MODELLING VOWEL ACQUISITION USING THE BIRKHOLZ SYNTHESIZER

Ian S. Howard[1] & Peter Birkholz[2]
[1]Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, PL4 8AA UK.
Email: ian.howard@plymouth.ac.uk
[2]Institute of Acoustics and Speech Communication, TU Dresden, 01062 Dresden, Germany.
Email: peter.birkholz@tu-dresden.de

Abstract: Human infants have a remarkable ability to learn to speak. To examine theories of some aspects of speech production development we previously developed Elija, a computational model of infant speech acquisition. Elija is an agent that can influence its environment by generating acoustic output by controlling an articulatory synthesizer as well as receiving somatosensory feedback from the environment. We first describe the Elija model more formally within the framework of reinforcement learning. Then we implement Elija's vocal apparatus using the more sophisticated 3-D articulatory Birkholz synthesizer instead of the Maeda model used previously. Here we focus on vowel learning and show that, despite the increase in synthesizer complexity, the Elija model agent can still learn to generate vocalic speech sounds unassisted. Subsequently, using a selection process by a caregiver, Elija can refine these utterances leading to a set of L1 vowels. We present examples of the discovered vowels and show that they compare favorably to standard vowel configurations made available with the Birkholz synthesizer.

1 The Elija Model

During the acquisition of speech production, infants progress through several identifiable developmental stages [1]. Elija is a computational model of infant speech acquisition [2]-[4], which previously ran in three such stage, as illustrated in Fig. 1. Elija has a motor system that drives a simulated vocal tract and a perceptual system that evaluates sensory input. Elija's motor system can generate exploratory articulations and a perceptual system can evaluate the value of actions based on their sensory consequences. To generate speech utterance, Elija needs to learn appropriate motor patterns, which take the form of values for the vocal tract synthesizer control parameters as a function of time.

The acquisition of speech production in the Elija model begins by means of an unsupervised active learning phase, in which Elija discovers how to produce potential speech sounds based on vocal self-exploration (Fig. 1.1). This is formulated as an optimization problem, setup to find motor patterns that generate salient and diverse sensory output. During this process, the act of generating a vocal action results in proprioceptive, tactile and auditory sensory consequences. Internal evaluation of the salience of this output is used to provides an estimate of the value of each motor action. This can be used to improve the next production attempt of the same utterance using gradient descent.

This babbling phase is followed by a stage involving reformative interaction with a learned caregiver of the target language L1 (Fig. 1.2). The reformulation process arises from the mirroring behavior on the part of the caregiver that is observed to natural arise in caregiver-infant interaction situations, and this phenomenon appears to be an instinctive behavior. This interaction selectively reinforces Elija's range of potential speech sounds. This causes Elija to retain motor patterns to which the caregivers responded. In addition, it enables Elija to associates his vocal actions with the speech sounds he hears in response to them. This results in Elija learning correspondences between his speech tokens and those of the caregivers. Importantly, the content of the correspondences is based on a judgment of sound similarity made by the caregiver rather than by Elija. Thus, reformulations allow association of adult linguistic form to infant's utterances, leading to the ability of Elija to imitate the caregiver.

In a final word learning phase (Fig. 1.3) we previously showed that caregivers speaking three different European languages were able to teach Elija by imitation to pronounce typical first words in English, French and German [4].

