Goal-Directed Exploration for Learning Vowels and Syllables: A Computational Model of Speech Acquisition

Anja Philippson¹

Received: 14 May 2020 / Accepted: 13 January 2021 / Published online: 27 January 2021
© The Author(s) 2021

Abstract

Infants learn to speak rapidly during their first years of life, gradually improving from simple vowel-like sounds to larger consonant-vowel complexes. Learning to control their vocal tract in order to produce meaningful speech sounds is a complex process which requires to learn the relationship between motor and sensory processes. In this paper, a computational framework is proposed that models the problem of learning articulatory control for a physiologically plausible 3-D vocal tract model using a developmentally-inspired approach. The system babbles and explores efficiently in a low-dimensional space of goals that are relevant to the learner in its synthetic environment. The learning process is goal-directed and self-organized, and yields an inverse model of the mapping between sensory space and motor commands. This study provides a unified framework that can be used for learning static as well as dynamic motor representations. The successful learning of vowel and syllable sounds as well as the benefit of active and adaptive learning strategies are demonstrated. Categorical perception is found in the acquired models, suggesting that the framework has the potential to replicate phenomena of human speech acquisition.

Keywords Goal babbling · Speech acquisition · Vocal learning · Imitation learning · Inverse model · Developmental learning

1 Introduction

Speech production is a complex motor task that requires the simultaneous coordination of dozens of muscles and extremely fast movements. In the light of these difficulties, human children acquire their first words remarkably fast during their first years of life. However, little is known about how they acquire speech [1, 2].

Machine learning systems learn in a vastly different way. Speech recognition and speech synthesis are active fields of research which are, nowadays, dominated by deep learning methods. State-of-the-art methods use large amounts of data to achieve remarkable performance [3, 4]. Such systems are trained on databases often containing hundreds of millions of words [5, 6], while infants are estimated to experience only around 20–40 million words in their first 3 years [7, 8]. Despite this effort, the generalization capability of machine

recognition and synthesis of speech is still limited, and far beyond what human beings are able to do. In particular, speech communication systems often lack flexibility and robustness against perturbations of the speech sounds that humans easily adapt to [9, 10]. Human children do not only experience less example utterances than such systems, but also cope with imperfectly labeled training data. Furthermore, they learn only from a limited number of speakers and still generalize later even to accented speakers of the language. How do children master this difficult task?

One central difference between speech processing in humans and machines is the way of how perception and production are treated. The human speech processing system has a strong coupling between perception and production which is considered to aid speech acquisition and to increase the robustness of perception [11]. Specifically, knowledge about how speech is produced helps to predict natural deviations of speech that is caused, for instance, by differences in the anatomy of the speaker.

Using such inspiration from how humans learn to speak also could benefit the development of speech recognition and production systems. Conversely, computational models

✉ Anja Philippson
anja@ircn.jp

¹ International Research Center for Neurointelligence, The University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo, Japan

of speech acquisition can also support research into the underlying mechanisms of speech acquisition in infants. Linguistic researchers nowadays mainly utilize observational approaches to gain an understanding of the methods and strategies that children use. Computational models could help to become more flexible in such analyses, for example, it is possible to modify specific parameters of learning and observe the effect on development.

The branch of developmental robotics aims to enable robots and, in general, artificial systems to acquire skills, starting with the learning of fundamental capabilities, and gradually extending to more complex tasks, similar to how human children learn [12–14]. Developmental learning methods have been extensively applied to standard robotic tasks, in particular, to reaching and grasping of objects. Acknowledging the fundamental role of speech acquisition for human children, in recent years, a growing number of computational approaches have been suggested to model speech acquisition in a more developmental way. A recent review was presented in [15].

This paper builds on these studies, proposing a computational model for acquiring a set of speech sounds by using goal babbling, an exploration strategy that is inspired by infants' learning. The target is to introduce a unified framework for acquiring either static or dynamic motor configurations for learning a set of vowel or syllable sounds. Static motor representations are a common learning target in previous computational speech acquisition models [15] and can be learned very efficiently. However, in contrast to dynamic motor representations, they lack the ability to represent full syllables including vowels as well as consonants. Therefore, this framework is designed to handle both, static motor representations (for the efficient learning of vowels), and dynamic motor representations (for learning syllables).

Furthermore, the effect of two acquisition strategies is discussed: (1) a flexible adaptation of articulatory exploration noise during learning and (2) a variant of active learning depending on current competence progress.

Finally, similarities to human learning are evaluated and discussed, in particular, considering categorical perception and the order in which the different sounds are acquired.

The source code for the framework, designed to support different types of vocal tracts, acoustic features and learning mechanisms, is available at GitHub: <https://github.com/philippippen/goalspeech>.

2 Related Models

Computational models typically define speech acquisition as the task to acquire sensorimotor coordination, i.e. learning about the relationship between articulatory movements of a vocal tract model and the corresponding speech sound that

this motor command produces [15]. To achieve this coordination, an exploration of the sensorimotor space is required.

The simplest way to approach this exploration is a direct random exploration in the space of motor commands, as it has been implemented, for example, in [16, 17]. By trying out new motor configurations, gradually more sounds are discovered. However, such motor babbling is not efficient in high-dimensional motor spaces: many motor configurations have to be explored before properly articulated speech sounds are discovered. Active exploration strategies may help to increase efficiency [18], by integrating, for instance, the saliency of the produced sound [19], or caregiver feedback [20, 21].

Such strategies can help to overcome sampling problems in the motor space, and might play a role in infants' learning. However, experimental evidence suggests that already young infants explore in a goal-directed way [22–26]. Thus, infants seem to be aware of the notion of goals very early in their development, and might use such goals actively for acquiring new skills.

Inspired by these findings, some recent studies propose to use the developmentally more plausible approach of goal-directed exploration [18, 27, 28]. The idea is to select goals for exploration in the space of outcomes which is typically lower dimensional and better structured than the motor space. This space is called goal space, and exploration in this space was introduced as goal babbling. Originally, goal babbling was proposed as an efficient algorithm for learning inverse kinematics for robots, where the challenge is to cope with the high-dimensional actuators which need exploration mechanisms that are feasible in real-world interaction [29–31]. The basic idea is to explore by trying to reach specific positions in the goal space. An inverse model is trained online during the babbling process. It maintains a mapping from desired goals to required motor configurations. New experience is integrated into the inverse model by updating it with newly discovered goal–action pairs, until it achieves proficiency in the desired tasks. Goals can be selected in a random manner, or using active exploration methods such as exploring goals primarily in regions where it is most beneficial for the system. In [27], an active variant of goal babbling was successfully applied for acquiring speech sounds based on the ideas of intrinsic motivation. The authors implemented a method that primarily explores goals for which a high recent increase in competence was observed. As a result, well articulated speech gradually emerged in a similar way to how it is observed in infants' babbling.

One issue in applying goal babbling for speech acquisition is that it is not obvious how the goal space of speech looks like. In robotic tasks such as reaching for an object, a low-dimensional goal space is naturally defined by three-dimensional space coordinates. In contrast, sound can be represented in various ways, and most acoustic feature

representations are high-dimensional. In previous studies, hand-crafted goal space representations based on formant features were used to make exploration feasible [27, 28]. As a more flexible solution, we previously developed a method of automatic goal space generation using high-dimensional acoustic features and dimension reduction methods [32]. The idea was to generate the goal space via dimension reduction methods from example speech sounds. This procedure was inspired by the fact that infants are exposed to the speech in their environment [1, 2, 33–35] and this information can help them to better structure the learning process. This paper builds on our previous study, and extends the framework developed in [32] to a more general formulation that can be applied not only for learning vowels, but also for acquiring short syllable-like sounds.

Additionally, the framework presented here extends previous studies in the field [27, 28] in three major ways: first, it models not the general emergence of articulated speech, but the bootstrapping of a set of concrete speech sounds. Second, learning is exemplified for VocalTractLab (VTL) [36], a vocal tract model which realistically simulates speech production in humans based on a three-dimensional geometric model, obtained from MRI data of a reference speaker [37]. Successful speech production with this model requires the coordination of about twenty articulatory parameters describing the movements of articulators such as tongue or lips (vocal tract parameters), and of the vocalization mechanism (glottis parameters). This realistic modeling of the speech mechanism enables VTL to produce a large variability of well distinguishable vowel and consonant sounds. Third, the goal space in which the system explores is not predefined, but derived from a set of speech sounds: these ambient speech sounds should reflect which sounds the system experienced in its synthetic environment. Here, as ambient speech only speech is utilized which the system is able to produce itself, namely, speech which was generated by the same vocal tract model. By modifying which vowels or syllables are contained in the set of ambient speech sounds, the mapping to the goal space changes. This mapping acts as a filter on the perception similar to how language exposure affects human perception of speech [38].

