's vocalization. Vocalizations from the 50 simulations in the same condition are superimposed. For each reinforcement condition, the targets of training are shown in black characters with circles delineating the 3 mel radius around each target.

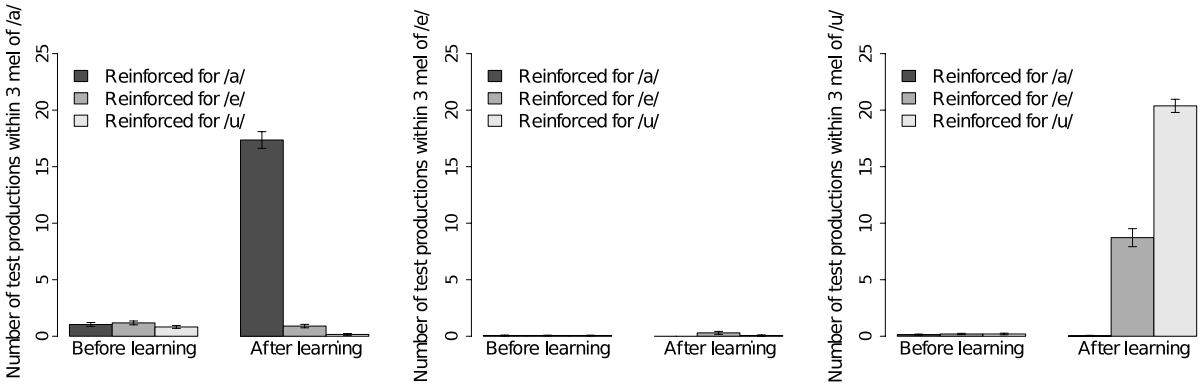

Fig. 7. Mean numbers of vowels within 3 mel of /a/, /e/, and /u/ for models trained on /a/, /e/, and /u/, before and after learning. Error bars indicate standard errors.

reinforced for /u/ produced more vowels resembling /u/ than the simulations reinforced for /a/ and /e/, $p < 0.001$ in both cases. The simulations reinforced for /e/ produced more vowels resembling /e/ than the simulations reinforced for /a/, $p = 0.02$, and marginally more than the simulations reinforced for /u/, $p = 0.10$. Overall, fewer productions were close to the /e/ target than were close to the /a/ or /u/ targets. Fig. 8 shows the relative formants of each model's productions after training. The plots confirm that while the model readily learned to produce precise /a/'s and /u/'s, it had more difficulty learning to produce /e/ as evidenced by the broad distribution of vocalizations produced in simulations where /e/ was the target vowel.

4. Discussion

We have presented a new neural network model wherein exploration and reinforcement are integrated with topographic self-organized learning. A layer of neurons is connected to the muscle inputs of a realistic human vocal tract synthesizer. The model explores its vocalization abilities by randomly activating neurons,

