INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Discovering articulatory speech targets from synthesized random babble

Heikki Rasilo[1], *Yannick Jadoul*[1]

[1]Artificial Intelligence Lab, Vrije Universiteit Brussel
hrasilo@ai.vub.ac.be, yajadoul@ai.vub.ac.be

Abstract

In several areas of speech research, articulatory models able to produce a wide variety of speech sounds, not specific to any language, are needed as a starting point. Such research fields include the studies of sound system emergence in populations, infant speech acquisition research, and speech inversion research. Here we approach the problem of exploring the possible acoustic outcomes of a dynamic articulatory model efficiently, and provide an entropy based measure for the diversity of the explored articulations. Our exploration algorithm incrementally clusters produced babble into a number of target articulations, aiming to produce maximally interesting acoustic outcomes. Consonant gestures are defined as a subset of articulatory parameters and are thus superposed on vowel context, to provide a coarticulation effect. We show that the proposed algorithm explores the acoustic domain more efficiently than random target selection, and clusters the articulatory domain into a number of usable articulatory targets.

Index Terms: articulatory exploration, speech synthesis, babbling, speech sound discovery

1. Introduction

Infants' vocalizations develop rapidly during the first year of their lives, from quasivocalic sounds to vowels, and further to canonical and variegated babbling, consisting of alternating consonant and vowel sounds [1,2]. The first words are produced around 12 months of age, but the fine tuning of the articulatory skills takes years of development to reach adult capabilities (e.g. [3]). Feedback by caregivers is seen to guide babbling [4,5], and vocal imitation by caregivers may teach infants the mapping between the acoustic outcomes of their own articulatory productions and the caregivers' vocal productions (e.g. [6,7]).

Computational modeling can be used as a tool to simulate infant language learning. If we manage to implement an artificial system that learns speech related phenomena similarly to a normal infant (given human-like learning environment), we can hypothesize cognitive processes that may underlie speech learning, possibly offering new insights to speech that can also be useful for technical solutions, such as speech recognition. The human-like learning approach often begins with a simulated articulatory model, that is able to produce human-like vocal sounds. Articulatory modeling has been used in a wide variety of speech related studies: studying the neural processes underlying speech production and perception (e.g. [8]), learning speech motor control ([9]), using articulatory representations to boost speech recognition (e.g. [10,11]), learning of speech imitation [12-15] and simulating emergence of sound systems in a population [16].

In numerous studies a method for articulatory exploration is needed, so the learner can discover articulations that lead to useful acoustic speech outcomes. This task is often very challenging due to the high dimensionality of the articulatory parameter spaces, and the non-linearity of the articulatory-to-acoustic mapping – in some articulatory regions, small changes in articulation produce large changes in the acoustic output, and in some regions, large articulatory changes produce small changes in the acoustic output (e.g. [17]).

Guenther's [9] DIVA model uses random babbling, but the work's purpose is not to explore the articulatory space autonomously. Rather the learning in this model is guided by an expert system that already knows the target speech sounds in the articulatory domain. Moulin-Frier, Nguyen and Oudeyer [18] and Najnin, and Banerjee [19] use dynamic articulatory exploration using the DIVA vocal tract model, to show that the complexity of vocalizations can increase and shift towards imitation of adult sounds intrinsically, without hard-coded goals. They do not try to discover phonetically realistic speech sounds automatically.

Many works where articulatory exploration is studied, concentrate only on vowel productions [12, 15, 20, 21]. In many of these works low dimensional acoustic features, such as formant frequencies, are used. When dynamic babbling and consonant sounds are also taken into account, both articulatory and auditory trajectories become continuous and the articulatory state space grows dramatically when compared to static vocal tract configurations. Moreover, high dimensional acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCCs), are often needed to discriminate between unvoiced consonant sounds, where formant estimation is unreliable or impossible.

In studies where babbling is allowed to be dynamic, the focus is often not on vocal exploration, but rather on modeling the acquisition of some specific aspects of speech production and perception or the underlying neural processes, and the babbling is aided by giving a restricted set of possible articulatory parameters (e.g. [8, 22-24]).

Howard and Messum [13, 14] use a vocal tract model to discover dynamic vocal patterns, including consonants and vowels, automatically. They use sensory salience (consisting of acoustic and touch sensations), diversity and articulatory effort to create a self-reward signal for the learner. After a discovery phase, motor patterns are clustered to a smaller number of categories and divided to their consonant and vowel components to recombine to a variety of syllables. To ensure efficient exploration of the complete vocalization space, they used separate optimization runs to discover vowels and consonants.