3 A Computational Model of Speech Acquisition

An overview of the components of the framework and their interplay is presented in Fig. 1. The process of learning how to speak is organized in two phases: The first phase, the perceptual learning phase (arrow with dashed line in Fig. 1), consists of the generation of a low-dimensional representation of goals, the so-called goal space, from the system's synthetic ambient speech. In the second phase (arrows with

Fig. 1 Overview of the model components: Exploration takes place in the goal space. Via inverse and forward model, speech sounds are generated which are mapped back into the goal space. The closed loop is the perception–action loop executed during babbling. Ambient speech is only utilized for the initial generation of the goal space

solid lines in Fig. 1), the sensorimotor learning phase, goals are drawn from this goal space and motor commands are explored in order to achieve these goals. In this way, an inverse model of speech production is bootstrapped.

In the following, the individual components and their roles in the framework are discussed before introducing to the goal babbling algorithm in Sect. 4.

3.1 Motor Representation

The motor representation depends on the vocal tract model that is used. In this study, the articulatory speech synthesizer VTL (version 2.1) [36] is employed. VTL has 30 controllable parameters: 24 articulatory parameters and 6 glottis parameters¹. Not all 24 articulatory parameters have to be learned, as some of them can be determined automatically from the other parameters [36]. Here, a subset of 18 articulatory parameters is used which is in line with the selection made in other studies using this vocal tract model [40, 41]:

- Hyoid position (HX, HY)
- Jaw position and angle (JX, JA)
- Lip protrusion and lip distance (LP, LD)
- Velum shape and velic opening (VS, VO)
- Tongue position parameters (TCX, TCY, TTX, TTY, TBX, TBY) and side elevation parameters (TS1–TS4).

For the first half of the experiments (Sect. 5), static articulatory configurations are used to represent vowel sounds. Glottis parameters can be set to default values and held constant while producing some airflow. Therefore, motor commands are expressed as 18-dimensional vectors of the articulatory parameters. When producing speech, the configuration is

¹ The number of glottis parameters depends on the selected glottis model, here, the triangular glottis model is used [39].

copied to 100 time frames (cf. “temporal rollout” in Fig. 1) which results in the production of a signal of 500 ms (with an articulatory sample rate of articulatory parameters fed into VTL of 200 Hz).

The second half of the experiments (Sect. 6) demonstrates babbling using a dynamic representation of motor parameters. In this way, vowels as well as syllables can be represented. Both articulatory and glottis parameters have to be adjusted to capture the dynamics required, for example, for forming plosives or fricatives. Therefore, the following 3 glottis parameters are added to the representation: fundamental frequency (F0), subglottal pressure and aspiration strength (cf. [42]). Other glottis parameters have a minor effect on the speech quality and can be disregarded. To represent changes in time in articulatory and glottis parameters, it would be possible to simply concatenate frames in each time step. However, this solution would generate a very high-dimensional representation for a single speech sound (number of time frames $\times$ number of parameters). Furthermore, such a representation does not account for the smoothness of articulatory parameter changes. For these reasons, articulatory trajectories are represented here with Dynamic Movement Primitives (DMPs) [43]. DMPs can represent trajectories by combining differential equations of point attractor dynamics with time-dependent perturbations. To define the level of the perturbation in each time step, K Gaussian basis functions are equidistantly spread over the time course of the trajectory. Each basis function has an associated weight θ_k . Then, the perturbation of the trajectory in a certain time step is determined by the distance in time to the basis functions and the corresponding weights θ_k . A trajectory, thus, can be described by θ_k . Additionally, an initial point and a target point for a trajectory can be specified. Owing to the point-attractor dynamics, the trajectory will converge to reach the target point. The implementation used in this study is the same as in [44, 45]. K is set to 4 which is the smallest value that allows for a comprehensible production of the target syllables used in this study². Additionally to the K basis function weights, the initial and target points are included in the representation, resulting in a dimensionality of $4 + 2 = 6$. The trajectories of all 21 motor parameters (18 articulatory plus 3 glottis parameters) are coded with separate weights, leading to a motor space representation of dimension $21 \times 6 = 126$. As the synchronous movement of multiple articulators is crucial for the successful production of consonants, a multidimensional DMP implementation is used: convergence of all 21 trajectories is assumed when

the absolute velocity (sum of individual components representing different dimensions) falls below a small threshold (0.001). By rolling out the DMP trajectory over time (cf. “temporal rollout” in Fig. 1), the DMP parameters are converted into articulatory trajectories which can be directly fed into the vocal tract simulator. Note that due to the above mentioned convergence criterion two *separate* DMP motor representations can yield articulatory trajectories of different lengths depending on how quickly or slowly they converge.

Example configurations for generating speech sounds with VTL can be obtained from the predefined configurations provided in speaker configuration JD2 in [36] for vowels, or online at [46] for syllables. In this study, a set of these examples (corresponding to the set of vowel or syllable sounds that the system should acquire in the given experiment) is used to normalize the articulatory parameters to the range $[-1, 1]$ (cf. Sect. 3.2). This normalization is performed parameter-wise (each articulatory parameter and each movement primitive parameter is normalized separately). The normalization ensures that the noise that is later applied to the articulatory parameters has a comparable effect on all parameters.

3.2 Ambient Speech

Infants are exposed to speech of their native language(s) extensively during their first months of life, and even before birth [1, 2, 33, 34]. It has been shown that very young infants are able to perceive human speech in a language-independent way. For instance, American and Japanese infants are similarly well able to distinguish /r/ and /l/ in speech [38]. However, the infant’s perception changes during the second half of their first year [47, 48] to focus on differentiating those sounds that are phonemes in their native language. The goal space used in this study is designed to function as such a language-dependent representation, which is likely to be formed already when infants achieve their first proto-syllables during the canonical babbling phase [35].

Communication requires an agent to use similar sounds as other agents in its environment. Ambient speech, thus, plays an important role for infants; it provides them with goals that they want to achieve. Inspired by the function of ambient speech for infants, the framework presented here is provided with a set of speech sounds, referred to as ambient speech.

Although, in principle, any kind of speech signals could be used as ambient speech, the requirement is that the acoustic processing extracts features which are sufficiently invariant to irrelevant differences in the sound such that the same vowel produced by different speakers projects to the same region in goal space. How infants cope with this correspondence problem, or how speech features can be created to overcome this problem is still an open field of research [49, 50], and might require the usage of semantic information and

² How the syllables /aa/, /baa/ and /maa/ sound when generated with DMPs with different values of K can be checked in the GitHub repository in the folder: https://github.com/aphilippsen/goalspeech/tree/master/data/results-data/syllables/DMP-basis-fct_comparison.

context. An analysis in a previous study of this framework [51] (p. 167) indicates that vowels of a human male speaker are perceived by the goal space in a similar way to the artificially generated speech, whereas a female speaker's vowels project to different goal space regions. In the present study, the correspondence problem is excluded by assuming that the tutor (providing ambient speech) and the learner have the same vocal generation mechanism.

In particular, the sounds that the system should learn in a given experiment are generated from default vowel and syllable configurations (cf. Sect. 3.1). By adding a small amount of Gaussian random noise to the articulatory parameters, a number of varying example sounds can be generated. For vowels, noise is drawn from a Gaussian distribution with variance $\sigma^2 = 0.1$ and added to the normalized articulatory parameters for hyoid, jaw and lip parameters. For tongue and velum parameters which are more sensitive, the noise variance is reduced to 0.01. For syllables, noise is analogously added to the normalized DMP parameters. For the additionally used glottis parameters, the noise variance is set to 0.05. Per speech sound that should be included in the ambient speech set, 100 noisy variants of the sound are generated. These generated articulatory configurations are only used for generating ambient speech, but discarded afterwards.

3.3 Auditory Perception

From the generated speech sounds, acoustic features are extracted in a frame-wise manner to represent their temporal-spectral properties. The most commonly used features in models of speech acquisition are formants [17, 18, 27, 52]. Formants refer to the characteristic frequency bands of a speech signal. The first two or three formants well represent differences in the speech spectrum of vowel sounds which are caused by the movement of the tongue. For the experiments in Sect. 5, the first, second and third formants are extracted via Praat [53]. The extraction is performed for each millisecond of the speech signal, and afterwards, the values are downsampled by taking the mean of each 10 extracted formants. To filter out erroneous formant values, the changes in the formants are monitored: if within 10 ms the formant value changes by more than 50 Hz, this value is considered unreliable and filtered out.