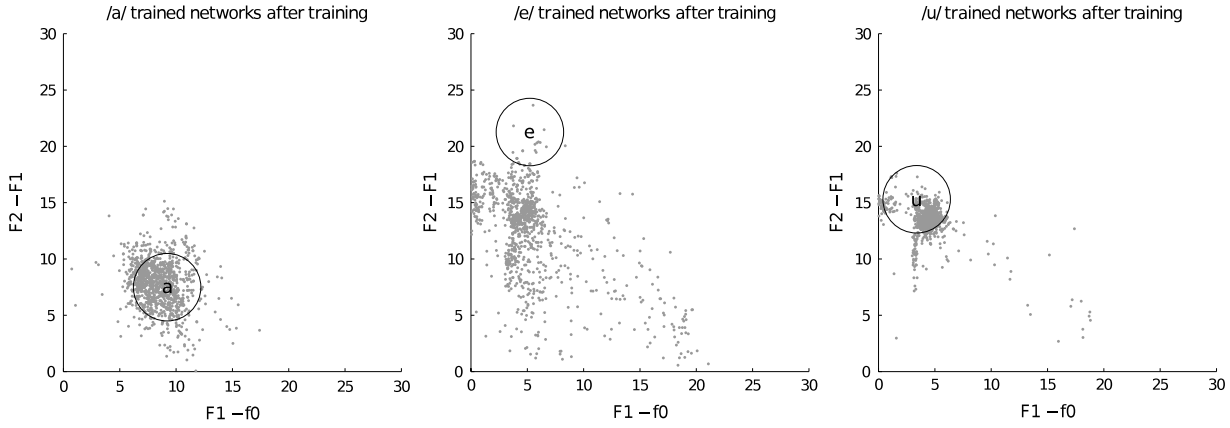

Fig. 8. Relative vowel formants produced when activating neurons in isolation from networks in trained on /a/, /e/, or /u/. Neurons from different simulations in the same reinforcement condition are superimposed. The targets of training are shown in black characters with circles delineating the 3 mel radius around each target.

with some noise added at the muscular level. When it receives reinforcement for a vocalization, it updates its neuromuscular connection weights so that similar motor commands become more likely to be produced in the future. We show that the model can learn at least two foundational speech-related skills: production of phonated sounds and production of specific vowel types.

One of the contributions of this work is that it specifies at a mechanistic level how reinforcement, which is known to play a role in speech development (Goldstein et al., 2003; Goldstein & Schwade, 2008; Gros-Louis et al., 2006), may be used by human infants as they develop sounds with speech-like characteristics. For example, it has been shown that when caregivers' vocal reinforcement is contingent on infants' production of vocalizations, the frequency of infant vocalization increases (Goldstein et al., 2003). In our model, reinforcement that is contingent on phonation (as measured by the sound having an identifiable f_0) signals the model to modify the connection weights from its motor neurons to its vocal tract muscles so that future neuronal activity will be more likely to result in phonated sounds. As discussed in Results, the laryngeal muscle activation levels produced after learning correspond to what would be expected based on previous physiological studies of speech production.

Note, however, that our model is agnostic regarding the source of reinforcement. Reinforcement could come directly from social sources, such as a mother vocalizing toward her infant. It is also possible for an infant to be reinforced intrinsically, for example by producing an appealing sound, where the appeal is based on auditory salience or similarity to sounds that the infant has previously heard other individuals produce. In future work, it would be good to model these distinct possible sources of reinforcement in more detail, for example incorporating extrinsic reinforcement that has contingencies similar to those observed in naturalistic parent–child interactions or in experiments with children. Intrinsic reinforcement could perhaps be modeled by adding an auditory system that perceives different sounds as having different levels of saliency, where reinforcement would increase as saliency increases. In addition to or instead of saliency, information content could be used. The auditory system could also learn from sounds produced by speakers of the target language, perhaps altering its sense of saliency or desirability. In either case, the extrinsic or intrinsic reward could be used to gate learning according to the model proposed here.

In addition to learning to phonate, the model also develops a propensity toward producing vowels like those for which it has been reinforced, whether that be the whole set of American English or Korean vowels or a single isolated vowel. A process of reinforcement-gated learning may be one of the mechanisms underlying babbling drift findings, i.e. shifting of vowels toward those that are most frequent in the infant's language environment (de Boysson-Bardies et al., 1989). Previous neural network models of speech production learning have all depended critically on learning sensorimotor correspondences in order to achieve ambient-language effects (Guenther et al., 1998; Heintz et al., 2009; Kanda et al., 2009; Warlaumont et al., 2011; Westermann & Miranda, 2004; Yoshikawa et al., 2003). None of those prior studies report data on spontaneous vocal productions, although those that include learning of connections between motor neurons and the vocal tract would be expected to exhibit ambient language effects on spontaneous productions. Our model, in contrast, requires no learning of sensorimotor correspondences, relying instead on reinforcement-gated learning of neuromotor connections and therefore illuminating an additional pathway through which the ambient language environment may shape spontaneous productions.

