Goal Babbling of Acoustic-Articulatory Models with Adaptive Exploration Noise

Anja Kristina Philippsen
Cognitive Interaction Technology Center (CITEC), Bielefeld University
aphilipp@techfak.uni-bielefeld.de

René Felix Reinhart
Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University
freinhart@uni-bielefeld.de

Britta Wrede
Applied Informatics Group, Bielefeld University
bwrede@techfak.uni-bielefeld.de

***Abstract*—We use goal babbling to bootstrap a parametric model of speech production for a complex 3D vocal tract model. The system learns to control the articulators for producing five different vowel sounds. Ambient speech influences learning on two levels: it organizes the learning process because it is used to generate a space of goals in which exploration takes place. A distribution learned from ambient speech provides the system with targets during exploration.**

Previous research with this vocal tract model showed that visual information have to be included for acquiring the vowel [u] via reward-based optimization. We model the learning process instead with goal-directed exploration where all targets are learned in parallel. As some vowels require more exploratory noise in the articulators than others, we propose a mechanism to adapt the noise amplitude depending on the system's competence in different regions of the goal space. We demonstrate that this self-aware learning leads to more stable results. The implemented system succeeds in acquiring vocalization skills for rounded as well as unrounded vowels using only a single modality.

I. Introduction

Learning how to speak can be interpreted as a motor coordination problem: Infants explore the capabilities of their vocal tract in order to discover articulatory trajectories that produce the desired speech sounds. Although it is still largely a mystery how infants achieve this, there is consensus that babbling plays a crucial role in early speech learning [1], [2]: By producing speech sounds and observing the outcomes, infants gradually learn to coordinate their articulators.

To model this development computationally, a system can be equipped with a vocal tract model. Executing this *forward model* the system can generate speech signals from articulatory configurations in a similar way as infants use their articulators. Acquiring articulatory control can then be defined as learning an *inverse model* to estimate from the acoustics which articulatory configuration is required to reproduce this sound.

Findings in developmental psychology suggest that infants explore the space of possible motor configurations not randomly, but with targets in mind [3], [4], [5]. Accordingly, many developmental models of speech acquisition implement vocal learning as an imitative process [6], [7], [8]. Applying active motor exploration, these studies acquire the articulatory configurations to successfully imitate a set of speech sounds in a babbling phase. But speech sounds are learned sequentially. Due to redundancies in the motor system (a number of articulatory trajectories might result in the same speech sound), this approach bears the risk that no inverse model can be trained from the collected acoustic-articulatory pairs [9]. The Elija model [10] removes redundancies after the babbling phase by consolidating the learned motor patterns based on the acoustic consequences. [6] and [7] connect articulation and acoustics via a map of acquired speech sounds.

Goal babbling is an exploration mechanism first introduced for kinematic motor control learning and resolves such redundancies by learning to achieve several targets in parallel and directly bootstrapping the inverse model during exploration [11], [12]. Goal babbling achieves high efficiency by organizing exploration in the so called goal space, the space of (here: acoustic) outcomes.

Studies by Moulin-Frier et al. [13], [14], [15], [16] have applied the idea of goal babbling to the speech domain. By using formant frequencies as goals, they could demonstrate the emergence of articulated speech sounds [16] and the bootstrapping of vowel sounds [13]. In a recent work [17], Liu and Xu used goal babbling to control F0 for Chinese language. A limitation in these works is the low-dimensional acoustic representation that is required to make goal babbling efficient. Actually, speech is a very high-dimensional signal, as it exposes high variability in the spectral as well as in the temporal domain.

In [18], we presented an approach to overcome this limitation by generating a goal space from high-dimensional acoustic features via dimension reduction. This method follows the idea that infants' learning is influenced by the ambient language which they perceive from their environment [19], [5]. Here, we extend this model and use it to learn articulatory control for imitating five vowel sounds with the 3D articulatory speech synthesizer VocalTractLab (VTL) [20]. In VTL, the vocal tract shape is determined by 20 articulatory (and additional glottis) parameters. It is physiologically more natural and produces better intelligible sound than the Diva or the Praat articulatory synthesizer that most speech acquisition models use [6], [10], [16], [17], [18], [21]. It causes also

a higher redundancy in the motor system, and therefore learning to control its articulators can be assumed to be closer to the problem infants face during their development.

Although it has been used for neurocomputational modeling of speech acquisition ([22], [7]), goal babbling has not yet been applied to this vocal tract model. Recently, Murakami et al. [8] applied reinforcement learning to learn the vowels [a], [i] and [u] with VTL. They observed that learning the rounded vowel [u] is difficult due to a plateau in the reward function, but including visual cues (as children might perceive from a caregiver) makes learning possible. However, also blind children learn to speak, so visual information is a support, but not crucial to the learning progress itself. Sighted children might have advantages in early consonant articulation [23], but Pérez-Pereira and Conti-Ramsden conclude in a recent research overview that "the development of the sounds of speech in blind children do not appear to differ greatly in terms of the pattern nor the rate of development found in sighted children, with the possible exception of early production of sounds that have clear visual articulation" (p. 71) [24].