In Sect. 6, a dynamic motor representation is used which can include the generation of consonant sounds. Thus, a more sophisticated feature representation is required. A common choice in the field of speech recognition are Mel-frequency cepstral coefficients (MFCCs) [54]. These features are computed by applying the discrete cosine transform on the log-mel-scaled power spectrum. 13 coefficients are

computed on windows of 20 ms, shifted by 10 ms, using the implementation³ of [55].

The time-varying acoustic features have to be reduced in the time dimension to map them to the goal space in which a single position corresponds to a full speech sound. Various approaches can be used to achieve this temporal integration (cf. Fig. 1). A simple approach would be to concatenate the high-dimensional acoustic features to one large vector. As the time series for different sounds may differ in length (in particular for syllable sounds due to the DMP representation), too long or too short time series would have to be cut or augmented to ensure an equal number of dimensions (which is required for mapping them to the goal space, cf. Sect. 3.4). A more elegant solution to perform temporal integration is to use a kernel that can represent time series data. Recent studies showed that a model-based kernel using a generative model can be useful for representing time series in a low-dimensional way [56, 57]. The basic idea is to describe a time series by the parameters of a generative model that would reproduce this time series. These parameters are then known as the model space representation of the time series [56, 57]. As a generative model that makes very little assumptions on the structure of the underlying time series, for example, an Echo State Network (ESN) [58] can be used. This recurrent neural network type has a recurrent layer (the reservoir) and has been previously demonstrated to be rich enough for modeling speech production and perception processes [41, 59]. In an ESN, the input weights and the recurrent connection weights are set to fixed random values which provide rich temporal dynamics. ESNs, thus, can be trained very efficiently by collecting the internal state sequence when presenting the input and determining the output weights via linear regression.

To generate the model space representation of an input signal (here, of the acoustic feature time series), the network is trained during run time to perform one-step-ahead prediction of the input time series (see Fig. 2). The determined output weights that minimize the prediction error then serve as a time-independent representation of the time series. Importantly, the same set of fixed recurrent weights have to be used for each processed speech sound to make the obtained parameters comparable.

Note that other generative models may be used instead of the ESN, however, the efficient training via linear regression makes it particularly attractive as the training is performed on every time series individually during run time.

³ Standard parameters of [55] are used, i.e. an FFT size of 512, 26 filters for filtering and 22 for liftering.

Fig. 2 Pipeline of temporal integration for processing an example syllable /baa/: MFCC features are extracted on windows of 20 ms of the signal. The ESN is trained to represent this time series by estimating

Here, an ESN with 10 neurons in the recurrent layer⁴ is utilized. The dimension of the input and output layer depends on the dimension D of the used acoustic features ($D = 3$ for formants and $D = 13$ for MFCC features). Each sequence, thus, is represented by a vector representation of size $10 \times D$. The full pipeline from a sound signal to the ESN model space representation is illustrated in Fig. 2.

Note that the procedure proposed here is not the only possible way to model the auditory perception. Specifically, when only vowels should be learned the ESN representation could be omitted. Results using alternative feature processing pipelines are made available as supplemental material in the GitHub repository⁵. These analyses indicate that static motor configurations can be learned when using formant features directly as goal space dimensions. In contrast, the acquisition of dynamic motor configurations succeeds only when using the model space representation.

3.4 Goal Space

Even after reducing the dimensionality of speech sounds via feature computation and model space representation, exploration directly in the space of acoustic features would be computationally inefficient. The sounds that a learner would experience constitute an extremely sparse subset of a high-dimensional space—exploration would suffer from the curse of dimensionality. Therefore, a low-dimensional representation is formed using dimension reduction techniques: the goal space. The advantage of exploring in this goal space is two-fold. First, exploration becomes more efficient, and second, the created representation is specific to the ambient speech from which it was generated. The system, thus, learns to achieve a set of speech sounds that is useful for communication in its synthetic environment.

the next time frame $x(t+1)$ from the previous one $x(t)$. The trained output weights are used as a time-independent representation of the signal

Here, simple dimension reduction techniques are employed to generate a two-dimensional goal space. Namely, Principal Component Analysis (PCA) [60] and Linear Discriminant Analysis (LDA) [61] are utilized to project the time-independent sound representation, obtained by the ESN, to two dimensions. After the projection, a full vowel or syllable sound is represented as a single point in the goal space. PCA extracts from high-dimensional data those dimensions that capture most of the variance in the data based on the eigenvectors. LDA is a supervised approach that additionally utilizes class information and, therefore, sharpens the contrast between different speech sound classes. Using such supervised information can be motivated by the remarkable sensitivity of young infants to speech contrasts [62, 63]. Furthermore, whereas infants have access to multimodal information that helps them to decide which sounds should be treated as different phonemes, this framework, being based only on acoustic information, does not have sufficient clues to know which classes it should separate. Using LDA can be considered as a measure to ensure separation of the desired target classes. Both dimension reduction methods are parametric dimension reduction techniques which directly yield a mapping to project from the high-dimensional to the low-dimensional space. Thus,

Fig. 3 Examples for goal space representations generated for a set of vowels (left, using formants and model space representation) or syllable sounds (right, using MFCCs and model space representation). Black circles show the covariance of the target clusters (cf. Sect. 4.3), /@/ is the *schwa* vowel (cf. Sect. 5)

⁴ The ESN's leak rate is set to 1, lambda regularization to 1 and connection density to 0.2. Weights are automatically initialized in range $[-1, 1]$ (scaled to spectral radius 0.95). As activation function tanh is used.

⁵ See in GitHub repository <https://github.com/aphilippsen/goalspeech/tree/master/data/results-data/result-overview.txt>.

the projections preserve linear relationships between the different sounds. Such linearities are useful during exploration because learning can be guided from one speech sound to another (see Sects. 5.2 and 6 for a discussion of linearity).

For vowel as well as for syllable sounds, data are first preprojected to 10 dimensions using PCA, followed by a LDA to 2 dimensions. Examples for goal spaces generated for vowel or syllable sounds are displayed in Fig. 3. It can be observed that all sounds are mapped to distinct clusters, but dependencies among the vowels are captured. For example, /e/ and /i/ as well as /o/ and /u/ lie close to each other in the goal space, indicating that they are perceptually similar.

Goal space generation is a crucial step in learning, as it determines the “perception” of the system. Two points that are projected to the same point in goal space cannot be distinguished by the system. The goal space, thus, provides a language-dependent representation of speech. When making a comparison to infant’s learning, the goal space would represent the perception of a child of about one year that already became attuned to the native language contrasts. However, infants are certainly still adaptable even in later stages of learning. To some degree even older children and adults might be able to adjust their “goal space” via training or speech therapy. In this regard, the presented model in its current form is less flexible and is only an approximation of how infant learning proceeds. Extensions of the system, however, could account for adaptability by making the goal space flexibly adaptable even after the initial tuning to ambient speech (see Sect. 7).

4 Learning by Babbling

The previous section described how we can derive a low-dimensional representation, the goal space, from raw speech sounds. The whole pathway, from articulatory parameters, over acoustic features, to a position in goal space, represents the *forward model* f of speech acquisition which maps articulatory parameters into a representation space in which they can be evaluated or compared to ambient speech⁶. Learning how to speak, then, can be defined as learning the inverse model g which reverses this process: g should estimate which motor action $\mathbf{q}$ is required in order to achieve a position $\mathbf{x}$ in goal space. The ultimate goal is that the inverse model proposes an appropriate motor action $\hat{\mathbf{q}} = g(\mathbf{x}^*)$ for each goal $\mathbf{x}^*$ of a set of goals X^* such that $f(\hat{\mathbf{q}})$ equals (or is very close to) $\mathbf{x}^*$.

⁶ The term forward model, here, is used in line with the goal babbling literature [29, 45, 64] to refer to the real forward model that is given by the world. In contrast to internal forward models which are typically acquired during the learning process, f here is assumed to remain fixed during the babbling process.

