Hyperarticulation Aids Learning of New Vowels in a Developmental Speech Acquisition Model

Anja Kristina Philippsen*, René Felix Reinhart†, Britta Wrede*, and Petra Wagner§
*Cognitive Interaction Technology Center, Bielefeld University, Germany
Email: {aphilipp, bwrede}@techfak.uni-bielefeld.de
†Fraunhofer Research Institute for Mechatronic Systems Design IEM, Paderborn, Germany
Email: felix.reinhart@iem.fraunhofer.de
§Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
Email: petra.wagner@uni-bielefeld.de

***Abstract*—Many studies emphasize the importance of infant-directed speech: stronger articulated, higher-quality speech helps infants to better distinguish different speech sounds. This effect has been widely investigated in terms of the infant's perceptual capabilities, but few studies examined whether infant-directed speech has an effect on articulatory learning.**

In earlier studies, we developed a model that learns articulatory control for a 3D vocal tract model via goal babbling. Exploration is organized in the space of outcomes. This so called goal space is generated from a set of ambient speech sounds. Similarly to how speech from the environment shapes infant's speech perception, the data from which the goal space is learned shapes the later learning process: it determines which sounds the model is able to discriminate, and thus, which sounds it can eventually learn to produce.

We investigate how speech sound quality in early learning affects the model's capability to learn new vowel sounds. The model is trained either on hyperarticulated (tense) or on hypoarticulated (lax) vowels. Then we retrain the model with vowels from the other set.

Results show that new vowels can be acquired although they were not included in early learning. There is, however, an effect of learning order, showing that models first trained on the stronger articulated tense vowels easier accommodate to new vowel sounds later on.

I. Introduction

Babbling is a crucial phase in infant's speech development, in which infants explore the possibilities of their vocal tract and learn from the articulatory-acoustic correspondences they experience.

Almost all computational models of speech acquisition incorporate a babbling phase where motor configurations are explored randomly [1], [2], [3] or actively [4], [5], [6]. Instead of exploring in the space of motor parameters, babbling can also be organized in the space of outcomes [7], [8]. In [9] it was shown that such goal babbling can bootstrap vowel sounds quicker than motor babbling. Goal babbling has subsequently been successfully applied for learning F0 contours [10] or for modeling the emergence of speech-like sounds in general [11], [12].

In [13], we presented a model for learning articulatory control for a 3D vocal tract model by goal babbling. Our model bootstraps a parametric model of speech production for perceiving and producing a number of different speech sounds: the five vowels sounds [a], [e], [i], [o] and [u] in [13]. In contrast to previous studies, speech percepts in our model are not represented with simple features such as formants and intensity, but with high-dimensional features as they are also used for speech recognition purposes. This opens up the possibility to distinguish between a large variety of speech sounds.

For efficient goal-directed exploration, however, we need a low-dimensional representation of speech sounds. Young infants might face a similar problem, as they are able to distinguish different phones very well. These high perceptual capabilities gradually reduce during their first year of life to those sounds that are phonemes in their mother tongue [14], [15]. Inspired by this finding, our model derives the space in which it explores from a set of ambient speech sounds.

It has been largely observed in a range of different languages that caregivers hyperarticulate when speaking with their infants (e.g. [16]). This infant-directed speech ("motherese") makes the vowels more distinct from each other which is why hyperarticulated speech is typically seen as higher quality speech. There is evidence that higher speech quality correlates with better speech discrimination performance in infants [17], [18]. Computationally, this has been confirmed in [19] by Adriaans and Swingley: they showed that a learning model based on a Mixture of Gaussians could better discriminate vowels when being trained with prosodically prominent vowel sounds.

We extend this view by testing the effect of higher quality speech sounds on articulatory learning in our model of speech acquisition. We model hyper- and hypoarticulated sounds with tense and lax vowel sounds, respectively. Tense vowels are stronger articulated than lax vowels and are acoustically better discriminable. First, we train models for producing either tense or lax vowels or both types of vowels simultaneously. After initial training, the models are retrained with all vowel sounds. We investigate (1) how the quality of vowels used as ambient speech in initial training affects the model's ability to learn to produce vowels from the other vowel set and (2) how these observations may be explained

based on the articulatory system.

We find that the model can acquire vowel sounds which were not included in initial learning. But a clear difference can be observed depending on the order in which the vowels are learned: If first the stronger articulated tense vowels are learned, a generalization to lax vowels is learned quickly. In the opposite direction, after the system adapted to lax vowels, tense vowels are acquired with significantly lower competence in the same time. Comparison of the learned articulatory configurations reveals that models first trained with lax vowels tend to articulate less also in later learning which might be the reason for the lower competence levels.

This suggests that caregiver's hyperarticulation not only enhances infant's speech discrimination ability [17], [18], [19], but also causes stronger articulation in the infant which might facilitate later articulatory learning.