We demonstrate in this study that visual features are not required for learning to produce rounded vowels, but rather, a sufficient amount of noise in the motor system helps to overcome nonlinearities in the learning process. Our system bootstraps an inverse model for producing five vowel sounds by applying skill babbling [25], a recently introduced goal babbling variant that makes exploration more efficient. After evaluating the system's performance with different levels of exploratory noise in the space of articulator configurations, we show that a competence-based decrease of motor noise causes a transition from global to more focused local exploration and leads to more stable learning results.

II. A model for learning articulatory control

Learning in our model is organized in two major stages: a *training phase* in which a low-dimensional representation of ambient speech sounds is derived, and a *babbling phase* in which the system explores how to acquire articulatory control to produce the speech sounds present in the ambient speech.

Fig. 1 depicts how the forward and inverse model are defined between the acoustic and articulatory speech representations. The forward model realizes the mapping from articulation to acoustics. The inverse model maps acoustic targets to articulatory configurations. The high-dimensional acoustic space P and articulatory space Q are not connected directly, but linked via a *goal space* X which is generated during the training phase (see Sec. II-B) as a low-dimensional representation of the acoustic space. As a result of the training phase, the forward mapping $\boldsymbol{x} = f(\boldsymbol{q})$ from an articulatory configuration $\boldsymbol{q}$ to a position in goal space $\boldsymbol{x}$ synthesizes the speech sound for the given articulatory configuration, computes its acoustic features and maps this representation into the goal space.

In the following babbling phase (see Sec. II-C) the system iteratively tries to reach targets within the goal space and retrains the inverse model according to this new experience. This bootstraps the inverse model $\boldsymbol{q} = g(\boldsymbol{x}, \theta)$ by adjusting the parameters θ such that all feasible positions of the goal space $\boldsymbol{x}$ can be achieved by a corresponding articulatory configuration $\boldsymbol{q}$.

Sec. II-A describes how the ambient speech sounds were generated, Sec. II-B and Sec. II-C explain the training and babbling phase, respectively. While ambient speech sounds were generated only once, the training and babbling phase are repeated in each experimental run. Sec. II-D and Sec. II-E explain algorithmic details required to adapt the original goal babbling to the problem of speech acquisition.

A. Generation of Ambient Speech

We generated a set of ambient speech sounds using the articulatory synthesizer VTL[1] itself to ensure that the system is able to achieve the targets derived from ambient speech. The correspondence problem that occurs if ambient speech from a different source is used (e.g. a human tutor), is not tackled in this study.

We generated 100 acoustic variations for each of the 6 vowel sounds (the sound produced by the "neutral" vocal tract shape "schwa" (denoted as [@]) and five vowels [a], [e], [i], [o] and [u]) by adding Gaussian distributed noise σ to the defined vocal tract shapes of the VTL default speaker (JD2) after normalizing the ranges for each articulator to $[-1, 1]$. Due to the different sensitivity of the articulators we set $\sigma = 0.1$ for hyoid (HX, HY), jaw (JX, JA) and lip (LP, LD) parameters and $\sigma = 0.01$ for the more sensitive velum (VO), tongue tip (TTX, TTY), tongue center (TCX, TCY), body (TBX, TBY) and tongue side elevation (TS1-TS4) parameters (for details about the parameters see [20]). Other vocal tract and the glottis parameters do not change for the given vowel set, so they were fixed to default values.

B. Training Phase: Goal Space Generation from Ambient Speech

While a low-dimensional representation of these ambient speech sounds could be achieved by extracting low-dimensional features such as formant frequencies, these features limit the amount of speech sounds that the system's perception can possibly discriminate. By learning an embedding from a set of speech sounds representing the ambient language, we profit from the possibly high discriminability and still create a low-dimensional space that can be used for goal-directed exploration. Additionally, this approach models that the system's perception is biased by ambient speech [18].

We extracted the first three formant frequencies via Praat (averaged over the course of the utterance) and 13 Mel-frequency cepstral coefficients (MFCCs) (including log energy) and the MFCC derivatives from the middle of the speech signal. The resulting 42-dimensional acoustic feature vectors were normalized dimension-wise, such that formant frequencies and MFCCs have equal influence, and then projected to a 2-dimensional space in two steps. First, it is projected to 10 dimensions via Principal Component Analysis

[1] VocalTractLab 2.1 Linux API (released in September 2014), see: http://vocaltractlab.de/index.php?page=vocaltractlab-download