The inverse model is trained in a supervised way during babbling using newly explored action–outcome pairs. To implement the inverse model, various machine learning methods are applicable. However, it is important that the learner can be trained in an online fashion and that it well extrapolates to unseen data. Here, a Radial Basis Function (RBF) network is used which clusters the goal space with basis functions $\mathbf{c}_i$ which are associated to readout weights $\mathbf{u}_i$ that correspond to articulatory configurations. When queried for a specific goal space position $\mathbf{x}$, the inverse model returns an interpolation of acquired articulatory configurations depending on the distance of the desired goal space position to the I clusters of the inverse model:

$$\mathbf{q} = \sum_{i=1}^I h_i(\mathbf{x}) * \mathbf{u}_i. \quad (1)$$

Here, $h_i(\mathbf{x})$ is the activation of the i -th basis function when input $\mathbf{x}$ is presented to the network:

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

With this formula, basis function i has higher activation when $\mathbf{x}$ is closer to its center $\mathbf{c}_i$. The distance is scaled with r , which can be imagined as the radius of the basis functions (here, $r = 0.15$). Softmax is applied in Eq. (2) to improve extrapolation properties of the inverse model.

Before babbling, the inverse model is initialized with a single $(\mathbf{x}_{\text{home}}, \mathbf{q}_{\text{home}})$ pair which correspond to the default goal space position and the corresponding action command (e.g., /@/ in Sect. 5).

4.1 Babbling Cycle

The goal babbling algorithm used in this study is based on the skill babbling algorithm suggested in [45], which is an extension of the original goal babbling algorithm [29–31]. Learning proceeds in iterations. In each iteration, the system tries to achieve a number of B different targets around a selected target seed—this constitutes the exploration step. In the subsequent adaptation step, the collected experience is evaluated and used to update the inverse model.

Formally, in the exploration step, a new target seed $\mathbf{x}_{\text{seed}}$ is drawn from the goal space (see Sect. 4.3 for details on target selection). B targets $\mathbf{x}_b^*$ are generated around this seed by adding Gaussian-distributed noise with variance σ_{goal}^2 :

$$\mathbf{x}_b^* = \mathbf{x}_{\text{seed}} + \mathcal{N}(0, \sigma_{\text{goal}}^2), \quad b = 1 \dots B. \quad (3)$$

Then, the inverse model is queried to propose actions for achieving these targets:

Fig. 4 How the inverse model (top row) and the workspace model (bottom row) cluster the goal space during the course of learning in the vowel learning task

$$\hat{\mathbf{q}}_b = g(\mathbf{x}_b^*) \forall b. \quad (4)$$

The inverse model is deterministic and returns only an interpolation of actions that it has already learned about. Therefore, an important factor to drive the discovery of new actions is the addition of noise in the action parameters. This exploration noise can be modeled as Gaussian-distributed noise with variance σ_{act}^2 that is added independently to each dimension i of the estimated action parameters:

$$\mathbf{q}_{b,i} = \hat{\mathbf{q}}_{b,i} + \mathcal{N}(0, \sigma_{\text{act}}^2). \quad (5)$$

The variance term σ_{act}^2 determines the amplitude of the exploratory noise, and therefore, is referred to as exploratory noise amplitude in the following.

The explored actions are tried out by executing the existing forward model:

$$\mathbf{x}_b = f(\mathbf{q}_b) \forall b. \quad (6)$$

In this way, B new action–outcome pairs $(\mathbf{q}_b, \mathbf{x}_b)$ are explored in each iteration.

In the subsequent adaptation step, the explored pairs are used to update the inverse model parameters. In particular, basis function centers $\mathbf{c}_i$ are added and adjusted, and the corresponding readout weights $\mathbf{u}_i$ are updated. As an underlying mechanism to cluster the goal space with the basis function centers of the RBF network, various types of clustering algorithms are applicable. One clustering algorithm that has been shown previously to be useful for goal babbling [29] is Instantaneous Topological Map (ITM) [65], a variant of the Self-Organizing Map (SOM) which can cope with correlated inputs. Therefore, an ITM

is used here to keep track of the basis function centers, following the implementation used in previous goal babbling literature [29].

An example of how the basis functions of the inverse model cluster the goal space during the learning process is shown in Fig. 4 (top row). From left to right, the inverse model gradually extends to include new regions of the goal space, until finally the relevant part of the goal space is clustered with basis functions.

The readout weights $\mathbf{u}_i$ are updated via gradient descent to minimize the error between the motor command $\mathbf{q}_b$ that was used for exploration and the motor command as it would be currently estimated by the inverse model $\hat{\mathbf{q}}_b = g(\mathbf{x}_b)$. Thus, $\mathbf{u}_i$ are updated as follows:

$$\mathbf{u}_i^{\text{new}} = \mathbf{u}_i + \lambda \cdot w_b \cdot h_i(\mathbf{x}_b) \cdot (\mathbf{q}_b - \hat{\mathbf{q}}_b). \quad (7)$$

The learning rate λ determines the size of the update step (here, set to $\lambda = 0.9$). Important is the term w_b which refers to the weight of the explored action–outcome pair. Naturally, not all of the discovered pairs provide useful information for learning. As random noise is added to the articulatory parameters in Eq. (3), also not well articulated sounds are produced from which the system should not learn. Therefore, weights w_b are determined for each action–outcome pair which measure the usefulness of the produced sound using a number of objective criteria.

These weights are general quality measures for the explored sounds. For instance, infants would evaluate the speech sounds they produce according to their saliency (loud and clear sounds will be more interesting to explore than non-articulated sounds). For vowels, two weighting schemes are utilized, which measure:

- How close the discovered goal space position is to the desired goal space position (target weighting scheme),
- And the general saliency of the sound, i.e. its loudness (saliency weighting scheme).

A detailed description of these weighting schemes is provided in [32]⁷.

For syllables, also the target weighting scheme is applied. Additionally, a syllable weighting scheme is used which assesses the babbled sounds by comparing them to the original sounds provided by ambient speech. In particular, the distance between the default speech sound and the discovered speech sound is measured via Dynamic Time Warping (DTW) [66] of the absolute values of the speech signals. To speed up the computation, the speech signal is downsampled beforehand by a factor of 100. The DTW distance d_b of speech sound b to the default sound is computed, and the weights for all sounds of one batch are determined as $w_b = 1 - \frac{d_b}{\max(d_b)}$. Thus, this weight evaluates the similarity of the general loudness contour of the speech sounds and, therefore, constitutes a sophisticated version of the saliency weighting scheme that is used for vowels.

All of these weighting schemes return values between 0 and 1, where 1 marks better babbling examples. The weight w_b for an action–outcome pair is determined by multiplying the weights obtained by the individual weighting schemes. Babbled examples which have a low weight w_b will contribute little to learning. Additionally, a weight threshold of 0.1 is set such that for action–outcome pairs with low weights the ITM algorithm does not generate additional clusters.

In all experiments presented here, babbling continues for a maximum of 500 iterations with a batch size of 10 (i.e. the system babbles 10 sounds in each iteration). Learning stops earlier if the inverse model is able to achieve all relevant goals, i.e. when the system’s reproduction of a goal $\mathbf{x} = f(g(\mathbf{x}^*))$ is similar to the original goal $\mathbf{x}^*$ for all relevant goals in a set of goals X^* . As a threshold, a Euclidean distance of 0.05 in goal space is used.

4.2 Workspace Model

In the skill babbling framework [45], it is suggested to maintain a so-called workspace model, containing clusters in regions of the goal space that have already been successfully achieved. Knowing which regions of the goal space can be achieved is useful to estimate the competence of the system for different types of tasks (where a task is expressed

as a position in goal space) and can be used to decide where to explore next.

The workspace model, thus, clusters the goal space similarly to the ITM of the inverse model. However, whereas the inverse model should have a low weight threshold such that newly discovered goal space positions are quickly integrated into the inverse model, the workspace model should only cluster a region in goal space when it can be reached with a certain proficiency. Therefore, a higher weight threshold is used (experimentally, 0.5 for vowels, 0.3 for syllables were selected). If the weight w_b of the newly discovered goal space position $\mathbf{x}_b$ is above this threshold, a new prototype with center $\mathbf{c}_{i+1} = \mathbf{x}_b$ and radius $\gamma = 0.1$ is introduced if $\|\mathbf{x}_b - \mathbf{c}_i\| > \gamma \forall i \in [1 \dots I]$.

In Fig. 4 (bottom row), the development of the workspace model is displayed in parallel to the inverse model clustering. It can be seen that clustering in the workspace model grows more slowly than in the inverse model: while in iteration 150, /a/ is already covered by the inverse model, the workspace model does not yet cover this region. Thus, the model has not yet reached proficiency in reaching this sound. Due to the smaller radius, the workspace model more densely clusters around relevant regions of the goal space, i.e. at positions where speech sounds are located.

In this study, the workspace model is utilized for an active selection of targets (Sect. 4.3) and for adapting the articulatory noise amplitude during learning (Sect. 4.4).

