INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Emergence of Vocal Developmental Sequences in a Predictive Coding Model of Speech Acquisition

Shamima Najnin[2], *Bonny Banerjee*[1,2]
[1]Institute for Intelligent Systems, and
[2]Department of Electrical & Computer Engineering
The University of Memphis, Memphis, TN 38152, USA
{snajnin, bbnerjee}@memphis.edu

Abstract

Learning temporal patterns among primitive speech sequences and being able to control the motor apparatus for effective production of the learned patterns are imperative for speech acquisition in infants. In this paper, we develop a predictive coding model whose objective is to minimize the sensory (auditory) and proprioceptive prediction errors. Temporal patterns are learned by minimizing the former while control is learned by minimizing the latter. The model is learned using a set of synthetically generated syllables, as in other contemporary models. We show that the proposed model outperforms existing ones in learning vocalization classes. It also computes the control/muscle activation which is useful for determining the degree of easiness of vocalization.
Index Terms: speech acquisition, babbling, predictive coding

1. Introduction

Emergence of syllables is one of the most significant developmental states in infant speech acquisition. The transition between a consonant and a vowel is fast in adult speech. In developing children, it takes time to learn this transition. Usually infants learn to produce syllables between 4 and 10 months of age [1]. Their speech production is influenced by feedback from their own production, besides native language and social interaction in their ambient environment [2]. An understanding of the underlying mechanism of syllable (developmental sequence) generation will be useful for rehabilitation of children with early risk of autism spectrum disorder.

Most of the existing computational models focus only on the development of vowels [3, 4, 5, 6, 7, 8]. Although some have addressed how syllables might be learned, they have a number of limitations. In one model [9], intrinsic interest was exploited to learn syllables. However, the model does not learn the temporal pattern among the consonant and vowels and it requires prior knowledge about acoustic-articulatory states. In [10], a reinforcement-based spiking neural network model is developed for generating canonical babbling which receives a reward when the produced sound is more salient than previously generated sounds. But the model cannot learn the sensorimotor mapping and it exploits only one motor degree of freedom for implementation simplicity. In [11], a neural network model is developed to learn the temporal relationship between consonant-vowel sequences to explain how babies learn to speak. But it requires both the initial forward and inverse model which contains the prior relationship between acoustic and articulatory states. Moreover, it lacks the control policy for vocalization. In reality, babies do not have access to any articulatory information. In [12, 13], a cognitively plausible neural network model learns spatiotemporal auditory and somatosensory target regions for different speech sounds stored in a map but it fails to explain how others' speech can be recognized as auditory activations. The Eliza model [14, 15] is based on non-imitative child and caregiver behaviors. It does not focus on the emergence of syllabic patterning. Various parameterizations are involved in computing the reward function. Also, the model requires an external caregiver for providing feedback.

Based on the predictive coding principle of perception, action and learning [16, 17], we propose a model for learning the articulatory-acoustic association, the temporal relation among the developmental sequences and the control policy of vocalization during the emergence of syllables. The model explores the acoustic space by perceiving the sounds from its own production as well as from the environment. It randomly chooses the acoustic goal from the perceptual space, infers the cause and performs the optimal action to explain away the prediction error. The aim of the model is to reproduce the acoustic goal.

The key features of this model are as follows. It learns temporal patterns among developmental sequences which is not considered in [14, 10, 9]. Action/motor command is not generated directly from the sensory state but as a result of explaining the sensory prediction error, which is consistent with mirror neuron theory [17]. Our model exploits nine degrees of freedom for controlling the vocal tract whereas the model in [10] exploits only one. Our model can compute the required action/motor activation to manipulate the articulators in order to identify which vocalization is easy/hard to learn. Such identification is necessary for the emergence of developmental sequences [9, 18].

