Journal of Integrative Neuroscience

J. Integr. Neurosci. **2023**; 22(5): 124
https://doi.org/10.31083/j.jin2205124

Original Research

The NEF-SPA Approach as a Framework for Developing a Neurobiologically Inspired Spiking Neural Network Model for Speech Production

Bernd J. Kröger[1,*]

[1]Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, 52074 Aachen, Germany
*Correspondence: bernd.kroeger@rwth-aachen.de (Bernd J. Kröger)
Academic Editor: Gernot Riedel
Submitted: 23 May 2023 Revised: 14 June 2023 Accepted: 3 July 2023 Published: 16 August 2023

Abstract

Background: The computer-based simulation of the whole processing route for speech production and speech perception in a neurobiologically inspired way remains a challenge. Only a few neural based models of speech production exist, and these models either concentrate on the cognitive-linguistic component or the lower-level sensorimotor component of speech production and speech perception. Moreover, these existing models are second-generation neural network models using rate-based neuron approaches. The aim of this paper is to describe recent work developing a third-generation spiking-neuron neural network capable of modeling the whole process of speech production, including cognitive and sensorimotor components. **Methods**: Our neural model of speech production was developed within the Neural Engineering Framework (NEF), incorporating the concept of Semantic Pointer Architecture (SPA), which allows the construction of large-scale neural models of the functioning brain based on only a few essential and neurobiologically well-grounded modeling or construction elements (i.e., single spiking neuron elements, neural connections, neuron ensembles, state buffers, associative memories, modules for binding and unbinding of states, modules for time scale generation (oscillators) and ramp signal generation (integrators), modules for input signal processing, modules for action selection, etc.). **Results**: We demonstrated that this modeling approach is capable of constructing a fully functional model of speech production based on these modeling elements (i.e., biologically motivated spiking neuron micro-circuits or micro-networks). The model is capable of (i) modeling the whole processing chain of speech production and, in part, for speech perception based on leaky-integrate-and-fire spiking neurons and (ii) simulating (macroscopic) speaking behavior in a realistic way, by using neurobiologically plausible (microscopic) neural construction elements. **Conclusions**: The model presented here is a promising approach for describing speech processing in a bottom-up manner based on a set of micro-circuit neural network elements for generating a large-scale neural network. In addition, the model conforms to a top-down design, as it is available in a condensed form in box-and-arrow models based on functional imaging and electrophysiological data recruited from speech processing tasks.

Keywords: speech production; speech processing; neural network model; leaky-integrate-and-fire-neurons; spiking neuron model

1. Introduction

Different neural approaches exist for modeling speech production. These modeling approaches can be classified as cognitive-linguistic models or sensorimotor models. The cognitive-linguistic neural production model WEAVER [1–3] represents the production pathway from the intention to produce an utterance down to the activation of its phonological form. The sensorimotor neural production models DIVA [4] and GODIVA [5] represent further components of the production pathway from the phonological form via activation of motor programs down to motor execution and include auditory and somatosensory feedback to allow motor corrections. These models use non-spiking rate-based neuron models (second-generation neuronal networks [6]). The basic operation unit here is a node, representing an averaged neural activity rate (time window, 5 ms to 25 ms) as it could be generated by an ensemble of neighboring spiking neurons. These nodes are connected by edges or links characterized by specific link weights. Time is introduced in these models indirectly by defining discrete time steps (up to 5 ms) and by defining neural activation decay constants for the nodes.

The architecture of the production-perception model WEAVER [1–3] comprises different neural layers (layers of nodes) for activating concepts, lemmata, phonological forms, syllables, and phonemes. The neural links between these neural layers allow activation spreading, while connections within specific neural layers allow inhibition processes so that here, exclusively, one node is activated per layer and per time step (winner node or winner neuron). The model simulates activation spreading top-down as well as bottom-up, e.g., from concept via lemma to phonological form level and from phonological form via lemma to concept level. These spreading activation processes are always active; top-down activation spreading thus also plays a role during speech perception and bottom-up activation spreading processes are also active during speech production.

The sensorimotor production models DIVA and GODIVA allow the generation of articulatory patterns (speech

articulator movements) from the phonological specifications of an utterance. The DIVA model [4] starts with activating motor programs followed by the generation of muscle activation patterns for direct articulatory execution (top-down motor signal generation), while the expected auditory and somatosensory states (sensory expectations) are compared with the actual sensory states in order to generate potential sensory error signals (difference between estimated and produced sensory states) followed by a potential activation of correction signals that are used as feedback signals for the top-down motor signal generation. The GODIVA model [5], in addition, provides a motor planning module for selecting bunches of phonological sound sequence for which motor programs are directly available (already learned) or can be generated quickly on the basis of available sensorimotor knowledge. This approach includes a cortico-basal ganglia-thalamus-cortical loop which is involved in motor planning and motor execution, but the model still uses non-spiking neuron models (nodes and links) for modeling neural activity (second-generation neural network [6]).

The nodes of the adaptive neural networks building the DIVA and GODIVA models are called cells or neurons. Both models are composed of modules, each exhibiting one central neural map, i.e., a layer or set of cells. Each of these layers represents a specific state, e.g., motor program state, auditory or somatosensory state, or primary motor state. The neural connections between modules are called mappings, and each mapping represents a specific transformation of one neural state activated in one map to another neural state activated in the next neural map. Both approaches (DIVA and GODIVA) include the modeling of time for (i) representing the time-flow of speech items at the auditory, somatosensory, and motor program level; (ii) executing a sequence of speech items, where a GO signal was used in early versions of the DIVA model [7], later replaced by the timing modeled within the cortico-cortical control loop approach including the basal ganglia and thalamus (motor loop, see [5], p. 1517); and (iii) processing motor planning using a time-dependent competitive queuing approach (planning loops).

Furthermore, the comprehensive model of speech processing [8,9] uses spiking neuron models and thus represents a third-generation neuronal network (e.g., [6,10]) developed by applying the Neural Engineering Framework and Semantic Pointer Architecture (NEF-SPA) modeling framework [11,12], which allows the modeling of temporal processing in a more straightforward way by using spiking neurons and not classical connectionist rate-based, node-edge models (second-generation neuronal networks). Spiking neurons control basic temporal parameters defining the rapidity for increasing membrane potential and thus the time span for reaching the firing threshold before a post-synaptic spike is generated. This defines the latencies and processing time intervals of all functional processes of neural micro-circuits and all neurofunctional processes of the large-scale neural network. Moreover, in contrast to other neural simulation toolboxes (e.g., NEURON [13], NEST [14], BRIAN [15]), the NEF-SPA approach provides modeling elements not only at the neuron-level for simulating the flow of neural activation within neural micro-circuits (as described in the Neural Engineering Framework (NEF) component of this approach, see [11]), but also provides modeling elements for higher-level cognitive processing (the Semantic Pointer Architecture (SPA) component of the approach, see [12]) for constructing large-scale neural models capable of modeling higher-level cognitive functions and higher-level sensorimotor functions, and thus developing brain-scale (also called large-scale) neural models. The SPA is built on top of the NEF and allows the modeling of higher-level cognitive processing, which is needed in speech processing for representing higher-level auditory, motor, or somatosensory states of syllables and cognitive states of phonological or semantic forms of words. The repertoire of neural modeling elements or building blocks for constructing a brain-scale neural model as it is needed for simulating speech production is described in the following sections.

In section 2, the repertoire of neural modeling elements is listed, and the neurofunctional modeling approaches of the NEF (section 2.1) and the SPA (section 2.2) are introduced. In section 3, the functional organization of the speech production network model is introduced (section 3.1), typical simulation scenarios are described (section 3.2), and the implementation of a new component of the model, i.e., the preparation and execution of syllable sequences, is introduced (section 3.3). Section 4 provides an overview of model performance during different speech production and some speech perception tasks. In section 5, the modeling approach is discussed with respect to other research endeavors in the fields of speech production and speech perception.

2. Method: Neural Elements of the NEF-SPA Approach for Constructing Neural Models

The NEF [11,16] allows the creation of complex brain-scale models. The approach uses a limited set of basic neural network elements (e.g., neuron ensembles, neural connections, neural oscillators, neural buffers, neural associative memories, neural gaiting elements, neural ramp function generators) for constructing the neural network modules (e.g., modules for cognitive processing, auditory processing, sensorimotor processing, action selection, etc.) of the brain-scale or large-scale neural network. The basic neuron model here is the leaky-integrate-and-fire (LIF) neuron model. This model allows the generation of neural spike trains dependent upon neural input activity (input spikes) reaching the neuron's synapses. While the NEF introduces basic neural elements for signal representation, signal transformation, and the generation of neural oscillations (mainly

neuron ensembles and neuron connections between ensembles), the number of neural elements is extended by the SPA in order to be able to model complex cognitive neural processes in a straightforward manner. Typical SPA elements are neuron state buffers, associative memories, and the elements constructing the action selection cortico-cortical loop. These NEF and SPA elements allow the construction of all modules needed for generating a large-scale network, e.g., for simulating a wide variety of tasks such as recognizing digits or letters, memorizing a sequence of digits or letters, decision-making, and reacting by performing hand-arm actions such as writing [11].

The NEF-SPA framework is implemented as a Python package called Nengo [Neural Engineering Objects [11,17], version 3.2.0, see (https://www.nengo.ai/)] that assembles the basic elements of each neural network model. All Nengo elements that are needed for constructing a speech-processing neural model are summarized in **Supplementary Material A**. The corresponing Python-Nengo source code is given in form of iPython-Notebooks in **Supplementary Material B**.

2.1 NEF Principles and NEF Elements

Three basic principles (representation, transformation, dynamics) and three basic construction elements (neuron, neuron ensemble, neural connection) characterize the NEF [11] (see **Supplementary Material A1**):

(i) Representation: signals (e.g., values of sensory parameters, such as the intensity of an auditory or visual signal, or motor parameters such as the intensity of a neuromuscular activity for the strengthening of a specific muscle) can be coded as (time-dependent) neural states and, vice versa, neural states can be decoded as signals. These neural states are represented by the neural activity (spike patterns) of a set of neurons, called neuron ensembles (see Fig. 1a). Each neural ensemble consists of a specific number (N) of leaky integrate-and-fire neurons (LIF neurons). Typically, neuron ensembles consist of N = 20–100 neurons to represent a signal with sufficient precision.

(ii) Transformation: a neural connection of a neuron ensemble ens_A with a neuron ensemble ens_B, by connecting each neuron from ens_A with each neuron from ens_B, allows a transformation (or modification) of neural states from ens_A to ens_B (see Fig. 1b). Neural connections between neuron ensembles exhibit N $\times$ N synaptic units with variable synaptic weight values. This allows the modeling of a wide range of transformations (i.e., mathematical functions y = f(x), where x and y represent the neural activity of the neuron ensembles ens_A and ens_B). Besides these neuron ensemble connections, single neural connections can also be defined from neuron 1 to neuron 2 (see **Supplementary Material A1**).

