EPJ Nonlinear Biomedical Physics
a SpringerOpen Journal

RESEARCH Open Access

Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception

Bernd J Kröger[1,2*], Jim Kannampuzha[1] and Emily Kaufmann[3]

* Correspondence:
bernd.kroeger@rwth-aachen.de
[1]Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany
[2]Cognitive Computation and Applications Laboratory, School of Computer Science and Technology, Tianjin University, Tianjin, China
Full list of author information is available at the end of the article

Abstract

Background: Quantitative neural models of speech acquisition and speech processing are rare.

Methods: In this paper, we describe a neural model for simulating speech acquisition, speech production, and speech perception. The model is based on two important neural features: associative learning and self-organization. The model describes an SOM-based approach to speech acquisition, i.e. how speech knowledge and speaking skills are learned and stored in the context of self-organizing maps (SOMs).

Results: The model elucidates that phonetic features, such as high-low, front-back in the case of vowels, place and manner or articulation in the case of consonants and stressed vs. unstressed for syllables, result from the ordering of syllabic states at the level of a supramodal phonetic self-organizing map. After learning, the speech production and speech perception of speech items results from the co-activation of neural states within different cognitive and sensorimotor neural maps.

Conclusion: This quantitative model gives an intuitive understanding of basic neurobiological principles from the viewpoint of speech acquisition and speech processing.

Keywords: Speech production; Speech perception; Speech acquisition; Babbling; Imitation; Associative learning; Self-organization; Neural maps; Self-organizing maps; Sensorimotor learning

Background

While a great deal of research has been carried out in order to investigate brain locations of different parts or modules which comprise the speech production and speech perception system (e.g. [1-3]), little is known about the *neural functioning* of these modules during speech acquisition, speech production, and speech perception. In order to fill this gap, quantitative functional neural models have been developed (e.g. [4-8]).

One model, the neuroanatomically grounded Hebbian-learning model [8], establishes highly specialized functional units called "Hebbian neuronal circuits" (HNCs, see also [9]). This model appears to be especially neurobiologically realistic since it learns to associate sensory and motor speech items in a similar way to the early phases of speech acquisition in children. In order to maintain balance between neurobiological realism

and computational tractability, this approach does not model single neurons and spike chains, but rather uses "cells" or "nodes" as basic neuron-like elements which represent a local set of neurons, and thus it realizes a lumped-type or mean-field type model in which the primary objects of modeling are the average activity rate of the neuron set (or cell) and cell exhibitory and inhibitory connectivity.

Self-organizing map approaches (SOM, Kohonen) belong to the group of lumped element rate based approaches as well, but it should be noted that the degree of abstraction is much higher in SOM models than in models such as the neuroanatomically grounded Hebbian-learning model of [8]. On the other hand, SOM approaches – as well as more neuroantomically grounded approaches – are capable of representing the basic principles of neural systems, i.e. self-organization, associative learning, Hebbian learning, adaptation, and neural plasticity.

Quantitative neural models of speech processing (i.e. speech production and speech perception) and speech acquisition which include the generation and/or perceptual processing of articulatory and acoustic speech signals are rare. One of the most cited approaches in this direction is the DIVA model [7,10,11]. The DIVA approach mainly concentrates on modeling the relationship between sensory feedback and speech articulation. That model has been successfully applied e.g. to exemplifying motor adaptation in speech production [7,10,12]. The approach introduced in the present paper concentrates on the questions of how speech knowledge, including knowledge concerning speech motor skills, is learned and how this knowledge is stored. In contrast, no assumptions concerning knowledge or skill storage are given in the DIVA approach. Thus, the goal of the present paper is to introduce a comprehensive model of speech acquisition, speech production, and speech perception, based on SOM theory, which includes knowledge and skill storage.

Methods

Structure of the model

Biologically-based neural models that describes complex behavior or complex human functions, such as speaking, separate functional structure and knowledge [13]. The *functional structure* of such a system is basically composed of neural maps and neural mappings [7].

A *neural map* is an assembly of model neurons which represents a specific *neural state*, i.e. a phonemic, phonetic, motor plan, or sensory state in the case of our modeling approach. These maps are located in specific cortical regions. A neural map, e.g. neural map A, comprises N_i model neuron n_i ($i = 1, \ldots, N_i$). Each of these model neurons may be activated to a certain degree $a_i(t)$ at each time instant t. The whole activation pattern $a_i(t)$ ($i = 1, \ldots, N_i$) of a neural map represents a specific neural state, e.g. a motor plan, a sensory state, or a phonemic state at a certain time instant. The strength of activation of each model neuron varies between zero (0, no activation) and one (1, full activation).

All model neurons n_i ($i = 1, \ldots, N_i$) of the neural map A and all neurons n_j ($j = 1, \ldots, N_j$) of the neural map B can be connected with each other (Figure 1). The entirety of N_i x N_j *neural links* or *neural connections* between a neural map A and B is called a *neural mapping*. The strength (or connectivity) of each neural link is called *synaptic*

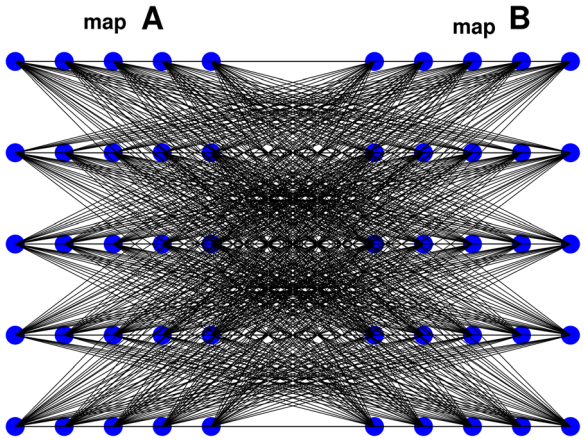

Figure 1 Neural map A, B and neural mapping between the two maps. Map **A** comprises 5 x 5 model neurons ($N_i = 25$); map **B** comprises the same number of model neurons ($N_j = 25$); neural mapping (black lines) comprises $25 \times 25 = 625$ synaptic connections.

link weight w_{ij}. Like neuron activity, each synaptic link weight is typically quantified on a scale between 0 and 1, where 0 represents no connection and 1 represents maximal exhibitory connection.

Additionally, neurons *within* map A or *within* map B can be interconnected. But in our modeling approach, this occurs only in the case of the self-organizing phonetic map and is included within the learning equation for the development of that map.

Only partial speech *knowledge* exists at birth or even prenatally [14]. Especially speech motor skills must be acquired step by step during the first years of life (ibid.). In our approach, speech knowledge is stored by *adjusting* (or changing) the *synaptic link weights* (w_{ij}). The way our neural model acquires speech knowledge is described in detail in Results Section.

In contrast to contemporary spiking neuron approaches for modeling brain functions (e.g. [15,16]), the *model neurons* defined in our approach represent an *ensemble* of natural cortical neurons which are spatially and functionally closely connected. Each model neuron may represent, for example, a cortical column. (This concept is used mainly in vision, e.g. [17,18]) Thus, the activation of a model neuron, and the forwarding of that activation by the axon of that neuron, (i) is the average activation of a bundle of natural neurons and (ii) the average over a specific time interval, i.e. the time interval of syllable processing in the case of our approach. This kind of averaging is widely used in approaches to modeling higher-level brain functions e.g. for *self-organizing maps* [19-21] and working memory [17]. These types of neural models can be summarized as *activation rate models*. In contrast to spiking neuron models, activation rate models do not focus on neurophysiological details such as neural spike trains of individual neurons. Rather, the simplification on the "microscopic" neural level allows us to model large-scale and higher-level brain functions and thus allows us to model "macroscopic" behavior (e.g. speech learning and speech processing) on the basis of neurofunctional principles.

The most important neurofunctional principle used in our modeling approach is *Hebbian learning*, i.e. a synaptic link between two model neurons is strengthened (w_{ij} increases over time) if both neurons are activated during the same time interval. A further neurofunctional principle, neural self-organization, results from Hebbian learning and is described in detail in Results Section.

Basic components of the model

In the present paper we will focus on the description of the higher-level aspects (cortical aspects) of speech processing, while the lower-level aspects (subcortical and peripheral processing) implemented in our model have already been discussed in other papers [22-24]. Figure 2 gives an overview concerning the cortical neural maps in our model. The blue colored boxes represent *state maps* (i.e. motor plan, phonemic, auditory, and somatosensory state map). These neural maps are assumed to be part of short-term memory because the neural states occurring in these maps – e.g. a *motor plan, a phonemic, or a sensory state* of a specific speech item – are activated only during a specific time interval, i.e. the time interval in which that speech item is processed by the model (e.g. up to a few hundred milliseconds for a syllable). Auditory and somatosensory states are summarized in this paper as sensory states. The green colored map (self-organizing phonetic map) and the neural mappings between the state maps and the phonetic map (green colored mappings) are assumed to be part of long-term memory. Here, the motor plan as well as sensory information is stored for each frequent syllable of the language being acquired (the target language).

The motor plan map and the sensory maps are also connected via the *sensorimotor processing loop* (Figure 2). The feedforward part of this loop (*execution feedforward pathway*) generates speech articulator movements at the level of the articulatory-acoustic vocal tract model on the basis of currently activated motor plan states for a syllable, word or utterance. (These are also referred to as "vocal tract action scores" in [24-26]). The vocal tract model generates geometrical data (i.e. a vocal tract shape and a set of articulator positions) for each time instant during the production of each

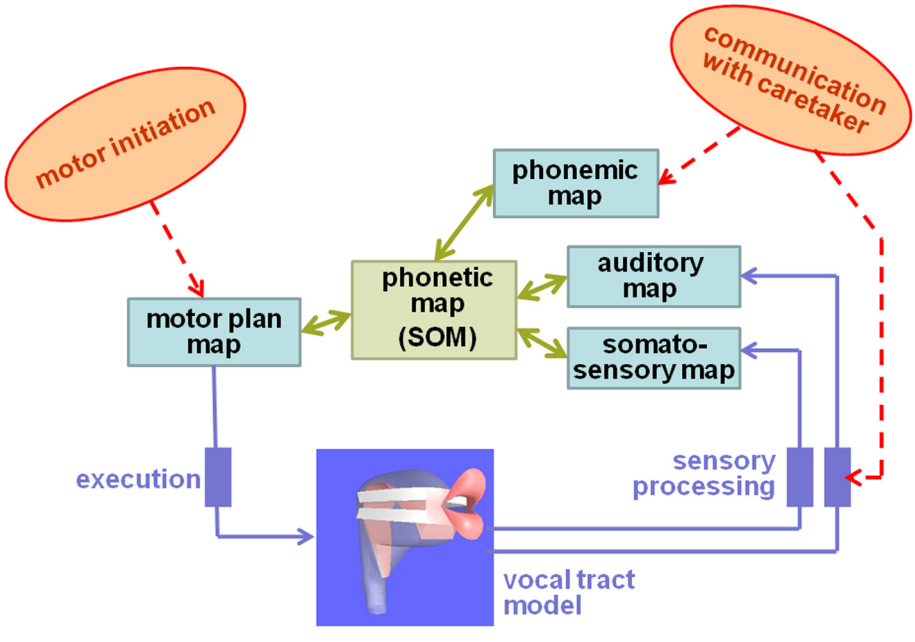

Figure 2 The structure of our neural model for simulating speech acquisition, speech production, and speech perception. Light blue and light green boxes represent cortical neural maps, while green double arrows indicate cortical neural mappings (see text). Dark blue arrows and boxes represent the cortical, subcortical and peripheral sensorimotor processing loops (see text). Red ovals represent babbling and imitation training sets which control the model during speech acquisition for implementing knowledge.

speech item as well as an acoustic speech signal for the speech item being produced [26,27]. Both signals are processed within the *sensory feedback pathway* or *sensory processing pathway* in order to generate an appropriate auditory and somatosensory neural state for the speech item being produced.

