5th International Conference on Development and Learning
and on Epigenetic Robotics
Aug 13-16, 2015. Providence, RI USA

Seeing [u] aids vocal learning: babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing

Max Murakami*§, Bernd Kröger†, Peter Birkholz‡, and Jochen Triesch*
* Frankfurt Institute for Advanced Studies (FIAS), Frankfurt, Germany
† Dept. of Phoniatrics, Pedaudiology, and Communication Disorders,
Medical School, RWTH Aachen University, Aachen, Germany
‡ Institute of Acoustics and Speech Communication
Technische Universität Dresden, Dresden, Germany
§ Email: murakami@fias.uni-frankfurt.de

***Abstract*—We present a model of imitative vocal learning consisting of two stages. First, the infant is exposed to the ambient language and forms auditory knowledge of the speech items to be acquired. Second, the infant attempts to imitate these speech items and thereby learns to control the articulators for speech production. We model these processes using a recurrent neural network and a realistic vocal tract model. We show that vowel production can be successfully learnt by imitation. Moreover, we find that acquisition of [u] is impaired if visual information is discarded during imitation. This might give sighted infants an advantage over blind infants during vocal learning, which is in agreement with experimental evidence.**

I. Introduction

Speech is an outstanding capacity of human beings. But how do infants learn how to speak? A crucial stage of speech acquisition is *babbling*, which denotes the productive speech development from the newborn's first cry to the first distinct words. During babbling, infants explore their vocal tracts and learn to associate specific muscle activations with the resulting acoustic signals.

We posit that babbling is an imitative process right from birth. This is supported by findings that newborns reproduce the stress pattern of their native language while crying [1]. Further evidence suggests that newborns are able to imitate facial gestures such as mouth opening and lip protrusion [2], [3]. This type of imitation might facilitate vocal learning by providing additional, visual information for the imitating infant. We explore this idea in the current study.

Several biologically inspired models of imitative vocal babbling exist. However, their imitation depends on factors that are not crucial in our view. The DIVA model [4], [5], [6] and related models [7], [8] as well as the model proposed by Moulin-Frier and Oudeyer [9], [10], [11] incorporate an imitation phase which utilizes knowledge gained from a prior (semi-) random self-exploration phase. Both the Elija model [12], [13] and the model proposed by Miura et al. [14], [15] depend on the presence of an imitative caregiver, which is not a necessity in our model. Recent work by Philippsen et al. [16] bases imitation on prior articulatory-acoustic supervised training, whereas our model starts without any prior articulatory knowledge.

Our proposed model is based on considerations found in the literature about birds' song learning, which is often viewed as a model system for human speech acquisition [17], [18]. For example, song learning in male zebra finches is thought to proceed in two phases [19], [20]:

- *Sensory learning*: The young zebra finch listens to his father's song and stores it as an auditory template.
- *Sensory-motor learning*: The young bird learns the song by matching his vocalizations with the stored template (memory) of his father's song.

We adapted this schema for the case of human vocal learning. We propose that infants are exposed to the ambient language (sensory learning) and thereby acquire imitation targets for imitative babbling (sensory-motor learning) [21]. Our simulations confirm that vowels can be acquired this way and indicate that visual guidance facilitates vocal babbling.

II. The Model

We consider a model architecture comprising the following components (Fig. 1):

- The *Reinforcement Learning agent* (*RL agent*) is the learner, i.e. the part of the infant brain that acquires skills by maximizing rewards.
- The agent interacts with the environment by manipulating the *vocal tract*. This interaction is the *motor control* that the infant exerts on her own articulators: The brain sends motor commands to the vocal tract, which produces a speech sound based on the articulators' configuration.
- The *auditory system* processes this speech sound and evaluates it with respect to the imitation target. The auditory system comprises the infant's inner ear, auditory memory, and the pathway in between.
- The auditory system's *evaluation* of the current speech sound is the reward signal. The higher the similarity