The model appears to exhibit not only learning effects but also biases with regard to the sounds the realistic vocal tract simulator can learn to reliably produce. These are (1) a bias against the vowel /e/ compared to the vowels /a/ and /u/ and (2) a bias toward better performance on American English vowel targets compared to Korean targets. Regarding the first bias, physiological vocal tract constraints are known to play a strong role in vowel development, as Oudeyer discusses with regard to his own model of speech sound learning and evolution (Oudeyer, 2005), and presumably play a role in the human system as well. In support of this, it is observed that /e/, /i/, and /u/ are less frequent in human infants' vocalizations than /a/ (de Boysson-Bardies et al., 1989; Ishizuka, Mugitani, Kato, & Amano, 2007). Thus, the model's weakness on /e/ and /i/ relative to /a/ fits with the human infant data. However, the model's strong performance on /u/ does not correspond to the pattern from human data. Furthermore, the synthesizer used in the present study models an adult female vocal tract and the acoustic vowel targets are based on average adult female productions from the literature, making the particular pattern of difficulty on mid and high front vowels such as /e/ in the present study even more surprising.

We suspect the difficulty with /i/ and /e/ reflects issues with our acoustic measure for evaluating vowel similarity. Although the geometry of the vocal tract model was intended to be similar to that of a typical adult female, there are likely still a number of differences from the vocal tracts of the adult females whose mean vowel fundamental and formant frequencies were used as targets. It is known that any differences across speakers' vocal tract shapes can affect vowel perception (Johnson, 2005). Additionally, the sounds produced after training in the /e/ target simulations to our ears tended to sound more similar to /e/ than the sounds produced after training in the /u/ target simulations sounded similar to /u/.

Thus, it may be that a better metric for comparing the vowels produced by the model to those in other languages is essential for performance that better reflects that of human infants. First and second formant frequencies (F1 and F2) are the most popular metric for quantifying vowel acoustics, which is consistent with the fact that they are the most prominent perceptual dimensions identified through multidimensional scaling (Johnson, 2005). It has been shown that perception of formant frequencies is sensitive to fundamental frequency (f_0), and there is support for the idea that measures that make formant frequencies relative to each other and to f_0, such as F1/f_0 and F2/F1 or F1–f_0 and F2–F1 (which become ratios rather than differences when log frequencies or approximately log frequencies such as in the mel or Bark scales are used), may prove to be better for vowel classification (Johnson, 2005). Here we have taken this approach, using F1–f_0 and F2–F1. We also tried using simply F1 and F2, without any subtraction of other frequencies, and found roughly the same pattern of results. It may be important that there is strong evidence from the human vowel acoustic literature that the fundamental frequency and formant frequencies, regardless of which of the above transformations are used, do not completely account for listener perceptions of vowel type (Heintz et al., 2009; Ito, Tsuchida, & Yano, 2001; Johnson, 2005; Zahorian & Jagharghi, 1993).

Future research should explore other less traditional acoustic correlates of vowel productions as well as human listener judgments to see if better results on /i/ and /e/ can be obtained. A particularly useful approach might be to replace the a priori choice of acoustic features defining the vowels with a system, such as a Hebbian network or a multilayer perceptron trained with backpropagation, that learns the mappings between acoustic features and human listener vowel labels for model-produced sounds.

The second bias, toward better performance on American English vowels compared to Korean vowels, could be due in part to the way the vowels are distributed in the different languages. The American English vowels tended to be clustered together more continuously in formant space whereas there were distinct gaps between groups of Korean vowels (see Fig. 6). The model's productions tended to cover a fairly continuous region of space regardless of which language it was trained on, and reinforcement for a particular language tended to shift and reshape this region, but without breaking the continuity of the vowel production space. Since we set a fixed 3 mel boundary for the vowels, the greater separation between vowel formant means with Korean than that within American English resulted in more vowel productions landing in the spaces between the Korean vowel circles. This bias might disappear if instead of evaluating the model based on the number of its vowels falling within a fixed 3-mel circle, the model's language-specific performance were evaluated based on the number of its vowels falling within the borders of the complete vowel space. Consistent with the observation of variability and overlap in model productions both before and after learning, the vowels produced by children at various ages from 3 months up through 5 years are observed also to occupy continuous regions of formant space (de Boysson-Bardies et al., 1989; Ishizuka et al., 2007; Kent & Murray, 1982), as are the vowels produced by adults during spontaneous speech (Harmegnies & Poch-Olivé, 1992; Nicolaidis, 2003). Thus, although eventually children's vowel productions shift over development toward those vowels characteristic of the language, a great deal of variability and overlap among vowels is always present in spontaneous productions. Interestingly, some of this variability in adult productions could potentially prove useful to infants during word learning (Rost & McMurray, 2009).

