Modeling Speech Production Using the Neural Engineering Framework

Bernd J. Kröger
Neurophonetics Group,
Department for Phoniatrics, Pedaudiology,
and Communication Disorders,
Medical School, RWTH Aachen University, Germany
Email: bkroeger@ukaachen.de

Trevor Bekolay and Chris Eliasmith
Centre for Theoretical Neuroscience,
University of Waterloo, Canada
Email: {tbekolay, celiasmith}@uwaterloo.ca

***Abstract*—A neurobiologically plausible model of speech production is introduced here using the Neural Engineering Framework (NEF). This approach allows detailed modeling of temporal aspects of action selection and action execution in speech production at the level of single spiking neurons. A preliminary architecture of our NEF speech production model is introduced and discussed in the second part of this paper. The first part focuses on an articulatory-acoustic model, generating acoustic speech signals on the basis of articulatory geometries. Our approach uses a small set of functional articulatory control parameters. Motor planning is based on the concept of speech or vocal tract actions (Kröger et al. 2010, Cognitive Processing 11: 187-205). A 2D-geometrical model is used and the acoustic speech signal is calculated using a reflection-type line analog model.**

I. Introduction

Different approaches exist for describing the speech production hierarchy. However, in general it is thought that cortical processing starts with conceptualization of the communicative act, followed by lexical retrieval of words, and its syntactic processing. Subsequently, a phonological sound sequence is encoded for the planned utterance (e.g. [1]). At lower levels of speech production, the phonological representation activates syllable-level motor plans mainly in premotor cortical areas, which is followed by motor execution, involving primary motor areas, cerebellum, basal ganglia, and the motor neuron system of speech articulators [2]. Subsequently, a temporal succession of vocal tract shapes and acoustic speech signals are generated by the peripheral vocal tract system.

In this paper, we first describe an articulatory-acoustic model that focuses on the lower levels of speech production. Then, we describe an architecture for the higher levels of speech production that can be realized using the Neural Engineering Framework [3], and can be used to generate signals that will drive the articulatory-acoustic model.

Input units for our articulatory-acoustic model are speech or vocal tract actions [4], [5]. These actions define a sequence of vocal tract shapes. The most important information extracted from each vocal tract shape, i.e. from each set of positions for the speech articulators, is the shapes of vocal tract cavities, which serve as the basis for the generation of the acoustic speech signal [4].

II. The Action-Based Approach for Controlling Speech Articulation

Based on previous work [6], [7], we assume *vocal tract action units* (also called speech actions or speech gestures) as the basic motor planning units in speech production [5]. From the viewpoint of speech learning, it is evident that infants babble a multitude of *gross vocal tract actions* in their first year, mainly leading to successions of vocal tract opening and closing actions, which can sound, for example, like [bɑbɑ], [dɑdɑ] or [gɑgɑ] [8], [9]. Thus, we can start by separating vocal tract *opening actions* and vocal tract *closing actions*. Vocal tract opening actions can also be called *vocalic actions*, because opening actions result in vowel-like sounds. Vocal tract closing actions are also called *consonantal actions*, because closing actions lead to local vocal tract constrictions as are part of consonant-like sounds, e.g. produced by the lips, by the tongue dorsum, or by the tongue tip. In addition, infants are capable of producing *velopharyngeal ab- and adduction actions* (lowering and raising the velum), which separate nasal from non-nasal sounds, and *glottal ab- and adduction actions*, which separate voiced from voiceless sounds [10]. Vocal sounds produced in these ways are not necessarily language specific but result from an infant's exploration of the vocal tract (i.e., babbling).

All types of vocal tract actions resulting from babbling are listed in Table I and it is the main goal of speech learning to 1) fine tune these (gross) vocal tract actions with respect to *spatial* as well as to *temporal intra-vocal tract action parameters* in order to later on produce different vowels and consonants of a specific target language, and to 2) fine tune the temporal coordination between different vocal tract actions by varying *inter-vocal tract action parameters* with respect to the specific prosodic characteristics of that target language. These parameters determine the temporal location of consonantal, velopharyngeal and glottal actions with respect to vocalic speech actions. An example for the temporal ordering of a monosyllabic word, "palm," produced by our control method is given in Figure 1. In our current model, the *intra-action temporal control* as well as the *inter-action timing* is defined by values for the beginning and ending of onset-, target- and offset-time interval for each vocal tract action. These time instants are elucidated in Figure 1 for all speech actions forming the word "palm". During the *onset time interval*, articulators move towards the spatial action target, e.g. to lip