(iii) Dynamics: A neuron ensemble that comprises recurrent neural connections (i.e., neural connections that start and end at neurons of the same neuron ensemble ens_A) is able to simulate dynamic neural processes such as (i) oscillation of neural activation (ensemble ens_A as neural oscillator, Fig. 1c) or (ii) building (integrating) and maintaining a neural activation pattern over a specific time period (ens_A as an integrator or short-term memory for ramp-signals, Fig. 1d). The type of neural transformation, oscillation, or integration depends on the synaptic weights of the recurrent connections.

2.2 Semantic Pointer Architecture and SPA-Elements

Additional principles for coding and transforming cognitive states lead to further neural elements or model construction elements such as state buffers, associative memories, binding and unbinding buffers, neural circuits for evaluating state similarity (dot-products), and S-pointer-networks, etc. (see **Supplementary Material A1**). While the NEF mainly regulates the processing of lower-level states (e.g., states describing parameters such as signal amplitude or frequency in case of a simple sensory input signal, e.g., an auditory signal), the SPA allows the processing of more complex items such as auditory processing of whole syllables or words, or such as visual processing of a figure or an object. Thus, the SPA is an important concept for modeling, especially the higher-level cognitive and sensorimotor processing of speech.

(i) Neural state buffers: for neural modeling of complex cognitive or higher-level sensory and/or motor processing, the NEF has been augmented by the SPA [11,12,18]. Neural states are defined here in the form of semantic pointers (S-pointers). The neural activity associated with an S-pointer can be realized in neural state buffers. S-pointers and their neural activity pattern represent or point to cognitive states or on mental objects, i.e., mental representations of concrete objects (e.g., things, persons, animals), abstract objects (e.g., thoughts, categories for things or creatures, intentions, emotions), or high-level sensory or motor patterns (e.g., a visually perceived object or creature, a motor act such as grasping, writing a letter, or articulating a word). In the case of modeling speech processing, cognitive-linguistic states play a major role. Thus, words as concepts or lemmata; syllables as phonological forms, gesture scores, or motor programs; and other linguistic items such as phonemes or features etc., can also be represented in the form of S-pointers. Mathematically, S-pointers are D-dimensional vectors (typically D = 512 for coding an entire lexicon of a particular language, see [19]). Therefore, a neural state buffer consists of D neuron ensembles, and each of the D values of the vector representing an S-pointer is coded as a neural activity within one neuron ensemble.

(ii) The (time-dependent) neural activity within a neural state buffer can be visualized by similarity plots. A similarity plot displays the dot-product of the current neural activity with each S-pointer and thus represents the neural activity of a buffer in terms of its degree of how strong the neural activity within a buffer is represented by each already

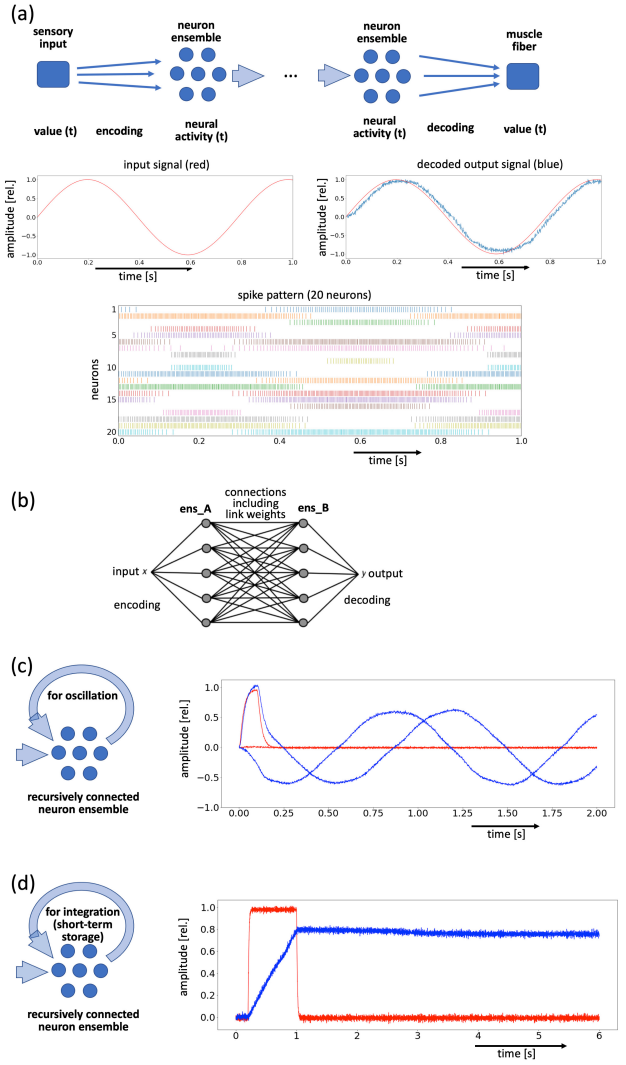

Fig. 1. Graphic representation of Neural Engineering Framework (NEF) elements. (a) Neuron ensembles (N = 20) for coding a sensory input signal (red) as a spike pattern and for decoding the spike pattern as an output signal (blue). (b) Neural connections between two neuron ensembles (ens_A and ens_B). (c) Recursively connected neuron ensemble for generating oscillatory signals (two values are coded within this neuron ensemble by dividing the ensemble into two neural subpopulations; the input signal pulse (red) acts on one subpopulation only and triggers the oscillator). (d) Recursively connected integrator neuron ensemble for generating a nearly constant hold signal; a ramp-signal appears as long as a (constant) input activation is feeding this integrator neuron ensemble (red: input; blue: decoded output). The input signal is also called a trigger pulse or trigger signal (see section 3.3: trigger pulse ensembles feed gesture ramp signal integrators).

defined S-pointer, i.e., by each S-pointer already representing a specific item or mental object (see Fig. 2a,b). The dot-product is a scalar value and can be calculated for each pair of S-pointers by using a specific small subnetwork (similarity between states network, also called similarity network) starting from two buffers carrying the neural activity representing the two S-pointers and leading towards a neuron ensemble carrying the resulting scalar. Such a subnetwork is not only used for decoding purposes, but these similarity networks, e.g., appear within the cortico-striatal component of the action selection network for evaluating the potential utility of all potentially applicable (pre-selected) actions (see below).

(iii) Short-term memories: if a neural state buffer is augmented by recursive neural connections, the buffer can represent a short-term memory (see also: recursively connected neuron ensembles). Here, a neural activity (an S-pointer) can be held for a longer time interval (Fig. 2c).

(iv) Knowledge-repositories: long-term knowledge is represented by a set of S-pointers in the SPA. Many S-pointers represent mental objects that are learned during the lifetime, e.g., during the speech and language acquisition period. These objects, as far as they can be concretized by a word, are stored as lexical items in their phonological form, lemma form, and semantic form in the mental lexicon. Thus, vocabularies can be defined directly in the SPA as sets of S-pointers for different levels of word representations in the mental syllabary. Therefore, the SPA concept allows a neurofunctional interpretable definition of lexical levels as introduced in [20]. While the associations between phonological form, lemma, and semantic form (also called concept) are established by associative memories (see below), the relations of items within each level, e.g., the relations of concepts such as “cat” and “dog” which is “animals”, or “red” and “green” which is “colors”, are defined by relation-S-pointers and the S-pointers, and the appertaining set of S-pointers needs to be saved within an S-pointer network. Thus, concepts, lemmata, and phonological forms build three S-pointer networks. In each network, relation-S-pointers are defined (concept level: “belong to concept category”, e.g., “dog” and “cat” belong to concept category “animal” while “red” and “green” belong to the concept category “colors”; lemma level: “belong to word category”, e.g., “dog” and “cat” belong to the word category “noun” while “to bark” and “to meow” belong to word category “verb”; phonological form level: “dog” and “dark” “belong to phonological forms starting with the same consonant /d/”). These three different S-pointer networks exist as part of the mental lexicon, i.e., concept network, lemma network, and network of phonological forms (see [21,22]). The degree of similarity of items within each S-pointer network (concepts, lemmata, phonological forms) can be estimated by calculating the dot-products between all pairs of S-pointers within each S-pointer network. As already stated above, the most active S-pointers in a neural state buffer and their similarity to other S-pointers can be visualized by using similarity plots (see Fig. 2a for a set of S-pointers without any S-pointer relations and Fig. 2b, which includes S-pointer relations).

(v) Neural connections can be realized in the SPA approach by direct connections between buffers if the neuronal information needs only to be passed from buffer to

buffer. This type of neural connection represents the simplest form of neural processing. Neural connections that transform neural states, e.g., through the conversion of an S-pointer A (e.g., concept or lemma of a word) into an S-pointer B (e.g., a phonological form of the same word), require the interposition of an associative memory (see Fig. 2d). While Fig. 2d shows an associative memory directly transforming concepts into phonological forms, normally, two associative memories realize the lexical transitions, i.e., firstly, a transformation from concepts to lemmata and, secondly, a transformation from lemmata to phonological forms as part of the lexical processing within the production pathway, and two further associative memories realize the vice versa transition from phonological forms to concepts as part of the lexical processing within the perception pathway [8,17,21].

Thus, associative memories (such as S-pointer networks) include long-term knowledge and can be labeled as long-term memories. It should be noted that while associative memories allow long-term storage of relations between different sets of S-pointers, i.e., the relations between different lexical levels, such as between concepts and lemmata or between lemmata and phonological forms, S-pointer networks allow long-term storage of relations between S-pointers or items within a lexical level, e.g., semantic relations (e.g., "is an animal" or "is a color"), word relations (e.g., "is a noun" or "is a verb"), or phonological relations (e.g., "starts with the sound /d/" or "starts with the sound /b/").

(vi) A further type of neural transformation that is used in the SPA context is the binding and unbinding of S-pointers. Binding and unbinding processes allow short-term storage of relations between S-pointers, e.g., binding a specific object with a specific color currently of interest: "blue ball" or "red ball" (e.g., SP_Red * SP_Ball -> SP_currentObject). Unbinding processes can also be implemented, e.g., to find the answer to questions such as "what is the color of a currently focused ball?" (e.g., SP_currentObject $*^{-1}$ SP_Ball -> SP_Red). Binding and unbinding processes are implemented by binding networks (see Fig. 2e) and can be used especially in S-pointer networks, e.g., to activate all objects belonging to a specific group of objects (e.g., all "fruits") or to extract a specific feature of an object, e.g., "apple is a fruit"; see [8]).