Here we propose – to our knowledge for the first time – an incremental vocal exploration algorithm, that discovers realistic articulatory consonant (C) and vowel (V) targets incrementally. The algorithm compresses dynamic babble consisting of a sequence of random consonants and vowels produced by a vocal tract model continuously into a small number of

http://dx.doi.org/10.21437/Interspeech.2020-3186

categories, based on their acoustic and articulatory characteristics. We assume that the movements between these targets are constrained by the vocal tract anatomy, and that the task of the speaker is just to execute these targets in a sequence, on certain time moments. Importantly, the consonant targets in this study are defined only for a subset of the 9 articulatory parameters. This characteristic leads automatically to context dependent consonant realizations, e.g. a closure performed only with the tongue tip parameters leaves the other vocal tract parameters untouched, and they can freely vary depending on the previous and following vowel sounds. The context dependency, or coarticulation, of speech sounds is a typical phenomenon of human speech, and an important aspect to take into account in good quality speech synthesis [25]. We also offer a solution of how the acoustic space spanning the very differing characteristics of consonants and vowels can be explored more efficiently than with a simple random sampling.

Our aim in this work is to show that realistic speech sounds can be discovered incrementally with a simple clustering algorithm with a reasonably small number of babble productions, and without external feedback or knowledge of the distribution of sounds in the learning environment or other speakers. This kind of vocal exploration can be used as a starting point when implementing more complicated computational models to study for example speech acquisition, acquisition of imitation skills or population level emergence of phonetic systems. The articulatory-acoustic trajectories of this unsupervised exploration algorithm also offer a starting point for training a rough universal (i.e. not language-specific) speech inversion mapping from acoustics to articulation, that can be then tuned towards specific characteristics of a given language in later learning phases.

2. Articulatory model

We use the LeVI (Learning Virtual Infant) acoustic model, implemented in MATLAB and described in detail in [26] and its supplementary material. The synthesizer produces dynamic trajectories of nine articulatory parameters, given their target locations and target time instances. The 9 parameters control the positions of tongue base and tip (4 parameters in total), hyoid bone position, velum opening, jaw angle, lip protrusion and lip length. The movements between points follow smooth minimum-jerk trajectories, known for example from human arm movements (see e.g. [27]).

In the simulations we use two different kinds of targets: vowel and consonant targets. Vowel targets have all the nine parameter values defined and provide the context onto which consonants are superposed to. Consonant targets have only a subset of the 9 targets defined. Thus, a vowel target might look like a vector [0.3, 0.7, 0.2, 0.2, 0.2, 0.5, 1, 0.2, 0.2] and a consonant target [nan, nan, 0.8, 0.9, nan, nan, nan, nan, nan]. If these two targets are placed subsequently in time, only the tongue tip x and y coordinates, defined for the consonant, move to the target positions at the given moment in time. Consonant targets are given a lookahead time of 150 ms, a hold time of 100 ms and a release time of 150 ms (see [26] for details. All the parameters take values in the range [0, 1].

3. The learning algorithm

We set the algorithm to search for $T = 200$ speech sound targets in total. The general idea of the algorithm is to produce babbled vocalizations one by one, and after every production, merge the obtained acoustic and articulatory vectors so that maximally T categories remain. The merging (or clustering) method used defines what kinds of target vectors are left when the babbling is terminated. We thus maintain a list, L_{art}, of T articulatory target vectors and a list, L_{acu}, of corresponding MFCC feature vectors representing the approximate acoustic outcome of the articulatory target. Additionally, a counter vector c is saved for all T targets, to keep track how many produced vectors have been merged into each target.

The model is set to choose two vowels and two consonants on each iteration, and produce a $V_1C_1V_1C_1V_2C_2V_2C_2$ vocalization on every iteration, where the interval between consecutive targets is set to 400 ms. The interesting articulatory targets are searched from a region of [t – 100 ms, t + 100 ms], where t is the time instance of the given target. Since the synthesizer outputs data every 10 ms, we get 21 acoustic and articulatory vectors per every babbled articulatory target. After every babble the output vectors are appended to L_{acu} and L_{art} correspondingly. When these lists end up with more than T vectors after a babbled utterance, vectors are iteratively merged until only T entries remain, before the next utterance is created.