978-1-4673-9319-5/15/$31.00 ©2015 IEEE

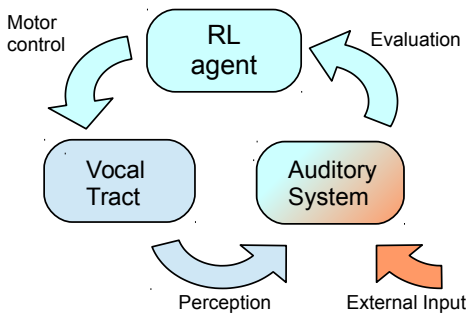

Fig. 1. The model in conceptual terms. It applies reinforcement learning to solve the imitation task. The model proceeds in two stages: (*red*) During target acquisition, the auditory system models the vowels that are fed in by external speakers. (*blue*) During actual imitation, the agent searches the motor space for the motor command producing the maximal reward.

between the current speech sound and the target, the higher the reward. Thus maximizing the reward means uttering the sound which is most similar to the target.

Successful imitation proceeds in two phases (Fig. 1):

(1) The imitation target is acquired. The auditory system is exposed to speech sounds from external speakers. It captures the statistics of those speech sounds and is later used for evaluating novel speech sounds. This stage models the perceptual part of speech acquisition, when the infant takes in the regularities of her mother tongue during the first months of life. The model assumes phonemes are learnt during this stage, i.e. the infant generalizes across different realizations of the same vowel. The phonemic distinction is crucial to gain vowel targets for imitation.

(2) Imitation occurs in an iterative fashion. Using an arbitrary motor command, the agent initiates the first own speech sound. This sound is then processed and evaluated by the auditory system. By sampling different motor commands, the agent explores the motor space and finds motor commands that yield higher rewards than others. Based on the auditory feedback, the agent then optimizes the motor command with respect to the reward signal. Eventually, the agent finds the motor command to reproduce the target and thus gain the highest reward. This motor command is then stored and becomes accessible for imitative and communicative utterances.

This stage models the vowel babbling phase of infants up to the point when they have learnt to utter the vowels of their mother tongue.

A. The Vocal Tract Model

The vocal tract is simulated using *Vocaltractlab 2.1* (VTL) [22], which is being used in various models of motor learning [7], [8], [16], [23]. VTL is based on a 3-dimensional model of the vocal tract, and it simulates the production of speech sounds based on articulator and vocal fold motion.

The acoustic properties of the produced speech sound depend on the articulators' configuration. This is captured by the *area function*, which describes the cross-sectional area of the airway between glottis and lips. VTL calculates the area function based on 20 articulator coordinates, two of which (velic opening and horizontal jaw position) are constant for the vowels under consideration ([@], [a], [i], [u]) and thus were discarded as variables (the symbol "@" refers to the mid-central *schwa* vowel, which is pronounced like "a" in the word "about"). Two more parameters (tongue root x- and y-coordinates) are determined automatically by VTL based on the other motor parameters to avoid unphysiological configurations. The remaining 16 degrees of freedom that constitute the motor space are:

- 10 tongue parameters: tongue body coordinates (TBX, TBY), tongue tip coordinates (TTX, TTY), tongue center coordinates (TCX, TCY), and tongue side elevations (TS1, TS2, TS3, TS4),
- 2 lip parameters: lip separation distance (LD) and lip protrusion (LP),
- 2 hyoid coordinates (HX, HY),
- jaw opening angle (JA),
- velum shape (VS).

By adjusting anatomical parameters in VTL, speech production of different speakers can be simulated. Apart from the default male adult speaker, VTL 2.1 offers a 1-year old infant speaker model based on [24]. For both speakers, vocal tract configurations are available for [@], [a], [i], [u], which were fitted to dynamic MRI data. These predefined motor parameters are termed "mentor parameters" and used for training the auditory memory and for analyzing the learning; they are not accessible to the RL agent during imitation learning (except [@], see below).