In the following, we first introduce our model of speech acquisition (Sec. II), then we present experimental results from the initial training (Sec. III) and from the retraining phase (Sec. IV). Finally we discuss the results in terms of the articulatory system (Sec. V).

II. A Model for Learning Articulatory Control

Learning to speak in our framework means to learn which articulatory movements are required for producing desired outcomes. Thus, the system needs to learn the *inverse mapping* $g(\boldsymbol{x})$ from acoustic outcomes $\boldsymbol{x}$ to motor commands $\boldsymbol{q}$ which here encode articulatory parameters of the vocal tract.

The *forward mapping* $f(\boldsymbol{q})$ captures the process of speech production: by executing the vocal tract model with some articulatory parameters an acoustic outcome is produced.

In kinematic learning tasks, forward model and inverse model directly connect the motor configurations (joint angles) with the outcomes (position coordinates). While 2D or 3D space constitute a good low-dimensional space for goal-directed exploration in kinematic learning, speech sounds do not have an inherent low-dimensional representation. We have to first generate such a space, e.g. from ambient speech, such that babbling can take place. Outcomes $\boldsymbol{x}$ in our model, thus, denote positions in this *goal space*: a low-dimensional representation of the acoustic feature space.

An overview of our model is given in Fig. 1. Sec. II-A describes how the goal space is learned from ambient speech (left part of Fig. 1). This forms the goal space and initializes the forward model (including speech synthesis, feature extraction and projection into the goal space). Sec. II-B describes babbling (right part of Fig. 1) which bootstraps the inverse model.

A. Goal Space Formation

Choosing good features is crucial for later learning ability: Sounds with identical feature encodings cannot be distinguished by the model. If we would design low-dimensional features by hand, e.g. by using formant frequencies, we artificially restrict the perception of our model.

A better way is to learn a low-dimensional acoustic representation from data. This is also developmentally plausible, as infant's perception of speech is highly affected by early exposure to their native language [20], [15]. By using statistic information and semantic cues, infants might be able to develop a representation to evaluate the speech sounds they perceive from their environment.

In [13], we proposed to use high-dimensional acoustic features (formants plus mel-frequency cepstral coefficients) and then apply dimension reduction on a set of ambient speech sounds to generate a low-dimensional space capturing the most important variation. First, a statistical dimension reduction was performed using Principal Component Analysis (PCA), reducing the dimensionality to 10 dimensions. Subsequently, we applied Linear Discriminant Analysis (LDA), taking the class information of the vowels into account. This second dimension reduction is motivated by the semantic information infants receive in interaction with their caregiver, as well as by the high sensitivity of young infants to speech contrasts [14]. The resulting 2-dimensional goal space is normalized to $[-1, 1]$ in each dimension. It captures the most relevant information from the acoustic space. This process is depicted on the left side of Fig. 1.

In this study, models trained on different vowel sets should be compared. For this purpose, additionally to the vowel set comprising tense vowels used in [13], we generate a second vowel set with the lax correspondences of these vowels (denoted with capitalized letters). Tense vowels (e.g. the [i] sound in "leap") generally are associated with higher muscle tension than lax vowels (e.g. the [I] sound in "lip"), hence the naming, but this difference does not consistently show in experiments (e.g. [21]).

The ambient speech was generated using the articulatory synthesizer VocalTractLab[1] [22] by executing the predefined articulatory shapes[2] for a duration of 500 ms. We generated 100 acoustic variations for each vowel sound by adding Gaussian distributed noise ($\sigma = 0.01$) to the vocal tract configurations defined for the default speaker (JD2) (cf. [13]).

B. Goal-directed Exploration

After the goal space has formed, the model explores the goal space during *babbling*. The aim is to be able to reproduce the sounds that are present in ambient speech. Thus, targets for the exploration are randomly drawn from a target distribution $P(\boldsymbol{x})$ that is modeled by a Mixture of Gaussians trained on ambient speech data (displayed with small circles in the goal space in Fig. 1).

The underlying algorithm for the exploration is goal babbling, a method for bootstrapping an inverse model for a motor coordination task [7], [23], [24], [8]. Goal babbling operates in the space of outcomes. The inverse mapping estimates which articulatory configurations achieve the desired targets. By adding noise in the space of motor commands and executing the forward mapping, new correspondences of motor commands and outcomes are collected. These are used to improve the inverse mapping.

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

[2] Articulatory shapes (or configurations) are defined with 18 vocal tract parameters describing e.g. lip shape and tongue posture (see [13]).