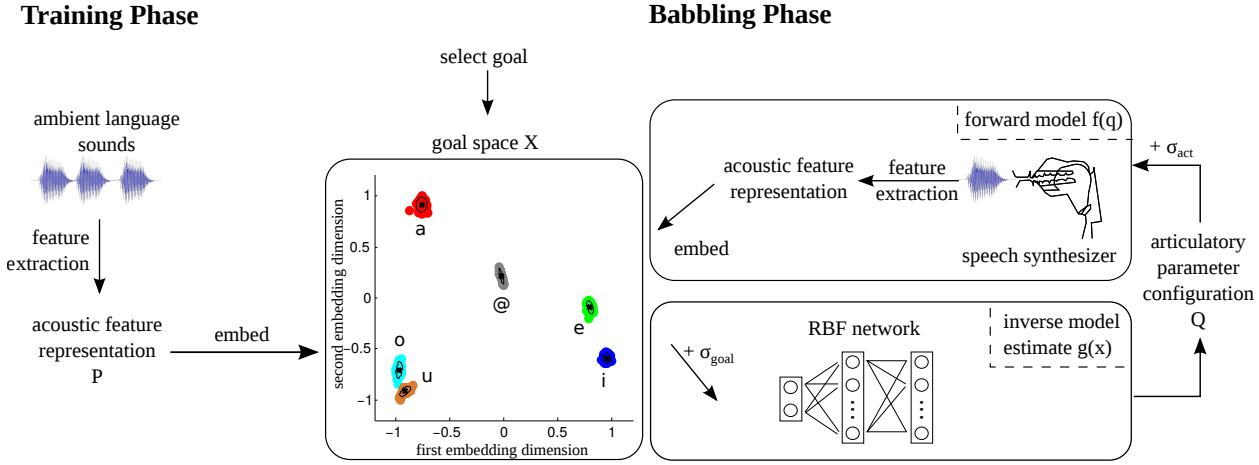

Fig. 1. Training phase: the goal space is generated from ambient speech. Babbling phase: the inverse model $g(\boldsymbol{x})$ is trained to estimate an articulatory configuration $\boldsymbol{q}$ for a desired target $\boldsymbol{x}^*$ in goal space such that the forward model $f(\boldsymbol{q})$ embeds the produced acoustics close to $\boldsymbol{x}^*$.

(PCA). Then, to achieve good separation between different sound classes, Linear Discriminant Analysis (LDA) is applied on the 10-dimensional features, taking into account the vowel class information. We motivate the usage of a supervised embedding technique by the remarkable sensitivity of young infants to sound contrasts [26]. An example of the resulting goal space is displayed in Fig. 1.

Using these mappings, the system can now project any acoustic signal to the two-dimensional goal space. In other words, the system is tuned to the ambient speech and perceives acoustics in terms of the variability present in the ambient speech. Targets for exploration in the babbling phase are drawn from the distribution of ambient speech in the goal space as described in Sec. II-D.

C. Babbling Phase: Bootstrapping Speech Sounds

Goal babbling is a method to bootstrap an inverse model for a motor coordination task. The idea is to organize the exploration in the goal space, i.e. in the space of outcomes. In each goal babbling iteration, there is an *exploration step* where the system tries to reach a new target within this goal space and an *adaptation step* where the recent experience is integrated into the inverse model.

The original goal babbling algorithm [11] (which we used in [18]) draws targets along linear paths in goal space. Redundancies are resolved during learning by using a weighted regression scheme. However, in the case that for each step a full motion has to be executed, this exploration method is relatively inefficient. Reinhart, therefore, recently proposed skill babbling [25] which combines goal-directed exploration with episodic learning in mini-batches, such that the inverse model can be updated more efficiently in each episode. Skill babbling proceeds similar to goal babbling ([11], [27], [28]) by iteratively exploring and adapting the inverse model, but instead of trying to reach a single target, it explores a batch of targets in each iteration.

Skill babbling also integrates an explicit and continuously updated model of the workspace in the learning process. This workspace model represents the part of the goal space where the system can already achieve targets. The workspace model tracks the learner's progress and could e.g. guide the exploration by selecting new targets.

In [18] we demonstrated that original goal babbling is applicable for learning articulatory control with the Diva synthesizer, but VTL is much more high-dimensional and causes more redundancy. Besides, the generation of sounds takes approx. 5 to 6 times longer than with the Diva synthesizer. For these reasons we use skill babbling in this paper which we describe in the following:

The inverse model is initialized with a default action $\boldsymbol{q}_0$ and the corresponding goal space position $\boldsymbol{x}_0 = f(\boldsymbol{q}_0)$ which correspond to the neutral sound [@]. Now, in each iteration t the system explores a batch of K new targets (exploration step) and adapts the inverse model accordingly (adaptation step).