4.3 Target Selection and Active Learning

In each babbling iteration, a target has to be selected for exploration. This target can be drawn randomly from the goal space, or actively considering competence progress [27] or the novelty of discovered regions in the goal space [45]. In the framework presented here, ambient speech is known, and per definition, clusters of speech sounds in the goal space correspond to speech sounds that the system should acquire, i.e. they are interesting for exploration. Targets, thus, can be drawn from the distribution of ambient speech in the goal space. Therefore, a target distribution is generated by fitting a Mixture of Gaussians model [67] via the k-means algorithm [68] on the ambient speech points in the goal space:

$$P(\mathbf{x}^*) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}^* | \mu_k, \Sigma_k), \quad (8)$$

where K is the number of clusters and π_k , μ_k and Σ_k are the prior probability, mean and covariance matrix for each target cluster. Here, the number of clusters was fixed according to the number of speech sounds in the ambient speech. Fig. 3 shows the covariances of the formed target clusters as ellipses.

⁷ In [32], the ambient speech weighting scheme was additionally used. This scheme is left out in the present experiments as it proved to be not necessary for learning.

From this distribution, targets can be drawn for exploration as $\mathbf{x}_{\text{seed}} \sim P(\mathbf{x}^*)$. In the following, two different exploration modes are discussed. *Random exploration* refers to drawing targets from $P(\mathbf{x}^*)$ with equal probabilities, i.e. $\pi_k = K^{-1} \forall k$ is set. *Active exploration* refers to drawing targets in a more sophisticated way, favoring targets which will provide more progress. Specifically, the probability π_k is adapted in each learning iteration based on the relative minimum distance of μ_k to the clusters $\mathbf{c}_i$ of the workspace model:

$$\pi_k = \frac{\min_j (||\mu_k - \mathbf{c}_j||)}{\sum_{l=0}^K \min_j (||\mu_l - \mathbf{c}_j||)}. \quad (9)$$

In this way, targets are more frequently drawn from regions which the system has not yet discovered. Thus, this strategy has the potential to speed up the learning process.

4.4 Adaptation of Exploration Noise

Generally, in goal babbling, exploratory noise with a fixed noise amplitude is used. However, while exploring, children are more flexible. For achieving a completely new sound, more variation in the exploration is required. In contrast, if an already discovered sound should be fine-tuned, it is better to reduce the amount of applied noise. To enable the system to adjust the amplitude of applied noise depending on which sound should be currently produced, we introduced an adaptive noise mechanism in [32]. By using information about already explored regions of the goal space, represented by the workspace model, articulatory noise can be adjusted according to how novel a task is for the system.

Similarly to how the prior probabilities for the target distribution are determined (cf. Eq. (9)), the distance of the current target seed $\mathbf{x}_{\text{seed}}$ to the workspace model (WSM) can be used to determine an appropriate noise amplitude:

$$\sigma_{\text{act}}^2 = \alpha \cdot (1 - \exp(-4 \cdot d_{\text{WSM}}(\mathbf{x}_{\text{seed}}))). \quad (10)$$

Here, α constitutes an upper threshold for the applied noise amplitude. The distance d_{WSM} to the current workspace model containing J clusters is determined as:

$$d_{\text{WSM}}(\mathbf{x}) = \frac{\min_j (||\mathbf{x} - \mathbf{c}_j||)}{\min_{k,l} (||\mu_k - \mu_l||)} \quad \forall j \in [1, \dots, J]. \quad (11)$$

The term in the denominator computes the smallest distance between every two cluster centers μ_k and μ_l ($k, l \in [1, \dots, J]$) of the target distribution. For example, in Fig. 3 (left), the two clusters that are closest to each other are /o/ and /u/. This distance acts as a normalization factor that ensures that the distance d_{WSM} only falls below 1 (resulting in a reduction of σ_{act}^2) when the distance to the closest workspace model cluster is smaller than the distance between /o/ and /u/. Thus,

even when /o/ is already acquired (i.e. a cluster is added to the workspace model located at /o/), the resulting σ_{act}^2 when exploring /u/ would still suffice to discover this vowel.

4.5 Evaluating Learning Progress

Learning progress is measured every 10 iterations by asking the system to reproduce the mean values μ_k of the target distribution (cf. Sect. 4.3). The desired goal space positions $\mathbf{x}^* = \mu_k$ are compared to the corresponding achieved goal space positions $\mathbf{x} = f(g(\mathbf{x}^*))$. The reproduction error is evaluated via Euclidean distance as:

$$\delta(\mathbf{x}, \mathbf{x}^*) = ||\mathbf{x} - \mathbf{x}^*||. \quad (12)$$

From these error values, a competence value for each learning cluster can be computed as suggested by [69]:

$$\text{comp}(\mathbf{x}, \mathbf{x}^*) = \exp(-\delta(\mathbf{x}, \mathbf{x}^*)). \quad (13)$$

This value lies between 0 and 1, where 1 corresponds to the highest competence.

5 Results: Vowel Acquisition with Static Motor Representation

In this section, it is demonstrated how the presented framework can be used for bootstrapping a set of vowel sounds, represented via static motor configurations. In particular, the five vowels /a/, /e/, /i/, /o/, and /u/ should be discovered.

Babbling starts from one known vowel sound which is used to initialize the inverse model. Specifically, the default position of the vocal tract simulator is used, $\mathbf{q}_{\text{home}} = \mathbf{q}_{/@/}$, which generates a sound that corresponds to the vowel *schwa* [70], here abbreviated as /@/. By executing the forward model, $\mathbf{x}_{\text{home}} = \mathbf{x}_{/@/} = f(\mathbf{q}_{/@/})$ is determined.

As introduced in the previous section, the learning process can follow different strategies. In this evaluation, the effect of two mechanisms on vowel acquisition is analyzed. First, targets are drawn during babbling either randomly, or using active exploration (cf. Sect. 4.3). Second, the exploration noise in the motor space is set either to a fixed amplitude of 0.55, or alternatively, exploration noise is adapted during the learning process with $\alpha = 0.55$ as the upper threshold of the noise amplitude (cf. Sect. 4.4)⁸. For each condition, 30 individual trials are performed.

⁸ Parameter choices adopted from our previous study [32].

Fig. 5 Competence of the individual vowels during the learning process, evaluated every 10 iterations. Mean (solid lines) and the 95% confidence interval (shaded areas) across 30 individual trials are displayed

5.1 Effect of Exploration and Noise Adaptation Strategy

The effect of the two mechanisms random vs. active exploration and fixed vs. adaptive noise on the acquisition of vowel sounds is displayed in Fig. 5. Learning succeeds in all conditions: the competence increases during the learning process and the majority of the babbled sounds after learning are comprehensible⁹. However, the highest competence for most vowels is achieved at the end of babbling when active exploration and an adaptive noise amplitude are applied. In contrast, using random exploration and a fixed noise amplitude leads to lower competence values and a larger variability of the results.

This finding indicates that both mechanisms facilitate the acquisition of vowel sounds. Our previous evaluation of the adaptive noise mechanism in [32, 51] suggested

that nonlinearities in the goal space are the reason why different noise amplitudes are required for different vowel sounds. This study demonstrates that as an alternative approach, active exploration can be used. In fact, both mechanisms follow a similar underlying idea: reducing the exploration of already discovered sounds. This idea

Fig. 6 Amplitude of articulatory noise over the course of learning when exploring randomly (left) or actively (right). The average applied noise value across ten subsequent iterations is displayed

⁹ See sound examples in GitHub repository in the folder https://github.com/aphilippsen/goalspeech/tree/master/data/results-data/vowels/formants_modelspace.

Fig. 7 How the trained models for vowels reproduce interpolated goal space positions between the /@/ sound and either of the five target vowels. The distance of the perceived goal space positions of the model's own reproductions to the target goal space position is shown

for the four different training conditions. All measures correspond to the mean square error (euclidean distance), and mean (solid lines) and the 95% confidence interval (shaded areas) across 30 individual trials per condition are displayed

can be either implemented by reducing the noise amplitude for acquired sounds, or by reducing the probability that these sounds are further explored. The results also demonstrate that both mechanisms can be used in conjunction with each other, which leads to a further improvement of the competence values.

Figure 6 shows how the adaptive noise amplitude is differently adapted in the random exploration (left) and in the active exploration (right) condition. The main difference between the two conditions is how quickly the noise decreases during learning. In the case of random exploration, the noise decreases in a different speed depending on the vowel sound. With active exploration condition, all vowels are acquired in parallel because more difficult vowel sounds are practiced more frequently. Thus, although both mechanisms lead to equally good performance, the acquisition proceeds in different ways.

