MODELING SPEECH PROCESSING USING NENGO: FIRST STEPS

Bernd J. Kröger[1,2]

[1]*Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany*

[2]*Cognitive Computation and Applications Laboratory, School of Computer Science and Technology, Tianjin University, Tianjin, P.R.China*

bernd.kroeger@rwth-aachen.de

Abstract: In our previous modeling of speech processing, SOM and GSOM approaches were used for building up a language specific syllable based speech action repository (SAR) within a neurobiologically plausible model of speech acquisition and speech processing. Because time as well as internal neural noise are not explicitly modeled in SOM and GSOM approaches, it will be difficult to simulate speech disorders in such a model. Therefore we now started using the NENGO neural simulator (http://nengo.ca/) in order to increase the neurobiological realism of our approach. It will be shown that this new approach allows modeling temporal aspects of action selection and action execution. Two examples will be given: (1) sequencing of syllable production and (2) switching between perceiving and reproducing words. Moreover the architecture of our NENGO speech processing model will be introduced and discussed in this paper.

1 Introduction: The NENGO Environment

NENGO (Neural ENGineering Objects; see http://nengo.ca/) is a neural simulation environment capable of modeling for example visual object recognition or copy drawing of manually drawn digits by performing visual perception, cognitive tasks, as well as motor tasks for controlling a two-joint arm [1, 2]. These cognitive and sensorimotor tasks are performed on the basis of a complex brain model which comprises many networks representing cortical circuits, basal ganglia, thalamus, as well as peripheral sensory processing and motor control. Each neural (sub-)network within this approach is based on spiking model neurons, mainly LIF (leaky-integrate-and-fire) neurons ([2], p. 35ff). Moreover this simulation environment comprises a cortex-basalganglia-thalamus-cortex loop capable of modeling action selection and action execution as is needed in order to simulate for example communication behavior (e.g. question-answering scenarios).

We believe that this approach is helpful for modeling speech acquisition and speech processing, because its action selection and execution mechanisms can be extended or modified for modeling speech acquisition (including face-to-face interaction between baby and caretaker; see [3]) as well as for adult speech processing (i.e. production and perception). In the next section of this paper, a preliminary architecture for speech processing is introduced. The third section of this paper describes two simulation experiments which highlight some key features of NENGO which are important for speech processing: (1) simulation of syllable sequencing and (2) simulation of action selection and execution in a listening and articulatory reproduction task.

2 The Architecture of the Speech Processing Model

A speech processing model should comprise a cognitive component (mental lexicon) as well as a sensorimotor component (speech action repository SAR and a production-perception loop, see [3-7]). It is a key feature of our ongoing work on modeling speech acqusition and speech processing that phonological representations arise during early phases of speech acquisition and are *not* predefined in the model at the beginning of speech acquisition [3]. Lexical items (semantic as well as phonological representations) as well as phonetic (i.e. hypermodal sensorimotor) representations of syllables within the speech action repository [6, 8, 9] can be represented in NENGO using the semantic pointer architecture (SPA, see [2], p. 77ff). The word "semantic" is not used in NENGO in a narrow linguistics sense. Thus, semantic pointers are not used exclusively to represent meanings of words, phrases or sentences but can represent motor states as well, e.g. motor state (motor plan) of a complete syllable or motor state (motor plan) of a target-directed hand-arm gesture, or can represent sensory states, e.g. auditory states of syllables, words or phrases, visual states, etc. Thus, a semantic pointer in NENGO 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 SAR [6, 8, 9]).

An advantage of the NENGO framework is that it connects cognitive, sensory, and motor sates. A comprehensive brain model including cognitive, sensory and motor modules, called SPAUN ([1] and [2], p. 247ff) has been developed based on NENGO. 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 manual 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 [10]. 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 1).

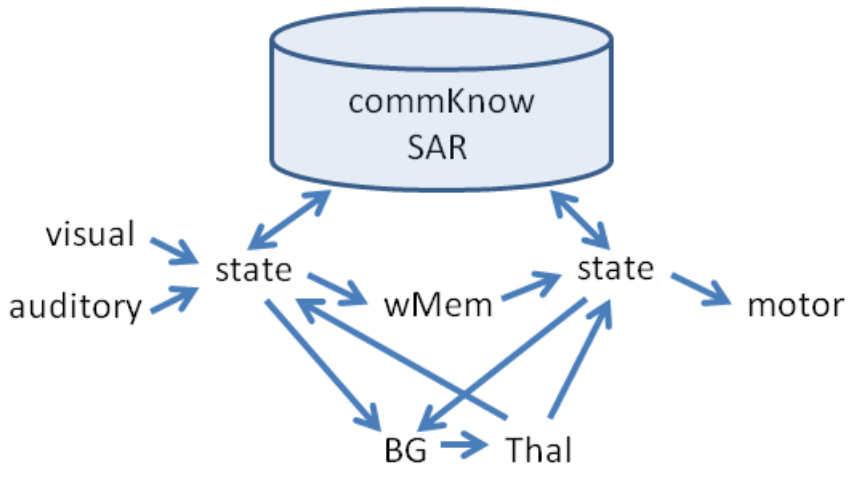

Figure 1. NENGO architecture for syllable and word processing. "commKnow" and "SAR" represent a neural long-term knowledge (communicative knowledge like mental lexicon and syllable action repository). "wMem" represents the working memory, "BG" the basal ganglia neural network, and "Thal" the thalamus neural network. Semantic pointers can be activated and processed in both state networks on the basis of audiovisual input (left state network: perceptual state network) or at the level of motor planning (right state network: action planning network; see also text).