2. Models and Methods

In this paper, we will assume that synthetically generated vocalizations from articulatory synthesizer constitute the agents perceptual space ($\boldsymbol{S}$). The agent's goal is to generate the syllable ($\boldsymbol{y}$) of sequences of length T. The task is to compute the cause ($\hat{\boldsymbol{x}}$) of generating $\boldsymbol{y}$. We construct the motor parameters as causes. The predicted auditory effect for $\hat{\boldsymbol{x}}$ is $\hat{\boldsymbol{y}}$. After inferring the causes, the model computes the optimal action/motor command ($\boldsymbol{a}$) to reach the inferred causal state ($\hat{\boldsymbol{x}}$) from the current state ($\boldsymbol{x}$) of the body. The consequence of action execution is perceived from the environment as $\boldsymbol{y}^e$. The overall objective of the model is to minimize prediction error ($\boldsymbol{E}$) where,

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{E_p} \\ \boldsymbol{E_i} \\ \boldsymbol{E_f} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x} - \hat{\boldsymbol{x}} \\ \boldsymbol{y} - \boldsymbol{y}^e \\ \boldsymbol{y}^e - \hat{\boldsymbol{y}} \end{bmatrix}$$

The architecture of the proposed model is shown in Fig. 1. It consists of three parts: a generative network for learning the

http://dx.doi.org/10.21437/Interspeech.2016-1126

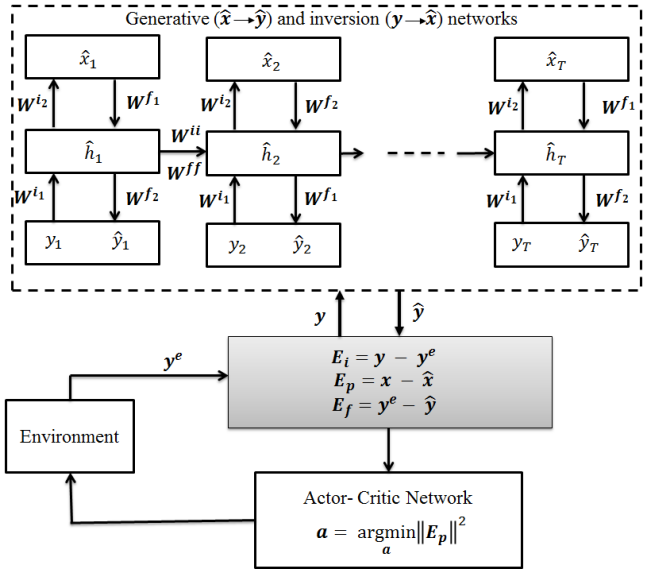

Figure 1: Architecture for proposed model.

association between $\boldsymbol{y}$ and $\hat{\boldsymbol{x}}$, an actor-critic network for computing the optimal action $\boldsymbol{a}$, and the environment for production of $\boldsymbol{y_e}$ from $\boldsymbol{x}$ and $\boldsymbol{a}$. The environment is modeled as:

$$\begin{aligned} y_t^e &= g^e(h_t^e, x_t^e, a_t) \\ h_t^e &= f^e(h_{t-1}^e, x_t^e, a_t) \end{aligned} \tag{1}$$

The agent's internal model/genrative network of the environment can be described as:

$$\begin{aligned} \hat{y}_t &= g(\hat{h}_t, \hat{x}_t) \\ \hat{h}_t &= f(\hat{h}_{t-1}, \hat{x}_t^) \end{aligned} \tag{2}$$

where g^e, f^e and g, f are continuous nonlinear functions of the states of the environment and internal model respectively. The states x_t, referred to as causes at time t, are responsible for generating the sensory data y_t, h is the hidden states. The difference between true generation (Eq. 1) and internal model's generation (Eq. 2) constitutes the sensory prediction error that action tries to fulfil. The internal model is learned by minimizing sensory prediction error in order to remain grounded to the changing environment, as in biological agents [19, 17]. The network is learned in a layer-by-layer manner, similar to [16, 20, 21].

The generative network is modeled using a 3-layered recurrent neural network whose sensory input is y, hidden state is h, and the cause x is the activation strength of a set of neurons with generative weights $\boldsymbol{W^f} = \{\boldsymbol{W^{f_1}}, \boldsymbol{W^{ff}}, \boldsymbol{W^{f_2}}\}$. An appropriate set of activations can be used to reconstruct y using $\boldsymbol{W_f}$. Another set of weights, $\boldsymbol{W^i} = \{\boldsymbol{W^{i_1}}, \boldsymbol{W^{ii}}, \boldsymbol{W^{i_2}}\}$, is required to compute the activation of the cause x that actually generated the input. $\{\boldsymbol{W^{f_1}}, \boldsymbol{W^{f_2}}\}$ are top-down or feedback weights, $\{\boldsymbol{W^{i_1}}, \boldsymbol{W^{i_2}}\}$ are bottom-up or feedforward weights, and $\{\boldsymbol{W^{ff}}, \boldsymbol{W^{ii}}\}$ are lateral weights. See Fig. 1. Unlike the model in [21] where tied weights are used, feedforward and feedback weights in our model are distinct.