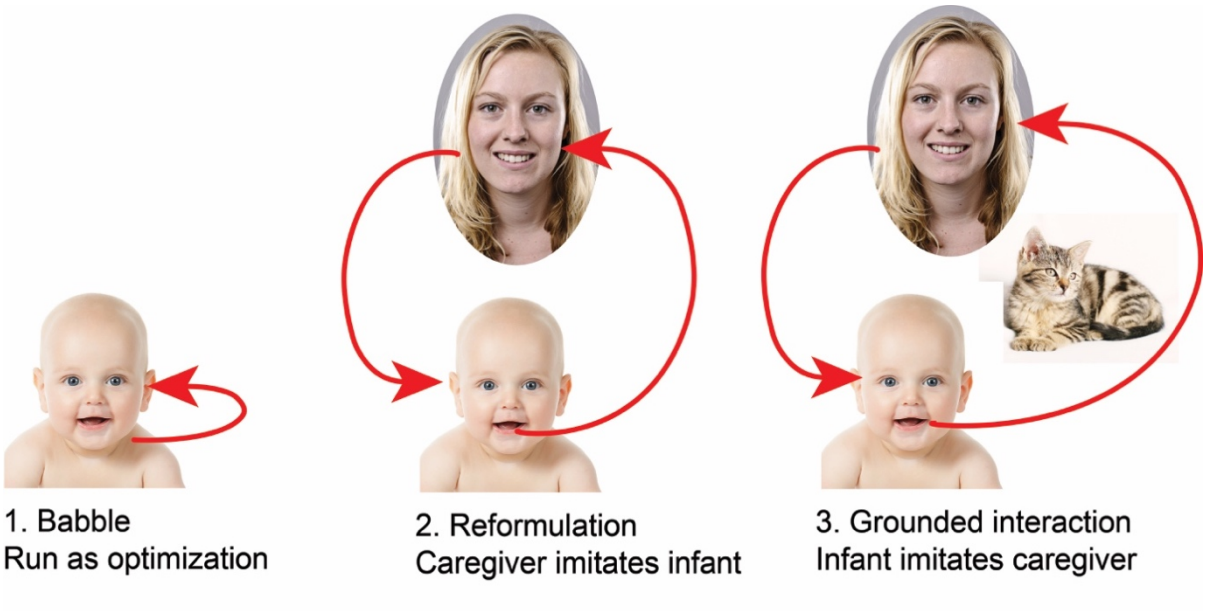

Figure 1 - Separate stages of interactions used in the original Elija model. 1 Elija begins by teaching himself to make sounds by exploration of his vocal apparatus. 2 Elija's repertoire of motor patterns is shaped through interaction with a caregiver. This has two effects: Reformulations reinforce his motor patterns and associate them with any caregiver responses 3. During the word imitation experiments, the caregiver teaching the infant to pronounce first words.

2 Reinforcement learning

The field of reinforcement learning (RL) is concerned with building agents that interact with the environment and can learn to generate appropriate state dependent actions in order to maximize cumulative reward, and thereby achieve a goal. Stated formally, at time t when a RL agent is in state S_t, the agent generates an action A_t. This results in a reward from the environment R_t. The action also leads to an observation of the environment O_t which the agent can use to update its state. This process is illustrated in Fig. 2. No supervisor is required to drive learning in this process, just reward from the environment. Of course, for the agent to operate effectively it is important that the agent's state adequately describes the circumstances the agent in, so it has enough information about the environment to appropriately generate future behavior.

A policy π defines how the agent behaves. It is the agent's probability of choosing a given action in a given state. In the case of Q-learning [5], learning Policy π involves learning a state-action pair function Q(s, a) to specify its behavior, which indicates how good actions are given the state. Alternatively, policy gradient methods try to learn the optimal policy π directly [6]. Initially the agent will not know policy π and consequently will not know which actions to generate in a particular state. RL agents can learn to behave optimally by exploring possible actions in a given state to find their long-term cumulative reward. After the agent has learned the policy π , it can then simply follow and exploit it to generate the optimum behavior.

Thus, reinforcement learning generally operates in two modes. In the first mode, the agent randomly explores its environment, generally only with a small probability, and learns the value of its actions by making use of a reward signal relating to the values of the actions. Using the information gained form the environment by exploration, in the second mode of operation it

can exploit this and generate the optimal action for a given state. Thus, action exploration is inherently a part of the RL framework, and the exploration mechanism plays an important part in learning. Conversely exploitation is the way to make use of the learned policy to behave optimally.

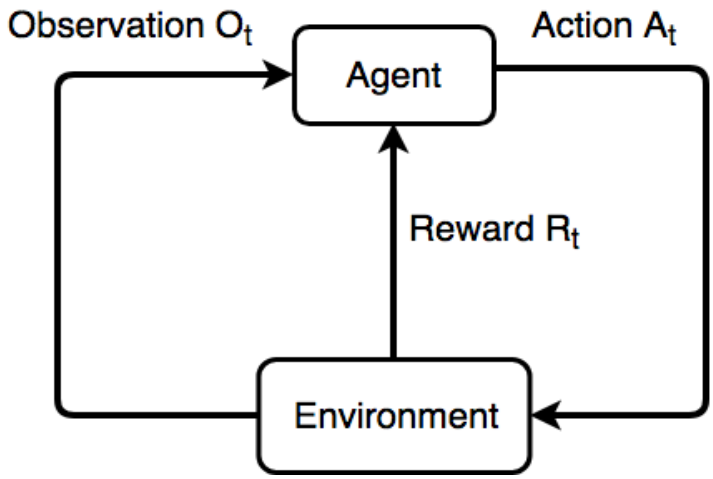

Figure 2 – Reinforcement agent interaction with its environment. At time t the agent in is state S_t, generates action A_t. On the basis of the action it received reward R_t and makes an observation O_t from the environment. The agent can then use the observation to update its state.