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 ([2], p. 163ff). We believe that the concepts introduced by [2] for control of visual-perception-manual-action tasks are applicable in a similar way for auditory-perception-articulatory-speech-action tasks.

Sensorimotor as well as semantic knowledge concerning the production and auditory state of syllables as well as of simple (mono-syllablic) words, semantic as well as basic behavioral knowledge for face-to-face communication in speech acquisition scenarios is stored in the SAR (speech action repository) and commKnow (communicative knowledge) module in form of predefined (i.e. learned) semantic pointers (Figure 1). Syllable state pointers (for example representing the syllables “ba”, “ga”, “ga”, …) as well as semantic pointers of communication scenarios (for example 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 1), 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 fulfill the tasks discussed in the next section (syllable sequencing in speech production and control of question-answering scenarios in speech communication), the size of each cortical state network is 3000 model neurons each, the size of the visual and auditory component as well as of the motor component is 300 model neurons each. The size of the recurrent network representing the working memory is 1000 model neurons. The basal ganglia comprises 5 subnetworks with 600 model neurons each (3000 model neurons in total, see [2], p. 164ff). The thalamus is represented by a network of 750 model neurons (see [2], p. 169ff).

3 Simulation Experiments:

3.1 Sequencing of syllables

For the first simulation experiment, five syllables were activated and sequenced as “bigibadaga” (nonsense word) using the NENGO architecture described in section 2. Each syllable can be represented using a semantic pointer at the level of the state networks. The sequence of syllable production is learned in advance (stored in long-term memory) and the timing of syllable sequencing and thus the activation of syllable pointers at the level of the motor state network is generated by the cortex-basal-ganglia thalamus-cortex loop of our model. Figure 1 displays the neural activation pattern at the level of the motor network. Each syllable state is represented here by one semantic pointer.

3.2 Listening and articulatory reproduction

For the second simulation experiment, a perception period followed by a production period and so on is simulated using the NENGO architecture described in section 2. Firstly, the semantic pointers for “listening” as well as the syllable “one” were activated at about 0.3 s (arrow “perc_1” in Figure 3) at the level of the perceptual state network (left side in Figure 1). These semantic pointers are selecting from long-term memory (“commKnow” part in Figure 1). No processing of visual input (in order to get the information “listen”) or of auditory input (in order to get the information “one”) is done at the current state of implementation. The auditory and motor information “one” is copied from long-term memory (SAR part in Figure

1) and hold in working memory. The semantic pointer for production of the word “one” is activated at the level of the motor state network (right state network in Figure 1) if the neural activity of the semantic pointer for “listen” decreases and the activity for the semantic pointer representing “reproduce” increases. This occurs at about 1 s (arrow “prod_1” in Figure 3) in our simulation example. The cortico-cortical basalganglia-thalamus loop is involved in this procedure as well. This two-state cycle (listening and reproduction) is activated for a second time for the syllable “two” (arrow “perc_2” and “prod_2 in Figure 3).

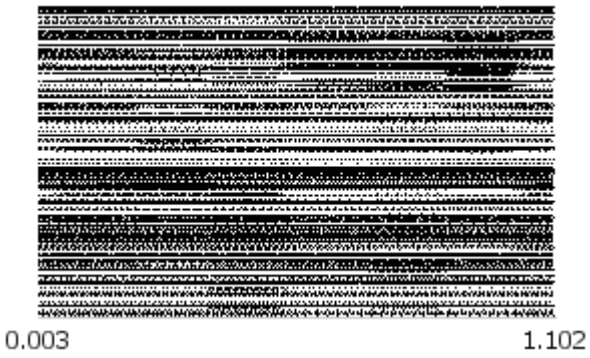

Figure 2. Neural spike raster occurring at the level of the motor network for a group of 100 model neurons during production of the syllable sequence “bigibadaga”. The time interval shown is about 1 s. Syllable “bi” starts around 0.3s. Duration of each syllable is about 150ms. The spike raster before first syllable and after fifth syllable represents neural activity for rest position of articulation.