All imitation simulations are initialized in the configuration of [@], which is produced in the most relaxed vocal tract state and thus can likely be uttered by newborn infants. Thereby the modeled stage of babbling is when the infant reliably utters [@] and explores the vocal tract to establish the so-called corner vowels [a], [i], [u].

Formally, VTL transforms a given set of motor parameters $\vec{m}$ into a sound $s(t)$:

$$\vec{m} \rightarrow s(t). \tag{1}$$

B. The Auditory System

1) The Auditory Processing: The first stage of the auditory system models the peripheral processing in the cochlea. In multiple steps, a sound $s(t)$ is transformed into nerve activations $\vec{n}(t)$ using the dual resonance nonlinear (DRNL) filter model as described in [25]:

$$s(t) \rightarrow \vec{n}(t). \tag{2}$$

The model is implemented in "BRIAN hears" [26], an extension of the BRIAN neural network simulator [27], [28] for auditory processing.

2) The Auditory Memory: The auditory memory transforms peripheral nerve activations $\vec{n}(t)$ into auditory memory responses $\vec{a}(t)$, which form the basis for sound evaluation during imitation:

$$\vec{n}(t) \rightarrow \vec{a}(t). \tag{3}$$

It is realized by an *Echo State Network*, which is a specific implementation of reservoir computing using analog units [29], [30]. It comprises a pool of non-plastic, recurrently connected units (*static reservoir*) and a set of output units (*linear readout*), which are connected to the static reservoir and are responsible for evaluating the input. Though simplistic, ESNs are widely used neural networks whose units correspond to neuron populations with variable firing rates.

Here, the static reservoir responds to the output of the cochlea $\vec{n}(t)$ and corresponds to a non-plastic version of biological circuits in the auditory cortex. The connection between the reservoir and the readout carries the auditory memory and is adapted during target acquisition. During imitation learning, the response of the linear readout $\vec{a}(t)$ is the basis for sound evaluation. Sound evaluation is crucial for imitation, as it provides the reward for the RL agent. It is important to note that target acquisition is realized by supervised learning, whereas imitation proceeds as reinforcement learning.

An ESN implementation in *Oger* (OrGanic Environment for Reservoir computing) was used, a Python toolbox for training and implementing various forms of reservoir computing [31].

3) Target Acquisition: Target acquisition was modeled by supervised training of the ESN-to-output weights. Using VTL, training samples were generated using both the adult male and the infant speaker: For each target vowel (/a/, /i/, /u/) and each speaker, white noise was added to the mentor coordinates with standard deviations 0.1 and 0.01, and 100 samples were created for each possible combination. These samples were manually categorized using four classes independent of speaker (/a/, /i/, /u/, null class (none of the previous)).

The samples and their corresponding labels were then used for training the auditory memory.

4) Softmax Classification: The scalar sensory reward is obtained from the ESN response after target acquisition. It is based on confidence levels (see below) and ultimately on the classifier output.

The response can be interpreted as a set of *confidence levels* that the current sample belongs to a certain class: Given the normalized and time averaged output activations $\langle \vec{a}(t) \rangle$ for all classes, we define the confidence c_v that the current sample belongs to class v as a softmax function:

$$c_v := \frac{\exp(\langle \vec{a}(t) \rangle_v)}{\sum_i \exp(\langle \vec{a}(t) \rangle_i)}. \quad (4)$$

The relevant entity for imitation is the confidence that the current sound belongs to the imitation target class (c_{target}). This confidence is finally selected as the current reward and passed to the RL agent.

C. The RL Agent

"Given my judgment of the sound I just uttered, which vocal tract setup will I try next in order to approach my imitation target? Which vowel do I imitate in the first place?" This kind of decision making is embodied by the RL agent and described in the following subsections.

1) The Search Algorithm: The reinforcement learning task is formulated as a black-box optimization problem: Maximize the reward by tuning parameters (sampling actions) that trigger the reward in an unspecified way. The agent can only access the motor parameters $\vec{m}$ and the sensory reward c_v, so the mapping between these entities (the environment) is treated as a black box.