The activations of the hidden states and causes necessary to generate a sensory input $\boldsymbol{y}$ starting at time t and of sequence length T, are given by:

$$\begin{aligned} \hat{h}_t^i &= W^{i_1} y_t + W^{ii} \hat{h}_{t-1}^i \\ \hat{x}_t &= W^{i_2} \hat{h}_t^i \end{aligned} \tag{3}$$

The predicted sensory input $\hat{y}_t$ is:

$$\begin{aligned} \hat{h}_t^f &= W^{f_1} \hat{x}_t + W^{ff} \hat{h}_{t-1}^f \\ \hat{y}_t &= W^{f_2} \hat{h}_t^f \end{aligned} \tag{4}$$

The internal model predicts the sensory (auditory) and its corresponding motor state. As in [20, 22], we define the predicted auditory sensory using the generative model as $\hat{\boldsymbol{y}} = \{\hat{y}_t\}_{t=1\ldots T}$ and the predicted proprioceptive sensory (also referred to as *sensation*) using the inverse generative model as $\hat{\boldsymbol{x}} = \{\hat{x}_t\}_{t=1\ldots T}$. Perception reduces prediction error by changing prediction and action reduces the prediction error by changing sensation. Current motor state of the body is $\boldsymbol{x} = \{x_{t-1}\}_{t=1\ldots T}$. So, the prediction error of the agent's internal model can be minimized either by performing action or by altering the internal model. The optimal parameters of the generative and inverse generative models can be determined by minimizing $\|\boldsymbol{E_f}\|_2^2$ and $\|\boldsymbol{E_i}\|_2^2$ respectively, written as:

$$\boldsymbol{W^f} = \arg\min_{\boldsymbol{W^f}} \|\boldsymbol{E_f}\|_2^2 \tag{5}$$

$$\boldsymbol{W^i} = \arg\min_{\boldsymbol{W^i}} \|\boldsymbol{E_i}\|_2^2 \tag{6}$$

In reality, the agent needs to execute the action to receive sensory input $\boldsymbol{y^e}$ via the environment. The optimal action can be computed by minimizing $\|\boldsymbol{E_p}\|_2^2$. Recent deep learning models can successfully learn optimal control policies in high dimensional data spaces with very little prior knowledge. The deep deterministic policy gradient (DDPG) [23, 24] is used in our model to learn the optimal policy (for other approaches, see [25, 26]). The inputs to the actor-critic network are the current motor state ($x|_{t=0}$) and the predicted motor state ($\hat{\boldsymbol{x}}$) which is also the goal state. The actor-critic network computes the optimal sequences of actions to reach the goal state.

The actor network, $\mu(x|\boldsymbol{\theta^\mu})$, takes the current state as input and generates action as output, while the critic network, $Q(x, a|\boldsymbol{\theta^Q})$, evaluates the action for that state. For learning the network, target values are generated using target critic and actor networks, $Q'(x, a|\boldsymbol{\theta^{Q'}})$ and $\mu'(x|\boldsymbol{\theta^{\mu'}})$. The parameters of the target network are learned very slowly as: $\boldsymbol{\theta'} \leftarrow \tau\boldsymbol{\theta} + (1-\tau)\boldsymbol{\theta'}$, $0 < \tau \ll 1$. The critic network is learned using error backpropagation, while the actor network is learned with sampled gradient as follows:

$$\nabla_{\boldsymbol{\theta^\mu}} \mu|_{x_m} \approx \frac{1}{M} \sum_m \nabla_a Q(x, a|\boldsymbol{\theta^Q})|_{x=x_m, a=\mu(x_m)} \nabla_{\boldsymbol{\theta^\mu}} \mu(x|\boldsymbol{\theta^\mu})|_{x_m} \tag{7}$$

Details of the algorithm can be found in [23]. After computing the optimal action, it is executed through the vocal tract to generate sound $\boldsymbol{y^e}$. If action fails to minimize sensory error $\boldsymbol{E_i}$, the model needs to be updated. The generative and inverse models are updated using backpropagation through time (BPTT) [27]. Algorithm 1 is the pseudo code for operation of the proposed model.