In simple tasks a lookup table can be used to map between state and value, or indeed to represent policy directly. Indeed, in the case of the vowel learning scenario we tackle here, such an approach is adopted, since a single pattern vector corresponds to the configuration of the vocal apparatus for a given vowel. However, is should be noted that as the number of states increases, it rapidly becomes intractable to directly model policy in this way. Therefore, in order to avoid the curse of dimensionality in real-world RL tasks, it becomes necessary to make use of generalization functions to implement policy. Indeed a recent impressive application of RL was the Atari game playing system developed by Minh et al. [7]. The latter application deals with high-dimensional sensory inputs using deep learning strategies implemented using in multi-layered feed forward convolutional networks and learnt to play video games to a better than human level of performance

One issue in RL is that most current methods focus on learn policies that generate discrete actions. However, many problems require continuous control. This is particularly the case in the control of dynamical systems. Fortunately, recent work indicate that progress in continuous control is now also being made by RL methodologies [8].

3 RL framework for Elija

We now formulate the Elija model formally within the framework of RL [6]. We first note that the motor pattern representing vocal tract configuration is not discrete, but rather a vector of continuous values, with different vectors corresponding to different vowel quantifies. Within the policy gradient formulation, to learn policy to attain maximum reward, motor parameters can to be modified by gradient descent to optimize utterance reward. We note that this is precisely what is implemented in the sound discovery phase in the Elija model, in which gradient descent is used to optimize the production of simple sounds. Thus, the self-organizing babbling stage in Elija is essential a policy gradient method. We note however in its current implementation in Elija, the gradient is computed within the optimization function on the basis of successive evaluations of performance of the vocal apparatus.

In the case of learning static sound qualities, such as constant vowels, the credit assignment problem is simple. In this case it is clear that the parameters responsible for the static configuration of the articulators resulted in the action and subsequent reward arising from sensory consequences. Such reward can either be generated internally on the basis of salience and diversity, or by a response signal from a caregiver.

However, RL is typically applied to more sophisticated problems that require the generation of sequential action. In such cases, the feedback of reward is generally delayed until a final goal is reached. This is the scenario for utterance generation that involves movement of the articulators, (i.e. in complex utterance generation made up of different articulator configurations) in which their temporal sequence of actions subsequently leads to the generation of dynamic sound qualities. In this case the problem needs to be solved as an RL problem that follows a Markov Decision Process (MDP) [6]. However, we do not concern ourselves further with these issues here, since we only investigate learning vowel qualities.

4 Birkholz synthesizer

The major contribution of this current work is to implement Elija's vocal apparatus using the Birkholz synthesizer [9]. This is based on a 3D geometric articulatory model fitted to the anatomy and articulation of a male reference speaker based on static and dynamic MRI data [9], [10]. The parameters used to describe the geometrical shapes and positions of the articulators are mainly based on the work of Mermelstein [11], which were extended to deal with a 3D model. The model uses 7 wireframe meshes; one for the palate and posterior larynx wall; one for the anterior side of the larynx, pharynx and lower jaw; one each for the upper and lower lips; one each for the upper and lower teeth; and one for the tongue. The velum, lip and tongue mesh are deformed as required by the articulation.

The velum has two parameters that define its shape (VS) and the degree of opening of the velopharyngeal port (VO). Lips are defined in terms of protrusion of lip corners (LP) and lip height, which is the distance between the upper and lower lip (LD). Lip deformation then follow the "law" for lips proposed by Abry et al. [12]. The position of the jaw is determined by JX and JA, which determine its protrusion and opening angle. The hyoid is defined by its horizontal (HX) and vertical positions (HY).

The tongue is somewhat more complicated. The mid-sagittal contour of the tongue is described by two circular arcs and two rational Bezier curves. The larger tongue body circle is specified by radius parameter TCR and its location by TCX and TCY. The smaller circle has a fixed radius of 0.4cm and location TTX and TTY. The section of the tongue contour from the hyoid to the tongue body circle and the section from the tongue body circle to the tongue tip circle are Bezier curves with the additional shape parameters (TRX, TRY) and (TBX, TBY).