In the **exploration step**, at first, a target seed $\boldsymbol{x}^{seed}$ is drawn from the target distribution (see Sec. II-D). Noise in the goal space ($\sigma_{goal} = 0.05$) is added to this target seed to generate a batch of K slightly varied targets that the learner should explore in this iteration:

$$\boldsymbol{x}_k^* = \mathcal{N}(\boldsymbol{x}^{seed}, \sigma_{goal}), \quad k = 1 \dots K \tag{1}$$

Then the loop depicted on the right side of Fig. 1 is executed: The inverse model tries to reach the targets $\boldsymbol{x}_k^*$ by estimating the corresponding actions (Eq. 2), adding exploratory noise σ_{act} in action space (Eq. 3) and producing and observing the outcomes via execution of the forward model (Eq. 4):

$$\hat{\boldsymbol{q}}_k = g(\boldsymbol{x}_k^*), \quad \forall k \tag{2}$$

$$\boldsymbol{q}_k = \mathcal{N}(\hat{\boldsymbol{q}}_k, \sigma_{act}), \quad \forall k \tag{3}$$

$$\boldsymbol{x}_k = f(\boldsymbol{q}_k), \quad \forall k \tag{4}$$

In the following **adaptation step**, weights w_k are computed (see Sec. II-E) for the action/outcome-pairs $(\boldsymbol{q}_k, \boldsymbol{x}_k)$.

These are used to update the parameters θ of the inverse model $g(\boldsymbol{x}, \theta)$. For learning the inverse model, we use incremental weighted regression as described in [27] and [25], but with local radial basis functions (basis function radius 0.15) instead of linear functions. Basis function centers are added in goal space according to vector quantization. When updating the inverse model in time step t with a new training sample k, the weighted square error $w_{t,k} \cdot \|\boldsymbol{q}_{t,k} - g(\boldsymbol{x}_{t,k}, \theta)\|^2$ is minimized by gradient descent (learning rate 0.9).

Additionally, a model of the explored workspace can be trained by clustering achieved targets in the goal space with prototypical hyper-spheres (see [29], [25]). We integrate this in Sec. IV to adapt σ_{act} during exploration.

Fig. 2 summarizes the overall babbling phase.

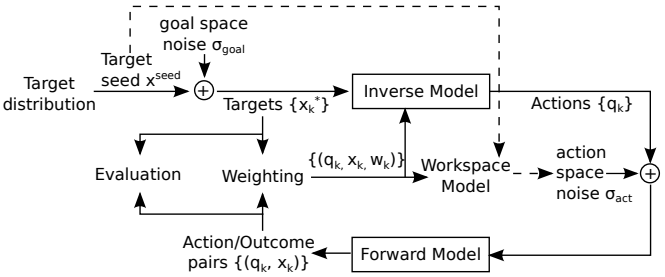

Fig. 2. Skill babbling algorithm for articulatory control as described in Sec. II with adaptive noise control from Sec. IV (dashed lines).

D. Selecting Targets from a Target Distribution

In original skill babbling, the system should discover the complete space of achievable goals. Targets are, therefore, provided by the workspace model that tracks the learner's progress in different regions of the goal space.

When babbling articulations, it is not necessary to reach every position in goal space, in fact the goal space is unlikely to be continuous and there might be regions between the target clusters that are undefined, i.e. not reachable. To represent the part of the goal space that is interesting for the learner, a Gaussian Mixture Model (GMM) [30] of the distribution of targets in ambient speech is trained as

$$P(\boldsymbol{x}) = \sum_{n=1}^{N} \pi_n \mathcal{N}(\boldsymbol{x}|\mu_n, \Sigma_n), \tag{5}$$

with prior probabilities π_n, the number of target clusters $N = 6$ and μ_n and Σ_n being the mean and covariance parameters of the Gaussian mixture components obtained from GMM training (displayed as ellipses in Fig. 1). This target distribution serves as a target selection mechanism.

In this paper we stick to fixed prior probabilities $\pi_n = 1/N$. A competence-based active selection of targets (as we demonstrated in [18]) can increase learning speed, but does not affect whether learning succeeds in general.

E. Weighting Scheme

For adapting the inverse model, the generated training examples from one batch are weighted according to three weighting schemes: $w_k = w_k^{tar} * w_k^{amb} * w_k^{sal}$, $w_k \in [0, 1]$.

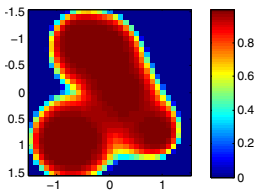

Fig. 3. Weight w^{amb} for a raster of points in the goal space with $M = 4$.