The search algorithm is realized by *Covariance Matrix Adaptation – Evolution Strategy* (CMAE-ES) [32]. CMA-ES is a stochastic optimization algorithm, which determines local optima by sampling and evaluating points $\vec{m}_i$ in the search space. These points are sampled by a multidimensional Gaussian distribution $\mathcal{N}(\bar{m}, \Sigma)$ with mean $\bar{m}$ and covariance Σ. This distribution is then modified based on the samples' elicited rewards $c_{v,i}$. The algorithm has found a local optimum whenever the sampled parameters have converged.

The *motor memory* contains the knowledge to produce speech sounds. In the model, it contains the sets of motor parameters the agent has learnt to associate with a vowel by successful imitation and that he can access any time. It contains only the mentor configuration of [@] in the beginning.

The search is initialized in the mentor configuration of [@], i.e. the search distribution's mean $\bar{m}$ is set to this configuration. N samples are drawn from the search distribution:

$$\vec{m}_1, \vec{m}_2, \ldots, \vec{m}_N \sim \mathcal{N}(\bar{m}, \Sigma), \quad (5)$$

whereby each sample $\vec{m}_i$ is synthesized by VTL. The produced sound $s_i(t)$ is processed and evaluated by the auditory sytem; the auditory system passes the reward $c_{v,i}$ to the agent, signifying the imitative value of the sample $\vec{m}_i$ regarding the current target v. The samples $\vec{m}_i$, their elicited rewards $c_{v,i}$, and the search parameters $\bar{m}, \Sigma$ yield the new search parameters $\bar{m}', \Sigma'$, which are then used in the next search iteration:

$$\vec{m}_1, \vec{m}_2, \ldots, \vec{m}_N, c_{v,0}, c_{v,1}, \ldots, c_{v,N}, \bar{m}, \Sigma \rightarrow \bar{m}', \Sigma'. \quad (6)$$

As soon as one of the elicited rewards is above the *reward threshold*, the search stops. The reward threshold was set empirically to indicate that the corresponding sound is indistinguishable from the target for human listeners. The motor parameters corresponding to this supra-threshold reward are added to the motor memory, and the current target is considered learnt. A new target is selected if any are left (see below). Otherwise, simulations are terminated, and the imitation phase is over.

Terminating the search when the reward threshold has been crossed follows from a pragmatic consideration: We hypothesize that infants learn a skill not until they have mastered it to perfection, but until they are "good enough". In the case of speech acquisition, the ultimate goal is communication. Infants engage in speech imitation – and thereby learn to speak – up to the point when their imitation is successful and they are understood. Imitation is successful if it is acknowledged as similar to the imitation target; thus oral communication unfolds. During development, time is precious, and perfection (convergence) is time-consuming. Especially when many skills

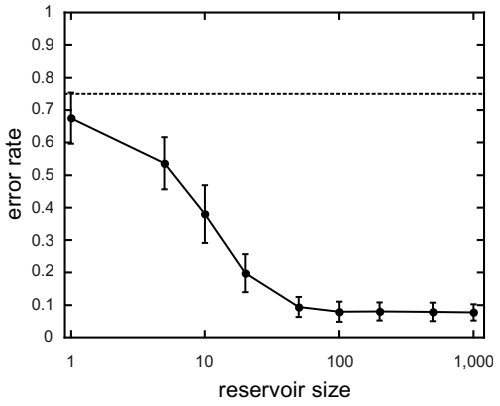

Fig. 2. The error rate of test samples depends on the reservoir size. Each data point is the mean error rate of 100 trained ESN readouts; the error bars are the corresponding standard deviations. The trials differed in the connectivity of the ESN and the test set. The dashed line indicates chance level.