It should be noted that a mental lexicon (as defined in [28,29]) is not included in our model and lies beyond the scope of our modeling approach. The only symbolic linguistic representation used in our approach is the *phonemic representation* (see Figure 2). Here it is assumed that each frequent syllable occurring in the target language is represented by one model neuron within the phonemic state map.

Cortical state maps and neural representations

Motor plan, auditory, and somatosensory state maps are cortical neural maps which represent higher-level unimodal (i.e. motor or sensory) representations for a currently processed speech item (in most cases a syllable).

The existence of a *higher-level motor representation* (i.e. a *motor plan state*) is postulated in our approach on the basis of [2,30,31]. At this higher motor level, the overall temporal arrangement of the speech actions, which constitute a speech item, is specified, while the concrete muscle activation patterns are generated at lower cortical and subcortical levels. This (higher-level) motor plan state representation is part of short-term memory because this representation comprises at least the motor organization of a complete syllable, the duration of which may be up to several hundred milliseconds.

An example of a neural representation of a motor plan state is given in Figure 3. Eight vocal tract actions are needed in order to produce the syllable "ist" (Standard German "is"). Up to three tiers (i.e. horizontal rows) represent up to three onset consonant actions for each syllable. Up to two neuron rows represent the vocalic syllable nucleus; both vocalic rows are needed e.g. for representing a diphthong in German. Up to three neuron rows represent the coda consonant actions because up to three initial and final consonants may occur in Standard German. Two neuron rows represent velopharyngeal opening and closing actions, which are needed for the production of nasals and obstruents (plosives and fricatives). Two neuron rows represent glottal opening and glottal closing actions, which are needed for the production of voiceless and voiced

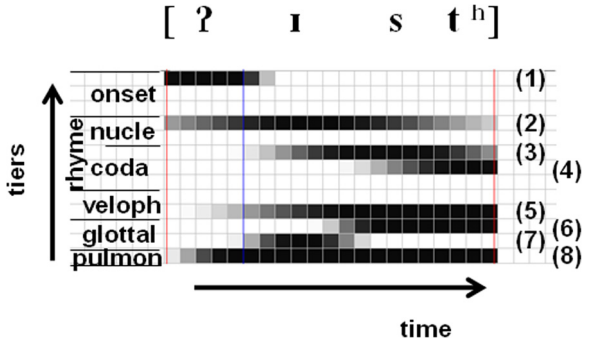

Figure 3 Neural representation of a motor plan state for the German word "ist" ("is"). Top line: phonetic transcription of the word. Below: a 21 x 13 model neuron map (motor plan state map) representing the motor plan of that word/syllable. Vertical columns: 12.5 ms time intervals; horizontal rows: temporal location and duration of each vocal tract action. The degree of activation (white = no activation; to black = full activation) of a model neuron represents the degree to which an articulator has approached the appropriate vocal tract action target.

speech sounds, respectively. The last neuron row represents the pulmonary pressure action which is needed as a basic power source for producing a speech item. The example syllable "ist" comprises eight vocal tract actions: (1) glottal tight-closing action for producing the initial glottal stop consonant [?]; (2) vocalic action for producing the German lax vowel [I]; (3) apical near-closing action for producing the fricative sound [s]; (4) apical full-closing action for producing the plosive sound [t]; (5) velopharyngeal tight-closing action in order to ensure a pressure build-up in the oral cavity during the production of the obstruent sounds [s] and [t]; (6) glottal opening action in order to ensure voicelessness during the production of [s] and [t]; (7) glottal closing action in order to ensure voicing during the production of the vowel [I]; (8) pulmonary pressure action in order to provide sufficient aerodynamic power for the production of the syllable. The neural representation for a further specification of each vocal tract action (e.g. articulator and target specification) is not shown here. A detailed description of our concept of vocal tract action scores (also called "gesture scores") is given in [25,32,33].

On the basis of this neural representation of a motor plan state (Figure 3), articulator movements controlling the vocal tract model can be calculated [25]. This represents the execution of a vocal tract action score by our vocal tract model. The neural representation of a motor plan state as introduced above is a *distributed neural representation*, because in principle all model neurons of the motor plan state map can contribute to the representation of a motor plan state. In contrast, the phonemic state map only comprises *local neural representations*, because here each syllable is represented by a specific neuron within the phonemic state map.

Evidence for a *higher-level auditory representation* is given by [34]. At this level, the sound impression of a whole syllable can be represented (i.e. how a speech item sounds). In our approach the higher-level auditory representation is assumed to be a "neural spectrogram" (see Figure 4). This auditory state representation is part of short-

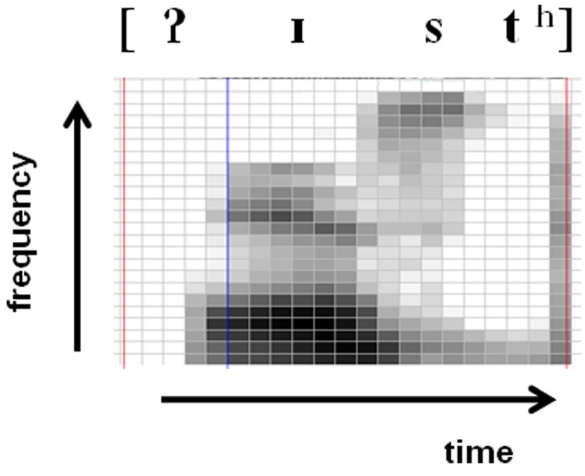

Figure 4 Neural representation of an auditory state within the auditory state map for the German word "ist" ("is"). Top line: phonetic transcription of the word. Below: a 21 x 24 neuron map (auditory state map) representing the auditory state of the word/syllable in the format of a Bark-scaled spectrogram. Horizontal rows: frequency in Bark groups from 1 to 24 (0 to 20000 Hz). Vertical columns: 12.5 ms time intervals. The degree of activation of a model neuron represents the acoustic intensity within that frequency-time-square. The blue vertical line represents the start time of the syllable nucleus, while the two red vertical lines represent the beginning and end of the syllable. (See also the motor plan state map, Figure 3.)

term memory because it comprises the auditory sound of a complete speech item (at least one syllable) and it must be activated over the whole time period of the production of at least one syllable in order to be able to check whether the syllable is produced correctly.

The calculation of the neural representation of an auditory state (i.e. auditory processing as a part of sensory processing, see Figure 2) is done by applying a spectral analysis (amplitude part of a Fourier analysis from 0 to 20000 Hz) to the acoustic speech signal (Figure 4). A 25 ms Hamming window is used. Thus, the spectral analysis of adjacent columns of model neurons overlaps in time. Each Bark group neural excitation is estimated by calculating the mean amplitude of all frequency bands occurring within that specific Bark group. The degree of neural activation (i.e. from 0 or white = no activation; to 1 or black = full activation) is proportional to the logarithm of spectral amplitude within that Bark group.

The existence of a *higher-level somatosensory representation* is postulated e.g. by [10,35]. Each somatosensory representation comprises a tactile and a proprioceptive aspect. A tactile higher-level representation in speech comprises at least the spatio-temporal pattern of contact between articulators (i.e. between the upper and lower lips and between the tongue and the palate). A higher-level proprioceptive representation comprises at least the spatio-temporal pattern of distance between articulators and intended targets for each vocal tract action and thus is comparable to a higher-level motor plan representation as defined in the present paper. Thus, for this preliminary version of our modeling approach we used the motor plan state representation also as a rough estimate for a higher-level somatosensory state representation. However, for a further refinement of our model, it would be possible to use geometrical data from the vocal tract model in order to estimate a detailed somatosensory representation (cf. [36,37]).

The existence of a *phonemic state representation* is postulated e.g. by [10], referred to as a "speech sound map" in that paper. In contrast to higher-level motor and sensory representations, which were implemented in our modeling approach as *distributed* neural representations, on the level of the *phonemic state representation*, each model neuron is assumed to represent exactly one phonemic state (i.e. a specific syllable or word; cf. [ibid.]). Thus, if a specific syllable is processed (produced or perceived), one model neuron becomes maximally activated at the level of the phonemic map. This is called a *local neural representation*.

Because speech processing is an ongoing flow of production and/or perception of syllables, it is assumed in our modeling approach that the activation of syllable-related model neurons within the phonemic map may overlap in time in order to allow the simulation of the succession of syllables over time.

Because motor plan and sensory state maps must be capable of representing all syllables (or at least all frequently occurring syllables) of a target language, the total number of model neurons n_i constituting the motor plan is set to $60 \times 13 = 780$ neurons, while the total number of model neurons composing the sensory state map is set to $60 \times 24 = 1440$ neurons. This allows for the modeling of a maximum syllable length of 750 ms, which occurs only if syllables are uttered in isolation. Typically, syllable length is much shorter, which – in our model representations – leads to many non-activated neurons at the temporal beginning and end of motor plan and sensory representations.

These non-activated neurons could be used for the representation of previous or subsequent syllables in a further version of our model. The beginning time of the syllable nucleus (i.e. the time of release of the last consonant within the syllable onset) is always set as a reference time and thus is always represented by the same neurons in the motor plan as well as in the sensory state maps (represented by the vertical blue line in Figure 3 and Figure 4). It should be noted that in addition to the required model neurons just described, 132 additional model neurons are needed at the level of the motor plan map for specifying each vocal tract action (for more detailed information see [24,25]). The phonemic map is of a comparably small size because the number of model neurons within this map directly reflects the number of syllables which occur in the target language at the phonemic level.

The phonetic map

The phonetic map in our modeling approach is a self-organizing map (SOM; for an introduction to SOMs, see e.g. [21]). The size of a self-organizing map increases during learning, which in our case means during speech acquisition. It results in an increasing amount of neural storage for speech knowledge within the mapping between the SOM and the state maps (see Growing-SOM approaches, e.g. [5,38]). In this preliminary version of modeling speech acquisition, we use phonetic maps of a fixed size, e.g. comprising 15×15, 20×20, or 25×25 model neurons.