All elements introduced here can be interpreted as building blocks for the development of neural network models and cover elementary neurobiological functions such as holding and forwarding neural states, transforming and processing neural states, and storing states and relations between states, etc. These basic functional bundles of neurons and functional micro-circuits are highly optimized, and thus it can be hypothesized that, somehow, similar natural bundles of neurons and similar neural micro-circuits fulfilling these neural functions as defined above could be a result of evolutionary processes, even if their specific anatomical organization in the brain is different to the highly idealized organization of neurons and neural micro-circuits currently proposed in the NEF-SPA approach.

(vii) Thus far, we have mentioned neural processes such as forwarding and modifying neural activity from buffer to buffer. These neural processes are implemented as network elements defining local networks (or neural microcircuits) consisting of input and output buffers as well as processing buffers such as associative memories or binding and unbinding buffers. The selection and activation of a neural process is an action. The action selection process is realized by the cortico-cortical feedback loop, including a model of the basal ganglia and thalamus. This control loop is a component of the SPA (see [11,12,23,24]). Action selection (e.g., calculation of utility values for available actions) is realized within a cortex to basal ganglia neural network and by further subprocesses appearing within the basal ganglia. The activation for the execution of an action is a neural disinhibition process taking place in the thalamus. Actions are coded by action S-pointers.

Depending on the scenario to be simulated, a set of actions needs to be defined (listen_and_hold, extract_meaning, form_answer, produce_answer, etc.; see [8,21,22]). At each point in time, i.e., for each internal and environmental state of the model, a utility value is estimated for each potential action by activating the dot product between specific S-pointers representing potential actions and the current state of the model. The action which exhibits the highest utility value at a specific point in time (i.e., in a specific situation the model is exposed to) is selected, i.e., disinhibited at the thalamus. It should be stated that the delay time that is needed for action selection appears as a natural byproduct in our spiking neuron model. It results from the detailed neurofunctional modeling of the basal ganglia and thalamus. This delay time (decision time) is normally approximately 50 ms, but can be longer if two utility values are of comparable magnitude, i.e., if the model benefits nearly to the same degree from two different actions in a specific situation.

3. Results I: The Implemented NEF-SPA-Model for Speech Production

3.1 Model Architecture

Speech processing is a component of the cognitive and sensorimotor system of humans and comprises speech perception (i.e., listening and comprehension) and speech production (i.e., utterance formulation and articulatory execution). The architecture of a neurobiologically based model of speech production comprises a cognitive-linguistic and a sensorimotor component (Fig. 3; and see [8,9]). Both components of the model include knowledge repositories, i.e., the mental lexicon and the mental syllabary. The mental lexicon contains the semantic, grammatical, and phonological information of all words already learned and known by the model and thus allows the processing of meanings to-

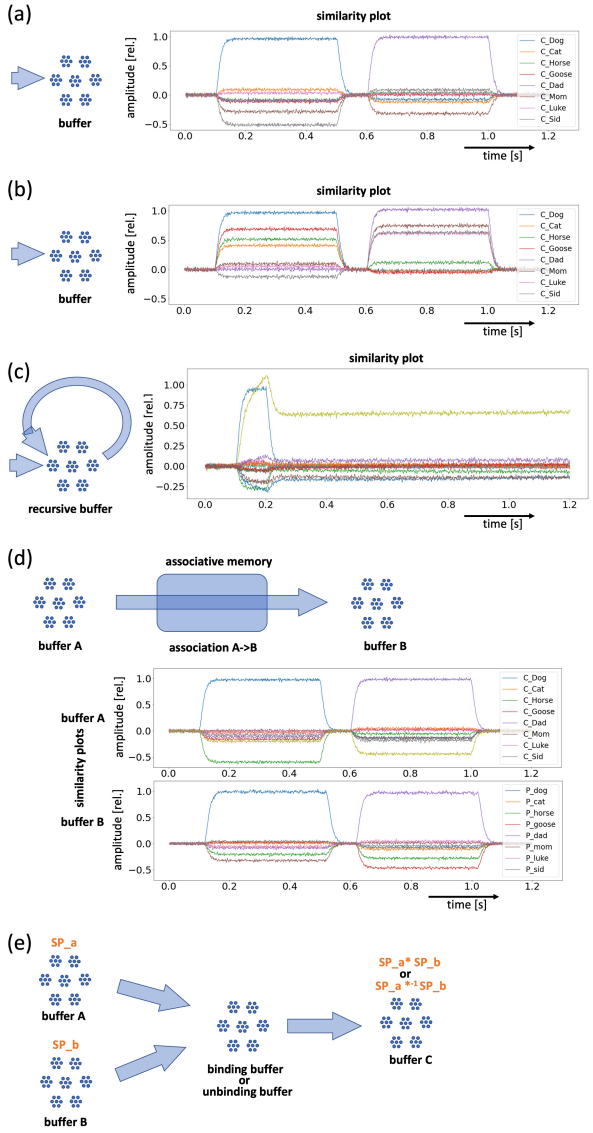

Fig. 2. Graphic representation of Semantic Pointer Architecture (SPA) elements. (a) Neural state buffer for hosting a neural activity pattern over time; the decoding of activity appearing within a neural state buffer can be performed using similarity networks and by displaying the result in similarity plots (the set of S-pointers used here comprises C_Dog, C_Cat, C_Horse, C_Goose, C_Dad, C_Mom, C_Luke, C_Sid; C_etc. indicates that these S-pointers represent the concept or meaning of the word). (b) Same as (a), but the set of S-pointers now includes S-pointer relations – the set of S-pointers is implemented here as S-pointer network including the categories C_Group_Animal and C_Group_Human). (c) Short-term memory (recurrently connected neural state buffer holding an input activity over a longer time interval – the input neural activity is displayed in blue). (d) Associative memory for transforming neural activity of buffer A (neural activity of word concepts C_Dog and C_Dad) into the neural activity of buffer B (neural activity of phonological forms of words P_dog and P_dad). (e) Binding network including input and result in buffers and binding buffer for binding process SP_a * SP_b -> SP_c (the same network structure is used for unbinding processes SP_a $*^{-1}$ SP_b -> SP_c; buffer activities are displayed in light red; here, SP means S-pointer).

wards phonological forms as a component of speech production for further sensorimotor processing within the production pathway. The vice versa process, i.e., processing of phonological forms towards meanings based on an activation of phonological forms from sensory processing, is part of the perception-comprehension pathway (here, simply called the perception pathway, see Fig. 3). The mental syllabary contains phonological forms, motor plans, and motor programs and thus allows the processing of phonological forms towards motor programs that are ready to be articulated. Furthermore, the sensorimotor component allows the processing of auditory forms, which results in activation of phonological forms in order to allow a higher level of cognitive-lexical processing within the perception pathway. The sensorimotor component of the perception pathway constitutes the feedback pathway during speech production and thus plays an important role, especially during speech learning [9]. A grammatical component for sentence generation is not part of our model thus far, and the tasks (scenarios) performed by the models based on this architecture are mainly word production, syllable production, and word comprehension tasks.

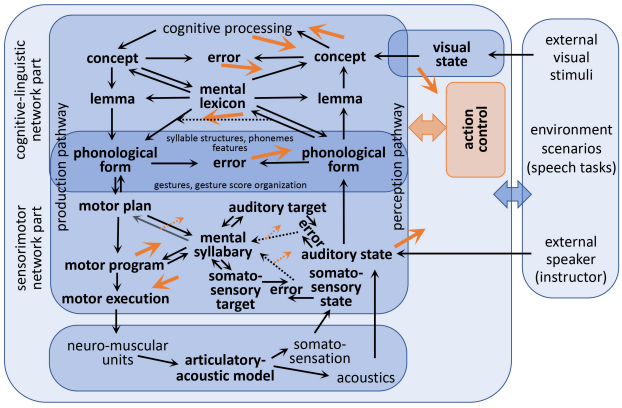

Fig. 3. Architecture of a neurobiologically based model of speech processing following and including additions from current further model development. Text between arrows indicate SPA-buffers for state representation. Solid black arrows indicate neural pathways for forwarding and processing states. Dashed black arrows in the sensorimotor network part indicate the transfer of learning results towards mental syllabary. Dashed black arrow in the cognitive-linguistic network part represents direct forwarding of phonological input on perception side to output on production side (shortcut). Orange arrows indicate neural pathways of cortico-cortical loop for action selection (including basal ganglia and thalamus). Reproduced with permission from Bernd J. Kröger, published by Frontiers [8,9].

The mental lexicon associates three different levels of word representations, i.e., the concept, lemma, and phonological form level. All words of a language are represented by an S-pointer at each of these three levels leading to three sets of S-pointers. Each concept, lemma, or phonological form of a word can be activated in a corresponding neural

state buffer (i.e., three buffers on the perception and three buffers on the production side of the model; see Fig. 3). Four associative memories allow the transformation of S-pointers from level to level on the production as well as on the perception side (indicated by arrows in Fig. 3). While these associative memories carry the knowledge of how lexical forms are associated with each other from level to level (i.e., between levels), the sets of S-pointers representing each of the three lexical levels is further laid out as an S-pointer network in order to include knowledge concerning associations of words within each level (by using relation S-pointers). Activation of lexical forms in speech production can start from the activation of auditory forms in the case of verbally presented words or word repetition tasks, or from the activation of visual forms in the case of picture naming tasks.

The phonological level interfaces the cognitive-linguistic component and sensorimotor components of the model. At this level, phonological forms of syllables can be activated basically in the form of phoneme sequences, but a coactivation appears in the form of a raw gesture score that is forwarded to the motor plan level (see Fig. 3) and, in addition, S-pointer relations exist for which the similarity relations are based on syllable type and on the type of phoneme or gesture that appear in a specific position of a syllable (for the definition of syllables and gestures see [9]).

The sensorimotor component of the model comprises the mental syllabary as a knowledge and skill repository concerning motor and sensory states of all frequent syllables, and it comprises different buffers for hosting motor programs, and auditory and somatosensory states of activated syllables. Sets of S-pointers exist for motor plans, motor programs, and auditory and sensory motor states for all learned syllables. Associative memories exist for the motor plan to motor program association and for the auditory state to phonological form association (arrows in Fig. 3). Moreover, at the motor plan level, each motor plan S-pointer is associated with S-pointers describing components of the syllable, i.e., syllable onset, syllable center, and syllable offset. In addition, each of these components of syllable S-pointers is associated with S-pointers describing the types of gestures (i.e., consonantal, vocalic, velopharyngeal, and glottal gestures; see Fig. 4). The buffers and associative memories describing this structure of each motor plan are the basis for quantifying gesture parameters at the level of motor programs. At the motor program level, gestures are no longer specified only phonologically as raw gestures, but now in a concrete phonetic way as articulator movement shapes (see [9]). This quantification allows the activation of motor plan execution (see Fig. 4). While the dimensionality of S-pointers and the number of neuron ensembles making up the associated neural state buffers is high (D = 512) at the cognitive-linguistic level (here, approximately 60,000 items need to be stored for a language), the dimensionality is low (D = 32) for gestures because the number of gestures appearing in a language is not more than approximately 50 to 100 gestures.