978-1-4799-7280-7/14/$31.00 ©2014 IEEE

TABLE I. TYPES OF VOCAL TRACT ACTIONS, THEIR SPATIAL INTRA-VOCAL TRACT ACTION PARAMETERS, AND EXAMPLES FOR POTENTIAL SPEECH SOUNDS (SYMBOLS IN PARENTHESES []) OR SPEECH SOUND FEATURES WHICH CAN BE PRODUCED AFTER LEARNING A CORRECT PARAMETER SPECIFICATION.

type of action	spatial parameters	examples (after fine-tuning of actions)
vocalic	high-low	high: [i, u, y]; low: [ɑ]
	back-front	back: [u]; front: [i, y]
	spread-rounded	spread [i]; rounded: [u, y]
consonantal	end-effector	lips: [b, p, f, m]; tongue tip: [d, t, s, n, l]; tongue dorsum: [k, g]
	degree (type) of constriction	full closure: [b, p, m, d, t, n, k, g] (plosives and nasals) near closure: [f, s, ʃ] (fricatives) lateral closure: [l]
	location of constriction	in case of tip: [s] in "saw" vs. [ʃ] in "show"
velopharyngeal	abduction	nasal speech sounds: [m, n]
	adduction	non-nasal speech sounds (vowels, plosives, fricatives, ...)
glottal	abduction	voiceless speech sounds: [p, t, k, f, s, ʃ]
	adduction	voiced speech sounds: [b, d, g, m, n, l] and all vowels
pulmonic	pressure (resulting from movements of chest and diaphragm)	one action per utterance; constant pressure; degree of that pressure determines speech intensity of whole utterance

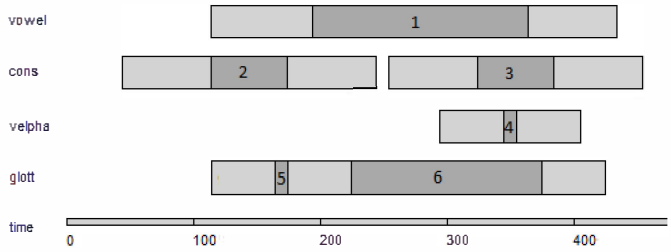


Fig. 1. Vocal tract action score of the word "palm" [pɑm]. Vocal tract actions are ordered with respect to four tiers (vocalic, consonantal, velopharyngeal, and glottal). Temporal location of onset (light gray), target (dark gray), and offset time intervals (light gray) for all six vocal tract actions of the word are shown. 1: vocalic action with low spatial target position; 2: labial full closing action; 3: labial full closing action; 4: velopharyngeal abduction action; 5: glottal abduction action; 6: glottal adduction action. The offset time interval for action 5 overlaps with the onset time interval of action 6. Time scale (bottom) is in milliseconds. Thus, actions 1 and 6 lead to [ɑ], actions 2 and 5 lead to [p], and actions 3, 4, and 6 lead to [m]. The temporal overlap of vocal tract actions can also be called coarticulation.

closure in the case of a labial closing action or to an open vocal tract shape in the case of the vocalic action in "palm". This (partial) vocal tract target shape is held during the *target time interval* (e.g. the lips are held closed during the target time interval of the [p] in "palm") and released at the beginning of the *offset time interval*.