5.2 Evaluating Smoothness and Linearity

The previous evaluation analyzed the performance of the model only at a number of specific target positions. But how do the trained models behave when queried outside of the desired target clusters? In particular, it is interesting to look at how transitions in the goal space between different vowel clusters are represented in the acquired models. If the model appropriately reflects the general ability to generate vowel sounds, it should be able to smoothly interpolate in the goal space between the acquired vowels (i.e. without sudden jumps). The motivation is that vowel sounds can continuously merge from one sound to another. For example, it is possible to gradually close the rounded lips in order to switch from an /o/ to a /u/ sound. Therefore, the capability of the trained models is tested to reproduce vowel sounds which lie in between target clusters. For this analysis, goal

Fig. 8 Competence of the individual syllables during the learning process, evaluated every ten iterations. Mean (solid lines) and the 95% confidence interval (shaded areas) across ten individual trials are displayed

space positions are linearly interpolated between the /@/ vowel and the five target vowels (in steps of 0.1). The trained system then produces the speech sounds corresponding to these goal space positions, and “perceives” its own productions: the sounds are mapped again to the goal space via the forward model. Figure 7 shows the distance of the “perceived” goal space positions to the goal space position of the target. Formally, it shows the distance $\|\mu_k - \hat{\mathbf{x}}_{ip}\|$ where μ_k is one target cluster mean and $\hat{\mathbf{x}}_{ip} = f(g(\mathbf{x}_{ip}))$, where $\mathbf{x}_{ip} = \beta \cdot \mu_k + (1 - \beta) \cdot \mu_{/@/}$ is the linear interpolation between the target cluster of /@/ and the target vowel cluster k with interpolation factor β .

The results show that in models trained with active exploration and noise adaptation a smoother transition between the /@/ cluster and the target cluster is acquired than in the other conditions. In the random–fixed condition, changes are not gradual and sometimes, sudden jumps occurred during the transition.

The transitions between vowels are mostly linear, however, also a small effect of nonlinearity can be observed: queried goal space positions which are closer to a target cluster are “perceived” as if they would be closer to the cluster than they are. This is particularly well visible in the active–adaptive condition when interpolating between /@/ and /i/. This effect is interesting because it can be related to categorical speech perception in humans (see Sect. 7 section for a discussion).

6 Results: Syllables Acquisition with Dynamic Motor Representation

The previous section limited learning to the acquisition of static motor configurations. However, a dynamic representation is required when syllable sounds should be acquired. In

this section, it is demonstrated that the presented framework can also acquire speech sounds which are represented with a dynamic motor representation. In particular, starting from the speech sound /aa/¹⁰, the syllables /baa/ and /maa/ are acquired.

Active exploration and exploratory noise adaptation is utilized, which produced the best results in Sect. 5. As exploratory noise is added to the DMP parameters, the maximum amount of noise was determined experimentally and set to $\alpha = 1$.

Figure 8 displays how the competence for /baa/ and /maa/ gradually increases during the babbling process.

How transitions between the acquired sounds are represented in the acquired models is evaluated analogously to the analysis in Sect. 5.2 by interpolating between /aa/ and the target clusters. The results are displayed in Fig. 9. The left graph shows the Euclidean distance measured between the target cluster and the reproduced goal space positions analogously to Fig. 7. The interpolated sounds smoothly change between /aa/ and the target sound, and the perception of the acquired model representation is nonlinear in the vicinity of the acquired syllables. In particular, all produced sounds with a $\beta \geq 0.8$ are perceived as equally close to the target syllable. This observation of nonlinearity can be confirmed when listening to the interpolated sounds¹¹. The perception of syllables, in particular of syllables which include consonant sounds, thus, is categorical, which resembles categorical perception of syllable sounds in human speech [71, 72].

The right graph of Fig. 9 displays the distances of interpolated sounds measured in the space of articulatory DMP configurations. Specifically, the distances are computed between the motor configurations generated by the inverse model for the interpolated goal space positions, and the articulatory configuration acquired for the target syllable. The Euclidean distance is measured between the normalized, flattened vectors of DMP parameters. It can be observed that there is some nonlinearity in the vicinity of /aa/, but not close to /baa/ or /maa/. This analysis indicates that not only the perception but also the production of speech may contribute to the categorical perception (see Sect. 7.1 for a discussion).

7 Discussion

This study proposes a framework for articulatory speech acquisition based on infant-inspired goal-directed exploration. The framework successfully bootstraps an inverse

¹⁰ The notation /aa/ is used instead of /a/ to indicate that the vowel is represented via DMP parameters and not as a single articulatory frame as in the previous section.

¹¹ See sound examples in GitHub repository in the folder https://github.com/aphilippsen/goalspeech/tree/master/data/results-data/syllables/mfcc_modelspace.

Fig. 9 How the trained models for syllables reproduce interpolated goal space positions. Left: distance of the perceived goal space positions of the model's own reproductions to the target goal space position. Right: distance of the DMP motor commands to the motor com-

mand generated for the target syllable. Mean (solid lines) and the 95% confidence interval (shaded areas) across 10 individual trials are displayed. All measures correspond to the mean square error of Euclidean distances

model for generating vowel and syllable sounds. Only a minimum amount of prior information is needed. Specifically, a set of target speech sounds is required for creating a low-dimensional embedding of ambient speech, and the initial vocal tract shape has to be known to initialize the inverse model. Both can be assumed to exist in a similar way in young infant when they are starting to babble approximately at the age of three months. Another source of prior knowledge in the framework are the weighting schemes which are used to evaluate the babbled sounds. To this end, the weighting schemes play a similar role like reward in other models of speech acquisition (e.g. [19, 21, 41]). The weighting schemes were designed to be as generic as possible: the target weighting measures the success of a produced sound based on the goal space representation and the saliency weighting or the syllable weighting (cf. Sect. 4.1) are based on parameters derived directly from ambient speech.

The remainder of this section discusses the findings of this study while drawing comparisons to human speech acquisition. Furthermore, the limitations of the current framework are discussed and future research directions are proposed.

7.1 Categorical Perception of Speech

Human speech perception is not linear but distorted by linguistic experience [73]. Patricia Kuhl described the tendency to perceive instances of a speech sound as if they were closer to the prototype than they physically are as the perceptual magnet effect [73, 74]. Furthermore, it is known that human speech perception is highly categorical [71, 73], i.e., oriented toward prototypes of phonetic categories. Such categorical perception is important for robust perception, for example, in the presence of environmental noise. Still, humans are also able to recognize continuous changes, in

particular, for isolated vowel sounds [71]. Similarly, Figs. 7 and 9 show that continuous changes can be produced and perceived by the system, but that nonlinear transitions occur, in particular, close to syllables that contain consonants.

Furthermore, Fig. 9 (left) suggests that nonlinearity is partially also present in the motor representations. In particular, around the starting syllable /aa/, the inverse model tends to propose articulatory configurations which are more prototypical, i.e. closer to the corresponding default motor configuration. This finding raises the question whether the perceptual magnet effect in human speech perception might also exist in the articulatory modality. Such an effect appears plausible considering the important role that motor representations seem to play for perception [11], and could be further investigated in future research.

Also, additional evaluations, available as supplemental material¹², indicate that the degree to which categorical perception occurs depends partially on the used acoustic feature representation. In particular, using Gabor filterbank features [75], an even stronger categorical effect has been found in goal space as well as in motor space.

7.2 Developmental Change During Learning

Developmental models commonly investigate whether the course of learning follows a trajectory that is similar to human learning. For example, a gradual increase of articulated in contrast to non-articulated sounds was found in [27]. In the present study, the developmental improvement can be observed as an increase of the competence during the babbling process.

¹² See in GitHub repository <https://github.com/aphilippsen/goalspeech/tree/master/data/results-data/result-overview.txt>.

Table 1 Percentage of 5000 randomly babbled sounds which are mapped to the vicinity of the vowel clusters in goal space

/a/ (%)	/e/ (%)	/i/ (%)	/o/ (%)	/u/ (%)	/@/ (%)
0.1	2.4	4.1	2.6	0.8	5.5

Another factor that can be looked at is the order in which different speech sounds are acquired. Regarding the order in which infants acquire different vowel sounds, little consistent experimental evidence is available. It can be assumed that high individual differences exist. The order might also depend on the frequency of the sounds in a specific language. Such differences were not modeled here (all sounds were equally likely), but could be tested in future studies.