It will be shown that phonetic maps – as well as self-organizing maps in general – exhibit *local* neural representations. We will see that a model neuron within a phonetic map represents a *specific phonetic realization* or *"exemplar"* of a phonemic state. This may be a specific phonetic realization of a syllable due to different contexts in which that syllable occurs. Thus, a model neuron within the phonetic map, as in the phonemic map, represents a syllable, but in the case of the phonetic map, it represents a concrete phonetic realization of a syllable. Thus, it is easy to see why a model neuron within the phonetic map must be directly connected with a motor plan and a sensory state (see Figure 2). This results from the fact that a fully activated model neuron at the level of the phonetic map (representing a specific phonetic realization of a syllable) directly activates the appropriate motor plan and sensory state for the realization of a syllable. Thus, the information concerning a motor plan and its appropriate sensory (auditory and somatosensory) states is stored completely within the synaptic link weights of the mappings between the phonetic map and the sensory state maps.

Modeling speech learning: babbling and imitation

The organization of the phonetic map is based on the link weight values of the neural mappings between the phonetic map and the sensorimotor state maps. These link weight values result from learning or training and directly reflect the acquired knowledge and the acquired speech skills. In our modeling approach it is assumed that the phonetic map is a self-organizing map (SOM; also called Kohonen-Map). Further, our approach assumes that the adjustment of synaptic link weights between the SOM and the state maps is established mainly within the early stages of speech acquisition, meaning in the first years of life, but that it can be modified and further developed over the lifetime.

Overview: babbling and imitation

It is well known that babies start to imitate facial expressions and communicative gestures within the first year of life (e.g. [39,40]). In parallel, they start to *imitate speech items*, which they hear from a caretaker, relatively early [14]. Later on, if a toddler is already capable of communicating with the caretaker, the toddler actively asks for specific words, e.g. by pointing at an object, and then looking to the caretaker (e.g. [41,42]). At this stage of learning to speak, the toddler receives (i) an acoustic realization of a speech item uttered by the caretaker and (ii) information concerning the meaning of the speech item. The first point is modeled in our approach by activating the auditory realization, while the second point is modeled by synchronously activating the phonemic representation for the relevant speech item (see red dashed arrows in Figure 2 in the area of "communication with caretaker"). Thus, the caretaker produces an acoustic realization which activates an auditory state within the auditory state map of the toddler (i.e. of the model). This auditory state can now be imitated by the toddler in order to generate an appropriate motor plan state.

But how does the toddler imitate a speech item produced by the caretaker? There are two major problems. Firstly, at the beginning of the learning procedure, the toddler does not know how to generate a motor plan state which could result in an auditory state similar to that produced by the caretaker. Secondly, the toddler does not possess the adult vocal tract and therefore cannot imitate the fundamental frequency or formant structures of a speech item produced by an adult: a toddler normally produces higher formants as well as higher fundamental frequency. This second problem is referred to as the "speaker normalization problem" (e.g. [43]).

These problems are not addressed in this modeling study for several reasons. Firstly, they may be solved in part by the caretaker, since caretakers are able to use a specific child-directed way of speaking ("motherese", see e.g. [14]). This way of speaking normally involves the adult trying to adapt the formant frequency and fundamental frequency targets of their speech sounds towards those targets which can be produced by the child. Additionally, the problem may be solved in part by the toddler herself: Even if the auditory result produced by the toddler differs from the original that was produced by the caretaker, the caretaker can reward the toddler for each correct or at least understandable realization of a word ("reinforcement learning"). Thus, the toddler is able to establish an initial association between his own realizations and adult realizations of the same word (e.g. the model stores two or more realizations for each speech item; one or more child versions and one or more adult versions).

A further, theoretically more important solution for these problems is that before the imitation state, the toddler (or the neural model) should already have some experience or knowledge of how to generate a motor plan state for a given auditory state. This knowledge is normally acquired by the toddler during the "babbling" phase [7,10,44]: The toddler (or the model) starts with randomly generated motor plan states (see red dashed arrow in Figure 2, "motor initiation"), then executes these motor plans (i.e. generation of an articulatory and acoustic speech signal by the vocal tract model), and finally performs the sensory processing of these signals, so that the appropriate auditory and somatosensory states are generated for all motor plan states (see Figure 2). Thus the model develops a set of "sensorimotor associations" by "exploring the acoustic and articulatory states of its own vocal tract". This set of associations is stored at the level

of the phonetic map and its mappings towards the motor plan and sensory maps. This learning or training is called babbling training. Babbling always precedes imitation, but the two can also be somewhat interwoven due to the reasons outlined in this section. Clear language-specific speech production starts at around 10 months of age). Indeed, the importance of babbling training in speech acquisition, particularly the sensory processing of the signals produced in babbling, is made clear by the fact that hearing children progress from babbling sounds to babbling syllables, while deaf children do not progress to the syllable stage (e.g. [45]).

Adjustment of synaptic link weights

Next we will describe how the adjustment of synaptic link weights between the state maps and the self-organizing phonetic map takes place in our model during training (learning). We assume that a *training data set* comprises an amount of D *training items*, i.e. D state representations d (d = 1, ..., D) with a model neuron activation pattern a_{id} (i = 1, ..., N_i), comprising a motor plan state, sensory state, and a phonemic state activation pattern, and that each training item is applied to the model C times (C = *number of training cycles*) where the succession of training items varies randomly per training cycle d(t). The *synaptic link weights* w_{ij} between *state map neurons* n_i and *phonetic map neurons* n_j were updated (i.e. were changed incrementally) from training step t to training step t + 1 (T = C*D, t = 1, ..., T training steps in total) by applying the following equation:

$$w_{ij}(t+1) - w_{ij}(t) = H_j(t)^*L(t)^*\left(a_{id}(t) - w_{ij}(t)\right), \tag{1}$$

where i = 1, ..., N_i is the index for all state map model neurons and where j = 1, ..., N_j is the index for all phonetic map model neurons. Here, $H_j(t)$ denotes the Gaussian neighborhood kernel around the best matching model neuron j (i.e. winner neuron), if the phonetic map is activated by the training item (see Equation 4), and where L denotes the learning rate. The learning rate decreases over time exponentially:

$$L(t) = L(0)^* \exp(-0.00001^*t), \tag{2}$$

where L(0) = 0.9 is the initial learning rate. The radius σ(t) of the Gaussian neighborhood kernel (Equation 3) also decreases exponentially during training (Equation 4):

$$H_j(t) = \exp\left\{-0.5^*\left(n_k - n_j\right)^2/(s(t))^2\right\}. \tag{3}$$

Here, n_k denotes any neuron within the self-organizing map and $|n_k\text{-}n_j|$ denotes the distance of a neuron n_k to the winner neuron n_i defining the center of the neighborhood kernel. σ(t) is defined here arbitrarily as the radius of the neighborhood kernel. The neighborhood factor H_j, which is 1 for n_j, declines below a value of 0.6 if the neuron distance exceeds σ(t) and below 0.14 if the neuron distance exceeds 2σ(t). The temporal decrease of the neighborhood radius is:

$$s(t) = s(0)^* \exp(-0.00001^*t), \tag{4}$$

where σ(0) = 5 neurons. Thus, 2σ(0) is 40% of the overall length of the phonetic map (the overall length is 25 neurons). The winner neuron n_j, specifying the center of the neighborhood kernel, is calculated by looking for the maximally activated neuron

j within the phonetic map. The activation pattern $a_j(t)$ of the phonetic map during the application of the training item $a_i(t)$ is calculated as

$$a_j(t) = \mathrm{sum}_{(i=1,\ldots,Ni)}\left\{w_{ij}(t)^* a_i(t)\right\}. \quad (5)$$

The incremental change of synaptic link weights, quantified by the equations above, describes *Hebbian learning*, because synaptic link weights w_{ij} mainly change for those synaptic connections for which phonetic map neurons as well as state map neurons are strongly activated. At least synaptic link weights w_{ij} increase towards a_i, i.e. link weights adapt to the currently applied activation pattern a_i ($i = 1, \ldots, N_i$). Thus, this learning is *non-supervised*: No ideal activation pattern is known in advance for the phonetic map. Learning always tries to approximate the activation pattern of the currently applied training item. Links with link weights w_{ij} grow in both direction, i.e. from state maps towards phonetic map as well as vice versa. This bidirectionality is compatible with the idea of Hebbian learning.

We started babbling training with a random link weights initialization for w_{ij}, taking values between 0 and 0.5 in order to model moderate neural interconnectivity at the beginning of training. Our criterion for stopping imitation training is defined as follows: A predefined amount of neurons of the *phonemic* map (e.g. 95%) should indicate a strong synaptic connection with one neuron of the phonetic map (e.g. $w_{ij} > 0.8$). That implicates that a predefined percentage of syllables is learned by the self-organizing network in a way that at least one phonetic realization exists for that syllable. From our experience, this indicates that the associated phonetic states already have a correct association between motor plans and sensory states for that syllable.

The situation is more complex in the case of babbling training, since in that case no simple criterion can be given for reaching a specific level in sensorimotor knowledge or sensorimotor skill learning. Since babbling training is always guided by imitation, in order to train appropriate regions within the multidimensional space of motor parameters, the criterion used in our modeling approach is that babbling training will be continued as long as a speech item under imitation is not reproduced "distinguishably". It should be noted that all babbling trials performed by our model for producing an imitation training item (this can be referred to as "guided babbling") are controlled aurally by a supervisor. It is up to this person to decide subjectively whether an item is "distinguishable" or not. In the case of this study, babbling knowledge was sufficiently acquired after 500 training cycles.

In our experience, babbling training requires a great deal more training cycles than imitation training, since the organization of the phonetic map emerges during babbling training and is only refined during imitation training (up to 500 training cycles are required during babbling training, while 50 to 150 training cycles are sufficient in many imitation training situations; see [44]). This decrease in training cycles towards imitation training may result from the fact that – based on our experience with our simulation experiments – the emergence of phoneme regions never indicates a reorganization of the sensorimotor structure of the phonetic map with respect to the phonetic features that have already been acquired.

The learning equations (1)-(5) given above are based on Kohonen's learning equations [21]. However, Kohonen used just one state map and thus did not subdivide state maps with respect to different modalities. Our model adapts and expands Kohonen's approach to the areas of speech acquisition and speech processing by considering the sensory, motor, and phonemic modalities.

Furthermore, Kohonen's approach calculates synaptic link weights only in one direction, e.g. from the sensorimotor state maps towards self-organizing map. In contrast, we assume that synaptic connections with strength w_{ij} arise not only from model neurons n_i of the state maps towards model neurons of the self-organizing phonetic map n_j, but also in the opposite direction, from the phonetic map towards the state maps with the same link weight values w_{ij}. This *bidirectionality* within the mappings of our model (Figure 2) is important especially in modeling not just speech acquisition and speech perception, but speech production as well. This modification is a further contribution of the present paper in adapting Kohonen networks to the areas of speech acquisition and speech processing.