Algorithm 1 Operation of the proposed model.

Initialize $\boldsymbol{W^f}, \boldsymbol{W^i}$.
for $k = 1\ to\ \infty$ **do**
  Get x from the current state of the body
  Choose $\boldsymbol{y} \in \boldsymbol{S}$
  Predict $\hat{\boldsymbol{x}}$ using Eq. 3
  Predict $\hat{\boldsymbol{y}}$ using Eq. 4.
  Compute $E_p = \boldsymbol{x} - \hat{\boldsymbol{x}}$
  Update $\boldsymbol{\theta^Q}$ using backpropagation.
  Update $\boldsymbol{\theta^\mu}$ using Eq. 7.
  for $t = 1\ to\ T$ **do**
    Compute $a_t = \mu(x_{t-1}|\boldsymbol{\theta^\mu})$
  end for
  Execute $\boldsymbol{a}$
  Get $\boldsymbol{y_e}$ from the environment
  if $|\boldsymbol{E_i}| > \epsilon$ **then**
    Update $\boldsymbol{W^i}$ by solving Eq. 6 using BPTT.
  end if
  if $|\boldsymbol{E_f}| > \epsilon$ **then**
    Update $\boldsymbol{W^f}$ by solving Eq. 5 using BPTT.
  end if
end for

3. Experiments

3.1. Simulation Conditions

For simulations, the generative model is a recurrent neural network consisting of one hidden layer with 20 units and linear activation function. Both the actor and critic networks are multilayered perceptrons with two hidden layers. In each hidden layer, there are 20 hidden units with $tanh$ activation function.

As part of the environment, we have used articulatory synthesizer of the DIVA model [12] based on Maeda's model [28]. Seven degrees of freedom constitute our motor space: jaw height (JH), tongue position (TP), tongue shape (TS), tongue apex (TA), lip area (LA), lip protrusion (LP), and larynx height (LH). During vocalization, infants learn to control the vocal tract by changing the muscle activations. For phonation controlling, we used glottal pressure and voicing parameter. So a vocalization of L ms duration is a trajectory in 9-dimensional motor space where seven are articulatory parameters and two are voice controlling parameters. The motor system is modeled as overdamped spring-mass system, as in [9]. The agent's optimal policy for the vocalization is learned by the actor-critic network using the DDPG algorithm [23].

On the perception side, we have used the first five formants, corresponding bandwidths and phonation as acoustic feature. The acoustic feature vector is normalized to zero mean and unit variance. If the phonation controlling parameters (glottal pressure and voicing parameters) have a value above a threshold (0.1) and area function of vocal tract is positive everywhere, the phonation occurs. If phonation occurs, depending on the level of phonation, the syllables are classified into one of three classes: vowels, consonants and none, as in [9]. If the phonation level I is less than 0.15, the generated sound is considered as none. For $0.15 < I < 0.9$, the produced sound is categorized as consonants, and for $I > 0.9$, it is classified as vowels.

Table 1: Percentage of vocalization classes produced in developmental stages of the generated vocalization sequence.

Vocalization classes	Proposed model			Model in [9]		
	Developmental Stages			Developmental Stages		
	I	II	III	I	II	III
NN	15.5	2.5	0.8	45.3	4.0	1.6
CN	10.1	15.2	3.9	13.4	26.9	3.7
NC	0.9	1.5	0.3	0.6	0.1	0.1
VN	17.5	20.1	6.9	18.9	62.2	12.1
NV	5.6	0.2	0.7	4.5	0.1	0.8
VV	43.4	50.1	72.1	9.9	3.4	67.5
CV	5.6	8.2	10.9	6.6	0.5	6.5
VC	0.9	2.1	4.9	0.7	2.5	6.8
CC	0.5	0.1	0.3	0.2	0.2	0.8