Still, in the experiments, it was found that different vowel sounds are discovered in a specific order: The vowels /@/, /i/ and /e/ are acquired first, whereas /a/ and /u/ are acquired more toward the end of the learning. This order, in particular, the late discovery of /a/ might be considered surprising as, intuitively, children typically acquire this sound early. In infants, the reason might be the similarity of /a/ to pre-speech sounds such as crying. Thus, using /a/ instead of /@/ as an initial vowel might be more adequate for modeling the developmental process of human learning.

In which order the sounds are acquired in the present framework may depend on many factors such as the used acoustic features and the nature of the goal space mapping. In previous studies, we demonstrated that nonlinearities in the forward mapping can slow down the acquisition of particular vowels such as /u/ [32]. Additionally, the redundancy in the mapping plays a role: naturally, discovering new speech sounds by random exploration is easier when there are multiple different articulatory configurations for achieving this goal. Table 1 shows the probability that a randomly babbled speech sounds is perceived close to a target cluster (within the radius of 0.2). For this analysis, 5000 sounds were generated from random motor configurations (in the normalized articulatory space), and then mapped to the vowel goal space. When comparing the values in Table 1 to the order of discovery of different vowels, it can be seen that vowels which are achieved with higher probability usually are learned earlier by the system. Future work is required to decide where this bias originates from.

The presented framework may also be used to model specific phenomena known from infant development. In our previous study [76] we tested the unverified hypothesis from infant development that infant-directed speech is beneficial for articulatory learning. The study modeled the presence of infant-directed speech in the early stage of development by using speech sounds which were either strongly articulated (tense vowels) or not (lax vowels). The results demonstrated that the model is better able to acquire novel speech sounds

later on when the goal space was generated from strongly articulated speech sounds. Thus, infant-directed speech which is characterized by stronger articulatory effort might indeed be beneficial for infants to increase their flexibility to adapt to new speech sounds in the later development. Similarly, the presented framework could be applied to test or generate other hypotheses on infant’s developmental learning of speech in the future.

7.3 Limitations and Future Research Directions

The main feature of the presented framework is the goal space which is automatically generated from ambient speech. Despite the advantages of this procedure compared to a fixed goal space representation that were demonstrated in this study, in its current form, the goal space representation also has some limitations. One point that might be seen as a shortcoming of the goal space is that it cannot differentiate between two sounds that are projected to the same position in goal space. On the one hand, this property of the goal space is desired because the model’s perception becomes adjusted to ambient speech. Also, it is developmentally plausible as infants make similar mistakes early in their development [77]. On the other hand, infants eventually are able to overcome such problems while the model cannot adjust its goal space during learning. Therefore, an important aspect for future research is to create a more flexible notion of the goal space. For instance, the goal space could be adaptable during the learning process, either gradually or organized in stages. Furthermore, more research is required to understand how well the goal space’s “perception” matches human perception of speech.

A gradual development of the goal space from coarse to fine-grained perception of speech is not only developmentally plausible [78], but could also help to overcome another problem: while it is possible to learn a large amount of vowel sounds in a single goal space [79], the attempt to acquire a larger amount of syllable sounds in parallel impairs the performance [51]. Higher dimensional goal spaces or a step-wise approach to learning could remedy this shortcoming.

Alternatively, new methods for generating the goal space could be explored. The current dimension reduction method extracts linear properties of the underlying ambient speech. Syllable sounds require drastic changes in a high-dimensional articulatory space which are only partially captured by this dimension reduction technique. The amount of syllables that can be learned in parallel, therefore, is naturally limited. Using deep neural networks such as variational autoencoders for extracting features of the goal space could improve the framework in this aspect.

The developed framework is not restricted to a specific vocal tract model, but can be adjusted to use other vocal tract models. Therefore, it would be interesting to test

the framework with different vocal tract models, either in software or also in hardware (using vocal robots) where efficiency in exploration is particularly important.

Finally, many phenomena in speech learning involve the aspect of social interaction [80] and require learning about the meanings of the produced sounds. A conceptualization of such aspects in the presented framework, for instance, by using multimodal goal spaces [76], is an important challenge for future work that could further increase the potential of the framework to investigate aspects of infant learning.

Acknowledgements Many thanks go to Britta Wrede and Felix Reinhart for their valuable support and advice during the supervision of my PhD project, which forms the foundation for this study.

Funding Part of this research has been conducted at CITEC, Bielefeld University, with the support of the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) of Deutsche Forschungsgemeinschaft.

Data Availability Statement All data generated for the experiments is available from the author on reasonable request.

Compliance with Ethical Standards

Conflict of interest The author declares that there is no conflict of interest.

Code availability The Python3 source code for the presented framework is available at: <https://github.com/aphilippsen/goalspeech>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Vouloumanos A, Werker JF (2004) Tuned to the signal: the privileged status of speech for young infants. *Dev Sci* 7(3):270–276
2. Werker JF, Yeung HH (2005) Infant speech perception bootstraps word learning. *Trends Cogn Sci* 9(11):519–527
3. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A et al (2014) Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint. arXiv:1412.5567*
4. Pratap V, Hannun A, Xu Q, Cai J, Kahn J, Synnaeve G, Liptchinsky V, Collobert R (2019) Wav2letter++: a fast open-source speech recognition system. In: ICASSP 2019–2019 IEEE international conference on acoustics. Speech and signal processing (ICASSP), IEEE, pp 6460–6464
5. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, Yu D, Zweig G (2016) Achieving human parity in conversational speech recognition. *arXiv preprint. arXiv:1610.05256*
6. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G, et al (2016) Deep speech 2: end-to-end speech recognition in English and Mandarin. In: International conference on machine learning, pp 173–182
7. Hart B, Risley TR (2003) The early catastrophe: the 30 million word gap by age 3. *Am Educ* 27(1):4–9
8. Cristia A, Dupoux E, Gurven M, Stieglitz J (2019) Child-directed speech is infrequent in a forager-farmer population: a time allocation study. *Child Dev* 90(3):759–773
9. Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint. arXiv:1903.12261*
10. Mitra V, Franco H, Stern RM, Van Hout J, Ferrer L, Graciarena M, Wang W, Vergyri D, Alwan A, Hansen JH (2017) Robust features in deep-learning-based speech recognition. New era for robust speech recognition. Springer, Berlin, pp 187–217
11. Schwartz JL, Basirat A, Ménard L, Sato M (2012) The perception-for-action-control theory (PACT): a perceptuo-motor theory of speech perception. *J Neurolinguist* 25(5):336–354
12. Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. *Connect Sci* 15(4):151–190
13. Schmidhuber J (2006) Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connect Sci* 18(2):173–187
14. Asada M, Hosoda K, Kuniyoshi Y, Ishiguro H, Inui T, Yoshikawa Y, Ogino M, Yoshida C (2009) Cognitive developmental robotics: a survey. *IEEE Trans Auton Mental Dev* 1(1):12–34
15. Pagliarini S, Leblois A, Hinaut X (2020) Vocal imitation in sensorimotor learning models: a comparative review. *IEEE Trans Cogn Dev Syst*. <https://doi.org/10.1109/TCDS.2020.3041179>
16. Tourville JA, Guenther FH (2011) The DIVA model: a neural theory of speech acquisition and production. *Lang Cogn Process* 26(7):952–981
17. Kröger BJ, Kannampuzha J, Neuschaefer-Rube C (2009) Towards a neurocomputational model of speech production and perception. *Speech Commun* 51(9):793–809
18. Moulin-Frier C, Oudeyer PY (2012) Curiosity-driven phonetic learning. In: IEEE international conference on development and learning (ICDL-EpiRob)
19. Warlaumont AS (2013) Saliency-based reinforcement of a spiking neural network leads to increased syllable production. In: IEEE international conference on development and learning (ICDL-EpiRob), pp 1–7
20. Howard IS, Messum P (2011) Modeling the development of pronunciation in infant speech acquisition. *Motor Control* 15(1):85–117
21. Warlaumont AS (2012) A spiking neural network model of canonical babbling development. In: IEEE international conference on development and Learning (ICDL-EpiRob), pp 1–6
22. Meltzoff AN, Moore MK et al (1977) Imitation of facial and manual gestures by human neonates. *Science* 198(4312):75–78
23. Von Hofsten C (1982) Eye-hand coordination in the newborn. *Dev Psychol* 18(3):450
24. Konczak J, Borutta M, Topka H, Dichgans J (1995) The development of goal-directed reaching in infants: hand trajectory formation and joint torque control. *Exp Brain Res* 106(1):156–168
25. Craighero L, Leo I, Umiltà C, Simion F (2011) Newborns' preference for goal-directed actions. *Cognition* 120(1):26–32