Previous studies involving other neural network models of infant vocal development have not reported quantitative results regarding ambient-language effects on spontaneous vocal productions and have not addressed the development of phonation. In the future, doing so would permit direct comparison of our results to those of the previous models discussed in Section 1. Additionally, more detailed comparison of the behavior of this and other models to the behavior of human infants and their caregivers will be helpful in further developing the work. Increased efforts to tie neural network modeling directly to neurophysiological findings, to anatomical changes across the lifespan, and to patterns of difference observed in clinically relevant groups, such as those with hearing impairment or those with autism, would also be expected to improve the models and therefore increase their scientific and clinical value.

Our mechanism and those of previous models are not mutually exclusive. Reinforcement-gated motor learning, perceptual learning, and sensorimotor associative learning are likely all involved in infant vocal development. The various mechanisms likely also interact with each other. For instance, changes in perceptual representations as a result of exposure to sounds from an ambient language may affect how the infant perceives sounds to be salient or otherwise intrinsically rewarding. A model that combines perceptual learning with reinforcement-guided motor learning would provide a more complete account of how infants come to produce the vowels of their native language, since it would not assume as much prior knowledge as the current model about what vowels should be reinforced. In the future, a more comprehensive model of vocalization development that combines these various mechanisms should be developed and evaluated. Additionally, all existing models of vocal development must be extended in the future to address problems of the development of fine-grained dynamic sequences, such as those required for the precise syllable timing that also emerges in the first year of life and is a critical prespeech skill. Finally, it is worth exploring the possibility that the same principles exemplified by our model may generalize to domains such as the development of gestures and reaching skills.

5. Conclusions

We have presented the first neural network model to address how reinforcement may play a role in human vocalization development. It introduces an approach that combines self-organization with selective reinforcement. The model exhibits several general characteristics of human infant vocal development, including sensitivity of vocal productions to reinforcement, development of phonatory skill, and development of a tendency of vowel production acoustics to be more consistent with the vowels in the ambient language than with vowels from other languages. These positive results warrant the further development and improvement of our model and others that address the role of reinforcement in vocal motor learning.

Acknowledgments

This work was supported by the Department of Energy Computational Science Graduate Fellowship Program of the Office of Science and National Nuclear Security Administration in the Department of Energy under contract DE-FG02-97ER25308, by the Plough Foundation, and by NIH DC011027. We would like to thank Robert Kozma and the anonymous reviewers for their helpful feedback on earlier versions of the paper.

References

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Blakemore, S.-J. (2010). The developing social brain: implications for education. *Neuron*, *65*, 744–747.

Boersma, P. (1998). *Functional phonology: formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