Results

Description of simulation experiments

Six groups of simulation experiments were performed in order to feed speech knowledge into our model (Table 1). Because imitation always requires some previous sensorimotor knowledge (see above), in our model, babbling training is performed before imitation training. But it is important to note that the language-specific speech items spoken by the caretaker normally guide babbling: the set of all possible motor plan states which could be generated randomly would be infinitely huge and thus could not be trained by the model. For this reason, babbling training is always directed towards the specific language being acquired (e.g. [46]).

In order to feed first sensorimotor knowledge into our model, we started with protovocalic babbling (Table 1). In a first group of simulation experiments, the tongue position of the vocal tract model was randomly positioned and the corresponding auditory state was calculated for each training item. After babbling training, the synaptic link weights between the motor plan map and the auditory map are tuned in a way that (i) the motor plan states and the appropriate sensory states (as produced by the sensorimotor processing loop) are *associated* and that (ii) these sensorimotor states are *ordered* with respect to two important *vocalic phonetic features* "front-back" and "low-high" [44].

In a second group of babbling experiments, motor plan states were initiated which start from different labial, apical, and dorsal closures and which end in different protovocalic states (protoconsonantal babbling, Table 1). After training, the *consonantal phonetic feature* "place of articulation" (labial, apical, dorsal) are learned. These are in addition to the vocalic phonetic features, which had previously been learned. Three compact regions appear within the resulting self-organizing map. These regions represent the three different places of articulation [44,47].

Because initial sensorimotor knowledge is gained after performing these babbling training simulations (i.e. the associations between the motor plan and the sensory states, along with the ordering of these sensorimotor states at the level of the phonetic map with respect to phonetic features), the motor plan states can now be roughly estimated based on the auditory states. Thus, the model now is ready for vowel and consonant imitation. We continued with imitation training of a model language comprising five vowels V =/i, e, a, o, u/(see third group of simulation experiments, Table 1) and 15 syllables comprising all combinations of these five vowels with the consonants/b, d, g/as CV-syllables (consonant-vowel-syllables, see fourth group of simulation experiments, Table 1). This imitation training does not increase the number of phonetic features, which have already been learned by the model during babbling, but rather achieves a labeling of specific sensorimotor states with respect to phonemic

Table 1 Groups of simulation experiments for speech acquisition

Group name	Specification of training items	Number of training items	Size of SOM	Phonetic features learned	Published
Protovocalic babbling (prelinguistic)	Variation of tongue position	1076	15 × 15	Front-back; low-high;	[44]
Protoconsonantal babbling (prelinguistic)	Variation of initial closure: labial, apical, or dorsal	279	15 × 15	Place of articulation;	[44,47]
Vowel imitation (part of model language)	5 vowels (V): /i, e, a, o, u/	500	15 × 15	Front-back; low-high;	[44]
Consonant imitation (part of model language)	15 CV syllables: 5 vowels (V) with 3 consonants (C): /b, d, g/	465	15 × 15	Place of articulation;	[44,47]
Syllable imitation (model language: V, CV, CCV)	70 syllables: 5 V + 5 V with 9 single C: /b, d, g, p, t, k, m, n, l/ as CV + 5 V with 6 clustered CC: /bl, gl, pl, kl/ as CCV	600	25 × 25	Place and manner; voiced-voiceless; type of syllable;	[36,48]
Imitation of frequent syllables (real language)	Standard German; 200 most frequent syllables;				
	Male speaker; one realization per syllable	200	15 × 15	Stressed-unstressed; vowel type;	[22]
			20 × 20	Same;	
			25 × 25	Same;	
	Female speaker; up to 27 realizations per syllable	703	25 × 25	Same;	This study

categories. Due to the inclusion of phonemic states during imitation training (in contrast to babbling training), here there is also an association established between the sensorimotor states and the phonemic states. Thus, many neurons of the phonetic map can be labeled phonemically. Because phonemic states usually represent similar sensorimotor states, *phoneme regions* occur for each phonemic vowel state and for each phonemic CV-syllable state at the level of the phonetic map. A model neuron within the phonetic map is a member of a phoneme region if there are strong synaptic connections between this model neuron and a specific model neuron within the phonemic map which represents that phoneme or phonemic state [44,47].

After imitation training of V- and CV-syllables with V =/i, e, a, o, u/and C =/b, d, g/, a fifth group of simulation experiments was performed. Within this group of simulation experiments, the set of consonants was extended (C =/b, d, g, p, t, k, m, n, l/) and first CCV-syllables (i.e. syllables with initial double-consonant clusters/bl, gl, pl, gl/) were allowed and consequently trained. Within this training step – which included guided babbling with the new groups of consonants undergoing imitation training – new consonantal phonetic features result from the further ordering of sensorimotor items at the level of the phonetic map, i.e. "voiced-voiceless", and "manner of articulation" (plosive, nasal, lateral; see Table 1). Also, an appropriate ordering of phoneme regions is achieved at the level of the phonetic map for all 70 phonemic syllable states (see [36,48]).

Finally, a sixth group of simulation experiments was performed in order to train a natural language, in our case Standard German (see Table 1). First, we assembled a corpus of Standard German on the basis of 40 children's books. These books were targeted for children up to six years of age. We transcribed 6513 sentences in total, leading to 70512 words in total, comprising 4763 different syllables [22]. The 200 most frequent syllables were extracted from this database and used for training the model. In contrast to model language training (simulation experiments 3, 4, and 5 in Table 1), where phoneme realizations within the training set were generated in a synthetic way through a rule-based version of our articulatory synthesizer [25], here we used natural acoustic realizations of syllables which were uttered by native speakers, one male (33 years old) and one female (27 years old), both without any known anomalies in speech or hearing.

In a first simulation experiment of this sixth group of experiments (Table 1), only one sentence was uttered by the male speaker for each frequent syllable. This syllable was extracted from the acoustic signal, and the appropriate auditory state was generated and imitated by the model. Thus, just one realization was trained per syllable, but the training already took into account the differences in frequency of occurrence for the 200 most frequent syllables (for details, see [22]). In the present paper, we report on a further simulation of imitating of these 200 most frequent syllables by using a refined training data set comprising more than one realization per syllable (see Results section).

We now provide a summary of the babbling training results. Babbling training always resulted in (i) an *association of motor plan and appropriate sensory states*, and (ii) an *ordering of these sensorimotor states with respect to sensorimotor or phonetic features*. Subsequent imitation training always resulted in an additional labeling of specific sensorimotor states – represented by specific model neurons within the phonetic map – with respect to phonemic categories. Thus, *phoneme regions* appear within the phonetic map, since similar sensorimotor states are usually represented by adjacent model neurons within the phonetic map (see e.g. Figure 8, Appendix). The emergence of

phoneme regions implies that a phonemic state can be represented by more than one neuron at the level of the phonetic map. These neurons at the level of the phonetic map, representing the same phonemic state, are referred to as *exemplars* in the context of our modeling approach; see also [49] for a comparable use of this term. Thus, a phonemic state can be represented by *different learned or trained phonetic realizations*, referred to as exemplars of that phonemic state. For example, the three model neurons in the upper left corner of the phonetic map shown in Figure 8 (see Appendix) represent three (stored or learned) exemplars of the Standard German syllable/mIt/("with"). A closer inspection of these three exemplars indicates that there are slightly different sensorimotor or phonetic features for each state (see motor plan states and auditory states of these exemplars, presented in Figures 9 and 10 in the Appendix).

Imitation training of frequent syllables

A female speaker (27 years old; with no known anomalies in speech or hearing) uttered up to 27 realizations of the 200 most frequent syllables of our children's book corpus [22]. Syllable realizations were produced in proportion to the frequency of occurrence of a syllable in the corpus (Table 2). Thus, 703 sentences were recorded in total (D = 703).

The syllable under consideration was marked within each sentence and a motor plan state was generated for each of these syllables using our resynthesis procedure [50]. Neural representations of all motor plan states and auditory states were calculated for each syllable. The resulting training items were applied to the neural model.

In total, 120 training cycles were sufficient in order to represent 95% of the 200 most frequent syllables of our corpus (T = C * D = 120 * 703 = 84360 training steps in total). The criterion for terminating the training was defined as follows: A strong synaptic connection ($w_{ij} > 0.8$) was established for 95% of the model neurons within the phonemic map (phonemic states) and at least one model neuron within the phonetic map for each of these phonemic states. Thus, at least one exemplar must exist for 95% of all phonemic states (190 of 200 syllables). The learning curve resulting from training our model is presented in Figure 5. It can be seen that a strong increase in syllable learning occurs above training cycle 60. A saturation effect is reached after approximately 120 training cycles. We checked to make sure that the 5% of syllables which were not learned were below rank 150 for the 200 most frequent syllables.

The neurons of the phonetic map which exhibit a strong synaptic link towards a phonemic state (i.e. all exemplars) are marked in Figure 8 (Appendix) by their phonemic transcription. The link weight distributions for all model neurons of the phonetic map towards the motor plan and auditory state map are given in Figures 9 and 10 (Appendix). From these displays of the organization of the phonetic map, which is

Table 2 Frequency of occurrence of a specific syllable in the children's book corpus and in the training set

Rank of syllable (with respect to frequency)	Frequency of occurrence in corpus	Number of training items
1	2367	27
20	692	8
50	390	4
100	193	2
200	88	1

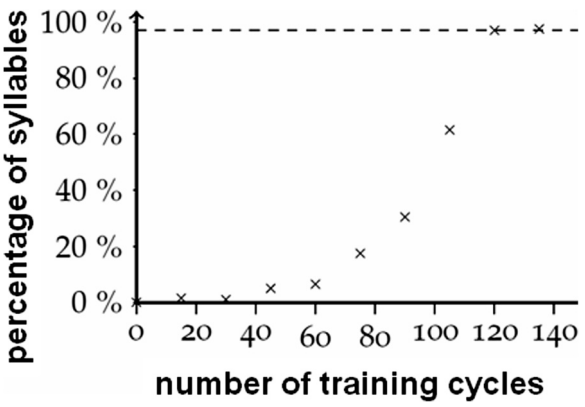

Figure 5 Learning curve resulting from simulation of imitation training for our children's book corpus. X-axis: number of training cycles; y-axis: percentage of syllables already acquired.