Standard 12-dimensional MFCC features are extracted from every vocalization. We perform whitening of the features also online, by calculating the mean μ and the standard deviation σ of the MFCC-vectors incrementally. L_{acu} stores the non-whitened MFCC-features, that are whitened using the updated μ and σ, before every distance calculation. We use two ways of selecting babbled targets, random selection and selection based on known targets, described below. Random selection is used before T target candidates are found, and after that, with 50% probability for each selected target.

Random selection: Vowel targets are chosen randomly from a uniform distribution over the allowed ranges of the 9 articulatory parameters. For consonants, for each articulatory parameter it is given a 40% probability that a given parameter takes a value from the edges of its allowed range, otherwise it

Algorithm 1. Pseudocode for the merging algorithm after each vocalization

While the number of entries in L_{acu} and L_{art} is larger than T:

1. Calculate pairwise distances between the entries in the acoustic and articulatory lists and sum them for total distance:
$$D_{i,j} = D_{i,j}^{art} + D_{i,j}^{acu}$$
2. Find the pair with the minimum distance: i_{min} and j_{min}
3. Calculate the new mean for the merged acoustic vector: $\mu_{new}^{acu} = \frac{\mu_{i_{min}}^{acu} \cdot c_{i_{min}} + \mu_{j_{min}}^{acu} \cdot c_{j_{min}}}{c_{i_{min}} + c_{j_{min}}}$
4. Calculate if the merge would increase acoustic diversity. If the acoustic novelty of μ_{new}^{acu} is larger than that of $\mu_{i_{min}}^{acu}$, acoustic diversity increases due to the merge. Only if acoustic diversity increases, update acoustic and articulatory means:
$$\mu_i^{acu} = \mu_{new}^{acu}$$
$$\mu_i^{art} = \frac{\mu_{i_{min}}^{art} \cdot c_{i_{min}} + \mu_{j_{min}}^{art} \cdot c_{j_{min}}}{c_{i_{min}} + c_{j_{min}}}$$
5. Update count:
$$c_i = c_{i_{min}} + c_{j_{min}}$$
6. Delete the entry from the position j_{min} from the acostic and articulatory vector lists L_{acu} and L_{art}

is chosen as for vowels. This is done in order to encourage vocal tract closures. Also, for context dependency of consonants, some articulator parameters are set to not-defined (NaN) values. From one to 8 parameters are set to (NaN) values, leaving at least one articulatory parameter defined to perform the consonant gesture.

Target based selection: The vowel or consonant articulatory target is selected from L_{art}. This is done based on weighted random selection over *acoustic novelty* scores given for each target. The acoustic novelty is calculated after every babble, for every member t of the list L_{acu}, as the median of its Euclidean distance to its 20 closest targets. The novelty measure is used to weight exploration to those regions in the acoustic domain that have little neighboring targets, and that might thus be worth exploring further to increase acoustic diversity. On the articulatory correspondent of the chosen target, uniformly distributed random noise of magnitude [-0.005, 0.005] is added, in order to aid exploration around the given target.

3.1. Merging of target articulations

During the merging phase, consonants and vowels are kept separate, so that vowel targets can merge only with vowel targets, and consonants with consonants. While merging, the targets that are close to each other both in the acoustic and articulatory domain are merged together by averaging. Using both representations in the distance measure encourages finding locally linear regions in the articulatory-acoustic domain, where a small shift in the articulatory target leads to a predictable shift in the acoustic target. Measuring distances in the acoustic domain only would lead to a problem in the merging: due to the many-to-one property of articulation, two very different articulations might lead to the same acoustic outcome, but the acoustic output of the averaged articulatory vectors could be something very different from the original output. The merging is done only if the acoustic diversity around the new target increases. This is done in order to encourage the targets moving to acoustically more diverse regions. Without this rule, the merging is seen to shift the targets towards more neutral, and more easily articulated vocalizations. Pseudocode of the merging procedure is visible in Algorithm 1.

The pairwise distance between the vectors in $L_{acu,}$ $D_{i,j}^{acu}$, is calculated simply as the Euclidean distance between the MFCC vectors. Similarly, for vowel sounds, the articulatory distance $D_{i,j}^{art}$ is calculated as the Euclidean distance over the full parameter vectors. In case of consonants, where only a subset of the articulatory parameters is defined, Euclidean distance is calculated over the common set of defined parameters per vector pair. In the latter case the Euclidean distance is corrected by scaling, so that missing vector elements do not show as relatively smaller distance between articulatory vectors. When articulatory vectors considering consonants are merged, only the common set of defined parameters are kept in the resulting merged vector. We are thus assuming that the similarity of the acoustic vectors was due to the movement of the common articulatory parameters.