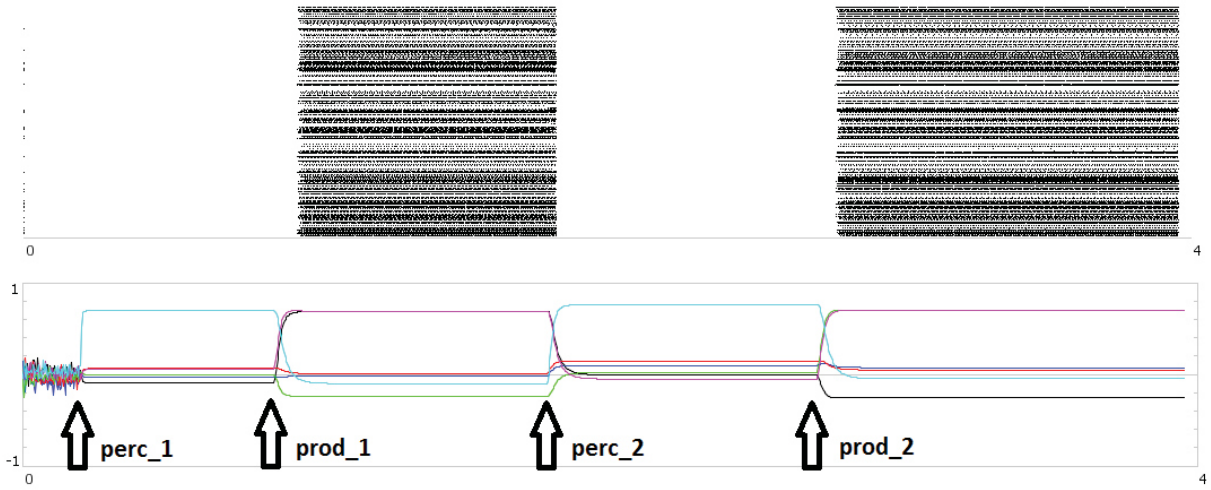

Figure 3. Neural spike raster occurring at the level of the motor state network (on top) and display of six semantic pointers (i.e. “listen”, “reproduce”, “one”, “two”, “three”, “four”) occurring at the level of the perceptual state network (below) over a time period of four seconds during performance of the listening and reproduction task (top: group of 100 model neurons from each state network). Mainly the semantic pointers for “listen” (cyan) and “reproduce” (magenta/pink) are represented within the semantic pointer display (for more information see text).

As already stated above, the switching of some network activity from listening to reproduction is mainly done by the corical-basal-ganglia-thalamus-cortical loop within our network. On the one hand this cortico-cortical loop inhibits the forwarding of neural activity towards the motor state network during listening (see neural connection of thalamus with motor state network in Figure 1) and on the other hand this loop inhibits forwarding of neural activity towards working memory from the sensory state network during production of a speech item (see neural connection of thalamus with perceptual state network in Figure 1).

4 Discussion and Conclusions

First steps in the direction of using NENGO (http://nengo.ca/) in order to model speech acquisition and speech processing are described in this paper. The NENGO framework seems to be advantageous for modelling speech acquisition and speech processing because this approach includes modelling of time as well as internal neural noise generation due to use of LIF neurons in a natural and straight forward way. This allows us to model aspects of speech acquisition and speech processing which are beyond the scope of our earlier SOM and GSOM based approaches. Especially aspects of face-to-face communication in speech acquisition due to perception-action routing in the brain and specific aspects of speech disorders due to different degrees of internal neural noise excitation can now be investigated now in more detail.

Literature

[1] Eliasmith C, Stewart TC, Choo X, Bekolay T, DeWolf T, Tang Y, Rasmussen D (2012) A large-scale model of the functioning brain. Science 338, 1202-1205

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

[3] Kröger BJ, Birkholz P, Neuschaefer-Rube C (2011) Towards an articulation-based developmental robotics approach for word processing in face-to-face communication. PALADYN Journal of Behavioral Robotics 2: 82-93

[4] Kröger BJ, Kannampuzha J, Kaufmann E (in press) Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. EPJ Nonlinear Biomedical Physics

[5] Kröger BJ, Kannampuzha J, Eckers C, Heim S, Kaufmann E, Neuschaefer-Rube C (2012) The neurophonetic model of speech processing ACT: structure, knowledge acquisition, and function modes. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller VC (eds.) Cognitive Behavioural Systems, LNCS 7403 (Springer, Heidelberg, Berlin), pp. 398-404

[6] Kröger BJ, Birkholz P, Kannampuzha J, Kaufmann E, Neuschaefer-Rube C (2011) Towards the acquisition of a sensorimotor vocal tract action repository within a neural model of speech processing. In: Esposito A, Vinciarelli A, Vicsi K, Pelachaud C, Nijholt A (eds.) Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues. LNCS 6800 (Springer, Berlin), pp. 287-293

[7] Kröger BJ, Kannampuzha J, Neuschaefer-Rube C (2009) Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793-809

[8] Eckers C, Kröger BJ, Sass K, Heim S (2013) Neural representation of the sensorimotor speech-action-repository. Frontiers in Human Neuroscience 7:121

[9] Eckers C, Kröger BJ, Heim S (2013) The speech action repository: Evidence from a single case neuroimaging study. In: Wagner P (ed.) Studientexte zur Sprach-kommunikation: Elektronische Sprachsignalverarbeitung 2013 (TUDpress, Dresden, Germany), pp. 128-135

[10] Kröger BJ, Kopp S, Lowit A (2010) A model for production, perception, and acquisition of actions in face-to-face communication. Cognitive Processing 11: 187-205