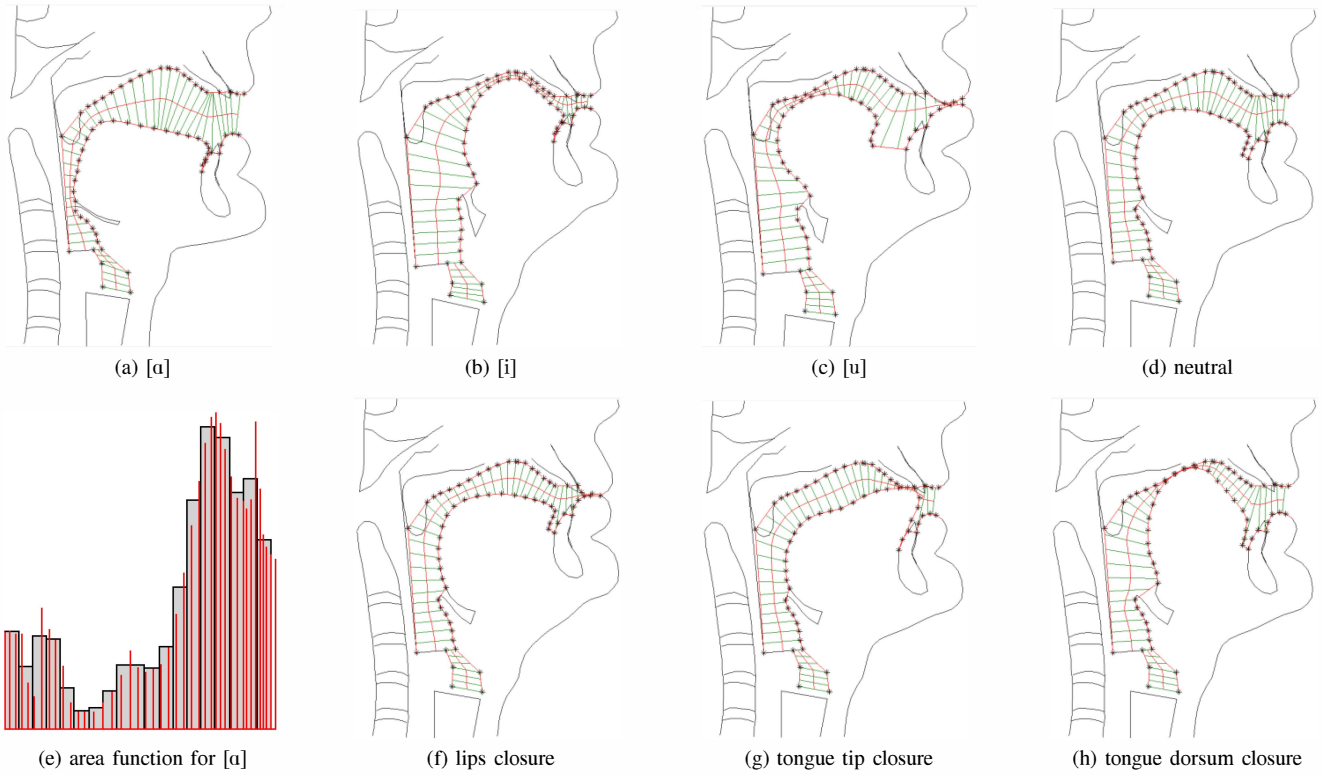
III. CONTROL PARAMETERS, VOCAL TRACT SHAPES, AND AREA FUNCTIONS

Our set of articulatory control parameters is small but *functional* from the viewpoint of speech production. Three vocalic parameters (*front-back*, *high-low*, and *spread-rounded*) control the overall shape of the vocal tract, as is needed for the production of all vowels. For example, *high-low* separates vowels like [i, y, u] from [ɑ] (vocal tract shapes for these vowels are given in Figure 2, top row). The consonantal parameters (*lip adduction*, *tongue tip elevation*, *tongue body elevation*) control *degree (or type)* of consonantal constrictions (e.g. full-closure in case of plosives or nasals, near closure in case of fricatives, lateral closure in case of laterals). Full-closure of the *lip adduction* parameter leads to production of [p, b, m]; full-closure of the *tongue tip elevation* parameter leads to production of [d, t, n]; and full-closure of the *tongue body elevation* parameter leads to production of [k, g] (see also Table I). Near-closure of the *lip adduction* parameter leads to production of [f]; near-closure of the *tongue tip elevation* parameter leads to production of [s, z, ʃ, ʒ]; and near-closure of the *tongue body elevation* parameter leads to production of [x]. Thus, while vocalic parameters control the overall shape of vocal tract, consonantal parameters control local parts of the vocal tract shape (e.g. the lip, tongue tip, or tongue dorsum regions; see Figure 2, bottom row). In addition, the parameters *glottal abduction* and *velopharyngeal abduction* control the position of the vocal folds and of the velum respectively.