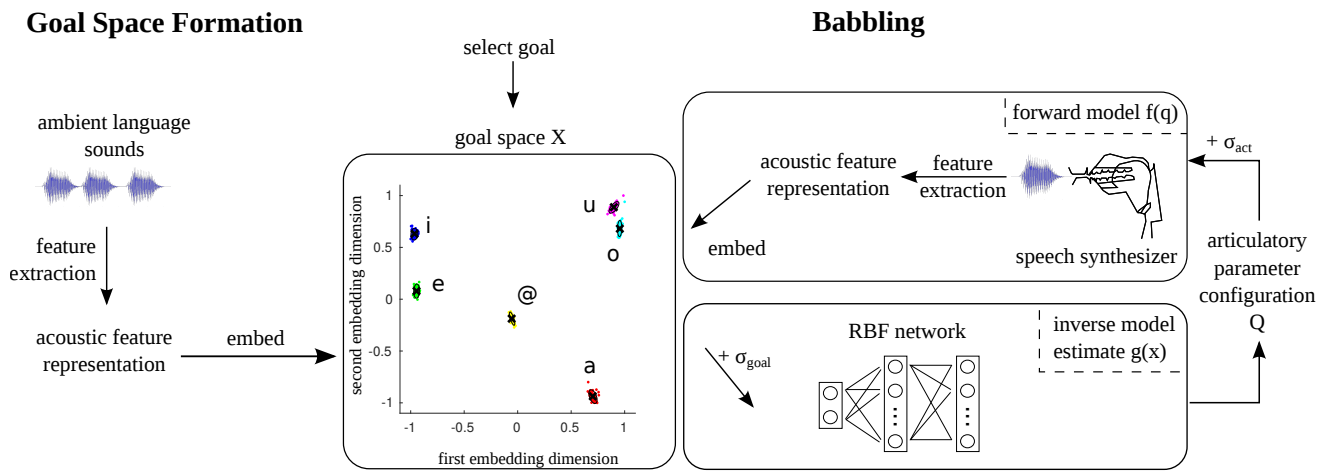

Fig. 1. Goal Space Formation: the goal space is generated from ambient speech. Babbling: the inverse model $g(\boldsymbol{x})$ is trained to estimate an articulatory configuration $\boldsymbol{q}$ for a desired target $\boldsymbol{x}^*$ in goal space such that the forward model $f(\boldsymbol{q})$ embeds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

In each goal babbling iteration, two steps are performed: In the *exploration step* the model tries to reproduce a target from the goal space (depicted on the right side of Fig. 1). In the subsequent *adaptation step* the inverse model estimate is updated with the new experienced correspondences of motor command and achieved goal space position.

We use the learning algorithm from [13], which is based on skill babbling [25], a recently proposed variant of goal babbling that combines goal-directed exploration [23] with episodic learning in mini-batches. This allows for a more efficient inverse model update.

We outline the general principle of skill babbling in the following, for further detail and formula please refer to [13].

The inverse model is initialized with the default action $\boldsymbol{q}_0$ (that produces the neutral sound [@]) and the corresponding goal space position $\boldsymbol{x}_0 = f(\boldsymbol{q}_0)$. Then, in each iteration the system explores around a new target seed $\boldsymbol{x}^{seed}$ (exploration step) and adapts the inverse model accordingly (adaptation step).

1) Exploration Step: A target seed $\boldsymbol{x}^{seed}$ is randomly drawn from the target distribution $P(\boldsymbol{x})$. Noise in the goal space σ_{goal} is added to $\boldsymbol{x}^{seed}$ to generate a number of K slightly varied targets. Then the loop on the right side of Fig. 1 is executed: The inverse model estimates actions $\hat{\boldsymbol{q}}_k$ to approximate the targets $\boldsymbol{x}_k^*$ and exploratory noise σ_{act} in the action space is added that drives the system towards the discovery of new articulatory configurations (cf. Sec. II-B3). Finally, the noisy articulatory configurations $\boldsymbol{q}_k$ are executed, and their actual positions $\boldsymbol{x}_k = f(\boldsymbol{q}_k)$ in goal space are observed.

2) Adaptation Step: The inverse model $g(\boldsymbol{x}, \theta)$ can be implemented with any supervised learning algorithm that minimizes the weighted error criterion and can be updated sequentially [25]. We use a radial basis function (RBF) network: basis functions with radius r are added according to data experience. This partitions the input space (goal space), gradually leading to a more local representation.

The parameters θ of the model are the basis function centers in goal space $\boldsymbol{c}_i$ and the corresponding readout weights $\boldsymbol{u}_i$ which constitute prototypes in action space.

The output of the network is computed according to

$$g(\boldsymbol{x}, \theta) = \sum_{i=1}^{I} h_i(\boldsymbol{x}) \cdot \boldsymbol{u}_i, \tag{1}$$

where $h_i(\boldsymbol{x})$ is the activation of basis function i when input $\boldsymbol{x}$ is presented. $h_i(\boldsymbol{x})$ is calculated with softmax (to ensure well-behaving extrapolation) and scaled with the radius:

$$h_i(\boldsymbol{x}) = \frac{exp(-\frac{1}{r} \cdot \|\boldsymbol{x} - \boldsymbol{c}_i\|^2)}{\sum_{j=1}^{I} exp(-\frac{1}{r} \cdot \|\boldsymbol{x} - \boldsymbol{c}_j\|^2)} \tag{2}$$

When updating the inverse model with a new action-outcome pair $(\boldsymbol{q}_k, \boldsymbol{x}_k)$, a new basis function with center $\boldsymbol{c}_{I+1} = \boldsymbol{x}_k$ is added if $\|\boldsymbol{x}_k - \boldsymbol{c}_i\| > r \ \forall i = 1 \dots I$.