Figure 6 25 × 25 phonetic map after training, representing phonemic link weights (see also Figure 8 in the Appendix). Separate regions emerge within this map, representing the syllable features stressed vs. unstressed and the vocalic features reduced vs. nonreduced. Nonreduced vowels can be short vowels, long vowels and diphthongs in Standard German.

present after imitation training, it can be clearly seen that neighboring states at the level of the phonetic map represent similar sensorimotor features. Thus, the distribution of states at the level of the self-organizing phonetic map can be labeled as quasi-continuous. This in part reflects the ordering of sensorimotor states with respect to phonetic features. The organization of the phonetic map for the experiment described here is discussed in more detail in Discussion section.

A careful inspection of the distribution of the phonemic transcriptions over the phonetic map (Figure 8, Appendix) indicates an ordering of syllables with respect to the syllable feature *stressed-unstressed* and with respect to the vocalic features *nonreduced-reduced*, where nonreduced vowels in Standard German can be short vowels, long vowels, or diphthongs (Figure 6). Also, a loose ordering with respect to other vocalic and consonantal phonetic features can be seen.

Because our training data set is designed in such a way that frequent syllables occur with more realizations than less frequent syllables (see Table 2), syllables are processed with different frequency during training. After training, this is reflected in the resulting organization of the phonetic map. Here, the number of neurons representing a syllable with strong synaptic connection to a phonemic state (i.e. the number of exemplars per syllable, or the size of the phoneme region) is proportional to the frequency of occurrence of that syllable in the training set (Figure 7). This means that the more often a syllable is processed by the model, the larger the area is within the phonetic map which represents this syllable. (Or, the more often a syllable is processed by the model, the higher the number of exemplars which are stored for that phonemic state is and the bigger the appropriate phoneme region is).

Moreover, we checked whether the phonetic variability of the exemplars for a specific frequent syllable, represented in the phonetic map, is comparable to the phonetic variability within the different realizations of that syllable within the training set. This was

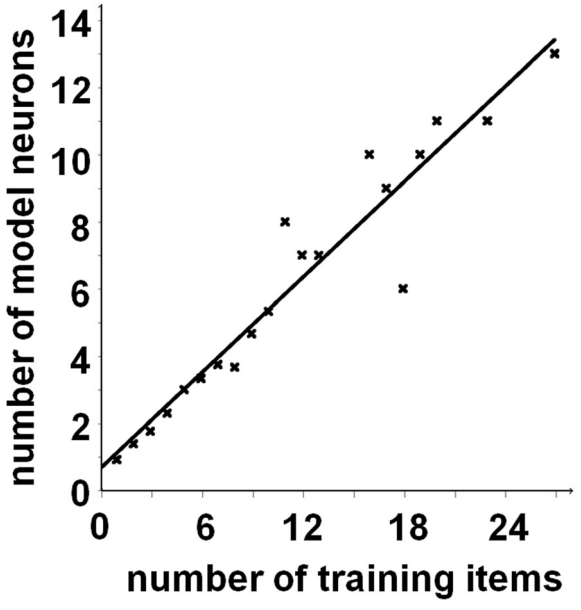

Figure 7 Number of model neurons representing a syllable at the level of the phonetic map as a function of the number of training items occurring for that syllable within the training set.

done by comparing the mean variance of the activation pattern of the neural auditory states resulting from the realizations of the training set for a specific syllable to the mean variance of the activation patterns of neural auditory states which are co-activated from neurons of the phonetic map, assuming that that neuron of the phonetic map represents an exemplar of the same syllable (exemplars which had already been learned, represented by link weights. It can be seen that the phonetic token variability within the training set is comparable to that of exemplars for some syllables, while it is smaller for most syllables (Table 3).

Speech processing

Because speech production and speech perception are already a part of speech acquisition (babbling and imitation are perception of a caretakers utterance, followed by several production trials of the toddler or model), the performance of speech production and perception increases during the continuously ongoing process of speech acquisition. In our modeling approach, we defined a specific landmark within the process of acquisition at which the performance in production and perception was tested. Thus, speech acquisition was defined as (nearly) complete if 95% of the syllables occurring in the training set have been acquired. It should be noted that this criterion is a direct reflection of the fact that our training sets and self-organizing maps are of fixed sizes. The remaining 5% of syllables are easily acquired if the training set is widened, e.g. if in an augmented training set and an augmented phonetic map are used for further training at this landmark for speech acquisition. Thus, testing the quality of performance of the model's production and perception capabilities at a landmark of speech acquisition is also a check on the quality of speech acquisition accomplished so far.

Bidirectional cortical mappings

It is important to keep in mind that self-organizing maps as defined by [19] only use mappings from that map, which represents the training items (i.e. the sensorimotor state maps in our case) towards the self-organizing map (the phonetic map in the case

Table 3 Training results for 10 most frequent syllables

Syllable	Number of model neurons (exemplars)	Mean variance of exemplars	Mean variance of training items
'?Unt	13	0,004114	0,004037
'di:	11	0,001899	0,004117
'zi:	10	0,002830	0,007564
t@	10	0,002761	0,003878
'dE6	10	0,001903	0,008438
'?E6	8	0,002896	0,013582
n@	7	0,002053	0,005684
g@	6	0,001418	0,003845
n@n	5	0,002184	0,005175
b@	3	0,001906	0,001881

First column: phonetic transcription (SAMPA) of the 10 most frequent syllables of our children's book corpus; second column: number of neurons representing all realizations of these syllables within the phonetic map (number of exemplars); third column: mean variance of the auditory states of all realizations by activating specific neurons within the phonetic map, representing exemplars for a syllable; fourth column: mean variance of the auditory states of all realizations activated by the training items directly.

of our modeling approach). But Hebbian learning in principle allows the emergence of neural connections between synchronously activated neurons in both directions. Thus, in our modeling approach we assume that the neural mapping between state maps and phonetic map is *bidirectional*. Thus, neural connections occur and increase in connectivity in both directions, i.e. from state maps towards the phonetic map and vice versa. The increase in connectivity directly reflects the degree of activation of the model neurons within each training step.

This existence of bidirectional mapping with the same link weight values for the synaptic neural connections in both directions is important for the other working modes of the model. Aside from speech acquisition, the model comprises the working modes of speech production and speech perception.

In the case of speech production, it can be assumed that the production of a word or utterance leads to a successive activation of phonemic syllable states at the level of the phonemic map. Thus, if the production of a specific syllable starts, we expect the full activation of that neuron, which represents that specific syllable at the level of the phonemic map. This directly leads to a co-activation of neurons at the level of the phonetic map, initiated by the neural mapping from the phonemic state map towards the phonetic map (Figure 2). Subsequently, the strongest co-activated model neuron at the level of the phonetic map leads to a co-activation of the appropriate motor plan and sensory states via the neural mapping from the phonetic map towards motor plan and sensory state maps. On the basis of the motor plan activation pattern, the syllable can be articulated by activating the feedforward execution path. Finally the resulting auditory state, activated by a subsequent sensory processing of the articulatory-acoustic vocal tract model output via the sensory processing pathway, can be compared with the auditory state of that syllable, already activated from the neural mapping between the phonetic map and the auditory state map. In this case, two states must be compared at the level of the auditory state map, which is not included in the version of our model introduced here. However, a concept for the comparison of sensory states and for the calculation of sensory error signals has already been introduced in [7].

In the case of speech perception (e.g. perception of the caretaker), an auditory state is activated at the level of the auditory state map (Figure 2). This leads to a co-activation of the phonetic map via the mapping from the auditory state to the phonetic state map and then to the selection of a winner neuron at the level of the phonetic map. Subsequently, a phonemic state is activated at the level of the phonemic map if the winner neuron at the level of the phonetic map is part in the relevant phoneme region.

It is possible that due to the context within a specific communication situation, competing candidates (competing syllable states) are already activated at the level of the phonemic map. This could lead to top-down effects in perception and influence the neural activation pattern at the level of the auditory map. These top-down effects in speech perception are not included in the version of our model presented in the present paper.

Speech production

The quality of the 50 most frequent syllables produced after speech acquisition (i.e. after the adjustment of synaptic link weights of the mappings between the phonetic map and all state maps as described in above) was tested by performing a combined

acoustic perception and vocal tract action identification test. A phonetic expert (BK, male, 53 years old, no known speech or hearing anomalies) performed this test. In this test, the 50 most frequent syllables, which were already acquired during simulation of speech acquisition, were generated by the model. A winner neuron was calculated at the level of the phonetic map by activating a specific phonemic state (a specific syllable). Subsequently, the appropriate motor plan state was co-activated for that winner neuron, and then the articulatory and acoustic signal was generated using the feedforward execution path including our vocal tract model (see Figure 2).

In a first step of this test, the expert was advised to transcribe all 50 acoustic speech stimuli. Stimuli were presented twice each in random order. In the case of differences between both transcriptions of the same acoustic stimulus (as was the case for 8 stimuli in this experiment), the stimulus was presented a third time and the expert was advised to choose one of both transcriptions as the nearest transcription.

An initial evaluation of the results of this transcription process was done by comparing the transcriptions with the phonemic transcriptions defining the phonemic states during the speech acquisition process. The overall rate of correct transcriptions was 78%. A close inspection of the resulting transcriptions showed that transcription errors mainly resulted from confusions in the place of articulation for nasals and for voiced plosives. Our impression is that this results primarily from the lack in acoustic quality of our vocal tract model rather than from errors in controlling articulation (errors in generating motor plan states).

Thus, in the case of those speech items which showed transcription errors, we checked in a second step whether the correct vocal tract action was produced at the level of the articulatory signal. This was done by inspecting the articulatory movement pattern generated by the vocal tract model. This resulted in a rate of correct syllable production of 94%, which we assume means that the model is capable of articulating well after speech acquisition.

The remaining errors (3 of 50 syllables) resulted from an incorrect perceptual categorization of the phonetic feature “stressed vs. unstressed”, which may result from the fact that the settings of the pulmonary and laryngeal actions was fixed within the current version of our model. Thus, the main acoustic cue for identifying the stressed vs. unstressed condition, which is available in our model, is the length of the whole syllable and the length of single sound segments within each syllable.

Speech perception

Speech perception (i.e. the speech recognition rate) was tested by calculating the recognition rate for the 200 most frequent syllables, spoken three times by the same person (female, 27 years old). This person was the same person whose voice was recorded for the training items. These test items were fed into the model and resulted in the activation of a winner neuron at the level of the phonetic map and subsequently to an activation of one neuron at the level of the phonemic map (the identification process). The rates for correct identification are listed in Table 4. It can be seen in Table 4 that the recognition rate decreases with decreasing frequency of occurrence of a syllable within the children’s book corpus.