26. Von Hofsten C (2004) An action perspective on motor development. *Trends Cogn Sci* 8(6):266–272
27. Moulin-Frier C, Nguyen SM, Oudeyer PY (2014) Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Front Psychol* 4:1006
28. Forestier S, Oudeyer PY (2017) A unified model of speech and tool use early development. In: 39th Annual conference of the cognitive science Society (CogSci (2017) Jul 2017. United Kingdom, London
29. Rolf M, Steil JJ, Gienger M (2010) Goal babbling permits direct learning of inverse kinematics. *IEEE Trans Auton Mental Dev* 2(3):216–229
30. Baranes A, Oudeyer PY (2010) Intrinsically motivated goal exploration for active motor learning in robots: a case study. In: International conference on intelligent robots and systems (IROS). IEEE/RSJ, IEEE, pp 1766–1773
31. Rolf M, Steil JJ, Gienger M (2011) Online goal babbling for rapid bootstrapping of inverse models in high dimensions. In: IEEE international conference on development and learning (ICDL-EpiRob)
32. Philippsen AK, Reinhart RF, Wrede B (2016) Goal babbling of acoustic-articulatory models with adaptive exploration noise. In: IEEE International conference on development and learning (ICDL-EpiRob)
33. DeCasper AJ, Spence MJ (1986) Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behav Dev* 9(2):133–150
34. Kisilevsky BS, Hains SM, Lee K, Xie X, Huang H, Ye HH, Zhang K, Wang Z (2003) Effects of experience on fetal voice recognition. *Psychol Sci* 14(3):220–224
35. Kuhl PK (2004) Early language acquisition: cracking the speech code. *Nat Rev Neurosci* 5(11):831–843
36. Birkholz P (2015) VocalTractLab—towards high-quality articulatory speech synthesis, used version: VocalTractLab 2.1 API for Linux (9 September 2014). <http://www.vocaltractlab.de/>. Accessed 20 Sept 2020
37. Birkholz P, Kröger BJ (2006) Vocal tract model adaptation using magnetic resonance imaging. In: 7th International seminar on speech production (ISSP'06), pp 493–500
38. Tsushima T, Takizawa O, Sasaki M, Shiraki S, Nishi K, Kohno M, Menyuk P, Best C (1994) Discrimination of English /r/ and /w/ by Japanese infants at 6–12 months: language-specific developmental changes in speech perception abilities. In: Third international conference on spoken language processing
39. Birkholz P, Kröger BJ, Neuschaefer-Rube C (2011) Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In: Interspeech, pp 2681–2684
40. Prom-on S, Birkholz P, Xu Y (2014) Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach. *EURASIP J Audio Speech Music Process* 1:23
41. Murakami M, Kröger B, Birkholz P, Triesch J (2015) Seeing [u] aids vocal learning: Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing. In: IEEE international conference on development and learning (ICDL-EpiRob), pp 208–213
42. Birkholz P (2013) Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS One* 8(4):e60–603
43. Schaal S (2006) Dynamic movement primitives: a framework for motor control in humans and humanoid robotics. *Adaptive motion of animals and machines*. Springer, Berlin, pp 261–280
44. Kulvicius T, Ning K, Tamosiunaite M, Wörgötter F (2012) Joining movement sequences: modified dynamic movement primitives for robotics applications exemplified on handwriting. *IEEE Trans Robot* 28(1):145–157
45. Reinhart RF (2016) Autonomous exploration of motor skills by skill babbling. *Auton Robots*. <https://doi.org/10.1007/s10514-016-9613-x>
46. Kröger B (2017) Speech acquisition: development of a mental syllabary. <http://www.phonetik.phoniatrie.rwth-aachen.de/bkroeger/research.htm>. Accessed 10 Oct 2017
47. Trehub SE (1976) The discrimination of foreign speech contrasts by infants and adults. *Child Dev* 47:466–472
48. Best CC, McRoberts GW (2003) Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Lang Speech* 46(2–3):183–216
49. Nehaniv CL, Dautenhahn K et al (2002) The correspondence problem. Imitation in animals and artifacts, vol 41. MIT Press, Cambridge
50. Messum P, Howard IS (2015) Creating the cognitive form of phonological units: the speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. *J Phon* 53:125–140
51. Philippsen AK (2018) Learning how to speak. Goal space exploration for articulatory skill acquisition. Dissertation, Bielefeld University
52. Westermann G, Miranda ER (2004) A new model of sensorimotor coupling in the development of speech. *Brain Lang* 89(2):393–400
53. Boersma P et al (2002) Praat, a system for doing phonetics by computer. *Glott Int* 5:341–345
54. Sahidullah M, Saha G (2012) Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech Commun* 54(4):543–565
55. Lyons J, et al (2020) Speech features library. Used version: 0.6. Zenodo. <https://doi.org/10.5281/zenodo.3607820>. https://github.com/jameslyons/python_speech_features. Accessed 3 May 2020
56. Chen H, Tang F, Tino P, Yao X (2013) Model-based kernel for efficient time series analysis. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 392–400
57. Aswolinskiy W, Reinhart RF, Steil JJ. Impact of regularization on the model space for time series classification. In: Machine learning reports, pp 49–56
58. Jaeger H (2001) The “echo state” approach to analysing and training recurrent neural networks—with an erratum note. *Bonn Ger Ger Natl Res Center Inf Technol GMD Tech Rep* 148(34):13
59. Philippsen AK, Reinhart RF, Wrede B (2014) Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In: IEEE international conference on development and learning (ICDL-EpiRob), IEEE, pp 195–200
60. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemom Intell Lab Syst* 2(1–3):37–52
61. Xanthopoulos P, Pardalos PM, Trafalis TB (2013) Linear discriminant analysis. *Robust data mining*. Springer, Berlin, pp 27–33
62. Shi R, Werker JF, Morgan JL (1999) Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition* 72(2):B11–B21
63. Werker JF, Tees RC (2002) Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behav Dev* 25(1):121–133
64. Rolf M (2013) Goal babbling with unknown ranges: A direction-sampling approach. In: IEEE international conference on development and learning (ICDL-EpiRob)
65. Jockusch J, Ritter H (1999) An instantaneous topological mapping model for correlated stimuli. In: International joint conference on neural networks (IJCNN), IEEE, vol 1, pp 529–534
66. Salvador S, Chan P (2007) Toward accurate dynamic time warping in linear time and space. *Intell Data Anal* 11(5):561–580

67. Calinon S, Guenter F, Billard A (2006) On learning the statistical representation of a task and generalizing it to various contexts. In: IEEE international conference on robotics and automation, pp 2978–2983
68. Hersch M, Guenter F, Calinon S, Billard A (2008) Dynamical system modulation for robot learning via kinesthetic demonstrations. *IEEE Trans Robot* 24(6):1463–1467
69. Moulin-Frier C, Oudeyer PY (2013) Exploration strategies in developmental robotics: a unified probabilistic framework. In: IEEE international conference on development and learning (ICDL-EpiRob), pp 1–6
70. Flemming E (2009) The phonetics of schwa vowels. *Phonological weakness in english*. MIT Press, Cambridge, pp 78–95
71. Repp BH (1984) Categorical perception: Issues, methods, findings. *Speech Lang Adv Basic Res Pract* 10:243–335
72. Schouten M, van Hessen AJ (1992) Modeling phoneme perception: categorical perception. *J Acoust Soc Am* 92(4):1841–1855
73. Kuhl PK, Iverson P (1995) Linguistic experience and the “perceptual magnet effect”. *Speech perception and linguistic experience*. York Press, New york, pp 121–154
74. Kuhl PK (1991) Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Percept Psychophys* 50(2):93–107
75. Schädler MR, Meyer BT, Kollmeier B (2012) Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *J Acoust Soc Am* 131(5):4134–4151
76. Philippsen A, Wrede B (2017) Towards multimodal perception and semantic understanding in a developmental model of speech acquisition. In: Workshop on language learning at IEEE international conference on development and learning (ICDL-EpiRob)
77. Locke JL (1980) The prediction of child speech errors: implications for a theory of acquisition. *Child Phonology*. Elsevier, Amsterdam, pp 193–209
78. Dobson V, Teller DY (1978) Visual acuity in human infants: a review and comparison of behavioral and electrophysiological studies. *Vis Res* 18(11):1469–1483
79. Philippsen A, Reinhart F, Wrede B, Wagner P (2017) Hyperarticulation aids learning of new vowels in a developmental speech acquisition model. In: IEEE international joint conference on neural networks (IJCNN)
80. Kuhl PK (2007) Is speech learning “gated” by the social brain? *Dev Sci* 10(1):110–120