The readout weights of each basis function i are updated with gradient descent (learning rate $\lambda = 0.9$):

$$\boldsymbol{u}_i^{new} = \boldsymbol{u}_i + \lambda \cdot w_k \cdot h_i(\boldsymbol{x}_k) \cdot (\boldsymbol{q}_k - g(\boldsymbol{x}_k, \theta)) \tag{3}$$

$w_k \in [0, 1]$ are weights that reflect how valuable the new experience $(\boldsymbol{q}_k, \boldsymbol{x}_k)$ is. Three weighting schemes are combined which rate the acquired training pairs according to:

- Accuracy: How close is the actual outcome to the desired target in the goal space?
- Relevance: Is the discovered goal space position in the range of the ambient speech?
- Sound quality: Is the speech sound articulated, i.e. above a given intensity threshold?

For details, see [13]. The product w_k of these three weights is incorporated into the gradient descent update (Eq. 3) in order to promote learning of "good" solutions.

3) Adaptive Articulatory Noise: The amplitude of action space noise σ_{act} is crucial for learning: A too low noise amplitude hinders the system from discovering new speech sounds. With a too high noise amplitude the system's performance is unstable.

In [13], we showed that learning with adaptive noise yields better and more stable results compared to learning with

a fixed noise amplitude. The adaptive noise amplitude is calculated from the distance of the desired target to the already discovered region of the goal space. We use this adaptive articulatory noise here, as well, but improve the metric to tackle a problem we discovered in [13]. The proximity of [o] and [u] in the generated goal space led to frequent confusion of these two vowels, especially in the adaptive noise condition. The reason is that the adaptive noise level reduces for both vowels as soon as one of the vowels is discovered. In this work, we overcome this issue by scaling the distance between desired target and discovered region with the distance between the two closest ambient speech clusters.

III. Experiment 1: Training with Different Sets of Ambient Speech Sounds

We generate goal spaces for three different sets of ambient speech as described in Sec. II-A:

(A) With tense vowels only: [a], [e], [i], [o], [u], [@]
(B) With lax vowels only: [A], [E], [I], [O], [U], [@]
(C) With tense and lax vowels: [a], [e], [i], [o], [u], [A], [E], [I], [O], [U], [@]

[@] is the "schwa" sound. It is included in all conditions and is produced by the vocal tract in its neutral configuration.

Fig. 2 depicts the goal spaces into which the ambient speech sounds were projected. We denote the resulting goal space from experiment 1A with X_T, from 1B with X_L and from 1C with X_{TL} in the following. In X_T, [o] and [u] are close to each other, in X_L, [E] and [@] are the most similar ones. In the plot of X_{TL} lax vowels seem to constitute intermediate clusters that lie in between tense vowel clusters.

Goal-directed exploration is executed for each condition (cf. Sec. II-B). Targets in this babbling phase are drawn from the ambient speech sounds from which the goal space was generated, i.e. the model's perception is specifically adapted to these vowel sounds. In each of the three experiments, we independently train 10 models. The parameters for learning were chosen like in [13]. That is, we train each model for a maximum of 500 iterations, or until the errors for reproducing all targets in goal space fall beyond a threshold (0.1). In each iteration, $K = 10$ targets are explored, generated with Gaussian distributed noise with $\sigma_{goal} = 0.05$ ($\approx 2.5\%$ of the goal space) around the target seed $\boldsymbol{x}^{seed}$. The radius of the inverse model basis functions it set to $r = 0.15$ which roughly fits the average expansion of the vowel clusters in goal space. Only the calculation of adaptive noise differs from [13] by a factor as described in Sec. II-B3.

The competence of the models is evaluated in each iteration by letting it imitate the cluster centers of the target distribution. In accordance with e.g. [11], we calculate the competence for a reproduction $\boldsymbol{x}$ of a target $\boldsymbol{x}^*$ as:

$$comp(\boldsymbol{x}, \boldsymbol{x}^*) = exp(-\|\boldsymbol{x} - \boldsymbol{x}^*\|) \tag{4}$$

Higher competence, thus, correspond to smaller euclidean distance in goal space.

Fig. 3 presents the results of experiments 1A and 1B. The graphs illustrate that all vowels are learned with high competence. Lax vowels generally are learned more accurately than tense vowels. The models learned in 1B also converge more quickly (in average after 174 iterations, opposed to 265 iterations in experiment 1A).