3.2. Emergence of syllables

Initially, the motor parameters of the DIVA model are randomly varied to generate 100,000 sounds which constitute the perceptual space, $\boldsymbol{S}$, of the agent. The environment plays an important role in the process of speech acquisition. As an environmental influence, 'ambient language' is modeled as set of two sequence of speech sounds. In order to make it coherent with human language and learning process observed in development, we have chosen speech-like sounds, such as vowel or consonant-vowel sounds, as in [9]. The environment consists of different articulatory sequences with different combinations of eight consonants, $\{/b/, /d/, /g/, /z/, /p/, /t/, /k/, /s/\}$ and eight vowels $\{/a/, /e/, /i/, /o/, /u/, /ae/, /oe/, /y/\}$ generated by the DIVA articulatory synthesizer.

The model is initially learned by choosing the acoustic goal from its own generated perceptual space by considering $T = 1$. Then the model randomly chooses the acoustic goal from the environment or from its own previously explored spaces with $T = 2$ (for syllables). Now the network starts learning the lateral weights and refine the generative and inverse weights, as explained in Section 2. The inverse generative model and actor-critic network were learned with learning rate 0.01, and the generative model is updated with very low learning rate of 10^{-4}.

After generation of 30,000 vocalizations using randomly selected goals, the percentage of vocalization classes are computed using the sounds produced by the model. This experiment is performed again for 150,000 and 250,000 vocalizations. The developmental stages are referred to as Stage I, Stage II and Stage III for 30,000, 150,000 and 250,000 vocalizations respectively. Table 1 shows the comparison between our model and the model developed in [9] in terms of percentage of vocalization classes. In the beginning, our agent cannot produce the sounds from the ambient language. Later during its development the agent gradually covers a wide range of sounds, which indicates it is learning to control its articulatory parameters. With reinforcement from the environment, it eventually starts to learn how to produce those sounds which are actually generated from the environment.

In one experiment, the syllable $/ba/$ is set as an acoustic goal for the model. The model inferred the cause of the motor state and predicts the reconstruction using the generative model. In order to reach the acoustic goal, the agent needs to execute action to reach from current motor state to desired causal/motor state which eventually generates the goal. The optimal trajectory of all the parameters and the required mus-