The weight w^{tar} measures how close the desired targets $\boldsymbol{x}_k^*$ and the actual outcomes $\boldsymbol{x}_k$ are in goal space.

$$w_k^{tar} = 1 - \frac{d_k}{max_{l=1\ldots L} d_l} \tag{6}$$

$$d_k = \begin{cases} \|\boldsymbol{x}_k^* - \boldsymbol{x}_k\| & \text{if } \|\boldsymbol{x}_k^* - \boldsymbol{x}_k\| \leqslant 1.5 \\ 1.5 & \text{otherwise} \end{cases} \tag{7}$$

w^{tar} is normalized across the K samples produced in one batch. A weight of 0 is always assigned to the worst example, better approximations receive higher weights. This emphasizes "relatively good" examples and leads to more efficient training especially in the beginning of the learning process. To account for cases where several productions are unsuccessful (e.g. if no phonation occurs), a threshold of 1.5 is defined in (7).

To make sure that learning concentrates on regions of interest (i.e. regions where ambient speech resides), we use a second weighting scheme w^{amb}, that gives a higher weight to regions that are close to the ambient speech.

For vowels, some degree of linearity in the goal space can be assumed, thus, we do not use the distance alone, but calculate a higher-order distance D from an ordered list to the M closest GMM target clusters $[d_1, \ldots, d_M]$ (from close to distant). M can be any value between 1 (taking only the closest cluster into account) and N (consider the distances to all target clusters).

$$D = \prod_{m=1\ldots M} d_m^{M-m+1} \tag{8}$$

A smaller D is better, as it means that the goal space position has a lower distance to the close targets. We calculate the weight accordingly as:

$$w^{amb} = max(0, 1 - D) \tag{9}$$

We use $M = 4$ here. As this measure does not depend on the currently explored target, Fig. 3 shows the values of w^{amb} for a raster of points in the goal space from Fig. 1.

Finally, w^{sal} measures the salience of the speech sound. Louder signals are preferred over sounds that are low in amplitude to avoid that unarticulated sounds are integrated into learning. We use a simple implementation that captures the signal's loudness:

$$w_k^{sal} = \begin{cases} min(1, \frac{s_k}{1.5 s_{ref}}) & \text{if } s_k \geqslant 0.5 s_{ref} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

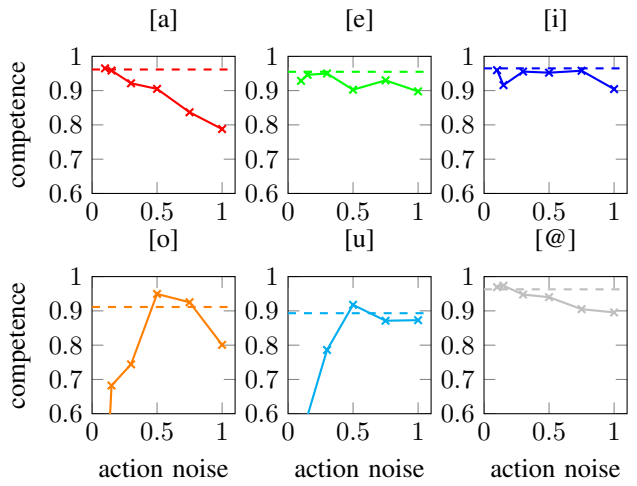

Fig. 4. Competences after training for different levels of articulatory noise. Dashed lines mark the competences achieved with adaptive noise (see Sec. IV).

Tab. I
AVERAGE OF PERFORMED ITERATIONS AND % OF UNARTICULATED PRODUCTIONS WITH FIXED EXPLORATORY MOTOR NOISE

noise level	0.1	0.15	0.3	0.5	0.75	1
∅ iterations	499	478	411	329	486	499
unarticulated	16.7%	8.3%	2.5%	0%	0.8%	0.8%

In this formula, s_k is the median of the absolute normalized amplitude of a speech sound corresponding to the observation x_k in goal space, and s_{ref} is a reference value for an articulated sound, calculated beforehand by averaging across the sounds from the ambient speech set.

III. BABBLING A SET OF SPEECH SOUNDS

We executed the training and babbling phases 20 times for different amplitudes of exploratory motor noise. In each trial the system learns for a maximum of 500 iterations. If the error falls below 0.1 for all targets for 5 subsequent iterations[2], we stop earlier to save computation time. In each iteration the system tries to achieve $K = 10$ targets, Gaussian distributed with $\sigma_{goal} = 0.05$ around the target seed (Eq. 1). Parameter values were set to the values specified in the previous section.

To evaluate the learning progress, we compute the learner's competence by letting it imitate the cluster centers μ_n of the target distribution $P(x)$ and points in the vicinity to evaluate how stable the learned mapping is:

$$\mu_n + \begin{pmatrix} 0.05 \\ 0 \end{pmatrix}, \mu_n + \begin{pmatrix} 0 \\ 0.05 \end{pmatrix}, \mu_n - \begin{pmatrix} 0.05 \\ 0 \end{pmatrix}, \mu_n - \begin{pmatrix} 0 \\ 0.05 \end{pmatrix}.$$

The competence for a reproduction $\boldsymbol{x}$ of a target $\boldsymbol{x}^*$ is computed in accordance with e.g. [14] as:

$$comp(\boldsymbol{x}, \boldsymbol{x}^*) = exp(-\|\boldsymbol{x} - \boldsymbol{x}^*\|) \quad (11)$$

[2]Due to instabilities there might be outliers where the error drops only for one iteration. To prevent the algorithm from stopping in these cases, a number of subsequent iterations are evaluated.