The elevation of the tongue sides is defined at 4 equidistant locations along the mid-sagittal line from the by parameters at the root (TS1), central dorsum (TS2), blade (TS3) and tip (TS4).

The Birkholz synthesizer comes supplied with pre-defined configurations to generate different vowels (as well as other sounds). These configurations were obtained from acoustic data from a single male subject, and estimated using an analysis-by-synthesis approach [13]. The availability of such data is very helpful, since such vocal tract configurations can be used as a benchmark in the investigation of other approaches to speech sound acquisition, like the one we adopt in this current work.

We note that using the Birkholz synthesizer offers new challenges to the Elija model, since it exhibits a larger number of degrees of freedom than the Maeda synthesizer [14] - in our implementation of the Maeda articulatory synthesizer, ten parameters were used to control the vocal apparatus. In addition, its control parameters are more anatomically inspired than those in the Maeda model, and are not statistically optimized to independently contribute to the vocal tract area function.

5 Unsupervised motor pattern discovery

Here we limit our investigation to vowel discovery. The vocal apparatus used is that of an adult male subject because standard vocal tract configurations were available for a range of vowel qualities (JD2) so that the results obtained by Elija could be compared with these pre-existing vowel qualities.

As before, Elija first discovers potentially useful articulations in an unsupervised manner by finding motor patterns that are solutions to an optimization problem [2]. In this simple case, reward relating to the performance of a motor pattern is defined as only a sum of its sensory salience, and diversity. The objective function is given by the expression for cost J (negated reward)

$$J = -\sum(salience + diversity)$$

To discover potentially useful articulator configurations, optimization of the objective function was carried out using gradient descent to find values of the motor pattern that minimize cost J. This was achieved for multiple runs to discover a range of vowels. The motor pattern was always started from a random initial position and a Quasi-Newton gradient ascent algorithm was used to find a solution, as implemented by the Matlab function *fmincon*. The control parameters values were constrained to their valid limits within the synthesizer.

The *diversity* term included in the objective function is very important and ensures the formation of a wide range of motor memories. It results in Elija performing active learning and encourages exploration of previously untried articulations that generate novel sensory feedback. It is computed by comparing the spectral representation of the current pattern's sensory consequences from those of previously discovered patterns in terms of Euclidian distance. Using this acoustic similarity metric, this leads to the discovery of vocalic sounds that are acoustically distinct. During this stage operation, no caregiver involvement is required.

Elija was implemented in Matlab (Mathworks Inc, Natick MA, USA) running on a Mac Book Pro. The vowel experiment discovery was run for 100 utterances and a wide range of vocalic sounds were generated, including many that did not constitute L1 in English. Processing took about 2 days. Caregiver reinforcement of L1 sounds by a native English speaker was used, in which desirable sounds are retained and ignored sounds rejected, and this interaction pruned Elija's vowel production towards L1.

6 Results

We compare eight different vowels found by Elija (using optimization and the subsequent caregiver selection interaction) with eight pre-set example vowels, for which the target configurations are supplied with the Birkholz synthesizer [13]. Acoustic listening tests by the author suggested a good range of vowels were discovered by Elija. Although acoustic comparison is a more desirable means of judgement, here we present results as wide band spectrograms. We show synthesized output relating to Elija's discovered vowels and the corresponding example vowels, to provide a means of visual comparison. In the presented results, the categorization of the vowels was carried out by a simple listening test by the author. Both Elija's output utterances and the reference utterances were annotated in this way.

From Fig. 3 it can be seen there is good correspondence between the Elija generated vowels and the corresponding reference values. Examination of the sounds by ear also confirmed the strong sound similarities. This demonstrates that although the Birkholz synthesizer is more complex and the previously use Maeda synthesizer, gradient descent can lead to the discovery of local optimal solutions of the motor patterns that represent good vocalic qualities simply on the basis of using a cost function based on acoustic salience and acoustic diversity. It is worth

pointing out that it's quite difficult to find the settings by hand, but a simple optimization is able to find these things quite easily.