Discussion and conclusions

A quantitative model is introduced in this paper which is capable of simulating the basic processes of speech acquisition, speech production, and speech perception. This

Table 4 Recognition rate of test items of 1 to N with N = 5 to N = 200 most frequent syllables

Number of most frequent syllables	Recognition rate in % without frequency correction	Recognition rate in % with frequency correction
5	100,0	100,0
10	97,2	97,4
20	91,1	93,2
50	90,2	92,7
100	88,9	92,3
200	85,8	91,8

The recognition rate without frequency correction (second column) shows the mean recognition rates for all N syllables, each syllable weighted equally. The frequency correction (third column) shows recognition rates for all N syllables, now weighted with respect to the frequency of occurrence of each syllable within the corpus (i.e. highly frequent syllables occur more often and also indicate higher recognition rates).

approach does not include higher-level linguistic processes such as the acquisition of semantic or higher-level grammatical knowledge. Thus, modeling of the mental lexicon (cf. [28]) is beyond the scope of our approach. Our model focuses on the sensorimotor aspects of speech acquisition, production, and perception. The speech knowledge, including speech production skills, acquired by our model may be comparable to the knowledge which is theoretically assumed to constitute the mental syllabary [51]. But in contrast to the model presented in [51], our model is strictly quantitative; it is capable of processing acoustic and articulatory data; and it is based on computational and theoretical neuroscience.

Thus our approach for modeling sensorimotor aspects of speech should be seen as a counterpart to the quantitative computational neurolinguistic approach developed by [5,38] for modeling the mental lexicon. Initial ideas as to how to combine these two approaches are discussed in [22,42]; see the discussion concerning P-Map and S-Map in these papers. Consequently, our modeling approach focuses on the early phases of speech acquisition (mainly first 18 months of life), while other approaches are focused on the rapid growth of the mental lexicon (the vocabulary spurt, starting at around 18 months; e.g. [52]).

As mentioned in the Introduction of the present paper, our approach should be seen as complementary and not as contradictory to the sensorimotor model introduced by Guenther (DIVA model [7,10,11,35]). On the one hand, while the DIVA model and our model both comprise phonemic, motor and auditory and sensorimotor neural state representations, our approach additionally introduces the phonetic map as an intermediate map between the sensorimotor and linguistic parts of the model. This intermediate level could be added to the DIVA model as well. On the other hand, a quantitative modeling of the processing of sensorimotor mismatch – as is done by Guenther and colleagues [ibid.] using specific "error maps" – is not currently implemented in our approach but could be added easily. A difference between the approaches is that our approach explicitly assumes a motor planning level (cf. [2]), while the DIVA approach directly associates its speech sound map with its primary motor map.

The modeling approach for speech acquisition and speech processing (production and perception) as introduced in the present paper bases the ability to simulate behavioral phenomena, such as the production and perception of speech items and learning to speak, upon brain-related principles such as Hebbian learning and self-organization.

This goal is achieved by using an activation rate model. In contrast, if we were to use fine-grained "neuromicroscopic" approaches such as spiking neuron models (e.g. [15]), we would not be able to simulate "macroscopic" large-scale behaviors such as speech acquisition, production and perception. Thus, our approach is a feasible compromise between simplicity in modeling neural "microfunctions" on the one hand and the ability to simulate complex "macroscopic" behavioral phenomena on the other. Moreover, our approach provides explanatory power for understanding the behavioral phenomena of speech processing from a neurofunctional point of view. At this point, it should be noted that our model not only gives quite good results in the performance of production and perception after training (see Discussion section), but also is capable of simulating complex behavior such as categorical speech perception [44].

A basic functional neural structure is hypothesized in this paper and is introduced as a preliminary neurofunctional model for speech acquisition, speech production, and speech perception. Three hypotheses form the basis of the model's structure: (i) We assume that there are three neural maps representing higher-level motor plans and higher-level auditory and somatosensory states. The activation patterns of syllabic speech items are manifested at this level during the time period of processing (i.e. producing or perceiving) of that speech item. (ii) In parallel, a phonemic (i.e. abstract linguistic) state representation is assumed to be activated for each speech item at the level of the phonemic map. The activation of a syllable state at the level of the phonemic map occurs at the beginning of the production process as well as at the end of the perception process. (iii) *Only one* intermediate neural map is assumed in order to associate motor plan, sensory, and phonemic states. This neural map is postulated to be a self-organizing map. This map is referred to as the phonetic map because it is supramodal, i.e. it is above the motor and sensory modalities. At the level of this map, phonetic features result from the ordering of sensorimotor speech states during neural self-organization.

Two hypotheses are assumed for speech acquisition, i.e. for training the model: (i) Hebbian learning (cf. [53]) occurs during speech acquisition: "Wire" two neurons if they are activated within the same time period, i.e. if these neurons "fire" within the same time interval. More specifically, one of these neurons must be part of the phonetic map, while the other neuron is part of the sensorimotor state maps. Because many neurons, especially at the level of the state maps, may be activated in parallel, the "wiring" process (i.e. the increase in synaptic link weights, also referred to as the increase in synaptic association) occurs in parallel for many neurons between the sensorimotor state maps and the phonetic map. (ii) Training can be subdivided into babbling training and imitation training, where babbling training comprises only motor and sensory information because the goal of babbling training is to learn sensorimotor relations. Learning sensorimotor relations means associating sensory and motor plan states. This is easily accomplished during babbling because sensory states are produced directly from motor plan states by using the peripheral vocal tract model, which is included in our model (Figure 2). Additionally, imitation training requires phonemic information in order to categorize sensorimotor states acquired previously at the level of the phonetic map. Imitation training leads to the emergence of phoneme regions.

The results of the imitation training experiment described in this study, which used the 200 most frequent syllables of Standard German on the basis of a children's book

sentence corpus, indicate that the phonetic map can be subdivided into two major areas: stressed and unstressed syllables. In addition, in the case of stressed syllables, some sub-areas can be found which reflect different types of vowels (long, short, diphthong; see Figure 6; frequent unstressed syllables in our corpus are mainly produced with a reduced vowel). Other imitation experiments indicate that sensorimotor states are ordered at the level of the phonetic map with respect to different phonetic features such as front-back, low-high, place and manner of articulation, etc. (see Table 1). This result is not affected if different random initializations for link weights are used or if motor and sensory representations are weighted in a different way. These modifications mainly cause rotations or a mirroring of the phonetic map, but they do not affect the resulting phonetic ordering.

Moreover, the simulation experiment discussed in this paper indicates that the more often a training item is processed by the model, the larger the resulting phoneme region is, i.e. the more model neurons represent exemplars (or realizations) of that phonemic state at the level of the phonetic map (see Figure 7). In addition, it should be noted that in our approach, imitation training requires fewer training cycles than babbling training. The number of training cycles is approximately 600 for babbling training and around 150 training cycles for imitation [44]. This reflects the fact that the phonetic map and especially the ordering of phonetic states is in large part established during pre-linguistic babbling training. This ordering is refined during imitation training with respect to language-specific demands. Even the emergence of phoneme regions does not lead to a reorganization of the phonetic map, which is already established during babbling training.

There are still some shortcomings in our approach which we will address by further refining the model in future studies. Firstly, simulation experiments for babbling and imitation were done step-by-step and not – as it would be more realistic – as one continuously ongoing simulation experiment. The main reason for that is that we did not use a growing self-organizing map (GSOM, see [5]) but rather SOMs of fixed sizes (see Table 1). A further reason is that we are currently not able to guide babbling by imitation. The current version of the model is not able to imitate a given speech item directly and thus is not able to generate motor plans which are related to a potentially successful motor plan for imitating the speech item. This work is done manually in the current version of the model. It is for this reason that we currently establish babbling and imitation training sets manually.

Thus, a further step in improving our model would be the implementation of a module for babbling training guided by imitation. In addition, the introduction of GSOMs is an important future step and will allow us to avoid predefining the size of the phonetic map and to allow an ongoing continuous acquisition of speech knowledge including sensorimotor speech skills. This augmented model should also include the acquisition of prosodic features, such as the basic intonation contours of the target language, as well as the acquisition of basic stress patterns of words and utterances.

A second shortcoming of the current version of our model is the problem of speaker normalization [43]. Even in the case of imitation training, the imitation trials produced by the model are evaluated aurally by an expert before they are used as training items. In the current version of the model, it would be possible to store both auditory realizations (the one pre-produced by the external speaker or caretaker) and is the one successfully imitated by the model or toddler. Both groups of auditory realizations could

be associated with the phonemic states as well as with motor plan states. In the case of imitating more than one external speaker, we potentially could integrate two or more "speakers" into the phonetic map into future versions of our approach, which would allow us to augment the phonetic map to include the phonetic dimension of "different speakers" or "different types of speakers" and thus to address the problem of adaptation to a new speaker in our framework.

A third shortcoming of our approach is that we start with sensorimotor (i.e. phonetic) information during babbling training, and directly thereafter we extend this to phonemic information for imitation training. From the viewpoint of natural speech acquisition, it would be more realistic to start imitation on the basis of face-to-face communication scenarios, where the model can associate words with semantic categories rather than directly with a (perfect) phonemic representation. There is some evidence showing that phonemic information is not available at the earliest developmental stages, but rather emerges during speech acquisition [42]. Thus, we will use sensorimotor and semantic information rather than sensorimotor and phonemic information for future training experiments. This information could be used for speech acquisition if our model comprised a semantic level