The specification of raw gesture scores towards phonetic gesture scores (fully specified gesture scores that are ready for execution) is carried out by activating syllable parameter values and gesture parameter values (one neuron ensemble holds one parameter; parameters for each gesture appearing in a syllable: gesture onset time, gesture offset time, both relative to syllable oscillation cycle, gesture target value; all parameters are scaled in relative values 0–1). The parameter values are learned and stored in the mental syllabary for each syllable trained during speech acquisition. If a motor program is not available, parameter values are assembled from subsyllabic units of phonologically similar syllables (see control action in the form of a dashed red arrow between a mental syllabary and motor plan level in Fig. 3). The motor program now allows a concrete activation of neuromuscular units for the articulatory realization of the activated syllable [25].

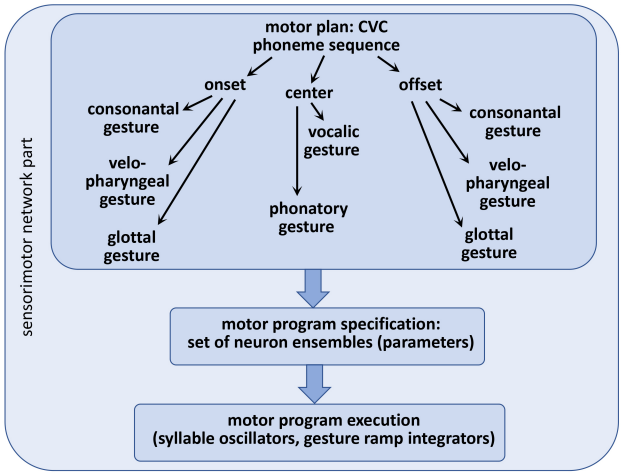

Fig. 4. Structure of neural subnetwork for motor preparation and execution: motor plans (case CVC syllables) and motor programs, including motor program execution. CVC, consonant-vowel-consonant.

The S-pointers representing auditory states comprise 24 × 64 dimensions (D = 1536), where 24 rows of 64 neurons represent the frequency scale (bark scaled center frequencies) and 64 columns of 24 neurons represent a time scale (time steps of 10 ms, [26]). The S-pointers representing somatosensory states comprise 4 × 64 dimensions (D = 256), where 4 rows of 64 neurons represent the relative articulator to articulator distance for lips, articulator to vocal tract wall distance for tongue tip and tongue dorsum, and a relative displacement value for the lower jaw [26]. The appropriate neural state buffers (Fig. 3) are of the same dimensionality. Two neural state buffers exist for activating learned auditory and somatosensory targets from mental syllabary. A dot-product comparison can be made for

S-pointers of sensory target and current state buffers (i.e., S-pointer similarity calculation) in order to estimate the current difference (error) appearing between stored auditory and somatosensory target states and those appearing from current articulatory execution (Fig. 3). The result of this sensorimotor level monitoring process is evaluated by the action control component and stops or halts further articulatory execution in case of a too large difference of states (a dot-product greater than 0.3 for auditory or somatosensory similarity; see dashed arrows starting from blue shaded error buffers within the sensorimotor component of the model in Fig. 3). These halts force the model to repeat the current syllable or word, allowing for improvement, thus updating the motor program parameters of the currently mispronounced syllable or word.

Visual processing, as well as auditory processing, is not included in the current version of the model. It is assumed that all words that are verbally presented in the form of an acoustic-auditory speech item or that are visually presented in the form of a picture are correctly recognized and correctly activate the word's phonological realizations.

Action control (red shaded box in Fig. 3) gets its input from many buffers or components of the cognitive-linguistic as well as the sensorimotor component of the speech processing model and, vice versa, from action control to these buffers of the cognitive-linguistic and sensorimotor component of the model (see red arrows in Fig. 3). One typical action selection process is the already mentioned process of interrupting the word or syllable production process by information based on S-pointer similarity calculation of feedback sensory states and stored sensory states. In this case, the execution of a syllable or word will be repeated with slightly modified parameter values. Action control also triggers the production process of a syllable or word as a result of the activation of auditory or visual input. The activation of lexical buffers in the perception and the production pathway appears in a straightforward way and results directly from the existing neural connections between these buffers. The only influence of action control here is a phonological and semantic monitoring process as implemented in our model for activating halt actions during speech production in case of discrepancies between phonological or semantic states between the production and feedback perception pathways [21]. Besides the linguistic-cognitive components, all sensorimotor components shown in Fig. 3 are cortical components. Subcortical neural activity (muscular, sensory, and other peripheral components) is included with the articulatory-acoustic model level in Fig. 3. Thus, the information needed for action control is generated in different cortical buffers or components. These buffers or components are marked by red arrows in Fig. 3, which are directed towards the action control component. Dashed red arrows indicate the initiation of learning actions.

Cognitive processing (see the top module in Fig. 3) depends on the task to be performed by the model (see below). The tasks mentioned below mainly forward the concept activated during perception towards the production pathway. Complex cognitive processing needs to be modeled only in some of the tasks mentioned below (e.g., a task involving verbal cues) [22,27].

The basic elements or building blocks defined in the NEF-SPA approach (see section 2 of this paper) can be used not only to develop speech processing models; in particular, the cognitive component of our large-scale neural network model could be extended in order to perform other tasks such as those described for the Semantic Pointer Architecture Unified Network (SPAUN) model [28]. This model is capable of performing tasks such as copy drawing, digit recognition, memorizing lists of objects, counting, question answering, and decision making, etc. Our model could be included as a part of the SPAUN model because the basic architecture of our model conforms to that of the SPAUN model. In that case we only have to add the mental lexicon and mental syllabary, as developed in the framework of our large-scale model, as well as an artificial ear, auditory input processing, and an artificial articulatory model for producing articulatory-acoustic speech output.

3.2 Simulating Different Speech Production Tasks

Different scenarios or tasks for speech production and speech perception (as far as it is closely connected with word production and word comprehension) are listed in Table 1 (Ref. [8,9,21,22,27,29,30]). These scenarios or tasks have already been simulated and discussed in earlier publications (see Table 1), but have never been collated in one paper together with an overview of the model architecture.

In the case of picture naming, the model is primed to name an object displayed on a picture that is exposed to the model (to the test person). During this task, the visual perception pathway and the speech production pathway are open. If the picture is exposed to the model, the model automatically selects a concept that is forwarded to the cognitive processing component. This automatically activates the top-down process of word production [8,21]. It should be kept in mind that the top-down production process from activation of concepts via lemmata to phonological forms, as well as the bottom-up perception process from phonological forms via lemmata towards concepts, are directly activated if a concept is activated in the production pathway or if a phonological form is activated in the perception pathway. No further action selection process is needed here from the action control component (if the model is primed on the task "picture naming" or "word repetition").

In the case of a picture naming task including distractor words, a distractor word is presented following onset of the picture presentation. The distractor word is phonologically or semantically similar, but not identical, to the target word displayed as a picture [21]. The distractor word

Table 1. Involvement of the action control component in the execution of different tasks.

Task (scenario)	Control module gets input from	Control module acts on	Reference
Picture naming (PN)	Visual state buffer	Cognitive processing: allow an already activated concept to pass	[8]
PN with distractor words (halt scenario)	Visual state buffer and auditory state buffer; phono and concept error estimation	Cognitive processing as in PN; motor execution: can be stopped if an error signal appears at phono or concept level	[21]
PN with retrieval aids (second/third trial scenario)	Visual state buffer and auditory state buffer	First trial: cognitive processing as in PN; second and third trials: cognitive processing by suppressing the direct cue (activating co-activated concepts only)	[8,22]
Single word repetition	Auditory state buffer	Cognitive processing as in PN	[8]
Single logatome repetition	Auditory state buffer	Cognitive processing as in PN; open the direct pathway at the phonological form level	[8]
Word list repetition (serial/free recall task)	Auditory state buffer	Cognitive processing: short-term storage, binding and unbinding processes	[27]
Syllable repetition (multiple times: diadochokinesis)	Motor program buffer (all production tasks)	Motor execution	[29,30]
Word comprehension (auditory input)	Auditory state buffer	Cognitive processing as in PN; in addition: stop with concept activation	[8]
Word learning	Auditory state buffer and visual state buffer	Cognitive processing: simultaneous activation of visual and auditory input	[9] (partial modeling)
Syllable learning	Auditory error buffer and somatosensory error buffer	Cognitive processing: simultaneous activation of motor, auditory, and somatosensory states	[9] (partial modeling)
Syllable execution (as part of all tasks)	Motor program buffer (all production tasks)	Motor execution	[9] and this paper

is verbally presented by the task instructor and thus is activated in the auditory perception pathway during the visually induced lexical retrieval process and word production process of the visually presented target word. It is known that such auditory input as produced by distractor words may stop target word production processes because an inner speech loop permanently monitors the speech production process by inner perception, i.e., by evaluating the similarity between neural activations within the production and perception pathways at different levels (conceptual level and phonological level, see [21,31]). Thus, distractor words may activate a stop or halt action during the word production process if the distractor word is identified as an incorrect auditory feedback signal of the current word production process [21].

A picture naming task, including semantic or phonological cues is performed as a sequence of word production tasks [8,22,27]. The first word production task or trial is initiated by picture presentation (normal picture naming task). The second and third task or trial, which directly follow the first trial, is the same task but, in addition, a phonological or semantic cue is given verbally by the task instructor (e.g., in order to find the target word "bike", the retrieval aids can be: "the word we are looking for starts with the sound /b/" or "the object we are looking for has two wheels"). The second and third trial are only initiated if word production does not lead to a correct result in the preceding task or trial. The activation of the cue in the second or third trial leads to a co-activation of specific phonological or semantic features of the target word, which now increases the activation of the target word for potential lexical selection at the beginning of the production pathway. Thus, in the case of this task comprising up to three trials, it is important to ensure that the concept and phonological S-pointer networks include all similarity relations [22,27].

Single word repetition is similar to the picture naming task, but now the input is activated at auditory input leading to phonological form activation and further to lemma and concept activation within the perception-comprehension pathway [8]. Visual perception directly leads to activation at the concept level of the perception-comprehension pathway in the case of our model.

Word list repetition starts with auditory input activation leading to concept activation at the end of the perception-comprehension pathway. Because the task now is word list repetition, the action control component is

primed in a different way by firstly keeping the list of words and then reproducing the list of words as far as the model was able to memorize the word list [27]. Thus, cognitive processing is more complex in the case of this task in comparison to a single word repetition task.