7 Discussion

Previously, Elija's vocal apparatus was implement using the Maeda articulatory synthesizer [14]. The Maeda model represents the vocal tract as a 2-D area function in terms of articulatory parameters found using statistical analysis, and this leads to synthesizer control using a relatively small number of independent parameters. The major contribution of this current paper has been to implement Elija's vocal apparatus using the more sophisticated 3-D articulatory Birkholz synthesizer [9]. This offered new challenges, since the Birkholz synthesizer exhibits a larger number of degrees of freedom than the Maeda synthesizer. In addition, its control parameters are more anatomically inspired than those in the Maeda model, and are not statistically optimized to independently contribute to the vocal tract area function. However, despite the increase in synthesizer complexity, we showed that the Elija model can effectively learn to generate speech sounds by self-organization which can then be refined by caregiver interaction.

Other methods have also been suggested for discovering of speech sound production. The active learning adopted in the Elija model has much similarity to the idea of intrinsic motivation [15] and curiosity [16]. Intrinsic motivation has been proposed as a mechanism to drive development in cognitive robotics and vocal development. There has also been recent interested in predictive models of speech development [17]. We noticed that there are similarities across all of these methods. Indeed, it is also been suggested that predictive models can also play an important role in the formulation of solutions based on reinforcement learning.

As a last note, we mention that formerly Elija learning ran in three distinct experimental stages using only simple mechanisms of active learning, reinforcement and association. This was primarily done to ensure all caregivers heard the same sounds and thereby enable cross-caregiver behaviors to be analyzed. However, this approach also had the benefit of reduce caregiver interaction time, since the babbling phase was completed before the participant acted as caregiver and they didn't have to sit around while Elija learned novel utterances before interaction became possible [4]. In a real infant, self-exploration and caregiver interactions will operate simultaneously. By formulating Elija in term of reinforcement learning it is easy incorporate this more realistic scenario of infant speech development, since the RL framework can just as well deal with internal reward for an utterance or reward arising from caregiver activity such as reformulation. However, we suggest that even with such natural autonomous operation, the former identifiable stages of development will still emerge naturally. This is because that behavior is driven by the development of the agent, which influence the subsequent behavior, and nature of interactions, of the caregiver.