mIt	'mIt	'hat	'hat		'?Ist	'?Ist	'?Es	'?Es	'?Es	'vas	'vas	'das	'das	'das	'das	'gants		'dEn	'de:n	'de:	'de:	'de:m	'kan	kOmt
'mIt		'hat	'hat		'?Ist	'?Ist	'?Es	'?Es		'bIs		'das	'das	'das			'dan	'dan		'de:n	'de:m	'de:n		'ta:k
	'gIpt	'rYk	'?Et		'?IC	'?I	'?ap	'?ap	d@t	'bEt	'dOx	'dOx		'za:k	'za:k	'vEn	'vEn	'man		'de:n		'mo:nt	mo:nt	
'nICt	'nICt		lIC		'?IC	'?IC	vas					've:k	'zi:t		'fra:kt		n@n	'man	'hant		'na:x	'ma:l		'SnEl
'nICt		'?Ist	lIC			'?IC	'di:	'di:	'du:	'di:		'ge:t		'So:n	'fOn	'fOn		'vIl		'ja	'ma:l	'da:		'tsvaI
	'mUs		n@s	'zIC	'zIC		'bE:	'di:	'di:	'di:		'gro:	'fe:		'tsUm	'tsUm		'vIl	'vOl	'da:	'da:	'kaI		
'jEtst	'maxt	'nOx		'zIC	'zIC		'di:	'di:	'di:	'di:	'vo:	'gro:		'fy6	'fraU		'vI6t		'va6	'va6	'dE6	'dE6	'ga6	baI
				'zIC		'vi:	'di:	'du:	'du:	'vi:	'vi:		'fo6	'fraU			'vI6t	'vi6	'dE6	'bE6	'dE6	'dE6	'dE6	'bE6	'dE6
'tsu:	'tsu:	'Spi:	'fi:		'le:	'li:				'zi:		'zi:	'fo6		'klaI			'dE6	'dE6	'dE6	'vi6	'm6		'dE6
	'tsu:	tsIm	'fo:	f6		'blu:	'zi:	'zi:	'zi:	'zi:	'zi:	'zi:		'klaI	'klaI	'fa:	'va6	'vE6	'vE6		'?i6		'?i6	'?E6
ts@	tsu:	ts@n		f@l	'fln		'zi:	'zi:		'zo:			'laN	'ja		'zaI	'zaI	'da:			'?E6	'?E6		'?E6
f@			f@n		C@n	C@n		'ze:	'zo:	'zo:	'zOn	'jUN	'mE:t	'me6	'hOI			'ra:	d6	'?E6	'?E6		'?aIn	
s@	S@		f@n	f@n	C@n	x@n	x@n				'kOn		ma		'ha:		'vaI			'?E6	'?E6	'?aIn	'?aIn	'?aIn
	fE6	z@		z@n	z@n			m@n	l@n	n@m	N@n		ma:l		'ha:	'?a:	'?a6	'?aI	'?aI	'?aI	'?aI	'?aIn	'?a	'?aIn
z@	fE6	z@				m@n	m@n			n@	N@n	g@n	dEl	'tE		'?aI	'?aI	'?a	'?aI	'?aI	'?aU	'?a	'?aUf	'?a
		l@		'ma	'nOk	'nOk		r@n	r@n	n@n	n@n	k@n	t@l	t6	pa:		'?a:	'?a:	'?a:	'?a	'?aU	'?aUf	'?aUf	'?aUf
n@		l@	ma:	'ma		m6		r@n	r@n	n@n	n@n		g@l	g6							'?aUx	'?aUx	'?aUs	'?aUs	'?aUf
n@	n@	n@		ma	ma:		n6		l@n	n@n		d@n	g@l		'?al		'ri:	'?i:	'ri:		'?a	'?alts	'?a	'haUs
	n@	n@		hE		m6	ma:	ma:		b@n	b@n	d@n		g@l	'?al		'ri:	'?i:		'?Ent	'?an	'?a	'?alts	
l@		m@	n@				l6		g@n	g@n	g@n	d@n	t@t	t@		'?o:	'?o:		'?Im	'?En	'?an	'?an	'?a	k@l
jo:	jo:	w@		ni:	di:	'da	'da	b6	b6	g@n	k@n	d@n	t@	t@		'?y:		'?Im	'?In	'hIn	'?an	'?an	'?am	'hIn
n@	l@	li	r@		r@	b6		d6		b@n	k@n		t@	t@			'?Un		'?In	'?I	'?I	'i:n	'?In	'?i:m
g@			d@	d@		b6		d6	g6		k@	t@	t@	t@	t@t	'kat		'?Unt	?Unt		'?Unt	?Um	'?Um	'?i:
g@	b@					d6		t6	t6		p@	k@		t@	ts@	'kOn	t@l		'?Unt	'?Unt	'?Unt	'?Unt	'?Unt	'?Unt	
g@	b@	b@	g@	g@	g@	d6		t6	t6	'pa	pi:	pi:	t@	t@		'kO	'kIn	'kIn		'?Unt	'?Unt	'?Unt	'?Unt	'?Ins	

Figure 8 25 × 25 phonetic map after imitation training as described in Results section. A model neuron is marked with a phonemic transcription if the synaptic link weight between that model neuron of the phonetic map and a model neuron of the phonemic map is strong (link weight > 0.8). An apostrophe at the beginning of the syllable indicates that the syllable is stressed. Transcriptions are given in SAMPA notation (Speech Assessment Methods Phonetic Alphabet).

in addition to the sensorimotor level described in this paper (ibid.). Additionally, this would expand our model towards word and utterance processing and thus more realistic performance tests for production and perception could be performed, especially considering that speech perception is difficult to check at syllable level.

A fourth shortcoming of our approach is that we have no explicit modeling of time as exists in spiking neuron models or more detailed rate models such as the neuroanatomically grounded Hebbian learning model developed by [8]. Time modeling would allow us to avoid the zero-paddings which currently occur in our syllable-based representations of motor plans and neural spectrograms. Furthermore, these approaches would allow the processing of whole utterances rather than isolated syllables. It is a major goal to introduce time as an explicit parameter into future versions of our model.

Appendix

Figures 8, 9, and 10 display the synaptic link weights of the *same* 25 × 25 phonetic map resulting from training the 200 most frequent syllables of our children's book corpus as

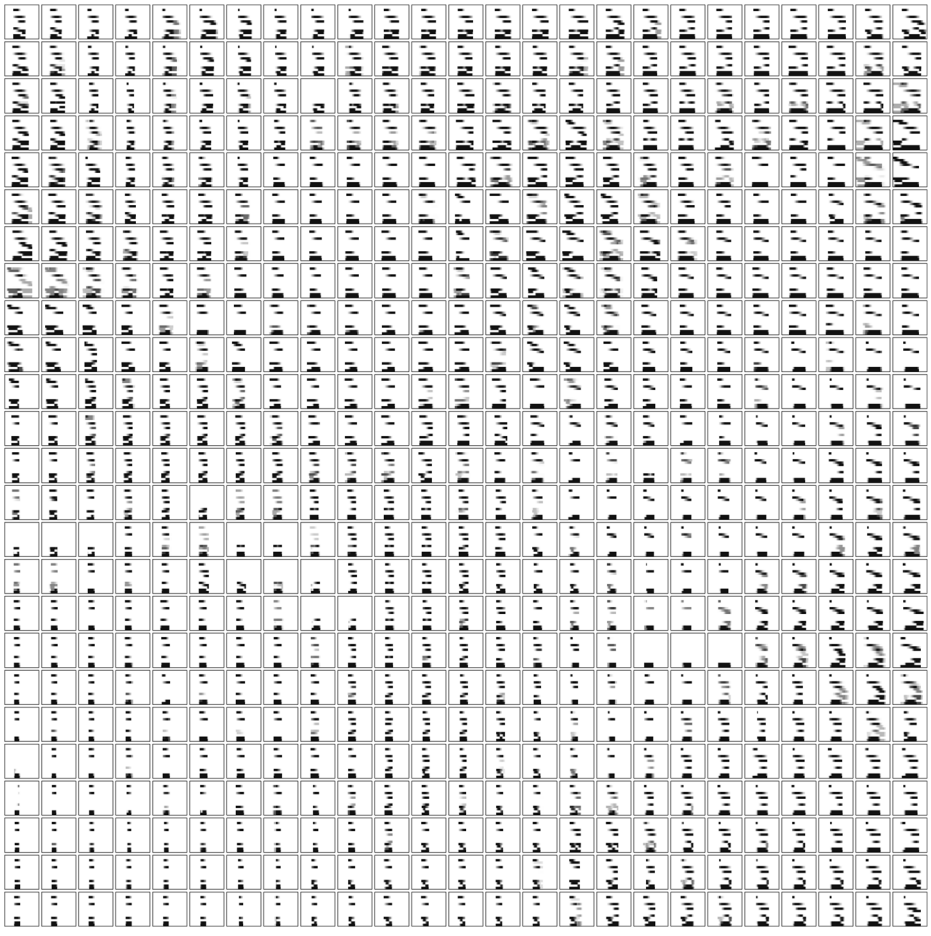

Figure 9 25 × 25 phonetic map after imitation training as described in Results section. Each model neuron within this phonetic map displays all synaptic link weights towards the motor state map (white: no connectivity, $w_{ij} = 0$; black; high connectivity, $w_{ij} = 1$). Thus, if one model neuron within the phonetic map is maximal activated, the resulting motor plan state activation pattern is displayed here in the box representing that model neuron.

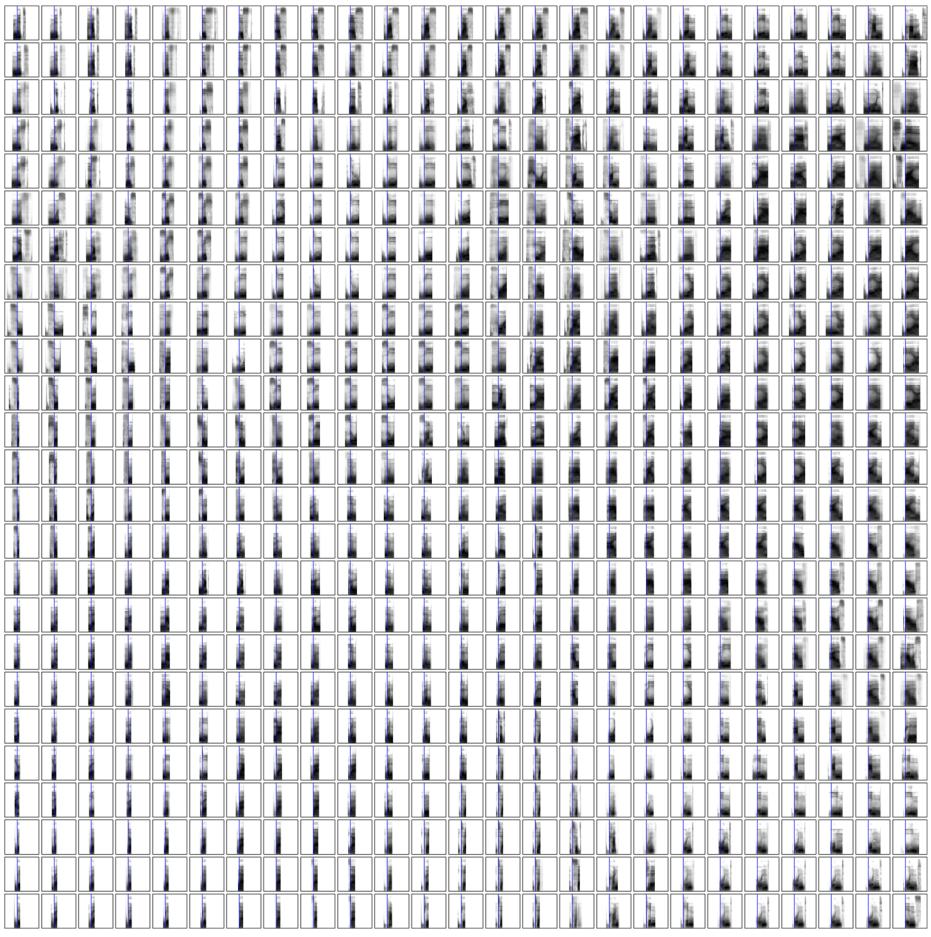

Figure 10 25 × 25 phonetic map after imitation training as described in Results section. Each model neuron within this phonetic map presents all synaptic link weights towards the auditory state map (white: no connectivity, $w_{ij} = 0$; black; high connectivity, $w_{ij} = 1$). Thus, if one neuron within the phonetic map is maximal activated, the resulting auditory state activation pattern is displayed here in the box representing that model neuron.

described in above. All three figures represent the speech knowledge stored as synaptic link weights between the phonetic map and the phonemic map (Figure 8), between the phonetic map and the motor plan state map (Figure 9), and between the phonetic map and the auditory state map (Figure 10). Thus, Figure 8 presents all model neurons within the phonetic map which represent a realization of a specific syllable (phoneme regions), Figure 1B presents the knowledge concerning the production of a phonetic realization (motor state for each realization) and Figure 1C indicates the knowledge of how that realization sounds (auditory state for each realization).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
BJK and EK developed the overall structure of the model and the design for the learning experiments. JK carried out the computer implementation and run the simulations. All authors read and approved the final manuscript.