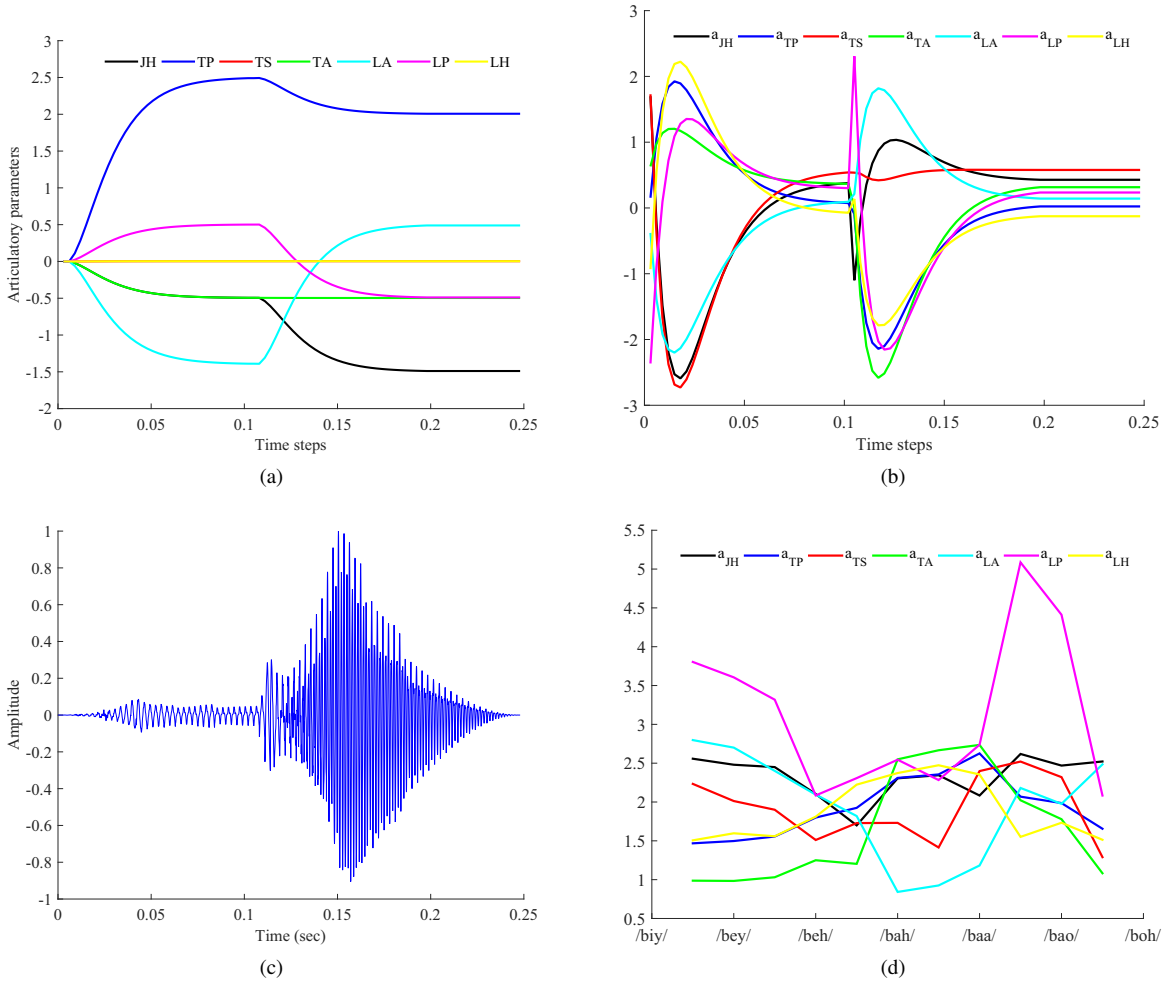

Figure 2: During the generation of syllable $/ba/$ (a) articulatory trajectories, (b) action required for each articulatory parameter, (c) reproduced sound;(d) action required for co-articulation of "b" and different vowels (best viewed in color).

cle activations to reach the desired motor state through the actor-critic network are shown in Fig. 2(a) and (b) respectively. The reproduced sound from the environment after action execution is shown in Fig. 2(c) using red line. The syllables $/ba/, /da/, /ga/, /ka/, /pa/, /ta/$ are given as acoustic goals and the model reproduced them by transition time steps of 96 ms,108 ms, 114 ms, 93 ms, 102 ms, and 114 ms. It requires less voice transition time to generate $/ba/, /pa/$ than $/ta/, /ga/, /ka/$ which is consistent with the findings in [29]. It is shown through clinical analysis in [30] that mean voice onset time (VOT) for $/pa/$ is 12 ms less than $/ta/$ and 22 ms less than $/ka/$. Using the proposed model, we found similar results; VOT for $/pa/$ is 12 ms less than $/ta/$ and 18 ms less than $/ka/$. Moreover, the average muscle activation required for seven articulatory parameters to produce $/biy/$, $/bey/$, $/beh/$, $/bah/$, $/baa/$, $/bao/$, $/boh/$, $/buw/$, $/biw/$, $/bew/$, $/boe/$ are calculated using the proposed model, and shown in Fig. 2(d). The muscle activation required for jaw height is greater for the CVs where the consonants are in conjunction with high vowels due to their effect of co-articulation. The tongue width is greater for high and front vowels as compared to other vowels [18]. In our model, the muscle activation of tongue shape turned out to be greater for consonants in conjunction with high vowels than those in conjunction with low vowels.

4. Conclusions

We proposed a computational model for speech acquisition that successfully explained the underlying mechanism for the emergence of syllables based on the principle of predictive coding. Unlike existing models, it learns the temporal patterns of developmental sequences exploiting nine degrees of freedom in motor space. It also computes the muscle activation for generating syllables which provides insight into the degree of easiness of the produced speech. By choosing random goals and minimizing prediction error relentlessly, the model was able to produce syllables successfully. The model can infer articulatory states through the inverse generative model to generate syllables without any prior knowledge of causal states, and computes optimal action which leads to the emergence of canonical babbling during infant speech development.