The results from experiment 1C, where tense and lax vowels are learned together, are presented in Fig. 4. Performances for tense and lax vowels are presented in separate graphs for the sake of clarity and to make it better comparable to Fig. 3.

We can observe that although all vowels increase to good competence levels, learning takes longer: on average 372 iterations were performed. Especially learning of the tense vowels [o] and [u] is slightly delayed. The presence of intermediate targets does not accelerate the overall learning process. Rather, learning progresses more slowly, because more possible targets are available, so each individual target is probed less often.

IV. Experiment 2: Retraining with New Vowel Sounds

Learning all vowels at once is more difficult for our model compared to learning only tense or only lax vowels. But the models trained in 1A and 1B are specialized: The goal space is organized to either represent tense vowels or lax vowels. In this second experiment we examine whether models trained on only tense or only lax vowels are still able to acquire new speech sounds that were not present in the set of ambient speech sounds.

Two experiments, 2A and 2B, are performed:

(A) Models trained on tense vowels (experiment 1A) babble additionally lax vowels.
(B) Models trained on lax vowels (experiment 1B) babble additionally tense vowels.

An important prerequisite for learning new sounds is that the models are able to discriminate the new sounds in terms of their goal space. So we first examined how the goal spaces X_T and X_L perceive vowels that were not included in the goal space formation. As Fig. 5 shows, there is more variance in the perception of new vowels (colored clusters) than in the perception of speech sounds from which the goal space has been derived (black clusters).

X_T perceives the lax vowels similarly to how lax vowels are perceived by X_{TL}. This suggests that the relationship in the acoustics learned from tense vowels captures important structures of lax vowels as well. Tense vowels perceived by X_L lie generally more outside of the original goal space. The inverse model has to extrapolate, thus, a higher reproduction error can be expected in experiment 2B.

Babbling proceeds similarly to the first babbling phase. Before babbling continues, the new vowels are perceived, i.e. projected to the goal space as depicted in Fig. 5, and the target distribution is extended to incorporate the new vowels in addition to the old ones. Then, 500 babbling iterations are performed, babbling old and new vowels randomly, and competence progress is computed analogously to experiment 1.

Fig. 6 shows the results for experiment 2A. The upper graph demonstrates that learning new sounds slightly affects the competence of the already acquired tense vowel sounds. In the graph at the bottom the competence growth of the

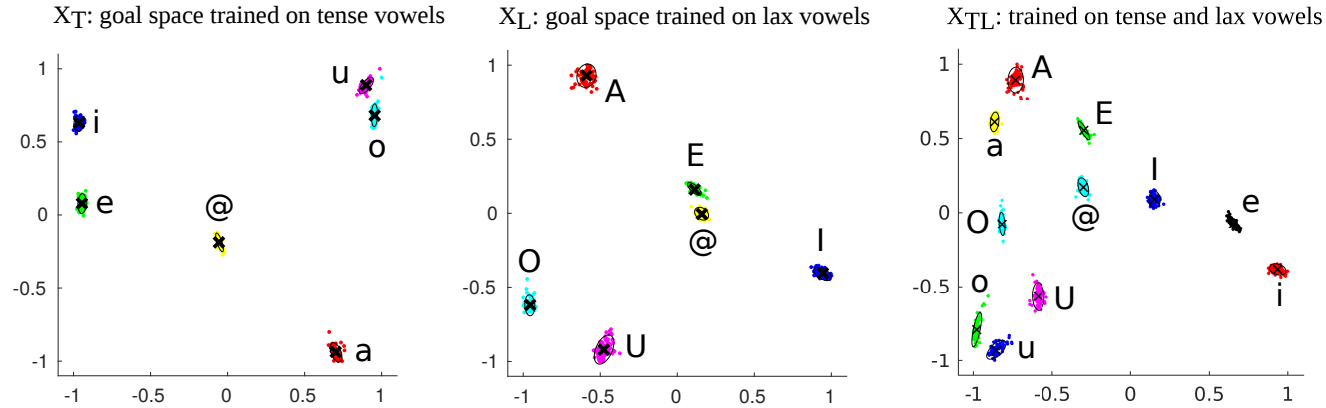

Fig. 2. Goal spaces generated via sequential application of PCA and LDA from three different sets of ambient speech, comprising only tense vowels (left), only lax vowels (middle) and tense and lax vowels (right).