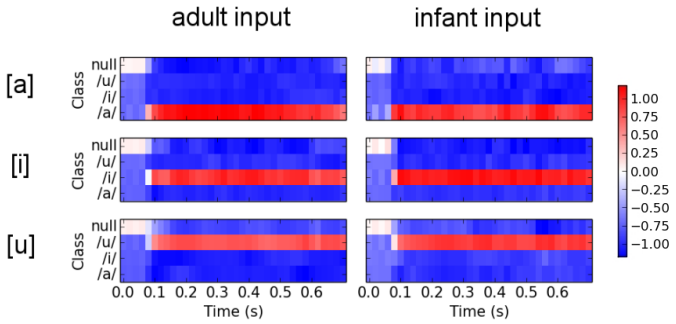

Fig. 3. Readout activations over time while prototypical vowel samples are fed into the auditory system. These prototypes were generated using the mentor configurations of both the adult (*left*) and infant speaker (*right*). The readouts were trained using reservoirs with 1000 units.

are to be learnt, perfecting one skill seems a waste of time when other skills are neglected that are similarly important on a pragmatic level.

Whenever the search algorithm has converged on a local optimum, the search distribution is reset to a random item of the motor memory, and the search is restarted. This is because convergence indicates failure to imitate, as convergence can only be reached for sub-threshold rewards.

2) Target Selection: When imitation starts and whenever the agent has learned a vowel, the agent needs to select a new imitation target. To this end, samples are generated based on the current search distribution. These samples are evaluated by the environment (VTL and auditory system) and assigned confidences. The highest confidence corresponding to a remaining target class determines the next target:

- All confidence values for remaining targets are considered.
- The target with the highest confidence is selected as the next imitation target.

This mechanism leads to developmental exploration: The agent learns targets first which are easy to reach in the motor space. After a vowel has been learnt, the agent is more likely to select as next target a vowel whose motor parameters lie close to the previously learnt ones, thus minimizing search time. Over time, the agent commits itself to targets that are harder to reach (further afar in the motor space). This is in line with pragmatic self-guided learning, as discussed in the previous section: Time is valuable during skill acquisition; the agent aims to learn as many skills "well enough" as fast as possible (well enough to accomplish specific goals, e.g. communication).

III. Results

A. Target Acquisition

Target acquisition was simulated using ESNs of varying size. Thus it was quantified how the classification depends on the reservoir size.

The test error for reservoir sizes N between 1 and 1000 units is shown in Fig. 2. The error rate is maximal for $N = 1$ (ca. 0.68) and decreases monotonically for increasing N up to $N = 100$, where it plateaus at around 0.08.

The readout responses to the infant and adult prototypes of /a/, /i/, and /u/ for $N = 1000$ (Fig. 3) show that readouts with large reservoirs are able to classify the unperturbed vowels with high specificity. Because these unperturbed vowels act as goals for the imitation phase, the readout is able to clearly signal once the agent reaches the target. Thus these readouts fulfill the model requirements and can be used during imitation.

In the following, we consider ESNs with $N = 1000$ because they display the smallest error rate.

B. Imitation

Using a trained readout with an ESN of size 1000, imitation learning was investigated. In two paradigms, the agent controlled different sets of motor parameters. Each of these paradigms represents a different dimensionality and complexity of the learning task.

- Visually Guided Learning: The agent controls all but the jaw and lip coordinates (JA, LP, LD) corresponding to a 13-dimensional motor space. This paradigm simplifies the full motor learning problem; here the agent knows the appropriate jaw and lip configurations from the start. The jaw and lips are the only articulators that are visible to the imitator. The imitator can thus assess the correct jaw and lips setup visually and doesn't need to rely solely on auditory information.
  This setting is realistic considering neonatal imitation of facial gestures [2], [3]. If neonatal imitation exists, it provides part of the learning problem's solution via orofacial (e.g. lips and jaw) imitation during articulation, thus simplifying vowel imitation. This reasoning is the basis of the visually guided learning paradigm.
- The Full Problem: The agent controls all parameters and searches for solutions in the 16-dimensional motor space. The agent is not guided in any way, which maximizes the difficulty and complexity of the learning problem. This