As opposed to complex models, in which the *shape of the vocal tract* is generated on the basis of modeling muscle activations and the tissue structures of speech articulators (e.g. [11]), our approach is purely geometrical. The contours of speech articulators are described by the 2D locations of 14 contour points for the upper lips, 17 points for the lower lips, 23 points for the tongue, 15 points for the hard palate, 34 points for the velum, and 23 points for the pharynx wall, larynx, and epiglottis. These 2D locations are obtained from static MRI scans for three contours representing the (extremal) cardinal vowels [i], [ɑ], [u] (see [12], [13]). It is assumed that all vowel shapes can be generated by interpolating between these extremal contours using the three *vocalic control parameters* introduced above. In addition, extremal contours are generated for maximal labial adduction, maximal elevation of the tongue tip, and maximal elevation of the tongue dorsum. Consonantal vocal tract shapes within the local regions of constriction are determined by interpolating between the current underlying vocalic tract shape and the current consonantal extremal contours. In addition to the *degree of constriction* (see above), one more parameter is needed for the tongue tip, which controls the *place of articulation*, in order to differentiate between alveolar and postalveolar tongue tip constrictions (e.g. [s] or [ʃ]; see Table I). The exact location for tongue dorsum closure is controlled indirectly by the current value of the vocalic front-back parameter.

While no coarticulatory corrections are needed for the interpolation of vocalic contours (i.e., lip rounding can be independently controlled from tongue positioning without causing any problems in our geometrical model), two coarticulatory corrections are needed in the case of consonantal articulation. (1) In the case of producing a labial constriction (increasing lip adduction), the front part of tongue needs to be elevated, because lip adduction implies an elevation of lower jaw. (2) In the case of producing an apical constriction (raising the tongue tip), spatial coarticulation needs to be high for high vowels. Here, the contour of consonantal closure needs to be influenced strongly by the current underlying vocalic vocal tract shape. In our model, this problem is solved by introducing different consonantal target shapes based on the current vocalic coarticulation (see also [13]).

For calculating the *vocal tract area function*, i.e. vocal tract cavity information from a vocal tract shape, a lower line,



(a) [ɑ] (b) [i] (c) [u] (d) neutral

(e) area function for [ɑ] (f) lips closure (g) tongue tip closure (h) tongue dorsum closure

Fig. 2. Top row: vocalic midsagittal views. These vocalic views reflect a change in the shape of the entire vocal tract. Bottom row: midsagittal views of consonantal closures. These views reflect a change in the shape of a local area of the vocal tract, and can therefore be combined with the global shape defined by the vocalic context.

upper line, and midline are defined (red lines in the midsagittal views of Figure 2). The lower line represents the front-low margin, and the upper line the back-high margin of the vocal tract cavity (vocal tract tube) from larynx to lips. The midline defines the midline for air to flow through the vocal tract tube from glottis to lips. While the lower and upper lines are defined by vocal tract contour points (black asterisks in Figure 2), the calculations of the midline and the green distance lines (see Figure 2) are done in a complex iterative procedure. A set of *42 distance lines* is defined; these lines are perpendicular to the midline and thus perpendicular to the airstream within the vocal tract tube (green lines in Figure 2). These distance lines are the basis for calculating the area function (an example of an area function is given in Figure 2e) for the vocal tract shape of [ɑ]. (cf. [14]).

IV. Modeling Vocal Tract Acoustics

Input for the calculation of the acoustic speech signal is the area function and glottal flow. The glottal flow is calculated using an LF-model derivate [15]. Flow and pressure values within the vocal tract are calculated using a reflection-type line analog [16]. Here, the geometry of the vocal tract tube is "digitized" into a succession equidistant cylindrical tubes (see the gray bars, representing the area function, in Figure 2e). The length of each tube segment is 0.875 cm in our case of 20 kHz sampling frequency for the acoustic signal. Flow and pressure values are calculated for each tube segment at each time instant. Acoustic and aerodynamic loss mechanisms including losses due to sound radiation at the mouth are included [16]. Because the length of the vocal tract and thus the number of tube segments varies with respect to overall tube length (e.g. longer tract length and thus more tube segments for [u] compared to [i] or [ɑ]), and because the reflection-type line analog cannot handle varying tube length easily, the vocal tract shape over time is evaluated only once per glottal pulse (glottal period) and is thus assumed to be constant for that time period. The resulting radiated sound signal (pulse response) is calculated over a time interval of two glottal periods (see Figure 3) and is overlayed pulse by pulse in time, in order to form the resulting speech sound signal.