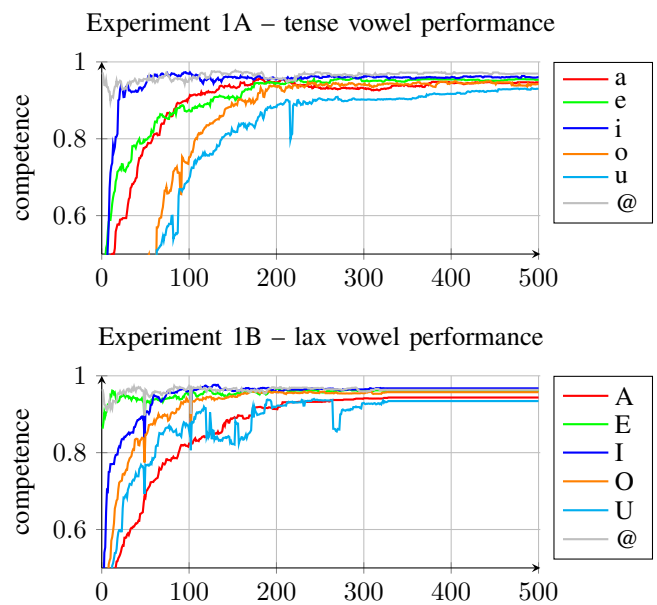

Fig. 3. Competence per vowel during exploration (averaged over 10 trials) for learning tense vowels with X_T (top) and for learning lax vowels with X_L (bottom).

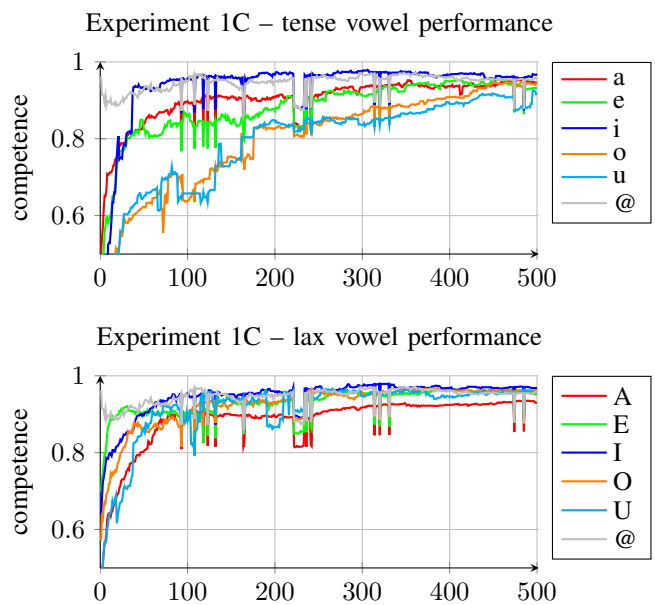

Fig. 4. Competences for tense (top) and lax (bottom) vowels during exploration (averaged over 10 trials) for learning tense and lax vowels with a combined goal space (X_{TL}).

new lax vowels is plotted. Some of the lax vowels can already be imitated quite well by the models in the beginning. This is possible because the inverse model is parametric: although exploration during experiment 1 mainly focused on regions of the goal space where ambient speech resides, the regions in between clusters might also be covered with basis functions while the model tries to achieve the desired speech sounds. Apparently, the linear interpolation between the tense vowels fits the lax vowel perception in most cases. Only [O] is reproduced inaccurately in the beginning, but during babbling its competence quickly increases.

In Fig. 7, we observe that for models trained on lax vowels only it is much harder to acquire tense vowels. The competence increases for all vowel sounds, but is significantly lower for [o] and [u].

A perceptual evaluation[3] of the results reveals that despite a higher reproduction error, vowels in both experiments are produced well enough for a human listener to recognize them. But while all sounds can be perfectly distinguished in experiment 2A, [o] and [u] as reproduced in experiment 2B can only be distinguished in 30% of the cases. However, the ongoing competence increase in Fig. 7 suggests that this might improve when babbling continues.

These results show that training our model on a set with higher articulatory variance first is beneficial for later learning of new vowel sounds.

[3]The auditory results from all experiments are available at https://techfak.uni-bielefeld.de/%7Eaphilipp/ijcnn17-results/

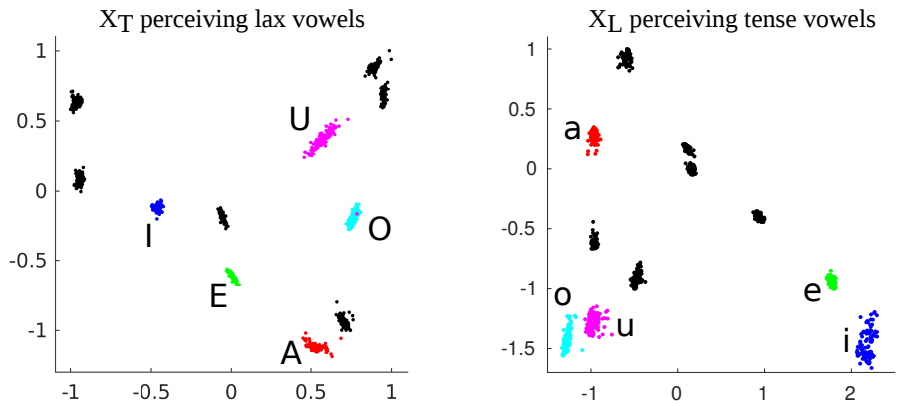

Fig. 5. How goal spaces trained only on tense or lax vowels (cf. Fig. 2) perceive lax and tense vowels, respectively. Black clusters show the perception of originally trained vowels, colored clusters correspond to the perception of new vowel sounds.