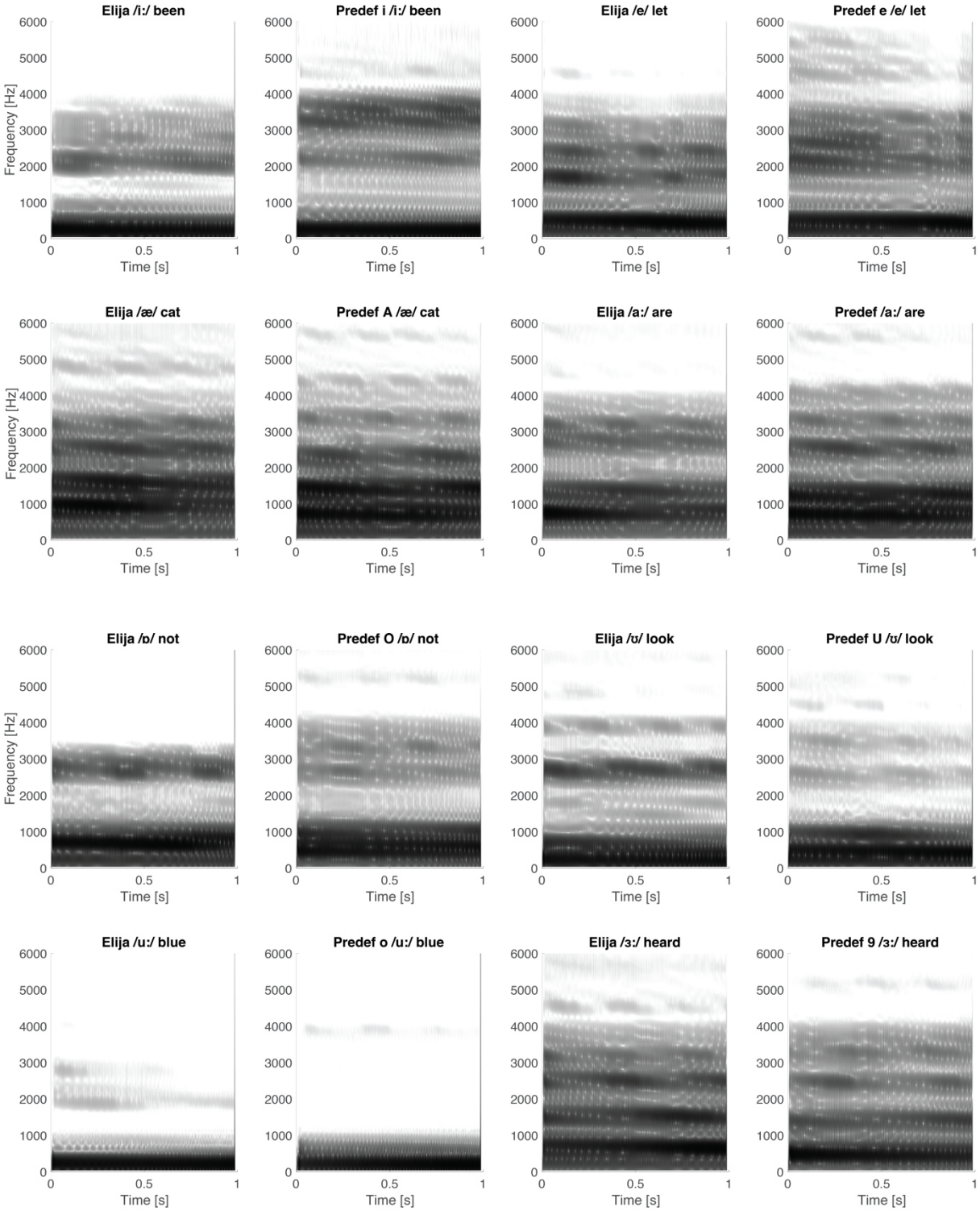

Figure 3 – Comparing vowels. The plots showwideband spectrographic analysis of 1 second of synthesised speech for 8 vowel qualities. The results compare vowels discovered by the Elija model after caregiver selection with comparable predefined values provided for the Birkholz synthesiser.

8 References

[1] D. OLLER, The emergence of the speech capacity. 2000.

[2] I. S. HOWARD AND P. MESSUM, "A computational model of infant speech development," presented at the XII International Conference "Speech and Computer" (SPECOM'2007), Moscow, 2007, pp. 756–765.

[3] I. S. HOWARD AND P. MESSUM, "Modeling the development of pronunciation in infant speech acquisition.," *Motor Control*, vol. 15, no. 1, pp. 85–117, Jan. 2011.

[4] 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.," *PLoS ONE*, vol. 9, no. 10, p. e110334, 2014.

[5] C. J. C. H. WATKINS AND P. DAYAN, "Q-learning," *Mach Learn*, vol. 8, no. 3, pp. 279–292, May 1992.

[6] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. 1998.

[7] V. MNIH, K. KAVUKCUOGLU, D. SILVER, A. A. RUSU, J. VENESS, M. G. BELLEMARE, A. GRAVES, M. RIEDMILLER, A. K. FIDJELAND, G. OSTROVSKI, S. PETERSEN, C. BEATTIE, A. SADIK, I. ANTONOGLOU, H. KING, D. KUMARAN, D. WIERSTRA, S. LEGG, AND D. HASSABIS, "Human-level control through deep reinforcement learning.," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[8] 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, 09-Sep-2015.

[9] P. BIRKHOLZ, D. JACKEL, AND B. J. KRÖGER, "Construction and Control of a Three-Dimensional Vocal Tract Model," presented at the 2006 IEEE International Conference on Acoustics Speed and Signal Processing, 2006, vol. 1.

[10] P. BIRKHOLZ, "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis," *PLoS ONE*, vol. 8, no. 4, p. e60603, Apr. 2013.

[11] P. MERMELSTEIN, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America 53 (4): 1070–1082.*, 1973.

[12] C. ABRY, "'Laws' for lips," vol. 5, no. 1, pp. 97–104, Mar. 1986.

[13] 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," *J AUDIO SPEECH MUSIC PROC.*, vol. 2014, no. 1, p. 23, May 2014.

[14] S. MAEDA, "An articulatory model of the tongue based on a statistical analysis," 1979.

[15] A. G. BARTO, "Intrinsic Motivation and Reinforcement Learning," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, no. 2, Berlin, Heidelberg: Springer, Berlin, Heidelberg, 2013, pp. 17–47.

[16] P.-Y. OUDEYER, F. Kaplan, and V. V. Hafner, "Intrinsic Motivation Systems for Autonomous Mental Development," *IEEE Trans. Evol. Computat.*, vol. 11, no. 2, pp. 265–286, Apr. 2007.

[17] S. NAJNIN AND B. BANERJEE, "A predictive coding framework for a developmental agent: Speech motor skill acquisition and speech production," *Speech Communication*, vol. 92, pp. 24–41, Sep. 2017.