4. Experiments

It is not a trivial task to measure how well an articulatory exploration algorithm performs in discovering interesting articulations. Here we propose that a good exploration algorithm discovers a limited set of articulatory targets that capture the diversity of the possible acoustic outputs of the vocal tract model maximally well. Optimally this set of targets should include sounds that resemble all the possible phones in human languages, a task that is very difficult to achieve with physically simplified vocal tract models. This *maximal acoustic space* is difficult to define by itself, since systematically exploring all articulatory combinations of a high-dimensional articulatory space would take millions of productions.

We begin the evaluation of the algorithm with producing 60,000 vocalizations with a random babbling algorithm, that does not perform clustering, target discovery, or novelty-based target selection, but just chooses consonant and vowel targets randomly (equally as in the *random selection* phase of the actual learning algorithm), on every vocalization. MFCC features are extracted from this set and their means and standard deviations are saved in order to perform the zero-mean and unit-variance normalization of the features in the first analysis.

In order to see how the extent of the discovered acoustic space grows when the number of exploratory babbles increases, we run the learning algorithm five times independently, for 10,000 vocalizations each. We take all the babbled audio files until the *N*th babble, extract and normalize the MFCC features from the audio files, reduce their dimensionality to three using principal component analysis, and investigate the volume of the convex hull of the acquired space. The development of this volume is compared to the volume of a random babbling algorithm (implemented equally to the description in the previous paragraph). In Figure 1 it can be seen that the acoustic space created by the learning algorithm stretches out faster than that of the purely random babble. The black cross shows that even after 60,000 random babbles, the volume of the acoustic space remains smaller than that of the learning algorithm. From these runs it appears that the learning algorithm is able to find novel acoustic outcomes faster than random babbling.

Now we want to find out how diversely the acoustic outcomes of our model are distributed in the vocal tract model's acoustic space. Using the final articulatory targets discovered during each learning run, we create five times 20 minutes of random speech (200 babbles of 20 articulatory targets each) by alternating vowel and consonant targets. The same amount of speech is created with random target selection, by choosing an equal number of vowel targets and consonant targets as found in the corresponding learning run. The MFCC-features of all the resulting speech is appended together, in order to have a maximally stretched out acoustic space.

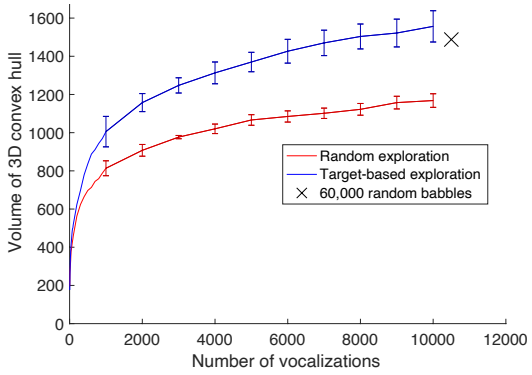

Figure 1. *The volume of the convex hulls of the MFCC-features gathered up to the Nth vocalization. The average and standard deviation of five runs of the learning algorithm (blue), and babbling based on random target selection (red) is shown. The black cross shows the convex hull volume of the initial 60,000 random babbles.*