V. The Neural Engineering Framework

The Neural Engineering Framework (NEF; [3]) provides three principles for *representing* and *transforming* information *dynamically* using feedforward and recurrently connected networks of spiking neurons. Nengo is a neural simulation environment that uses the NEF to build large-scale brain models [17]. The NEF and Nengo have been used to create models of visual object recognition and copy drawing of manually drawn digits by performing visual perception, cognitive tasks, and motor tasks with networks of spiking neurons [18], [19]. These cognitive and sensorimotor tasks are performed by complex brain models which are made up of many networks representing cortical circuits, basal ganglia, thalamus, as well as peripheral sensory processing and motor outputs. Each of these models uses a simulated spiking neuron approximation, usually LIF (leaky-integrate-and-fire) neurons (see [19], p. 35ff). Moreover, many of the models created with the NEF use

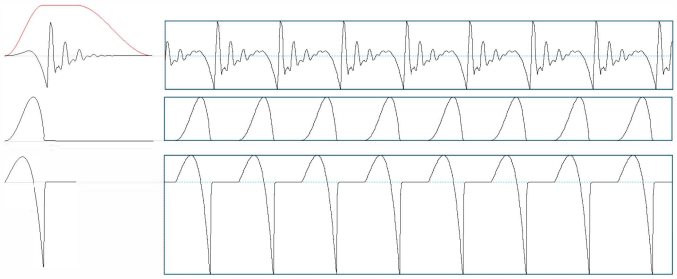


Fig. 3. Left side: single glottal pulse (middle: glottal flow; bottom: its time derivative for exactly one glottal period) and pulse answer for an [ɑ], radiated from mouth (top). Right side top: acoustic speech signal, resulting from overlayed pulse answers; middle: glottal flow waveshape for 9 glottal cycles; bottom: first time derivative of glottal flow. The Hanning window for temporal overlay of pulse responses is asymmetric. Onset ends at the time instant of maximum glottal excitation (i.e. negative peak of time derivative of glottal flow) and offset interval begins at end of glottal cycle. Right side signals are displayed using the PRAAT software to visualize the WAV file generated by our synthesizer; left side signals are displayed from the synthesizer software directly.

a cortex-basal ganglia-thalamus-cortex loop that is capable of modeling action selection and action execution, as is needed in order to simulate communication (e.g. question-answering scenarios).

We believe that this approach can be used to model speech production, because its action selection and execution mechanisms can be extended or modified and thus can meet the demands occurring in face-to-face interactions; see [20]. In the next section of this paper, a preliminary architecture for speech production is introduced.

VI. The Architecture of the Speech Production Model

A speech production model should be made up of a cognitive component (mental lexicon and general communicative knowledge) as well as a sensorimotor component (speech action repository, SAR, and a production-perception loop; see [5], [8], [9], [21]–[24]). It is a key feature of our ongoing work on modeling speech production that phonological representations arise during early phases of speech acquisition and are not predefined in the model at the beginning [8], [9], [20]. Lexical items (semantic as well as phonological representations) and phonetic (i.e. hypermodal sensorimotor) representations of syllables within the speech action repository [5], [21]–[24] can be represented in the NEF using the semantic pointer architecture (SPA; see [19], p. 77ff). The word "semantic" is not used in the SPA in a narrow linguistics sense; semantic pointers do not exclusively represent meanings of words, phrases or sentences but can represent motor states (e.g. the motor plan of a complete syllable or the motor plan of a target-directed hand-arm gesture) or sensory states (e.g. auditory states of syllables, words or phrases, visual states). Thus, a semantic pointer in the SPA can be used to describe discrete cognitive processing units as well as sensory and/or motor states (e.g. phonetic states of syllables as are defined in the SAR [5], [21]–[24]).

An advantage of the SPA is that it connects cognitive, sensory, and motor states. A comprehensive brain model including cognitive, sensory and motor modules called Spaun