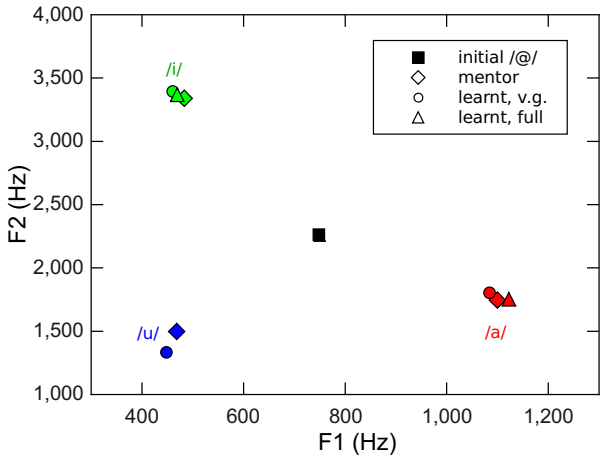

Fig. 4. Imitation learning in the space of the first two formants (F1-F2 space). The agent's initial speech sound is [@] (*square*). Starting from there, the agent imitates the vowels of the mentor (*diamonds*) and learns to produce acoustically similar versions (*circles* and *triangles*). The indicated mentor vowels are based on the mentor configurations of the infant speaker. *v.g.*: visually guided 13-dimensional learning paradigm. *full*: full 16-dimensional learning paradigm.

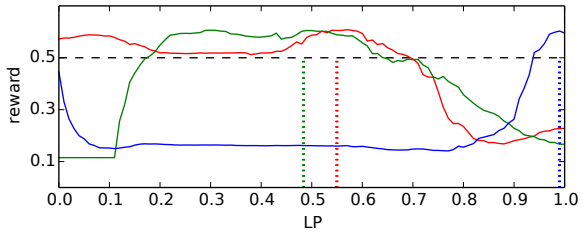

Fig. 5. The relationship between the reward and one-dimensional motor perturbations: The elicited reward is plotted as a function of the normalized lip protrusion (LP) when the other parameters take on the values of the mentor configuration of the imitation target (*red*: [a], *green*: [i], *blue*: [u]). The dashed horizontal line denotes the reward threshold for accepting a sampled vowel. The dotted vertical lines mark the mentor coordinates of the corresponding imitation target. The agent controls the LP only during the full learning problem.

scenario corresponds to vocal learning without neonatal imitation of facial gestures, e.g. in blind infants.

In the visually guided learning paradigm, the agent explored a 13-dimensional motor space and successfully imitated all target vowels. However, using all degrees of freedom for imitation learning, the agent failed to imitate [u]. This judgment is based on the fact that [u] was not successfully imitated during 10 days of simulation time, which is orders of magnitude larger than the needed time to learn the other vowels (between 10 and 20 minutes).

The agent was able to reproduce the acoustic features of the learnt target vowels, which is reflected in the formant space, where the proximity between learnt and prototypical vowels indicate acoustic similarity (Fig. 4).

However, considerable differences exist between the learnt vocal tract configurations and the mentor configurations. Overall, the agent used more extreme positions in the individual degrees of freedom than the mentor, yet the learnt target sounds were matched well.

To find the reason why [u] was learnt during visually guided learning but not in the full learning paradigm, we investigated how the reward depends on perturbations of motor parameters. This consideration reveals the most difficult parameter to learn: the lip protrusion (LP) for imitating [u] (Fig. 5). For this case, the reward is minimal in most of the coordinate range (ca. from 0.1 to 0.8) and peaks above threshold in a narrow interval in an extreme position, i.e. within about 0.95 to 1.0. Also, the peak lies apart from the target ranges of [a] and [i], which makes it difficult to discover [u] right after another vowel has been learnt. At the other end of the extreme (at LP=0), there is even a sub-threshold peak, which enables the search to get stuck in this local optimum. Considering the fact that these dependencies only reflect the simplified case of one-dimensional perturbations from an assumed optimum, it seems likely that the LP-supra-threshold range is even smaller during simulations, when few parameters are optimal. The difficulty of sampling lies not only in the width of the target ranges; it also increases exponentially with the number of dimensions. So in the full 16-dimensional problem, imitation of [u] by exploration alone is possible, yet highly unlikely.