Acknowledgements
This work was supported in part by German Research Council DFG, project number KR 1439/15-1, and in part by EU-COST action 2102. We thank Peter Birkholz for providing us with an articulatory-acoustic vocal tract model, and we thank Cornelia Eckers for lending her voice for the corpus.

Author details
[1]Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany. [2]Cognitive Computation and Applications Laboratory, School of Computer Science and Technology, Tianjin University, Tianjin, China. [3]Education and Rehabilitation of the Deaf and Hard of Hearing, Department of Special Education, Faculty of Human Sciences, University of Cologne, Cologne, Germany.

Received: 12 September 2013 Accepted: 18 December 2013
Published: 10 February 2014

References

1. Wise RJS, Greene J, Buechel C, Scott SK: **Brain regions involved in articulation.** *Lancet* 1999, **353:**1057–1061.
2. Riecker A, Mathiak K, Wildgruper D, Erb M, Hertrich I, Grodd W, Ackermann H: **fMRI reveals two distinct cerebral networks subserving speech motor control.** *Neurology* 2005, **64:**700–706.
3. Hickok G, Poeppel D: **The cortical organization of speech processing.** *Nat Rev Neurosci* 2007, **8:**393–402.
4. McClelland JL, Elman JL: **The TRACE model of speech perception.** *Cogn Psychol* 1986, **18:**1–86.
5. Li P, Farkas I, MacWhinney B: **Early lexical development in a self-organizing neural network.** *Neural Netw* 2004, **17:**1345–1362.
6. Westermann G, Miranda ER: **A new model of sensorimotor coupling in the development of speech.** *Brain Lang* 2004, **89:**393–400.
7. Guenther FH: **Cortical interaction underlying the production of speech sounds.** *J Commun Disord* 2006, **39:**350–365.
8. Garagnani M, Wennekers T, Pulvermüller F: **A neuroanatomically grounded Hebbian-learning model of attention-language interactions in the human brain.** *Eur J Neurosci* 2008, **27:**492–513.
9. Wennekers T, Garagnani M, Pulvermüller F: **Language models based on Hebbian cell assemblies.** *J Physiol Paris* 2006, **100:**16–30.
10. Guenther FH, Ghosh SS, Tourville JA: **Neural modeling and imaging of the cortical interactions underlying syllable production.** *Brain Lang* 2006, **96:**280–301.
11. Guenther FH, Vladusich T: **A neural theory of speech acquisition and production.** *J Neurolinguistics* 2012, **25:**408–422.
12. Perkell JS: **Movement goals and feedback and feedforward control mechanisms in speech production.** *J Neurolinguistics* 2012, **25:**382–407.
13. Arbib MA, Erdi P, Szentagothai J: *Neural Organization.* Cambridge, MA: The MIT Press; 1998.
14. Kuhl PK: **Early language acquisition: cracking the speech code.** *Nat Rev Neurosci* 2004, **5:**831–843.
15. Gerstner W, Kistler W: *Spiking Neuron Models.* Cambridge, UK: Cambridge University Press; 2002.
16. Kasabov N: **To spike or not to spike: A probabilistic spiking neuron model.** *Neural Netw* 2010, **23:**16–19.
17. Oberauer K, Lewandowsky S: **Modeling working memory: a computational implementation of the Time-Based Resource-Sharing theory.** *Psychon Bull Rev* 2011, **18:**10–45.
18. Bednar JA, Kelkar A, Miikkulainen R: **Scaling self-organizing maps to model large cortical networks.** *Neuroinformatics* 2004, **2:**275–301.
19. Kohonen T: **The self-organizing map.** *Proc IEEE* 1990, **78:**1464–1480.
20. Kohonen T: **Things you haven't heard about the self-organizing map.** In *Proceedings of IEEE International Conference on Neural Networks.* USA: ICNN; 1993:1147–1156.
21. Kohonen T: *Self-Organizing Maps.* 3rd edition. Berlin: Springer; 2001.
22. Kröger BJ, Birkholz P, Kannampuzha J, Kaufmann E, Neuschaefer-Rube C: **Towards the acquisition of a sensori-motor vocal tract action repository within a neural model of speech processing.** In *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues (LNCS 6800).* Edited by Esposito A, Vinciarelli A, Vicsi K, Pelachaud C, Nijholt A. Berlin, Germany: Springer; 2011:287–293.
23. Kröger BJ, Kopp S, Lowit A: **A model for production, perception, and acquisition of actions in face-to-face communication.** *Cogn Process* 2010, **11:**187–205.
24. Kröger BJ, Birkholz P, Kannampuzha J, Eckers C, Kaufmann E, Neuschaefer-Rube C: **Neurobiological interpretation of a quantitative target approximation model for speech actions.** In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011.* Edited by Kröger BJ, Birkholz P. Dresden, Germany: TUDpress; 2011:184–194.
25. Kröger BJ, Birkholz P: **A gesture-based concept for speech movement control in articulatory speech synthesis.** In *Verbal and Nonverbal Communication Behaviours (LNAI 4775).* Edited by Esposito A, Faundez-Zanuy M, Keller E, Marinaro M. Berlin: Springer; 2007:174–189.
26. Birkholz P, Jackel D, Kröger BJ: **Construction and control of a three-dimensional vocal tract model.** In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.* USA: ICASSP; 2006:873–876.
27. Birkholz P, Jackel D, Kröger BJ: **Simulation of losses due to turbulence in the time-varying vocal system.** *IEEE Transactions on Audio, Speech, and Language Processing* 2007, **15:**1218–1225.
28. Levelt WJM, Roelofs A, Meyer A: **A theory of lexical access in speech production.** *Behav Brain Sci* 1999, **22:**1–75.
29. Elman JL: **An alternative view of the mental lexicon.** *Trends Cogn Sci* 2004, **8:**301–306.
30. Cisek P: **Neural representations of motor plans, desired trajectories and controlled objects.** *Cogn Process* 2005, **6:**15–24.
31. Saltzman E, Munhall KG: **A dynamical approach to gestural patterning in speech production.** *Ecol Psychol* 1989, **1:**333–382.
32. Kröger BJ: **A gestural production model and its application to reduction in German.** *Phonetica* 1993, **50:**213–233.

33. Kröger BJ, Schröder G, Opgen-Rhein C: **A gesture-based dynamic model describing articulatory movement data.** *J Acoust Soc Am* 1995, **98:**1878–1889.
34. Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF: **Reconstructing speech from human auditory cortex.** *PLoS Biol* 2012, **10:**e1001251. doi:10.1371/journal.pbio.1001251.
35. Golfinopoulos E, Tourville JA, Guenther FH: **The integration of large-scale neural network modeling and functional brain imaging in speech motor control.** *Neuroimage* 2010, **52:**862–874.
36. Kröger BJ, Birkholz P, Lowit A: **Phonemic, sensory, and motor representations in an action-based neurocomputational model of speech production (ACT).** In *Speech Motor Control: New developments in basic and applied research.* Edited by Maassen B, Van Lieshout P. New York: Oxford: University Press; 2010:23–36.
37. Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C: **Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer.** In *Proceedings of the 9th International Conference on Spoken Language Processing.* USA: ICSLP & Interspeech 2006; 2006:565–568.
38. Li P, Zhao X, Mac Whinney B: **Dynamic self-organization and early lexical development in children.** *Cognit Sci* 2007, **31:**581–612.
39. Meltzoff AN, Moore MK: **Explaining facial imitation: a theoretical model.** *Early Development and Parenting* 1997, **6:**179–192.
40. Knapp ML, Hall JA: *Nonverbal Communication in Human Interaction.* 7th edition. Wadsworth, USA: Cengage Learning; 2010.
41. Tomasello M: *Origins of Human Communication.* Cambridge, MA: The MIT Press; 2008.
42. Kröger BJ, Birkholz P, Neuschaefer-Rube C: **Towards an articulation-based developmental robotics approach for word processing in face-to-face communication.** *PALADYN Journal of Behavioral Robotics* 2011, **2:**82–93.
43. Johnson K: **Speaker normalization in speech perception.** In *The Handbook of Speech Perception.* Edited by Pisoni DB, Remez RE. Oxford, UK: Blackwell; 2008:ch15.
44. Kröger BJ, Kannampuzha J, Neuschaefer-Rube C: **Towards a neurocomputational model of speech production and perception.** *Speech Comm* 2009, **51:**793–809.
45. Oller DK, Eilers RE: **The role of audition in infant babbling.** *Child Dev* 1988, **59:**441–449.
46. De Boysson-Bardies B, Sagart L, Durand C: **Discernible differences in the babbling of infants according to target language.** *J Child Lang* 1984, **11:**1–15.
47. Kröger BJ, Kannampuzha J, Lowit A, Neuschaefer-Rube C: **Phonetotopy within a neurocomputational model of speech production and speech acquisition.** In *Some Aspects of Speech and the Brain.* Edited by Fuchs S, Loevenbruck H, Pape D, Perrier P. Berlin: Peter Lang; 2009:59–90.
48. Kröger BJ, Miller N, Lowit A, Neuschaefer-Rube C: **Defective neural motor speech mappings as a source for apraxia of speech: Evidence from a quantitative neural model of speech processing.** In *Assessment of Motor Speech Disorders.* Edited by Lowit A, Kent R. San Diego, CA: Plural Publishing; 2011:325–346.
49. Pierrehumbert JB: **Exemplar dynamics, word frequency, lenition and contrast.** In *Frequency Effects and Emergent Grammar.* Edited by Bybee J, Hopper P. Amsterdam: John Benjamins; 2001:137–158.
50. Bauer D, Kannampuzha J, Kröger BJ: **Articulatory Speech Re-Synthesis: Profiting from natural acoustic speech data.** In *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions (LNAI 5641).* Edited by Esposito A, Vich R. Berlin: Springer; 2009:344–355.
51. Levelt WJM, Wheeldon L: **Do speakers have access to a mental syllabary?** *Cognition* 1994, **50:**239–269.
52. Plunkett K: **Lexical segmentation and vocabulary growth in early language acquisition.** *J Child Lang* 1993, **20:**43–60.
53. Hebb DO: *The Organization of Behavior.* New York: Wiley and Sons; 1949.

doi:10.1140/epjnbp15
Cite this article as: Kröger *et al.*: Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics* 2014 **2**:2.