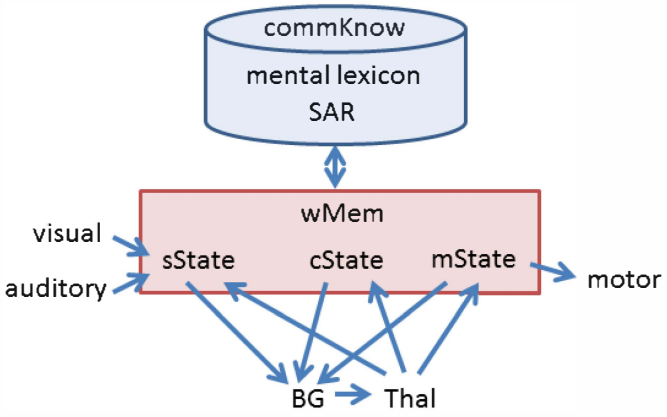


Fig. 4. High-level NEF model architecture for syllable and word processing. "commKnow", "mental lexicon" and "SAR" represent a neural long-term knowledge (communicative knowledge, mental lexicon, and syllable action repository). "wMem" represents the working memory, "BG" the basal ganglia network, and "Thal" the thalamus network. The working memory includes current high level sensory, cognitive, and motor plan states ("sState, cState, mState"). High level sensory states can be activated from sensory input (bottom-up) or from mental lexicon (top-down). Cognitive states like phonemic representations of syllables may be activated from visual input (e.g. reading letters) via BG-Thalamus and mental lexicon. Motor plans may be activated from phonemic representations in working memory via BG-Thalamus and SAR as well as a learned auditory states of the currently activated syllables.

has been developed with the NEF and Nengo (see [18] and [19], p. 247ff). Spaun uses visual information (an image of a hand-written digit) to do eight tasks, including copy drawing, pattern completion, and a reinforcement learning task similar to gambling. It provides its output by producing motor commands that drive a simulated three-link arm to write digits. We believe that the motor system of Spaun can be augmented by a speech motor component, i.e. by a speech articulator system, which can be implemented in parallel to the already existing arm motor component. A discussion of similarities and differences of controlling hand-arm motor system, articulator motor system and facial motor system (in face-to-face communication scenarios) is given in [5]. Moreover, the perceptual system within Spaun can be augmented by an auditory perceptual system in order to allow speech acquisition and speech perception. This auditory perceptual component can be implemented in parallel to the already existing visual perceptual component (see Figure 4). Upon augmenting Spaun's sensory and motor modules, similar cognitive tasks as are currently performed by Spaun through seeing and writing digits can instead be done by hearing and speaking the words corresponding to those digits.

One further advantage of Spaun is the neurobiological representation of the cortex-basal ganglia-thalamus-cortex loop in order to model action selection and control of perception-action tasks (see [19], p. 163ff). We believe that the concepts introduced by [19] for control of visual-perception-manual-action tasks are applicable in a similar way for auditory-perception-articulatory-speech-action tasks.

Sensorimotor knowledge concerning the motor and auditory state of syllables as well as a reference pointer towards the phonemic representation of each syllable is stored in the "SAR" (speech action repository) in form of predefined

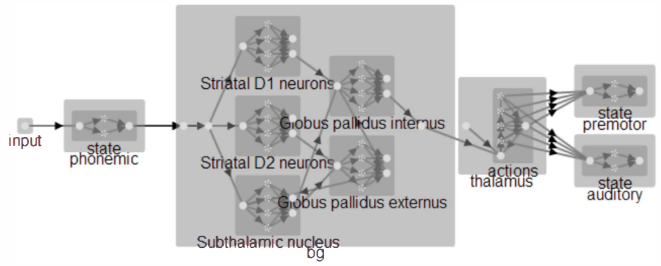


Fig. 5. Structure of the syllable sequencing network. "bg" refers to the basal ganglia. The input semantic pointer is represented by the phonemic state network. That semantic pointer is communicated to premotor and auditory networks through a basal ganglia-thalamus network. See text for more details.

(i.e. learned) semantic pointers (Figure 4). Basic behavioral knowledge for face-to-face communication in speech production scenarios is stored in the "commKnow" (communicative knowledge) module in form of predefined (i.e. learned) semantic pointers. The mental lexicon as well is part of "commKnow". Syllable state pointers (e.g., for representing the syllables "ba", "da", "ga") as well as semantic pointers of communication scenarios (e.g., for representing actions like "listen to a communication partner", "produce a syllable, word or phrase") can be activated at the level of the state networks (Figure 4) based on neural representations stored in long term memory and based on actual audiovisual input (for example from a communication partner / interlocutor). This information is processed in working memory as well as in the cortex-basal ganglia-thalamus-cortex loop in order to generate and activate motor plans (right state network) and in order to directly control motor execution for articulation.