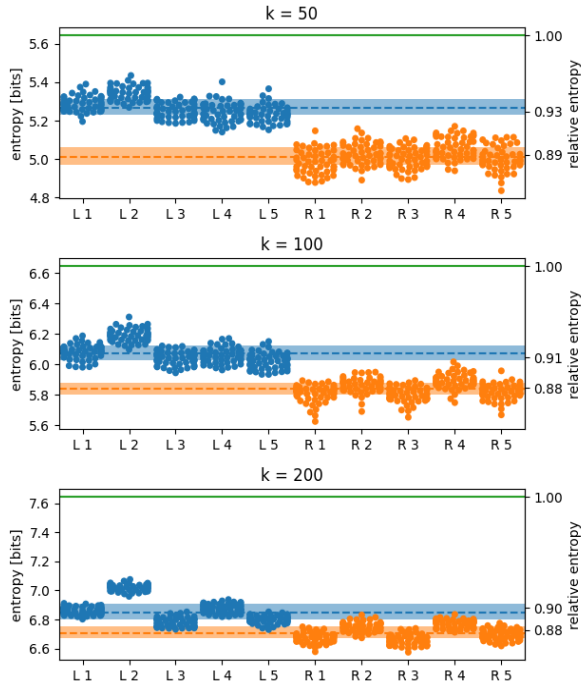

Figure 2. The clustered MFCC-vectors of the 5 runs of random speech based on the learned targets (L, blue) are more diverse than the 5 runs based on randomly selected targets (R, orange). The point swarms illustrate the effect of the variance due to the k-means clustering.The green line at the top indicates the maximum amount of entropy for k clusters (log 2 k bits), and is also represented on the right y-axis, rescaling the entropy relative to this maximal entropy. The median (dashed horizontal line) and first and third quartiles (shaded area) of all learned vs. random runs' data points (blue vs. orange) also show the overall increase in diversity achieved by the learning algorithm.

We combine the 10 independent speech fragments (i.e., the 5 based on learned targets, plus 5 based on random ones) and cluster the extracted 2.5 million MFCC-vectors into *k* clusters using k-means clustering. The cluster centers provide an approximation of the acoustic diversity of the common acoustic space. After the clusters have been created in the shared acoustic space, we classify the MFCC-vectors of the individual runs in the cluster centers. Finally, we calculate the entropy of the distribution of a fragments's MFCC-vectors over the clusters. A high entropy indicates that the output speech is evenly distributed to all cluster centers, and thus evenly occupies the acoustic space that was reached by the babbling runs with high diversity. Low entropy values mean that the output is unevenly distributed, and some parts of the acoustic space have relatively more activity than others.

We run the k-means clustering algorithm 50 times for three different values of *k* (50, 100, and 200), as to estimate and compensate the stochastic effects of the k-means clustering on our measure of diversity. Figure 2 shows the entropy distribution for the 10 independent 20-minute fragments of babbling: the five fragments based on the speech targets learned through our exploration algorithm are more diverse than the five fragments randomly selecting articulatory targets. Figure 3 shows the extent of the vowel spaces of the speech fragments, where formant frequencies are extracted using Praat [28] through the Python library Parselmouth [29]. It can be seen that the learning algorithm finds a wider variety of vowel targets.

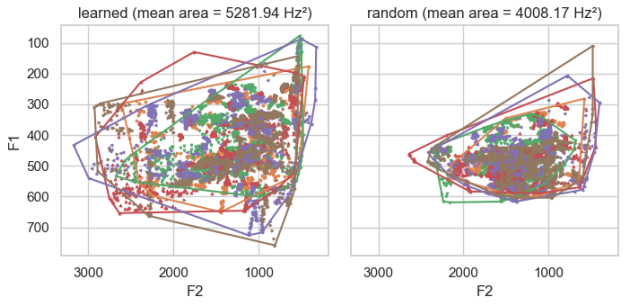

Figure 3. *All F1-F2 values and their convex hulls, extracted from the vowels of the final speech fragments.*

Averaging the entropy-based measure over all 50 clusterings for each independent speech fragment, using the learned targets results in a significantly higher entropy than babbling based on randomly selected articulatory targets (independent samples t-test). This holds for all numbers of clusters *k* we used to evaluate the entropy, 50 ($p < 0.001$), 100 ($p < 0.001$), and 200 ($p = 0.011$).

The entropy-based measure used does not tell us about the actual quality of the synthesized sounds, nor how much they resemble sounds in human languages, but only how diversely they appear in the acoustic domain. Subjectively listening to the final generated speech (available in [30]) reveals that generating speech based on the learned targets produces more closures, plosives and nasal sounds, and that the general variety of speech sounds appears larger. In the future, evaluation of the discovered articulations could include using automatic speech recognition to classify sounds automatically in phonemes found in human languages.

Such automatic vocal exploration, without biasing it with knowledge of human phoneme recognition or pre-defined phonetic gestures or constraints, is of importance for example in research of emergence of sound systems. In such research it is important to let the natural dynamics of the population refine the sound systems to an optimal direction, rather than imposing constraints manually, often based on our knowledge of the already emerged phonetic systems of our own.