Logatome repetition does not activate the lemma and concept level of the mental lexicon because a logatome is defined as a nonsense word or non-word (i.e., a syllable or a sequence of syllables that is not included at the lemma and concept level of the mental lexicon of the target language). In this case, the action control component directly activates a neural connection (a shortcut connection) from the perception to the production pathway of the model at the phonological level [8] (see the dashed line in Fig. 3 at the phonological form level of the mental lexicon). The performed task now is to repeat a phonological form without involving higher levels of the mental lexicon (i.e., to repeat an auditory impression, processed only up to the phonological level of the phonological system of the target language learned by the model).

The multiple repetition task of syllables—also called diadochokinesis—likewise does not involve the mental lexicon. This results from the task's priming procedure because the model is only instructed to repeat syllables and not words of a specific language. The simulation here starts with pre-activation of the motor program of one syllable or of a sequence of syllables (such as /badaga/ or similar, see [29,30]). The action control component of the model is primed here on production of multiple repetitions of one syllable or of a sequence of syllables.

The word comprehension task (which is also a component of word repetition) is based on the auditory activation of the syllable sequence representing a target word (word is acoustically presented by the test supervisor). The selected and activated concept at the end of the perception-comprehension pathway is decoded by similarity calculation processes. Here, that concept is chosen (i.e., is comprehended) which exhibits the highest dot product, i.e., which exhibits the highest activation in the concept buffer (see similarity plot of the concept buffer [17]).

The following "tasks" are special because there are no direct performance tasks like those discussed above. Rather, here we illuminate speech and language learning processes in order to increase the language and speech competence of the model. So far, speech and language learning processes are simulated using rate-coding neuron models (the spatio-temporal activity averaging [STAA] approach, see [32] and for a speech processing model see [9]), but the mechanism of action control in learning scenarios can be modeled in our spiking neuron NEF-SPA approach. In the case of word learning (lexical learning), specific communication scenarios are needed. For example, the model (i.e., the speech learning child) points to an object (e.g., a ball) and looks at the communication partner to motivate him to utter the word. Thus, a simultaneous activation of concept and auditory form appears. If the model has learned to convert auditory forms into phonological forms (see below: syllable learning), it is now capable in addition to co-activate the lemma form and the concept of that new word. This simultaneous activation of word forms at different levels of the mental lexicon now allows lexical learning, i.e., the synaptic weight adjustments within associative memories between phonological forms, lemmata, and concepts.

In the case of syllable learning (to increase the number of learned syllables within the mental syllabary), the specific learning scenario is a production-and-(self-)perception event for a new syllable realization. Thus, we have a simultaneous activation of phonological form, motor plan, motor program, as well as of auditory state and somatosensory state, activated by the feedback sensory loop. If the production trial is judged as "successful", i.e., if the produced syllable is labeled as correct, the neural states currently and simultaneously activated in the phonological form buffer, motor plan buffer, motor program buffer, and auditory and somatosensory state buffer are used to adjust the synaptic weights acting in the associative memories connecting these state buffers with each other. These associative memories form the core knowledge of the mental syllabary.

3.3 Modeling Neural Premotor Activations for Syllable and Gesture Execution

Besides selecting and executing task specific actions such as listening to auditory items (e.g., generated by a task supervisor), visually perceiving and identifying objects from a picture, memorizing a list of items (words or visual objects), (re-)producing and articulating syllables or words, etc., a further complex neural functioning that needs to be processed by the action control component is the sequencing, premotor preparation, and execution of the syllables of a word or utterance under production. Word or utterance production and execution becomes complex if more than one syllable needs to be executed, i.e., if syllable sequencing and its temporally ordered execution come into play. The simulation of the execution of single syllables has already been described [9,33]. In this paper, we present for the first time the neural preparation and execution processes for a succession of four syllables. Here, a syllable oscillator is implemented for each learned syllable, and the syllable oscillator triggers gesture activation and execution of all gestures building the word or utterance [9,25]. However, it is not possible to simulate a repetition of the same syllable with this older version of the model because syllable oscillators for a sequence of syllables need to be independent of each other due to the inherent overlap of the life-time cycle of temporally adjacent syllable oscillators during syllable sequence production and execution. Starting a syllable oscillator means preparing the syllable at a premotor level before the first articulatory gesture is activated and thus always starts within the lifecycle of the syllable oscillator of the preceding syllable.

It can be seen from our new simulation results shown in Fig. 5 that syllable oscillators (i) generate exactly a single oscillation cycle per syllable (the lifecycle of a syllable including syllable preparation at premotor levels and execution of its inherent articulatory gestures (see also [9,25]) and (ii) overlap in time for nearly half of the oscillation cycle (new simulation results). This overlap in time of syllable lifecycles can even become stronger with a rapid speaking rate. In that case, the oscillation cycle or lifecycle of up to three subsequent syllables may overlap in time. The oscillator of each syllable defines the timing of all articulatory gestures appearing in a syllable as a type of premotor temporal setting for all gestures of a syllable that is activated and, thus, that will be executed. This defines premotor movement estimates for all speech gestures appearing in that syllable (row "speech actions"; middle row in Fig. 5), and these movement estimates or movement patterns are activated in NEF ensembles called gesture ramp signal integrators (see **Supplementary Material A2**). These ramp signal integrators are triggered by syllable oscillators (see Fig. 6). It can be seen from Fig. 5 that the movement velocity (i.e., the speed with which a gesture executing an articulator reaches its target region; target regions are indicated by thick bold black horizontal lines in Fig. 5, row c; the thick upper line represents the beginning of the target region for vocalic and consonantal gestures, and the thick lower line for glottal and velopharyngeal gestures) is lower for vocalic gestures (sa_vow_I, sa_vow_A, sa_vow_U activation patterns in a row c in Fig. 5) in comparison to consonantal constriction gestures for labial, apical, and dorsal consonants (sa_lab_constr, sa_api_constr, sa_dor_constr; Fig. 5). Moreover, the movement velocity of velopharyngeal abduction gestures (sa_vel_abduc; Fig. 5) is lower than the movement velocity of glottal abduction (opening) gestures (sa_glott_abduc). Furthermore, in comparison to the ramp signal generation produced by other neural integrators, which are used as short-term memories for signals (see Fig. 1d), here the ramp signal decreases if the input signal, i.e., the trigger pulse or trigger signal feeding the integrator ensemble, ends (trigger signal pulses are shown in rows d and e of Fig. 5). The parameters for adjusting the ramp signal, i.e., the gesture movement estimates, can be adjusted by the parameters of the recurrent neural connections of the neural integrator ensemble itself, together with the amplitude parameter of the neural connection between gesture trigger signal ensemble and gesture ramp signal ensemble for each articulatory gesture, also called speech action (see **Supplementary Material A2**). The gesture trigger signals for the vocalic and consonantal gestures of each syllable processed by each syllable oscillator (constr-, vow-, constr-gesture, i.e., lab_constr, vow_A, lab_constr for syllable /bap/) are displayed in rows d and e of Fig. 5.

The difference between these two last rows in Fig. 5 is that the activation of gesture trigger signal ensembles for each gesture is given in the last but one row, while the trigger signals of the last row reflect the connection of each gesture to syllables. While the neural activities shown in the last but one row (row iv of Fig. 5) are no longer reflecting the origin of a gesture with respect to a specific syllable (trigger pulse ensembles for "gestures only"; see Fig. 6), the neural activities in the last row (last row of Fig. 5) reflect the trigger signal pulses generated by each syllable oscillator for each gesture as part of a syllable (trigger pulse ensembles for "gestures*syllables" in Fig. 6). A typical difference between the activations displayed in the last but one and in the last row of Fig. 5 can be seen at the end of the second from last syllable /dip/ and last syllable /bap/. Because the labial consonantal constriction gesture of /p/ in /dip/ temporarily overlaps with the first consonantal constriction gesture of /b/ in /bap/, these two gestures melt into one labial constriction gesture (see blue trigger pulse signals in the second last row of Fig. 5). This is modeled in our neural architecture by introducing two levels of trigger pulse ensembles (see Fig. 6 and **Supplementary Material A2**).

Because the start and end times of gesture trigger pulses are stored in values of a relative time scale, i.e., relative to the time scale defined by the oscillation cycle of a syllable oscillator, it seems reasonable to define a syllable oscillator ensemble for each learned syllable (approximately 2000 syllable oscillators for languages such as English or German). However, as already mentioned above, this would still cause problems if a syllable were repeated in a near temporal neighborhood (repetition of the same syllable in a direct neighborhood, as, e.g., in "bye bye" or in an indirect neighborhood as, e.g., in "my arms, my legs" if the speech rate is not too low). Thus, it can be assumed that syllable oscillators need to be defined from the viewpoint of utterance production as a sequence of oscillator ensembles that can be filled by syllables generated by the cognitive-linguistic component of the model. However, this concept causes another problem because now each syllable oscillator could be activated by each syllable, and thus, each syllable oscillator needs to be equipped with connections towards syllable trigger pulse ensembles for each type of syllable. This can be modeled using neural connection ensembles (also called gaiting ensembles), which normally are all inhibited (not activated and not capable of forwarding any neural activity) despite those connection ensembles that reflect connections towards gesture trigger pulse ensembles for all articulatory gestures that are needed to produce a specific syllable (i.e., disinhibited connection ensembles; for details see **Supplementary Material A2**). This inhibition/disinhibition process is controlled by the action control component, which thus realizes the assignment of a specific syllable and its parameters stored in the mental syllabary to a specific syllable oscillator as part of the syllable oscilla-