IV. Discussion

In this preliminary study we demonstrated that vowel acquisition by imitation is possible if imitation is preceded by auditory learning. We compared the imitation problem including all articulator positions with a visually guided imitation paradigm in which the agent has access to ideal positions of the visually accessible lips and jaw. The agent was able to reproduce all imitation targets ([a], [i], [u]) in the visually guided paradigm and failed to acquire [u] in the full learning problem. We showed that this failure is due to a sensitive dependency of the sound evaluation on the lip protrusion if [u] is the imitation target.

Using auditory feedback alone, it seems hardly possible to infer the exact shape of the lips needed to articulate [u]. Yet the lips' shape is a salient visual feature in the face of a speaker, which could be easily picked up by orofacially imitating infants.

This raises the question how blind infants acquire vowels like [u]. Experimental evidence indicates that missing visual information may impair the quality of speech acquisition [33], [34], [35]. According to Mills, "visually-impaired children clearly follow a different and slightly slower path in their early phonological acquisition", which suggests a significant role of visual feedback during babbling [36].

These results motivate further extensions of our model. We plan to use more diverse speech items as imitation targets, i.e. to model the acquisition of realistic vowel systems and and consonant-vowel syllables. We suspect that the inclusion of other rounded vowels in natural languages – such as [o] – might help infants to acquire [u].

We also plan to address the problem of speaker normalization. During imitation, our model relies on acoustic

matching between the produced sounds and the target vowels. Because infants' utterances differ acoustically from those of adults, voice-invariant acoustic matching of infants' and adults' speech sounds is non-trivial. We circumvented this problem by including prototypical vowels produced by the infant speaker in the training data during auditory learning, which is somewhat unrealistic. To tackle the problem of speaker normalization, we plan to diversify the speakers of the infant's environment. In a more realistic setting the infant receives input from many different proficient speakers during auditory learning. Moreover, their infant-directed speech simplifies acoustic matching by mimicking the high pitch of the infant. Modeling these aspects might facilitate generalization in the auditory system by the learning of voice-invariant features and enable correct evaluations of the infant speaker's speech samples without including the infant speaker in the auditory learning.

Finally, we plan to replace the ESN by an unsupervised training method for modeling auditory learning. So far, the use of ESNs places unbiological constraints on our work, such as the need for manual labeling of the training samples and the predetermination of the number of speech items to be learnt. By adopting the previous extensions as well as a biologically plausible unsupervised learning method, we hope to achieve a higher level of realism and gain new insight into vocal babbling.

ACKNOWLEDGMENT

This work was supported by the Quandt foundation.

REFERENCES

[1] B. Mampe, A. Friederici, A. Christophe, and K. Wermke, "Newborns' cry melody is shaped by their native language," *Current Biology*, vol. 19, 2009.

[2] A. Meltzoff and M. Moore, "Imitation of facial and manual gestures by human neonates," *Science*, vol. 198, no. 4312, 1977.

[3] E. Simpson, L. Murray, A. Paukner, and P. Ferrari, "The mirror neuron system as revealed through neonatal imitation: presence from birth, predictive power and evidence of plasticity," *Philosophical Transactions of the Royal Society B*, vol. 369, 2014.

[4] F. Guenther, S. Ghosh, and J. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language*, vol. 96, 2006.

[5] F. Guenther, "Cortical interaction underlying the production of speech sounds," *Journal of Communication Disorders*, vol. 39, 2006.

[6] F. Guenther and T. Vladusich, "A neural theory of speech acquisition and production," *Journal of Neurolinguistics*, vol. 25, 2012.

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

[8] B. Kröger, J. Kannampuzha, and E. Kaufmann, "Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception," *EPJ Nonlinear Biomedical Physics*, vol. 2, 2014.

[9] C. Moulin-Frier and P.-Y. Oudeyer, "Curiosity-driven phonetic learning," in *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE, 2012.