The size of the network components depends on the tasks which need to be performed. In order to perform a speech production task (e.g. syllable sequencing) as well as a more complex task including listening to a communication partner (e.g. a question answering task), the size of each cortical state network is 3000 LIF neurons, and the sizes of the visual, auditory, and motor components are 300 LIF neurons each. The size of the recurrent network representing working memory is 1000 LIF neurons. The basal ganglia is comprised of 5 subnetworks with 600 LIF neurons each (3000 neurons in total; see [19], p. 164ff). The thalamus is composed of a network of 750 LIF neurons (see [19], p. 169ff).

The structure of a syllable sequencing subnetwork as is visualized by the Nengo GUI[1] is shown in Figure 5. Here, the input stimuli are predefined sequences of semantic pointers which can be interpreted as visual input. This input sequence directly activates the phonemic representation of the syllable. This cortical state information forms the input signal for the basal ganglia-thalamus part of the network. This part of the network subsequently activates the premotor and auditory state of the syllable sequence. It can be seen that due to the basal ganglia-thalamus loop, the activation of the premotor and auditory states are delayed by around 50 ms with respect to phonemic input (see Figure 6).

[1]The Nengo GUI visualizes networks created in Nengo [17]. It is currently under development at https://github.com/ctn-waterloo/nengo_gui

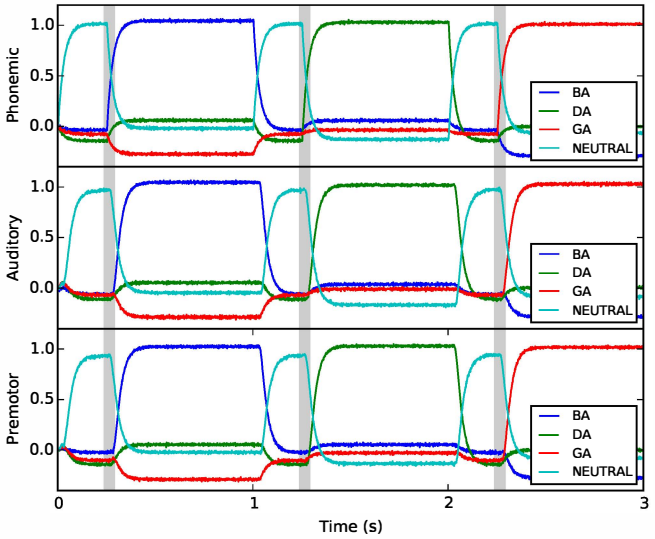


Fig. 6. Information flow through the syllable sequencing network shown in Figure 5. Information in the phonemic (top), auditory (middle), and premotor (bottom) populations are shown. Colored lines represent the similarity between the representation in the population and the target semantic pointer; the target semantic pointer is "BA" for the blue line, "DA" for the green line, "GA" for the red line, and "NEUTRAL" (i.e., no speech) for the cyan line. The shaded grey regions are 50 ms wide and show the delay between the phonemic and auditory/premotor representations.

VII. Discussion and Further Work

First steps in the direction of using the NEF, SPA, and Nengo in order to model speech production and thus to contribute to cognitive information communication issues [25] are described in this paper. These tools are advantageous for modeling speech production because this approach operates in continuous time, and is robust to the noise introduced by manipulating information with spiking neurons. This allows us to model aspects of speech production which are beyond the scope of other approaches. In particular, aspects of face-to-face communication in speech due to perception-action routing in the brain and specific aspects of speech disorders due to different degrees of neural noise can now be investigated in more detail.

We have also presented a functional articulatory-acoustic model. This model is capable of modeling the processes of varying intra- and inter-speech action parameters, i.e. for fine-tuning of action targets, for action onset-, target-, and offset-interval lengths, and for establishing the temporal relation between different speech actions involved in forming a syllable or word (cf. babbling and imitation training [8], [9]). Because these simulations of speech learning demand the generation of an abundance of speech items, our model is designed for generating speech items near real time. Integrating nasal tract and noise sources for the generation of nasals and fricatives respectively are the next tasks to be completed. Then, we will conduct a study on the perceptual evaluation of speech items produced by our articulatory model and we will start to implement associative sensorimotor learning using NEF, SPA, and Nengo.

REFERENCES

[1] W. J. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behavioral and Brain Sciences*, vol. 22, no. 01, pp. 1–38, 1999.