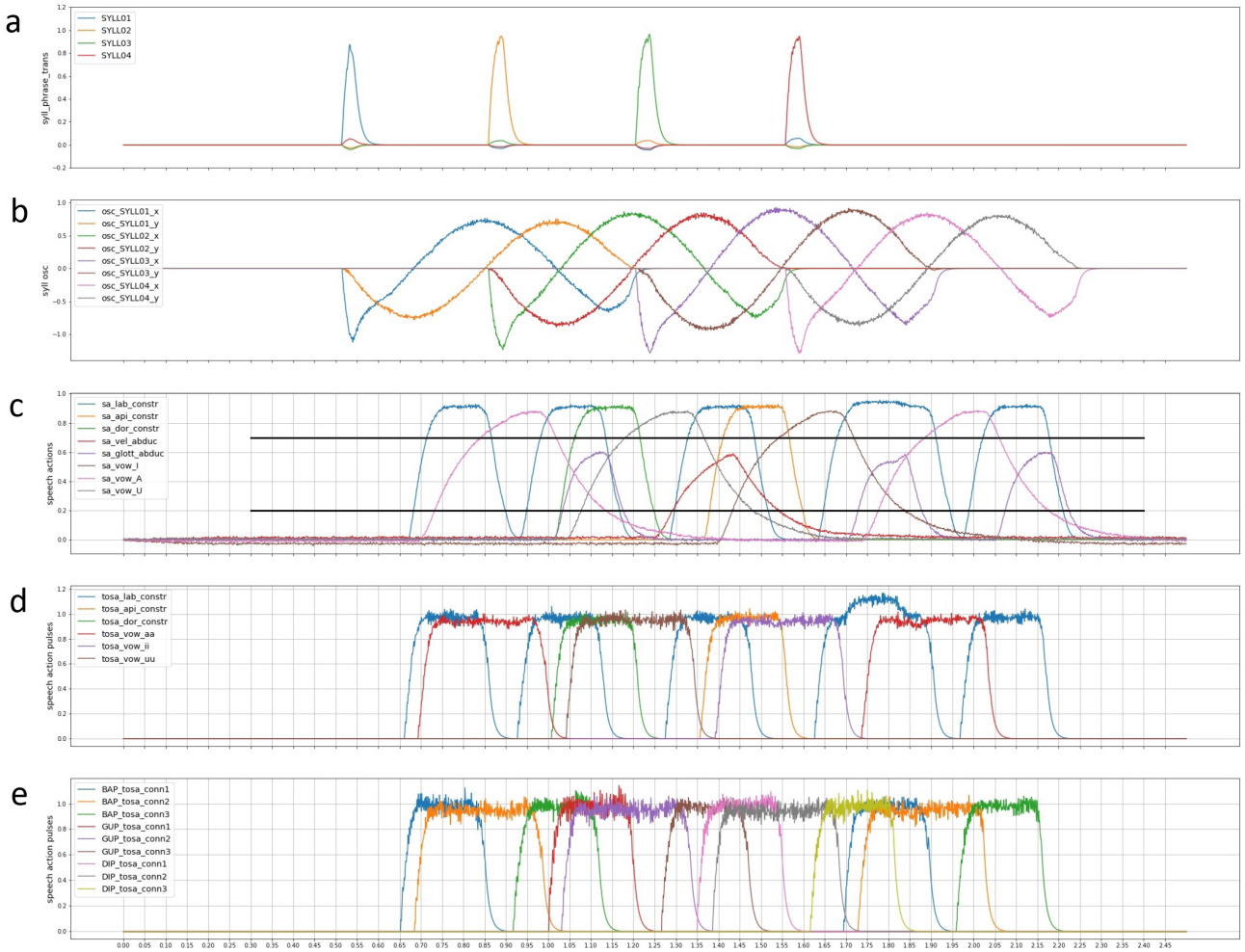

Fig. 5. Neural activation patterns for syllable sequences and articulatory gestures at the premotor stage (new simulation results). (a) Trigger signals for a sequence of four syllable oscillators (SPA buffer: syll_phrase_trans; see **Supplementary Material A2**). (b) Four NEF ensembles for syllable oscillators: activation is displayed for each oscillator ensemble for its two signal components (x and y-signals) – syllable oscillators are displayed in one graph (osc_SYLL0i_x, osc_SYLL0i_y, i = 1 to 4) representing the sequence of four syllables: /bap/, /gum/, /dip/, and /bap/ again. (c) Overlayed activation patterns for the neural ramp integrator signals of three vocalic gestures, three consonantal gestures, and one velum and one glottal abduction gesture (also called speech actions: sa_vow_aa, sa_vow_ii, sa_vow_uu, sa_constr_lab, sa_constr _api, sa_constr _dor, sa_abduc_vel, sa_abduc _glott). Neural ramp signal integrator ensembles are triggered by the activation pattern of trigger pulse ensembles displayed in rows (d) and (e) of this figure, i.e., "speech action pulses". The difference between the activation patterns displayed in these rows is explained in the text.

tor ensemble (sequence of syllable oscillators, see Fig. 6). This neural architecture realizes the activation of syllable articulation at the premotor level. The forwarding and processing of gesture movement estimate towards the primary motor level for direct control of muscle group activation (see [25]) is beyond the scope of this paper.

Despite the high number of potential connections between each syllable oscillator and each type of syllable stored in the mental lexicon, the number of neurons and the number of neural connections that are needed here to implement the potential connections between each syllable oscillator and all syllables stored in the mental lexicon remains in a convenient range and can be easily handled by the premotor component (located in the frontal cortex if the brain scale model components are compared with box-and-arrow models generated from neurophysiological data, e.g., [34,35]). An estimation of the number of neurons needed for implementing this syllable programming component at the premotor level is given in Fig. 7. Some 2500 LIF neurons are needed for implementing a syllable oscillator that guarantees a stable oscillation cycle. If we assume that 50 types of syllables (e.g., C1VC2 with C1 = voiced plosive, C2 = voiceless plosive, V = vowel, characterizes one type of syllable, representing approximately $10 \times 3 \times 3 = 90$ syllables such as /bap/ or /dut/ or /gik/ in the case of 10 vowels, three initial and three final consonants) with a mean of eight gestures per syllable is sufficient to represent the variety of syllables occurring in a language, then 400 gesture trigger pulse ensembles are needed. Trigger pulses are modeled with sufficient precision by ensembles comprising 50 neu-

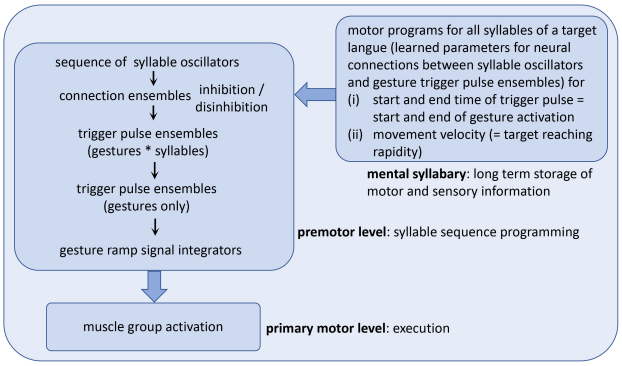

Fig. 6. The process of syllable sequencing (see main text). "*" means: a gesture as part of a syllable; all gestures can be concatenated in principle with all syllables.

rons. The same holds for the inhibiting/disinhibiting connection ensembles as well as for all other trigger pulse ensembles needed before the gesture ramp integrator ensembles can be connected. The gesture ramp signal integrator ensembles need 1000 neurons each in order to guarantee a stable generation and representation of a gesture movement estimate. The number of 30 different gestures is sufficient in order to model all needed gestures appearing in a language (e.g., approximately 10–15 vocalic gestures, two glottal gestures, two velopharyngeal gestures, 10–15 consonantal gestures). In total, the premotor component of our model is sufficiently equipped with approximately 111 thousand neurons (approximately 10^5) and approximately 201 million neural connections (approximately 2×10^8). If we assume that a cortical neuron is connected with approximately 10^{3-5} other cortical neurons (the total number of cortical neurons is approximately 10^{11}, e.g., as in [11]), our model is in the correct magnitude of connections to neurons and, at the premotor level, the within-module neural connections are still at the lower edge (approximately 2×10^3 connections per neuron).

3.4 Number and Setting of Basic NEF-SPA Elements, Neurons, and Their Connections

The main modules of this model are the cognitive-linguistic component (CL), the sensorimotor component (SM), and the action selection control loop, including the basal ganglia and thalamus (AS). A neural pathway connects buffers and/or associative memories, and each pathway comprises many neuron-to-neuron connections. Buffers include short-term memories (i.e., recurrently connected buffers). The dimension of buffers and associative memories is D = 64 in the case of a vocabulary of 1200 words [22]. The number of dimensions needs to be extended up to approximately 500 (e.g., 512) in the case of modeling a complete vocabulary of a language (approximately 100,000 items on each lexical level). All parameters defining the synapses and the built-up membrane voltage of LIF neurons; the size and organization of neuron ensembles and neuron buffers; the detailed structure of neuron pathways (of connections) between ensembles, buffers, or memories; and other elements of our large-scale neural network are set to the default values given by [11,12,17]. The total number of buffers and associative memories etc., for constructing our whole large-scale model for speech production is listed in Table 2. This listing includes all buffers, memories, and connection pathways of the model, including the cognitive processing for complex picture naming with cues (see [22,27]) and including all clean-up processes at all levels of the model (not all indicated in Fig. 3). The number of neurons will increase above the number of neurons stated in Table 2 if the vocabulary of the model increases and thus the number of dimensions for each S-pointer needs to be increased above the level of D = 64. Our current model represents the vocabulary of a 5–6-year-old child (approx. 1200 words). Thus, the model represents the vocabulary of a child in the middle of his/her word acquisition process [22].

The model (cognitive, sensorimotor, and action selection components) comprises 331,500 neurons plus 111,500 neurons for the syllable preparation and execution component (443,000 neurons in total) and approximately 58,000,000 normal neural connections plus further 200,000,000 connections for different gaiting paths in the syllable production component, from which only approximately 2,000,000 connections are permanently connected during the syllable production process.

A more efficient coding strategy for syllable preparation and syllable execution (in order to reduce the number of neural connections here) is under development. The syllable preparation and execution processes are currently not integrated into the main simulation model in order to hold the real-time factor for calculating a simulation at a factor of approximately 600:1 (i.e., 10 minutes processing time for calculation of one second of simulated behavior on standard personal computers).

4. Results II: Performance Features of the Model in Different Speech Processing Scenarios

A typical word production task is picture naming. This task can be performed by the model, in principle, for all words that are already included in the mental lexicon. Because learning is still a challenge for spiking neuron networks, word learning has been modeled in earlier versions of our model of speech production using an STAA neuron model ([36] and see [25]). The mental lexicon, including semantic and phonological word associations, is modeled in our current spiking-neuron based NEF-SPA version of the model of speech production in the form of hand-written S-Pointer networks [21,22]. The performance of word production is nearly 100% in the case of a model version representing a healthy subject (no ablated neurons in any buffer of the model, i.e., no insertion of any neural dysfunction in