5. Acknowledgements

This research was supported by NSF grant IIS-1231620.

6. References

[1] D. K. Oller, *The emergence of the speech capacity*. Psychology Press, 2000.

[2] P. Kuhl, "Early language learning and the social brain," *Cold Spring Harbor Symposia on Qu antitative Biology*, vol. 79, pp. 211–220, 2014.

[3] G. Westermann and E. R. Miranda, "A new model of sensorimotor coupling in the development of speech," *Brain and language*, vol. 89, no. 2, pp. 393–400, 2004.

[4] I. Heintz, M. E. Beckman, E. Fosler-Lussier, and L. Ménard, "Evaluating parameters for mapping adult vowels to imitative babbling." in *INTERSPEECH*, vol. 9, 2009, pp. 688–691.

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

[6] K. Miura, Y. Yoshikawa, and M. Asada, "Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver," *Advanced Robotics*, vol. 26, no. 1-2, pp. 23–44, 2012.

[7] C. Moulin-Frier and P.-Y. Oudeyer, "Curiosity-driven phonetic learning," in *IEEE International Conference on Development and Learning and Epigenetic Robotics*, 2012, pp. 1–8.

[8] M. Murakami, B. Kroger, P. Birkholz, and J. Triesch, "Seeing [u] aids vocal learning: babbling and imitation of vowels using a 3d vocal tract model, reinforcement learning, and reservoir computing," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, 2015, pp. 208–213.

[9] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer, "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation," *Frontiers in psychology*, vol. 4, 2013.

[10] A. S. Warlaumont, "Salience-based reinforcement of a spiking neural network leads to increased syllable production," in *IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics*, 2013, pp. 1–7.

[11] A. K. Philippsen, R. F. Reinhart, and B. Wrede, "Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model," in *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics*, 2014, pp. 195–200.

[12] F. H. Guenther and T. Vladusich, "A neural theory of speech acquisition and production," *Journal of neurolinguistics*, vol. 25, no. 5, pp. 408–422, 2012.

[13] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, no. 9, pp. 793–809, 2009.

[14] P. Messum and I. S. Howard, "Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation," *Journal of Phonetics*, vol. 53, pp. 125–140, 2015.

[15] I. S. Howard and P. Messum, "Learning to pronounce first words in three languages: An investigation of caregiver and infant behavior using a computational model of an infant," *PlosOne*, 2014.

[16] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects," *Nature Neurosci.*, vol. 2, pp. 79–87, 1999.

[17] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[18] Q. Fang, S. Fujita, X. Lu, and J. Dang, "A model-based investigation of activations of the tongue muscles in vowel production," *Acoustical Science and Technology*, vol. 30, no. 4, pp. 277–287, 2009.

[19] A. K. Stuart, *The origins of order: Self organization and selection in evolution*. Oxford University Press, USA, 1993.

[20] K. J. Friston, J. Daunizeau, J. Kilner, and S. J. Kiebel, "Action and behavior: a free-energy formulation," *Biological cybernetics*, vol. 102, no. 3, pp. 227–260, 2010.

[21] W. Muhammad and M. W. Spratling, "A neural model of binocular saccade planning and vergence control," *Adaptive Behavior*, vol. 23, no. 5, pp. 265–282, 2015.

[22] H. Brown, K. Friston, and S. Bestmann, "Active inference, attention, and motor preparation," *Frontiers in psychology*, vol. 2, 2011.

[23] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[25] I. Mordatch and E. Todorov, "Combining the benefits of function approximation and trajectory optimization," in *Robotics: Science and Systems*, 2014.

[26] Y. Tassa, N. Mansard, and E. Todorov, "Control-limited differential dynamic programming," in *IEEE International Conference on Robotics and Automation*, 2014, pp. 1168–1175.

[27] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[28] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.

[29] K.-Y. Chao and L.-m. Chen, "A cross-linguistic study of voice onset time in stop consonant productions," *Computational Linguistics and Chinese Language Processing*, vol. 13, no. 2, pp. 215–232, 2008.

[30] P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: a review," *Clinical Linguistics & Phonetics*, vol. 14, no. 2, pp. 131–150, 2000.