[2] A. Riecker, K. Mathiak, D. Wildgruber, M. Erb, I. Hertrich, W. Grodd, and H. Ackermann, "fmri reveals two distinct cerebral networks subserving speech motor control," *Neurology*, vol. 64, no. 4, pp. 700–706, 2005.

[3] C. Eliasmith and C. H. Anderson, *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press, 2003.

[4] B. J. Kröger, "A gestural production model and its application to reduction in german," *Phonetica*, vol. 50, no. 4, pp. 213–233, 1993.

[5] B. J. Kröger, S. Kopp, and A. Lowit, "A model for production, perception, and acquisition of actions in face-to-face communication," *Cognitive Processing*, vol. 11, no. 3, pp. 187–205, 2010.

[6] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989.

[7] L. Goldstein, D. Byrd, and E. Saltzman, "The role of vocal tract gestural action units in understanding the evolution of phonology," *Action to Language via the Mirror Neuron System*, pp. 215–249, 2006.

[8] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, vol. 51, no. 9, pp. 793–809, 2009.

[9] B. J. Kröger, J. Kannampuzha, and E. Kaufmann, "Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception," *EPJ Nonlinear Biomedical Physics*, vol. 2, no. 1, pp. 1–28, 2014.

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

[11] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 853–870, 2004.

[12] B. J. Kröger, J. Gotto, S. Albert, and C. Neuschaefer-Rube, "A visual articulatory model and its application to therapy of speech disorders: a pilot study," *Speech Production and Perception: Experimental Analyses and Models. ZAS Papers in Linguistics*, vol. 40, pp. 79–94, 2005.

[13] B. J. Kröger, P. Hoole, R. Sader, C. Geng, B. Pompino-Marschall, and C. Neuschaefer-Rube, "Mrt-sequenzen als datenbasis eines visuellen artikulationsmodells," *HNO*, vol. 52, no. 9, pp. 837–843, 2004.

[14] P. Perrier, L.-J. Boë, and R. Sock, "Vocal tract area function estimation from midsagittal dimensions with ct scans and a vocal tract castmodeling the transition with two sets of coefficients," *Journal of Speech, Language, and Hearing Research*, vol. 35, no. 1, pp. 53–67, 1992.

[15] R. Veldhuis, "A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 566–571, 1998.

[16] J. Liljencrants, *Speech synthesis with a reflection-type line analog*. Royal Institute of Technology, 1985, vol. 85, no. 2.

[17] T. Bekolay, J. Bergstra, E. Hunsberger, T. DeWolf, T. C. Stewart, D. Rasmussen, X. Choo, A. R. Voelker, and C. Eliasmith, "Nengo: A python tool for building large-scale functional brain models," *Frontiers in Neuroinformatics*, vol. 7, no. 48, 2014.

[18] C. Eliasmith, T. C. Stewart, X. Choo, T. Bekolay, T. DeWolf, Y. Tang, and D. Rasmussen, "A large-scale model of the functioning brain," *Science*, vol. 338, no. 6111, pp. 1202–1205, 2012.

[19] C. Eliasmith, *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford University Press, 2013.

[20] B. J. Kröger, P. Birkholz, and C. Neuschaefer-Rube, "Towards an articulation-based developmental robotics approach for word processing in face-to-face communication," *Paladyn: Journal of Behavioral Robotics*, vol. 2, no. 2, pp. 82–93, 2011.

[21] B. J. Kröger, J. Kannampuzha, C. Eckers, S. Heim, E. Kaufmann, and C. Neuschaefer-Rube, "The neurophonetic model of speech processing act: structure, knowledge acquisition, and function modes," in *Cognitive Behavioural Systems*. Springer, 2012, pp. 398–404.

[22] B. J. Kröger, P. Birkholz, J. Kannampuzha, E. Kaufmann, and C. Neuschaefer-Rube, "Towards the acquisition of a sensorimotor vocal tract action repository within a neural model of speech processing," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*. Springer, 2011, pp. 287–293.

[23] C. Eckers, B. J. Kröger, K. Sass, and S. Heim, "Neural representation of the sensorimotor speech–action-repository," *Frontiers in Human Neuroscience*, vol. 7, 2013.

[24] C. Eckers, B. J. Kröger, and S. Heim, "The speech action repository: Evidence from a single case neuroimaging study," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 128–135, 2013.

[25] P. Baranyi and A. Csapo, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, pp. 67–83, 2012.