[10] ——, "The role of intrinsic motivations in learning sensorimotor vocal mappings: a developmental robotics study," in *Interspeech*, 2013.

[11] C. Moulin-Frier, S. 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, 2014.

[12] I. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, 2011.

[13] ——, "Learning to pronouce first words in three languages: An investigation of caregiver and infant behavior using a computational model of an infant," *PLOS One*, vol. 9, 2014.

[14] K. Miura, M. Asada, K. Hosoda, and Y. Yoshikawa, "Vowel Acquisition based on Visual and Auditory Mutual Imitation in Mother-Infant Interaction," in *International Conference for Development and Learning, ICDL 2006*, 2006.

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

[16] A. Philippsen, R. Reinhart, and B. Wrede, "Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model," in *2014 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 2014.

[17] P. Marler, "Birdsong and speech development: Could there be parallels? there may be basic rules governing vocal learning to which many species conform, including man," *American Scientist*, vol. 58, 1970.

[18] A. Doupe and P. Kuhl, "Birdsong and human speech: common themes and mechanisms," *Annual Reviews in Neuroscience*, vol. 22, 1999.

[19] L. Baptista and L. Petrinovich, "Social interaction, sensitive phases, and the song template hypothesis in the white-crowned sparrow," *Animal Behaviour*, vol. 32, 1984.

[20] M. Konishi, "From central pattern generator to sensory template in the evolution of birdsong," *Brain and Language*, vol. 15, 2010.

[21] J. Triesch, "Imitation learning based on an intrinsic motivation mechanism for efficient coding," *frontiers in Psychology*, vol. 4, 2013.

[22] P. Birkholz, "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis," *PLoS ONE*, vol. 8, 2013.

[23] S. Prom-on, P. Birkholz, and Y. Xu, "Identifying underlying articulatory targets of thai vowels from acoustic data based on an analysis-by-synthesis approach," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 23, 2014.

[24] P. Birkholz and B. Kröger, "Simulation of vocal tract growth for articulatory speech synthesis," in *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007.

[25] E. Lopez-Poveda and R. Meddis, "A human nonlinear cochlear filterbank," *Journal of the Acoustical Society of America, JASA*, vol. 110, 2001.

[26] B. Fontaine, D. Goodman, V. Benichoux, and R. Brette, "Brian hears: online auditory processing using vectorization over channels," *frontiers in Neuroinformatics*, vol. 5, 2011.

[27] D. Goodman and R. Brette, "Brian: a simulator for spiking neural networks in Python," *frontiers in Neuroinformatics*, vol. 2, 2008.

[28] ——, "The Brian simulator," *frontiers in Neuroscience*, vol. 3, 2009.

[29] H. Jaeger, "Echo state network," *Scholarpedia*, 2007.

[30] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science*, vol. 304, 2004.

[31] D. Verstraeten, B. Schrauwen, S. Dielemann, P. Brakel, P. Buteneers, and D. Pecevski, "Oger: Modular Learning Architectures For Large-Scale Sequential Processing," *Journal of Machine Learning Research*, submitted.

[32] N. Hansen, "The CMA evolution strategy: a comparing review," in *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, J. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, Eds. Springer, 2006, pp. 75–102.

[33] A. Mills, "Acquisition of speech sounds in the visually-handicapped child," in *Language acquisition in the blind child*, A. Mills, Ed. Croom Helm, London, and College Hill, San Diego, 1983.

[34] ——, "The development of phonology in the blind child," in *Hearing by eye: The psychology of lip-reading*, B. Dodd and R. Campbell, Eds. Erlbaum, London, 1987.

[35] A. Mills, C. Meinecke, and H. Hattig, "Die rolle der visuellen information im spracherwerb," University of Tbingen, DFG Abschlussbericht, 1983.

[36] A. Mills, *Language Development in Exceptional Circumstances*. Lawrence Erlbaum Associates, Hove, Hillsdale, 1993, ch. Visual handicap, pp. 150–164.