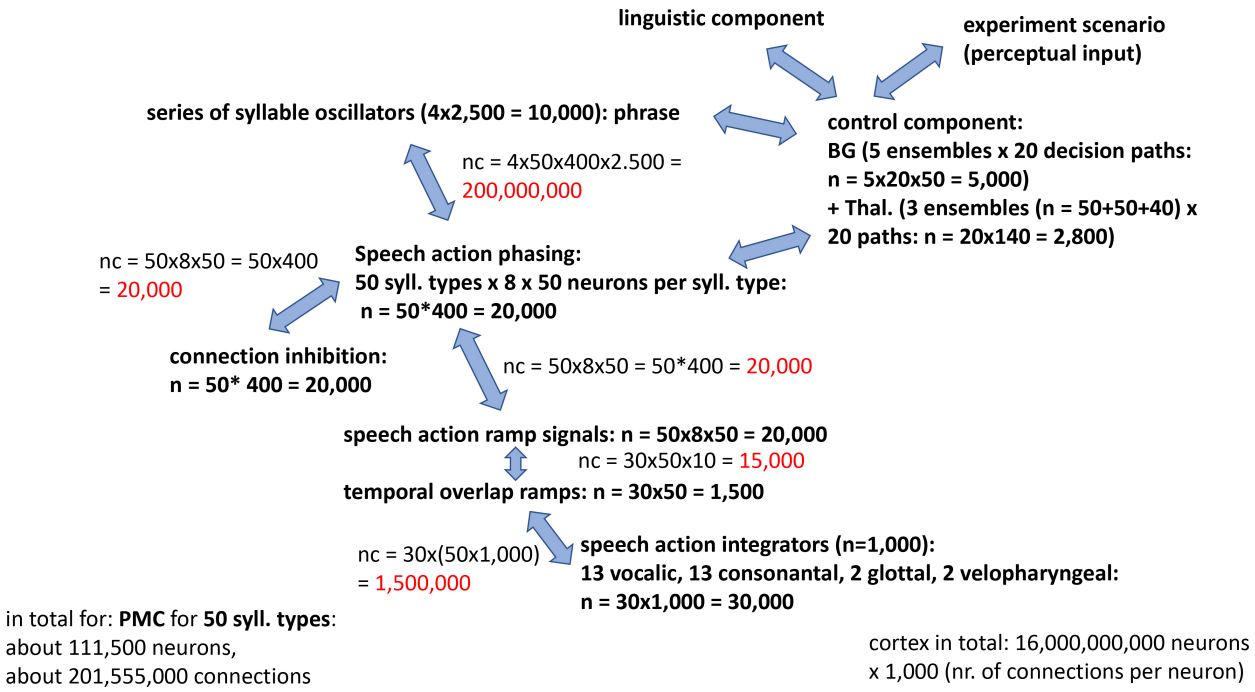

Fig. 7. Estimation of the number of neurons (red) needed for the mental syllabary and premotor component of speech production.

the model). The production of no word (no answer) or of an incorrect word or syllable happens in 0.6% of all production trials (approximately 1000 trials in total) and is caused by a too low word form activation at the phonological or semantic buffer level in the production pathway of the neural model [21,22]. In these rare cases, a second picture naming trial can be initiated by the action selection loop of the model (in the case of no answer) or by the task supervisor (in the case of production of an incorrect word), which then leads to an increase of up to an almost 100% correct production rate. These picture naming trials that need initiation of a second production process would appear as a correction process or, in the case of no word production, as a production process with a longer latency period. Word production in healthy subjects normally leads to an error rate of approximately 0.1% [37].

Because semantic and phonological similarities of words are included in the model in the form of S-Pointer-networks, the model is capable of simulating picture naming (visual input) in the presence of auditorily presented distractor words. These distractor words were chosen as phonologically similar, semantically similar, phonologically and semantically similar, or not at all similar to the visually presented target word. Simulations produced qualitatively and quantitatively similar results as those from experiments conducted on humans [21]. More speech errors were generated in the case of semantically similar distractor words compared with phonologically similar distractor words, but most errors were produced in the case of semantically and phonologically similar distractor words. This result is due to the existence of a phonological and semantic feedback or monitoring loop, monitoring the results of the production pathway on the perception pathway side [21].

With regard to simulating the speech production of children suffering from lexical access dysfunctions or lexical acquisition problems as a result of delayed or disordered speech acquisition processes, the percentage of speech errors is much higher even in normal picture naming tasks. Here, the delay in lexical knowledge acquisition is modeled by including only approximately 50% of the typical vocabulary of normally developed 5;6 to 6;5-year-old children [22]. Furthermore, neural dysfunctions were introduced in the form of ablation of a defined percentage of all neurons at the level of semantic, lemma, or phonological buffer, as well as for the neural connections modeled in the form of associative memories between these buffers, to model the lexical access disorder. Simulation experiments indicated that these neural distortions reduced the rate of correct word productions even more. At the same time, the adding of semantic and/or phonological cues increased the rate of correct word production in a qualitative and quantitative similar way to that in experiments on humans [22].

Word production, word comprehension, and logatome repetition tasks (logatome = nonsense word) were simulated in cases of different degrees of neuron ablation in different model buffers allocated to different levels of the mental lexicon at the production or perception pathway (concept, lemma, or phonological form level) or to the neural associations between these levels (associative memories) in the production or perception pathways [8]. Ablation of neurons was inserted in different buffers or associative memories to model different subtypes of aphasia. Fur-

Table 2. Number of network elements, neurons and neural connections used in large-scale model of speech production in case of a model vocabulary of approximately 1,200 words.

Model component	Name of NEF-SPA model element	Number of elements	Number of neurons or connections per element	Total number of neurons or connections
CL	Buffer	12	64 × 50 = 3250 neurons	39,000 neurons
CL	Associative memories	14	64 × 100 = 7500 neurons	91,000 neurons
CL	Neural pathways (between buffers)	20	approx. 1,500,000 connections	30,000,000 connections
SM	Buffer	10	64 × 50 = 3250 neurons	32,500 neurons
SM	Associative memories	12	64 × 100 = 7500 neurons	78,000 neurons
SM	Neuron ensembles	1500	50 neurons	75,000 neurons
SM	Neural pathways (between buffers)	18	approx. 1,500,000 connections	28,000,000 connections
SM	Neural pathways (between ensembles)	2000	approx. 2500 connections	5,000,000 connections
AS	Neuron ensembles	8	2000 neurons	16,000 neurons
AS	Neural pathways (between ensembles)	16	approx. 25,000 connections	400,000 connections

CL, cognitive linguistic component; SM, sensorimotor component; AS, action selection control loop; NEF-SPA, Neural Engineering Framework and Semantic Pointer Architecture; approx., approximately.

thermore, different degrees of ablation (different percentages of ablated neurons) were inserted for each subtype of neural dysfunction to model different levels of severity for each subtype of aphasia. Thus, a group of model instances modeling 11 different degrees of severity (0 to 100% ablation) for six different subtypes of aphasia (Broca, Wernicke, transcortical motor, transcortical sensory, conduction, and mixed aphasia) were simulated for testing the model performance of 66 instants of the model in the case of three different production-perception tasks for 18 target words (and 18 target logatomes respectively). Simulation results agree with results found by testing natural speakers suffering from these different types of aphasia as well as by simulation results reported from another neural modeling approach [3]. As expected, the performance rate decreases for word comprehension in the case of increasing severity (increasing percentage of ablated neurons) of Wernicke, transcortical sensory and mixed aphasia, and word production in the case of Broca, transcortical motor, and mixed aphasia. The performance of logatome repetition decreases as expected in the case of Broca, Wernicke, and conduction aphasia [8].

In addition, we were able to model a syllable sequencing task in the case of varying dopamine levels affecting the D1 and D2 synaptic receptors of the striatal neurons within the basal ganglia model of the action selection module of our neural model [29,30]. It was shown that in the case of low dopamine levels, syllable sequencing becomes erroneous, and effects such as syllable freezing appear as reported in patients suffering from dysarthria as a result of reduced dopamine levels, as seen in Parkinson's disease [29,30].

To summarize, the main goal of our modeling efforts was to deliver performance rates in the simulated tasks comparable to natural data. For all the production and perception tasks listed in Table 1, the performance rate was nearly 100%, as in natural speech production. When introducing distractor stimuli in picture naming tasks [21], our simulations produced error or halt rates up to approximately 50% of all word productions in accordance with natural data [21]. In the case of neural dysfunctions at different levels of the model (see word production tasks with additional cues [22] and production and perception tasks for screening patients with aphasia [8]), we were able to simulate error rates up to 100% if the degree of ablated neurons in a specific buffer was above 30%, depending on the exact functional location of the buffer in the production or perception pathway [8,22].

5. Discussion

A neurofunctional model of speech production has been described herein, which is based on a limited set of neurobiologically well-grounded construction elements such neural ensembles, neural buffers, associative memories, and neural connections. The model has been developed within the NEF-SPA framework, which has already proved capable of modeling a general purpose neurobiologically inspired model of the brain [11].

Modeling of different production-perception tasks such as picture naming, picture naming in the presence of verbally inserted distractor words, picture names including phonological and semantic cues, word comprehension (as it is needed in word production tasks such as word repetition), and syllable sequencing showed high performance rates in a model instance without insertion of any neural

dysfunctions. Neural dysfunction modeling typical speech disorders led to a decrease in task performance rates, and the model qualitatively and quantitatively showed results that were comparable with results generated by experiments in normal speaking subjects and patients suffering from speech and language disorders such as lexical access disorders, delayed or disordered speech acquisition (leading to reduced lexical word range), different subtypes of aphasia, and dysarthria of speech as a result of Parkinson's disease. Thus, the neurobiologically grounded modeling of speech production introduced here by using a spiking neural network approach, including the NEF-SPA concept, allows for not only an explanation of the processes appearing in normal (healthy) speech production but also the identification of the potential underlying neural dysfunctions appearing in specific modules, buffers, and associations between buffers, helping to explain, and thus elucidate, the etiology of different types of speech disorders.

While the overall organization and architecture of the neural model have already been described in detail in other papers (e.g., [9]), this does not hold for the complex neural processes appearing during syllable and gesture execution, which is described in this paper for the first time. We implemented a hierarchical structure for representing motor command generation, including a comparably small number of neural elements, model components, neurons, and neural connections, to be able to fulfill the required task of syllable generation and syllable sequencing.

The neural (construction) elements to construct our neural model of speech production are based on a limited set of neurobiologically grounded neural functional principles and their corresponding neural realization: (i) ensembles for neural representation of simple states (sensory or motor signals) and buffers for neural representation of higher-level sensory, motor, or cognitive states or items; (ii) neural connections between single neurons or between ensembles and buffers for forwarding neural states; (iii) recurrent neural connections between ensembles for ramp signal generation, short-term memory generation for simple states (signals), and the generation of neural oscillators and recurrent neural connections between buffers for the generation of short-term memories of states; (iv) associative memories and neural connections between ensembles or buffers for transforming neural states; and (v) neural modules for action selection and action execution. This limited set of neurofunctional elements is sufficient for constructing a neural model of speech production, as described in this paper.

Both a quantitative and qualitative comparison with other models of speech production (e.g., [3–5]) is difficult because one of these approaches mainly models the cognitive-linguistic component of speech production [3] while the two other approaches mainly model the sensorimotor component of speech production. Moreover, all these models are second-generation neural network models in terms of [6], while our approach uses spiking neurons (a third-generation neural network [10]) and thus comprises straightforward modeling of all temporal relations in our large-scale neural network.