Boersma, P., & Wennink, D. (2010). Praat: Doing phonetics by computer (Version 5.1.31) [Software]. Available from http://www.praat.org.
Buder, E., Chorna, L., Oller, D. K., & Robinson, R. (2008). Vibratory regime classification of infant phonation. *Journal of Voice*, *22*, 553–564.
Callan, D. E., Kent, R. D., Guenther, F. H., & Vorperian, H. K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research*, *43*, 721–736.
de Boysson-Bardies, B., Halle, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language*, *16*, 1–17.
de Boysson-Bardies, B., & Vihman, M. (1991). Adaptation to language: evidence from babbling and first words in four languages. *Language*, *67*, 297–319.
Domjan, M. (2010). *The principles of learning and behavior* (6th ed.). Belmont Calif.: Wadsworth.
Ellis, D.P.W. (2007). PLP and RASTA (and MFCC, and inversion) in MATLAB. Retrieved from http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/.
Fagan, M. K., & Iverson, J. M. (2007). The influence of mouthing on infant vocalization. *Infancy*, *11*, 191–202.
Goldstein, M. H., King, A. P., & West, M. J. (2003). 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.
Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*, 515–523.
Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, *30*, 509–516.
Grossmann, T., & Johnson, M. H. (2007). The development of the social brain in human infancy. *The European Journal of Neuroscience*, *25*, 909–919.
Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*, 280–301.
Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*, 611–633.
Harmegnies, B., & Poch-Olivé, D. (1992). A study of style-induced vowel variability: Laboratory versus spontaneous speech in Spanish. *Speech Communication*, *11*, 429–437.
Heintz, I., Beckman, M., Fosler-Lussier, E., & Ménard, L. (2009). Evaluating parameters for mapping adult vowels to imitative babbling. In *Proceedings of the 10th annual conference of the international speech communication association*, INTERSPEECH. Brighton, UK.
Ishizuka, K., Mugitani, R., Kato, H., & Amano, S. (2007). Longitudinal developmental changes in spectral peaks of vowels produced by Japanese infants. *The Journal of the Acoustical Society of America*, *121*, 2272–2282.
Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *The Journal of the Acoustical Society of America*, *110*, 1141–1149.
Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex*, *17*, 2443–2452.
Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni, & R. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Oxford: Blackwell Publishers.
Kanda, H., Ogata, T., Takahashi, T., Komatani, K., & Okuno, H. (2009). Continuous vocal imitation with self-organized vowel spaces in recurrent neural network. In *2009 IEEE international conference on robotics and automation* (pp. 4438–4443). Kobe.
Kent, R., & Murray, A. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *The Journal of the Acoustical Society of America*, *72*, 353–365.
Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*, 1464–1480.
Koopmans-van Beinum, F. J., & van der Stelt, J. M. (1986). Early stages in the development of speech movements. In B. Lindblom, & R. Zetterström (Eds.), *Precursors of early speech* (pp. 37–50). New York: Stockton Press.
Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, *100*, 2425–2438.
Lewis, J. M., Deák, G. O., Jasso, H., & Triesch, J. (2010). Building a model of infant social interaction. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 278–283). Austin, TX: Cognitive Science Society.
Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.
Max, L., Guenther, F., Gracco, V., Ghosh, S., & Wallace, M. (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: a theoretical model of stuttering. *Contemporary Issues in Communication Science and Disorders*, *31*, 105–122.
Molina-Luna, K., Pekanovic, A., Röhrich, S., Hertler, B., Schubring-Giese, M., Rioult-Pedotti, M.-S., et al. (2009). Dopamine in motor cortex is necessary for skill learning and synaptic plasticity. *PLoS ONE*, *4*, e7082.
Nicolaidis, K. (2003). Acoustic variability of vowels in greek spontaneous speech. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 3221–3224).
Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates.
Oller, D. K., & Lynch, M. P. (1992). Infant vocalizations and innovations in infraphonology: toward a broader theory of development and disorders. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: models, research, implications* (pp. 509–536). Timonium, MD: York Press.
Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, *233*, 435–449.
Papoušek, M., & Papoušek, H. (1989). Forms and functions of vocal matching in interactions between mothers and their precanonical infants. *First Language*, *9*, 137–157.
Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., et al. (2007). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonentics*, *28*, 233–272.
Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*, 339–349.
Stark, R. E. (1980). Stages of speech development in the first year of life. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology, vol. 1: production* (pp. 73–92). Academic Press.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.
Titze, I. R. (1994). *Principles of voice production*. Englewood Cliffs, NJ: Prentice Hall.
Titze, I. R. (2008). Nonlinear source–filter coupling in phonation: theory. *The Journal of the Acoustical Society of America*, *123*, 2733.
Vihman, M. M. (1993). Variable paths to early word production. *Journal of Phonentics*, *21*, 61–82.
Warlaumont, A. S., Westermann, G., & Oller, D. K. (2011). Self-production facilitates and adult input interferes in a neural network model of infant vowel imitation. In D. Kazakov, & G. Tsoulas (Eds.), *AISB 2011 Computational models of cognitive development* (pp. 8–12). York, UK: Society for the Study of Artificial Intelligence and the Simulation of Behaviour.
Westermann, G., & Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, *89*, 393–400.
Yang, B. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonentics*, *24*, 245–261.
Yoshikawa, Y., Asada, M., Hosoda, K., & Koga, J. (2003). A constructivist approach to infants vowel acquisition through mother-infant interaction. *Connection Science*, *15*, 245–258.
Zahorian, S. A., & Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *The Journal of the Acoustical Society of America*, *94*, 1966–1982.