The amplitude of articulatory exploration noise σ_{act} was set to $0.1, 0.15, 0.3, 0.5, 0.75$ and 1. Tab. I lists the average number of performed iterations and the percentage of unarticulated productions that occurred after training. Unarticulated productions are unsuccessful productions where incorrect vocal tract configurations cause click or hissing noise or no phonation occurred due to a vocal tract closure. Fig. 4 displays which competences were achieved with these noise levels for the six vowel sounds. Unarticulated productions (mostly [o] or [u]) were left out in these statistics, because its acoustic features cannot be computed reliably. Apparently, with a lower noise level the learner well achieves [a], [e] and [i], but cannot learn [o] and [u]. Increasing the noise level to up to 0.5 improves the system's competences for these more difficult vowels. On the contrary, a decrease can be observed for the competences of [a], [e] and [@]. The best results were obtained with a noise level of 0.5, where the algorithm also converges the fastest and all produced vowels were articulated.

To see why a higher noise level is required for learning [o] and [u], we analyzed the learning problem by determining the degree of nonlinearity in the goal space. More precisely, we measured how the acoustic representation changes when interpolating linearly between the articulatory configurations. We defined 20 equidistant linearly interpolated vocal tract shapes between [@] and each vowel and calculated the according goal space positions by executing the forward model. Fig. 5 depicts these goal space positions (top left: full goal space, top right: zoom on the region around the neutral shape). Apparently, when trying to achieve [o] and [u] the learner cannot take the direct way to the goal, but has to make a "detour". The bottom plot shows how the distance decreases between the goal space position of the interpolated sound and the goal space position of the target vowel shape. This distance decreases monotonically for [a], [e] and [i], but for [o] and [u] the error increases slightly before decreasing. "Jumping" over this hill requires a sufficient amount of motor noise. This explains why a higher noise amplitude leads to better learning of the rounded vowels.

However, with a higher noise level, the system babbles and constantly tries out new ways to achieve the targets; it does not fine-tune the results. This negatively affects the learning of easy vowels like [a] and also the imitation of the neutral sound [@].

IV. ADAPTIVE NOISE CONTROL

The results from Sec. III suggest that different targets require different levels of motor noise. But how to determine an appropriate noise level?

Our system organizes learning in a goal space, so it can monitor its own progress to a certain degree: it can observe whether it succeeds in reaching goals in different regions of the goal space [3]. The amplitude of motor noise can be adapted with this information. A reasonable schema would be to be more "adventurous" (i.e. explore within a wider range) if

[3]Note that the system's perception is not perfect, but limited by the accuracy of the learned forward model.

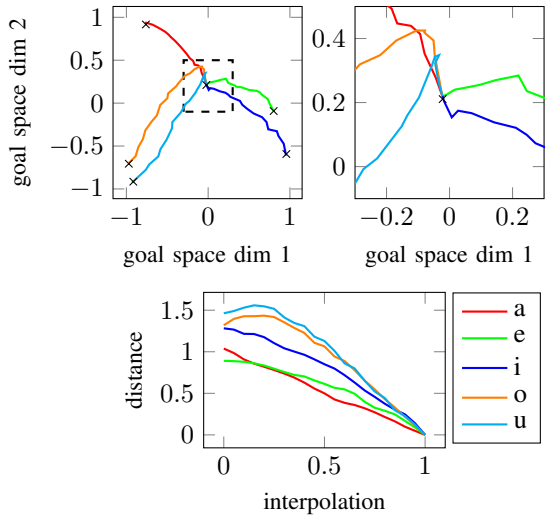

Fig. 5. Positions in goal space (top) and distance to target (bottom) for linearly interpolated articulatory configurations between the neutral (0) and the five vowel shapes (1).

new targets should be achieved. For reaching targets that can already be mastered to some extent, it is better to stay closer to the estimated motor command and focus on a more fine-grained, local exploration. Infants might use such information during learning as well, at least there is evidence that they are already aware of targets, as they perform goal-directed movements very early in their development [3], [4], [5].

Tracking the competence progress in different regions of the goal space has been implemented by Oudeyer et al. to model intrinsic motivation (e.g. [12], [16]): the system selects a target where it expects the highest competence progress. Instead of influencing the target selection mechanism, we here adapt the amplitude of exploratory motor noise.

The workspace model (introduced in Sec. II-C) represents our system's internal model of the already achieved part of the goal space. It can be used to measure how confident the system is about successfully reproducing a sound [25]. Using this information, the amplitude of motor noise can be adapted depending on the desired target.