The main limitations of our current version of the neural model of speech production are its limited vocabulary, lack of capability for sentence processing, and the incomplete implementation of front-end modules. Moreover, the simulation of learning in a third-generation neural network is still a challenge (c.f. [38]); first- and second-generation neural networks are typically favored for simulating learning and these types of neural networks are mainly used for modeling speech acquisition (c.f. [4,36]). Thus, currently, the generation of the whole set of lexical items (concepts, lemmata, and phonological forms), as well as all within and between relations, needs to be transcribed and implemented manually in the form of three S-Pointer networks in our spiking neural network model and the vocabulary used for simulating word production is still limited to approximately 1200 items (see [22]). The development of learning scenarios mimicking specific situations of language acquisition for constructing the mental lexicon automatically is now necessary, and the development of algorithms for modeling learning in a wider range within the NEF-SPA context, e.g., learning and storing a whole set of lexical items from simulations of supervisor-model interactions, is therefore also necessary (see [9]).

The simulation scenarios implemented so far comprise word production and, to some extent, word comprehension. The model still needs to be augmented by a sentence comprehension and sentence production module, including syntactic and semantic processing. This module should be implemented beside the mental lexicon because it interacts closely with the mental lexicon buffers on the production and perception side by using a lemma and concept information for each word of a sentence.

Speech perception is not included in our current version of the model due to the complex lower-level acoustic and phonetic processing, i.e., acoustic and phonetic feature recognition as well as for sound, syllable, and word recognition in a neurofunctional plausible way. On the perception side, our modeling currently starts at the phonological level. Acoustic input is generated in an artificial way by defining an auditory S-pointer for each syllable or word and by directly forwarding and transforming these auditory forms into phonological forms. On the production side, the articulatory front-end is already developed [25] and can now be coupled with the speech production neural network, which is now capable of generating and executing motor programs (see section 3.3 of this paper).

6. Conclusions

Due to the successful functioning of our neuronal simulation model in the case of simulating both normal and disordered speech production, it cannot necessarily be concluded that the neuronal architecture in this model also

reflects the neuronal architecture that has evolved in humans. However, the neural (construction) elements available in the NEF-SPA context form a minimal set that allows the construction of neurofunctional "large-scale" or "brain-scale" models. Therefore, as we attempted to build an architecture with a minimal number of neuronal components and as this attempt includes the effort of using only a minimal number of neurons and neuronal connections, which is able to fulfill the required behavioral functions, we can assume that the neural architecture created here represents a neurobiologically probable architecture.

All model components (e.g., mental lexicon or word selection, mental syllabary for phonological form generation, premotor component for motor command preparation, primary motor component for articulatory execution, cortico-cortical loop including basal ganglia and thalamus for action selection) are motivated from available neurophysiological data (e.g., the components of our model are based on components of typical box-and-arrow models derived from imaging data such as functional magnetic resonance imaging (fMRI) or from electroencephalography (EEG) and magnetoencephalography (MEG) data; see [20,34,35,39]).

Thus, in contrast to modeling efforts which try to copy the brain structure in a still more detailed fashion (e.g., [40]) our approach already delivers a practicable model which allows us to understand how a complex behavioral model can be constructed based on elementary neurofunctional elements. Moreover, our model enables a quantitative testing of its behavioral results with respect to human behavioral data and thus our model allows an evaluation of its reality by evaluating its performance.

Availability of Data and Materials

All data and materials are included in this published article and supplementary materials.

Author Contributions

Author BK designed the research study, performed the research, conducted the simulation experiments, and analyzed the data. Author BK contributed to editorial changes in the manuscript, read and approved the final manuscript. Author BK have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

I would like to thank Trevor Bekolay, senior research scientist at Applied Brain Research (ABR) for helping me with writing the source code and for his valued discussions concerning the structure and design of early versions of the speech production neural network model published here.

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.jin2205124.

References

[1] Roelofs A. A spreading-activation theory of lemma retrieval in speaking. Cognition. 1992; 42: 107–142.

[2] Roelofs A. The WEAVER model of word-form encoding in speech production. Cognition. 1997; 64: 249–284.

[3] Roelofs A. A dorsal-pathway account of aphasic language production: the WEAVER++/ARC model. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior. 2014; 59: 33–48.

[4] Guenther FH, Ghosh SS, Tourville JA. Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language. 2006; 96: 280–301.

[5] Bohland JW, Bullock D, Guenther FH. Neural representations and mechanisms for the performance of simple speech sequences. Journal of Cognitive Neuroscience. 2010; 22: 1504–1529.

[6] Maass W. Networks of spiking neurons: the third generation of neural network models. Neural Networks. 1997; 10: 1659–1671.

[7] Guenther FH. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychological Review. 1995; 102: 594–621.

[8] Kröger BJ, Stille CM, Blouw P, Bekolay T, Stewart TC. Hierarchical Sequencing and Feedforward and Feedback Control Mechanisms in Speech Production: A Preliminary Approach for Modeling Normal and Disordered Speech. Frontiers in Computational Neuroscience. 2020; 14: 573554.

[9] Kröger BJ, Bekolay T, Cao M. On the Emergence of Phonological Knowledge and on Motor Planning and Motor Programming in a Developmental Model of Speech Production. Frontiers in Human Neuroscience. 2022; 16: 844529.

[10] Yamazaki K, Vo-Ho VK, Bulsara D, Le N. Spiking Neural Networks and Their Applications: A Review. Brain Sciences. 2022; 12: 863.

[11] Eliasmith C. How to Build a Brain: A Neural Architecture for Biological Cognition. Oxford University Press: New York, NY, USA. 2013.

[12] Stewart TC, Eliasmith C. Large-scale synthesis of functional spiking neural circuits. Proceedings of the IEEE. 2014; 102: 881–898.

[13] Carnevale NT, Hines ML. The NEURON Book. Cambridge University Press: Cambridge, MA, USA. 2006.

[14] Gewaltig MO, Diesmann M. Nest (neural simulation tool). Scholarpedia. 2007; 2: 1430.

[15] Goodman DFM, Brette R. The brian simulator. Frontiers in Neuroscience. 2009; 3: 192–197.

[16] Eliasmith C, Anderson CH. Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems. MIT Press: Cambridge, MA, USA. 2003.

[17] Bekolay T, Bergstra J, Hunsberger E, Dewolf T, Stewart TC, Rasmussen D, *et al*. Nengo: a Python tool for building large-scale functional brain models. Frontiers in Neuroinformatics. 2014; 7: 48.

[18] Gosmann J, Eliasmith C. Optimizing Semantic Pointer Representations for Symbol-Like Processing in Spiking Neural Networks. PLoS ONE. 2016; 11: e0149928.
[19] Crawford E, Gingerich M, Eliasmith C. Biologically Plausible, Human-Scale Knowledge Representation. Cognitive Science. 2016; 40: 782–821.
[20] Levelt WJ, Roelofs A, Meyer AS. A theory of lexical access in speech production. The Behavioral and Brain Sciences. 1999; 22: 1–75.
[21] Kröger BJ, Crawford E, Bekolay T, Eliasmith C. Modeling Interactions between Speech Production and Perception: Speech Error Detection at Semantic and Phonological Levels and the Inner Speech Loop. Frontiers in Computational Neuroscience. 2016; 10: 51.
[22] Stille CM, Bekolay T, Blouw P, Kröger BJ. Modeling the Mental Lexicon as Part of Long-Term and Working Memory and Simulating Lexical Access in a Naming Task Including Semantic and Phonological Cues. Frontiers in Psychology. 2020; 11: 1594.
[23] Stewart TC, Choo X, Eliasmith C. 'Dynamic behaviour of a spiking model of action selection in the basal ganglia'. Proceedings of the 10th international conference on cognitive modeling. 2010.
[24] Stewart T, Choo X, Eliasmith C. 'Symbolic reasoning in spiking neurons: A model of the cortex/basal ganglia/thalamus loop'. Proceedings of the Annual Meeting of the Cognitive Science Society. 2010.
[25] Kröger BJ, Bekolay T. Producing syllables: motor planning, motor programming and execution. In Niebuhr, O., Lundmark, M.S., Weston, H. (eds.) Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung (pp. 1–8). TUDpress: Dresden, Germany. 2022.
[26] Kröger BJ, Bafna T, Cao M. Emergence of an Action Repository as Part of a Biologically Inspired Model of Speech Processing: The Role of Somatosensory Information in Learning Phonetic-Phonological Sound Features. Frontiers in Psychology. 2019; 10: 1462.
[27] Stille CM, Bekolay T, Blouw P, Kröger BJ. Natural Language Processing in Large-Scale Neural Models for Medical Screenings. Frontiers in Robotics and AI. 2019; 6: 62.
[28] Eliasmith C, Stewart TC, Choo X, Bekolay T, DeWolf T, Tang Y, *et al*. A large-scale model of the functioning brain. Science (New York, N.Y.). 2012; 338: 1202–1205.
[29] Senft V, Stewart TC, Bekolay T, Eliasmith C, Kröger BJ. Reduction of dopamine in basal ganglia and its effects on syllable sequencing in speech: a computer simulation study. Basal Ganglia. 2016; 6: 7–17.
[30] Senft V, Stewart TC, Bekolay T, Eliasmith C, Kröger BJ. Inhibiting Basal Ganglia Regions Reduces Syllable Sequencing Errors in Parkinson's Disease: A Computer Simulation Study. Frontiers in Computational Neuroscience. 2018; 12: 41.
[31] Postma A. Detection of errors during speech production: a review of speech monitoring models. Cognition. 2000; 77: 97–132.
[32] Kröger BJ, Bekolay T. Neural Modeling of Speech Processing and Speech Learning. An Introduction. Springer International Publishing: New York, NY, USA. 2019.
[33] Kröger BJ, Bekolay T, Blouw P, Stewart TC. Developing a model of speech production using the Neural Engineering Framework (NEF) and the Semantic Pointer Architecture (SPA). Proceedings of the International Seminar on Speech Production ISSP2020. Haskins Press: New Haven, CT, USA. 2021.
[34] Hickok G, Poeppel D. The cortical organization of speech processing. Nature Reviews. Neuroscience. 2007; 8: 393–402.
[35] Indefrey P. The spatial and temporal signatures of word production components: a critical update. Frontiers in Psychology. 2011; 2: 255.
[36] Kröger BJ, Kannampuzha J, Kaufmann E. Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. EPJ Nonlinear Biomedical Physics. 2014; 2: 1–28.
[37] Garnham A, Shillcock RC, Brown GDA, Mill AID, Cutler A. Slips of the tongue in the London-Lund corpus of spontaneous conversation. Linguistics. 1981; 19: 805–818.
[38] Jang H, Simeone O, Gardner B, Gruning A. An introduction to probabilistic spiking neural networks: Probabilistic models, learning rules, and applications. IEEE Signal Processing Magazine. 2019; 36: 64–77.
[39] Friederici AD. The brain basis of language processing: from structure to function. Physiological Reviews. 2011; 91: 1357–1392.
[40] Markram H. The Blue Brain Project. Nature Reviews Neuroscience. 2006; 7: 153–160.