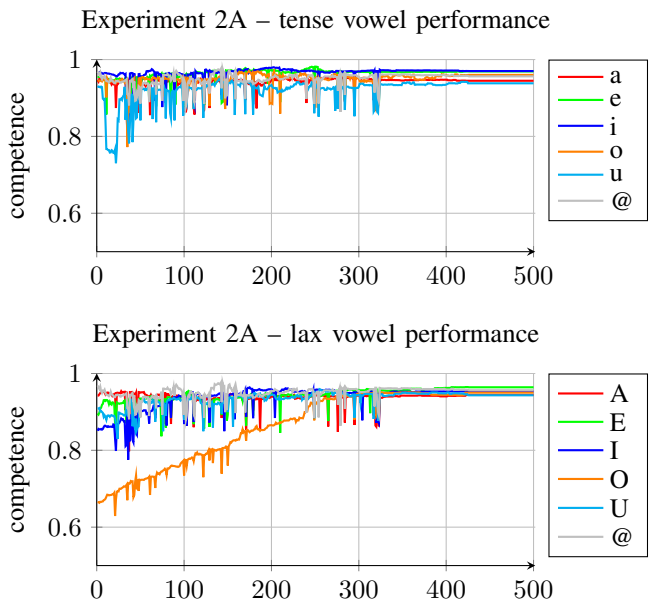

Fig. 6. Competences for tense (top) and lax (bottom) vowels (averaged over 10 trials) during retraining with models trained only on tense vowels.

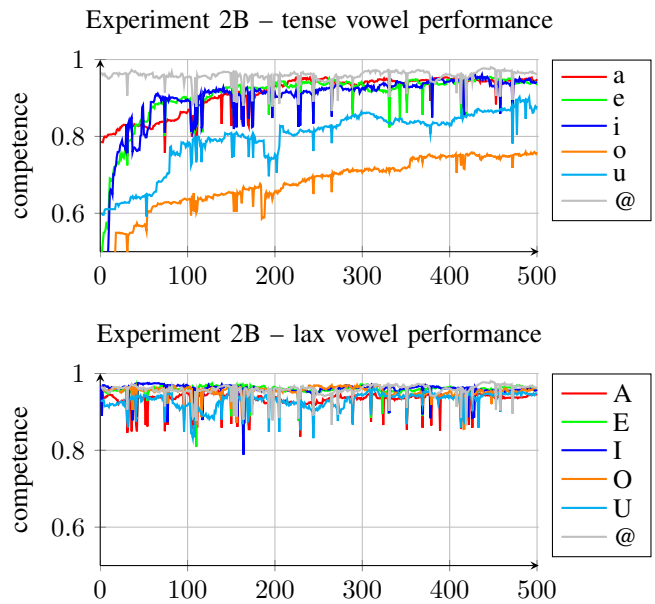

Fig. 7. Competences for tense (top) and lax (bottom) vowels (averaged over 10 trials) during retraining with models trained only on lax vowels.

V. Articulatory Effort of Acquired Vocal Tract Configurations

Why do models from experiment 2A learn faster and achieve a higher competence level than models from experiment 2B? A possible explanation is that different articulatory configurations are discovered in both experiments. Exploratory noise is added to the actions estimated by the inverse model, thus, the basis functions that the inverse model establishes during babbling in experiments 1A or 1B play a crucial role: Which articulatory configurations the system can reach depends on which articulatory configurations it already "knows" in terms of its inverse model.

To shed light on the articulatory configurations that the inverse model has learned, we measure the articulatory effort of the obtained articulatory configurations. We do this by computing their deviation from the neutral shape (which produces [@]) as the mean euclidean distance in the articulatory space of VocalTractLab. The neutral shape was used as home posture $\boldsymbol{q}_0$ for babbling in all experiments.

The results are presented in Tab. I. A difference between tense and lax vowels can already be found in the original articulatory configurations as defined in VocalTractLab: Tense vowels deviate from the neutral shape by approx. 3.6 cm, lax vowels by 2.3 cm, i.e. for tense vowels around 59% more articulatory movement is required. The babbled articulatory shapes from experiments 1A and 1B reflect this proportions.

When both vowels are included in a single goal space (1C), the articulators generally move more to produce the same vowels. Specifically, models trained on both vowel sets tend to over-articulate the lax vowels, perhaps to sharpen the contrast that the model is sensitive to due to the variability of

ambient speech in the goal space formation phase. Assuming that less movement in the articulators is favorable due to lower effort, the specialized models in this respect have an advantage over the models trained on both vowel sets: they achieve higher competence with less articulatory effort in their domain.