Prototypical hyper-spheres with centers $\boldsymbol{c}_i$ and identical radii ($\gamma = 0.1$) cluster the goal space. The distance of a target $\boldsymbol{x}$ to the workspace model is measured as the distance to the closest prototype:

$$d_{WS}(\boldsymbol{x}) = min_i(\|\boldsymbol{c}_i - \boldsymbol{x}\|) \tag{12}$$

We update the workspace model during the adaptation step of goal babbling similar to [25]: If a target $\boldsymbol{x}_k^*$ is achieved, a new prototype with $\boldsymbol{c}_i = \boldsymbol{x}_k$ is added if $d_{WS}(\boldsymbol{x}_k) > \gamma$ and the reproduction was successful, i.e. the weight w_k is above a threshold (0.5). In each iteration, the level of articulatory noise is now determined depending on the task seed $\boldsymbol{x}^*$:

$$\sigma_{act} = \alpha \cdot (1 - exp(-4 \cdot d_{WS}(\boldsymbol{x}^*))) \tag{13}$$

In this way, the articulatory noise amplitude increases for targets far away from the discovered part of the goal space and decreased for targets that are close to already achieved goal space positions. The factor α controls the maximum amount of noise.

Varying α in steps of 0.05 between 0.45 and 0.7, we repeated the experiments from Sec. III with adaptive articulatory noise. The best results could be achieved with $\alpha = 0.55$ and are displayed with dashed lines in Fig. 4. With this configuration, a high competence is achieved for all vowels. No unarticulated productions occurred and the average number of performed iterations was with 262 much lower than in the previous experiments (cf. Tab. I).

Fig. 6 compares the mean competence increase in the adaptive model (bottom) with the model with a fixed noise amplitude of 0.5 (top). How the articulatory noise is adapted during the course of learning is displayed in Fig. 7. The adaptive mechanism achieves results similar to a fixed noise amplitude of 0.5, but maintains a higher competence level for easy vowels like [a] and [@]. With adaptive noise control there is also less variation between subsequent iterations, so the stop criterion (goal space distance < 0.1 for 5 subsequent iterations) is fulfilled earlier. This might be the reason why in average 20% less iterations were performed.

We confirmed these observations in a perceptive evaluation by listening to the vowel sounds that the system can produce after learning[4]. With adaptive noise of $\alpha = 0.55$ and with fixed noise of $\sigma_{act} = 0.5$, all vowels were intelligible except for some cases where [o] and [u] were confused, likely due to the proximity of their target clusters in goal space. This confusion is more frequent in the case of adaptive noise, probably because fewer iterations were conducted.

V. Conclusion & Outlook

We presented an approach to acquire vocalization skills with goal-directed exploration. By adapting the amplitude of exploratory motor noise, all vowels can be well learned: high noise in the beginning promotes the learning of [o] and [u], a competence-based noise reduction ensures that high competence is maintained for already acquired sounds. The level of motor noise is determined in dependency of the already achieved goal space and the desired target.

In future studies we examine whether extending the goal space to a 3D representation makes learning of [o] and [u] more accurate. We also plan to extend the system to the learning of syllables by learning trajectories instead of articulatory configurations.

Acknowledgement

This research has been supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG), and by the European Project CODEFROR (FP7-PIRSES-2013-612555).

[4]All learned vowel sounds are available at: https://techfak.uni-bielefeld.de/%7Eaphilipp/icdl16-results/

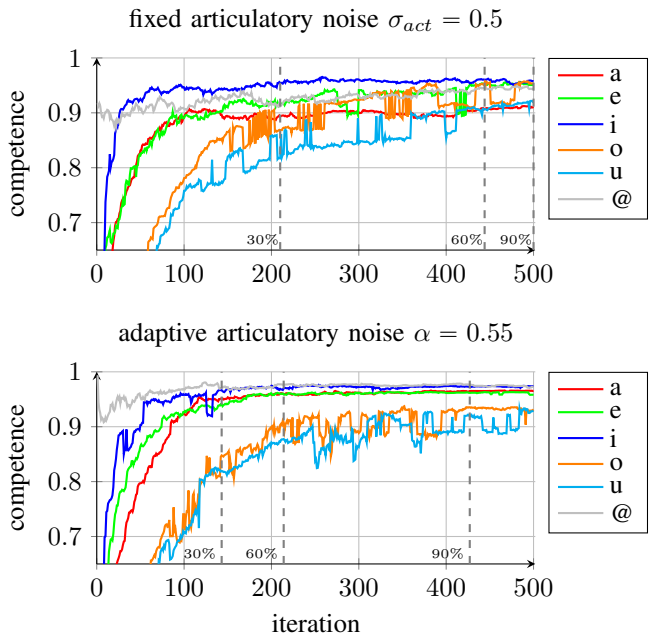

Fig. 6. Competence per vowel class during exploration (averaged over 20 trials) for learning with fixed (top) and adaptive (bottom) noise, measured by Eq. 11 after each iteration. The vertical lines indicate when 30%, 60% and 90% of the learning trials converged.