5. Conclusions

In this study we have shown that an incremental articulatory exploration algorithm can explore the possible acoustic space efficiently, and discover a number of articulatory targets, whose acoustic outputs are diverse in the acoustic domain. Consonant gestures, affecting only a subset of the total number of articulatory parameters, are superposed to a vowel context showing the coarticulatory effect that is known from human speech, and important for good quality speech synthesis [25]. The clustering of consonant gestures is performed so that the minimum shared set of articulatory parameters are used to produce the desired acoustic outcome. An entropy-based diversity measure, as well as the increase of the volume of the range of created acoustic outcomes show that the algorithm explores the articulatory-acoustic domain faster than an algorithm based on purely random selection of articulatory targets.

6. Acknowledgements

This research was funded by Ulla Tuominen Foundation for author HR. The authors would like to thank prof. Bart de Boer for valuable input to the research.

7. References

[1] P. K. Kuhl and A. N. Meltzoff, "Infant vocalizations in response to speech: Vocal imitation and developmental change," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2425–2438, 1996.

[2] D. K. Oller, *The emergence of the speech capacity*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2000.

[3] A. Smith and H. N. Zelaznik, "Development of functional synergies for speech motor coordination in childhood and adolescence," *Developmental Psychobiology*, vol. *45,* no. 1., pp. 22–33, 2004.

[4] M. H. Goldstein and J. A. Schwade, "Social feedback to infants' babbling facilitates rapid phonological learning," *Psychological Science*, vol 19, no. 5, pp. 515–523, 2008.

[5] J. Gros-Louis, M.J. West, M. H. Goldstein and A. P. King, "Mothers provide differential feedback to infants' prelinguistic sounds," *International Journal of Behavioral Development,* vol. 30, no. 6, pp. 509–516, 2006.

[6] S. Pawlby, "Imitative interaction," *In H.R. Schaffer (Ed.), Studies in mother- infant interaction*, London: Academic Press Inc., 1977.

[7] S. S. Jones, "The development of imitation in infancy," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1528, pp. 2325–2335, 2009.

[8] 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.

[9] F. H. Guenther, "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production," *Psychological Review*, vol. 102, no. 3, pp. 594–621, 1995.

[10] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, and M. Tiede, "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Communication*, vol 89, pp. 103-112, 2017.

[11] H. Rasilo, "Phonemic learning based on articulatory-acoustic speech representations," forthcoming, 2020.

[12] 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.

[13] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, no. 1, pp. 85–117, 2011

[14] 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, 2014.

[15] H. Rasilo and O. Räsänen, "An online model for vowel imitation learning," *Speech Communication*, vol. 86, pp. 1-23, 2017.

[16] B. De Boer, "Self-organization in vowel systems," *Journal of phonetics*, vol. 28, no. 4, pp. 441-465, 2000.

[17] K. N. Stevens, "On the quantal nature of speech," *Journal of phonetics*, vol. 17, no. 1, pp. 3-45, 1989.

[18] 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, pp. 1006, 2014.

[19] 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, 2017.

[20] A. K. Philippsen, R. F. Reinhart and B. Wrede, "Goal babbling of acoustic-articulatory models with adaptive exploration noise," *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2016, pp. 72-78.

[21] J. M. Acevedo-Valle, V. V. Hafner, and C. Angulo, "Social reinforcement in intrinsically motivated sensorimotor exploration for embodied agents with constraint awareness," In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, IEEE, 2017, pp. 255-262.

[22] B. J. 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, no. 1, 2014

[23] A. S. Warlaumont and M. K. Finnegan, "Learning to produce syllabic speech sounds via reward-modulated neural plasticity," *PloS one*, vol. 11, no. 1, 2016.

[24] H. Nam, L. M. Goldstein, S. Giulivi, A. G. Levitt and D. H. Whalen, "Computational simulation of CV combination preferences in babbling," *Journal of phonetics*, vol. 41, no. 2, pp. 63-77, 2013

[25] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, 2013.

[26] H. Rasilo, O. Räsänen and U. K. Laine. "Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion," *Speech Communication*, vol. 55*,* no. 9, pp. 909-931, 2013, https://doi.org/10.1016/j.specom.2013.05.002

[27] T. Flash, N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," The Journal of Neurosciences, vol. 5, pp. 1688–1703, 1985.

[28] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/, 2018

[29] Y. Jadoul, B. Thompson, B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, no. 1-15, 2018. https://doi.org/10.1016/j.wocn.2018.07.001.

[30] H. Rasilo, Y. Jadoul, Audio files of generated speech, https://doi.org/10.6084/m9.figshare.12316928