In experiment 2A, models trained on tense vowels were retrained with tense and lax vowels. There is no change for tense vowels, while the effort for producing lax vowels is higher than in 1B, but comparable to the effort of lax vowels in the original articulatory configurations.

Significantly lower articulatory effort for producing the vowels can be observed in experiment 2B. The lower competence that models trained on lax vowels exhibit for learning tense vowels might, thus, be caused by not enough articulatory movement.

Apparently, which vowel sounds the model learns in its early stage is important. It influences not only the model's capability to acquire new vowel sounds, but also affects how much effort the model puts into articulation later on.

Tab. I
MEAN DEVIATION (IN CM) OF ARTICULATORY SHAPES ACQUIRED IN EXPERIMENTS 1 AND 2 FROM THE NEUTRAL SHAPE.

	tense vowels	lax vowels
original shapes	3.6	2.3
1A	3.2	-
1B	-	1.8
1C	3.4	2.7
2A	3.2	2.4
2B	2.6	2.1

VI. CONCLUSION

We investigated the effect of hyperarticulation on articulatory learning in a developmentally inspired model of speech acquisition. In particular, we evaluated whether models trained only on tense or lax vowels are able to acquire sounds from the other vowel set. Results show that this generalization is possible. Although learning succeeds in both conditions, a significantly higher competence is achieved when tense vowels are learned first. Higher articulatory variance in early vowel learning in our model, thus, is beneficial for later accommodating additional vowel sounds.

We also observed that depending on the quality of early learned speech sounds (hyper- or hypoarticulated), the model acquires different articulatory configurations in the later learning phase, i.e. early language learning influences later articulatory learning. This supports findings related to infant-directed speech and might even have analogies to the appearance of an accent as observed in adult's foreign language learning.

ACKNOWLEDGEMENT

This research was 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 is related to the European Project CODEFROR (FP7-PIRSES-2013-612555).

REFERENCES

[1] 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.
[2] 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.
[3] 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.
[4] 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.
[5] A. S. Warlaumont, "Salience-based reinforcement of a spiking neural network leads to increased syllable production," in *IEEE Intern. Conf. on Development and Learning*, 2013.
[6] 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.
[7] 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.
[8] 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.
[9] C. Moulin-Frier and P.-Y. Oudeyer, "Curiosity-driven phonetic learning," in *IEEE Intern. Conf. on Development and Learning*, 2012.
[10] 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.
[11] C. Moulin-Frier and P.-Y. Oudeyer, "Exploration strategies in developmental robotics: a unified probabilistic framework," in *IEEE Intern. Conf. on Development and Learning*, 2013.
[12] 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.
[13] A. K. Philippsen, R. F. Reinhart, and B. Wrede, "Goal babbling of acoustic-articulatory models with adaptive exploration noise," in *IEEE Intern. Conf. on Development and Learning*, 2016.
[14] 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.
[15] N. Ferjan Ramírez, R. R. Ramírez, M. Clarke, S. Taulu, and P. K. Kuhl, "Speech discrimination in 11-month-old bilingual and monolingual infants: a magnetoencephalography study," *Develop. science*, 2016.
[16] P. K. Kuhl, J. E. Andruski, I. A. Chistovich, L. A. Chistovich, E. V. Kozhevnikova, V. L. Ryskina, E. I. Stolyarova, U. Sundberg, and F. Lacerda, "Cross-language analysis of phonetic units in language addressed to infants," *Science*, vol. 277, no. 5326, pp. 684–686, 1997.
[17] H.-M. Liu, P. K. Kuhl, and F.-M. Tsao, "An association between mothers' speech clarity and infants' speech discrimination skills," *Developmental Science*, vol. 6, no. 3, pp. F1–F10, 2003.
[18] J. Y. Song, K. Demuth, and J. Morgan, "Effects of the acoustic properties of infant-directed speech on infant word recognition," *The Journal of the Acoustical Society of America*, no. 128, pp. 389–400, 2010.
[19] F. Adriaans and D. Swingley, "Distributional learning of vowel categories is supported by prosody in infant-directed speech," in *Annual Conference of the Cognitive Science Society*, 2012, pp. 72–77.
[20] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831–843, 2004.
[21] L. J. Raphael and F. Bell-Berti, "Tongue musculature and the feature of tension in english vowels," *Phonetica*, vol. 32, no. 1, pp. 61–73, 1975.
[22] P. Birkholz, "VocalTractLab – Towards high-quality articulatory speech synthesis," http://www.vocaltractlab.de/, 2014.
[23] 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.
[24] M. Rolf, "Goal babbling with unknown ranges: A direction-sampling approach," in *IEEE Intern. Conf. on Development and Learning*, 2013.
[25] R. F. Reinhart, "Autonomous exploration of motor skills by skill babbling," *Autonomous Robots*, pp. 1–17, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10514-016-9613-x