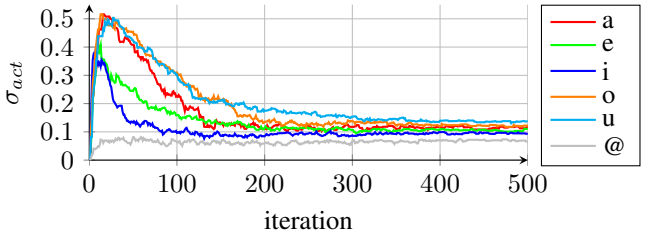

Fig. 7. Adaptive noise allotted to targets from the six target distribution clusters in each time step, averaged over 20 trials.

References

[1] D. K. Oller and R. E. Eilers, "The role of audition in infant babbling," *Child development*, pp. 441–449, 1988.

[2] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831–843, 2004.

[3] A. N. Meltzoff, M. K. Moore *et al.*, "Imitation of facial and manual gestures by human neonates," *Science*, vol. 198, no. 4312, pp. 75–78, 1977.

[4] A. Van der Meer, F. Van der Weel, D. N. Lee *et al.*, "The functional significance of arm movements in neonates," *SCIENCE-NEW YORK THEN WASHINGTON-*, pp. 693–693, 1995.

[5] B. Mampe, A. D. Friederici, A. Christophe, and K. Wermke, "Newborns' cry melody is shaped by their native language," *Current biology*, vol. 19, no. 23, pp. 1994–1997, 2009.

[6] J. A. Tourville and F. H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and Cognitive Processes*, vol. 26, no. 7, pp. 952–981, 2011.

[7] 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, pp. 1–28, 2014.

[8] M. Murakami, B. Kröger, 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 *IEEE Intern. Conf. on Development and Learning*, 2015.

[9] M. Jordan and D. Rumelhart, "Forward models: Supervised learning with a distal teacher," *Cognitive Science*, vol. 16-3, pp. 307–354, 1992.

[10] 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.

[11] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 216–229, 2010.

[12] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, vol. 61, no. 1, pp. 49–73, 2013.

[13] C. Moulin-Frier and P.-Y. Oudeyer, "Curiosity-driven phonetic learning," in *IEEE Intern. Conf. on Development and Learning*. IEEE, 2012.

[14] ——, "Exploration strategies in developmental robotics: a unified probabilistic framework," in *IEEE Intern. Conf. on Development and Learning*, 2013.

[15] S. M. Nguyen and P.-Y. Oudeyer, "Socially guided intrinsic motivation for robot learning of motor skills," *Autonomous Robots*, vol. 36, no. 3, pp. 273–294, 2014.

[16] 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.

[17] H. Liu and Y. Xu, "Learning model-based F0 production through goal-directed babbling," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2014, pp. 284–288.

[18] A. Philippsen, F. Reinhart, and B. Wrede, "Efficient bootstrapping of vocalization skills using active goal babbling," *International Workshop on Speech Robotics at Interspeech 2015, Dresden, Germany*, 2015.

[19] B. de Boysson-Bardies, L. Sagart, and C. Durand, "Discernible differences in the babbling of infants according to target language," *Journal of child language*, vol. 11, no. 01, pp. 1–15, 1984.

[20] P. Birkholz, "VocalTractLab – Towards high-quality articulatory speech synthesis," http://www.vocaltractlab.de/.

[21] 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.

[22] 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.

[23] A. Mills, *Language development in exceptional circumstances, Chapter 9: Visual handicap*. Psychology Press, 1993.

[24] M. Pérez-Pereira and G. Conti-Ramsden, *Language development and social interaction in blind children*. Psychology Press, 2013.

[25] R. Reinhart, "Autonomous exploration of motor skills by skill babbling," *Autonomous Robots*, 2015, under review.

[26] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant behavior and development*, vol. 7, no. 1, pp. 49–63, 1984.

[27] M. Rolf, J. J. Steil, and M. Gienger, "Online goal babbling for rapid bootstrapping of inverse models in high dimensions," in *IEEE International Conference on Development and Learning (ICDL)*, 2011.

[28] M. Rolf, "Goal babbling with unknown ranges: A direction-sampling approach," in *IEEE Intern. Conf. on Development and Learning*. IEEE, 2013.

[29] R. F. Reinhart and J. J. Steil, "Efficient policy search in low-dimensional embedding spaces by generalizing motion primitives with a parameterized skill memory," *Autonomous Robots*, vol. 38, no. 4, pp. 331–348, 2015.

[30] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 2, pp. 286–298, 2007, source code available at: http://www.calinon.ch/sourcecodes.php (GMM-GMR).