MODELING MOTOR PLANNING IN SPEECH PRODUCTION USING THE NEURAL ENGINEERING FRAMEWORK

Bernd J. Kröger[1], Trevor Bekolay[2] & Peter Blouw[2]

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

[2]Centre for Theoretical Neuroscience, University of Waterloo, Canada

bernd.kroeger@rwth-aachen.de, tbekolay@gmail.com, pblouw@gmail.com

Abstract: Background: Currently, there exists no comprehensive and biologically inspired model of speech production that utilizes spiking neuron. Goal: We introduce a speech production model based on a spiking neuron approach called the Neural Engineering Framework (NEF). Using the NEF to model temporal behavior at the neural level in a biologically plausible way, we present a model of the temporal coordination of vocal tract actions in speech production (i.e. motor planning) with neural oscillators. Method: Neural oscillators are postulated in our model at the syllable and vocal tract action level. They define relative or intrinsic time scales for each vocal tract action as well as for each syllable and thus allow intrinsic timing or phasing of speech actions. Results: The model is capable of producing a sequence of syllable-sized motor plans that generate muscle group activation patterns for controlling model articulators. Simulations of syllable sequences indicate that this model is capable of modeling a wide range of speaking rates by altering individual syllable oscillator frequencies. Conclusions: This approach can be used as a starting point for developing biologically realistic neural models of speech processing.

1 Introduction

Only a few biologically inspired neural models of speech production are available (e.g. [1-6]). None of these models use spiking neuron models and only one of these models [4-6] includes the sensorimotor repository in speech production, i.e. the *mental syllabary* (see [7-9]). Thus, there is a need for further efforts in modeling speech production using spiking neuron models and an implementation of the mental syllabary.

Different entities need to be represented as neural states in speech production (e.g. concepts, words, syllables vocal tract actions, muscle group activation levels for speech articulator movements, etc.). Syllable states occur in different domains, i.e., in the phonological, motor, auditory, and somatosensory domains. The corresponding neural state representations in each of these four domains establish the mental syllabary. The *processing* of these representations – e.g. the establishment of speech production from concept activation via the activation of lexical and syllable items – is done by implementing connections between different neuron ensembles. The Neural Engineering Framework (NEF; see [10-12]) allows state representations and transformations of these representations to be implemented in biologically plausible neural models. Specifically, we use leaky integrate-and-fire *neuron ensembles* to represent both cognitive and sensorimotor states (though neuron models other than the LIF model can be used in the NEF).

The NEF is comprised of three principles concerning representation, transformation and dynamics [10]. The principle of *representation* establishes mechanisms for encoding and decoding signals or states from *activity patterns* occurring in neuron ensembles. These neural activity patterns can be thought of as neural representations of signals or states. The principle of *transformation* specifies how to connect one neural ensemble to another so as to compute an arbitrary function of the state or signal represented by the first ensemble. The principle of

dynamics specifies how to use recurrently connected neuron ensembles to implement *neural buffers* or *neural memories*. These buffers and memories can be thought of as repositories for storing neural representations. A further important feature of recurrently connected neuron ensembles is that they can be used to implement *neural oscillators*.

On the basis of task dynamics and coupled oscillator theory within the framework of articulatory phonology [13, 14], it has been hypothesized that vocal tract actions are intrinsically timed by the behavior of harmonic oscillators whose states reflect the state of vocal tract actions. This intrinsic timing allows for a relative timing or "phasing" of different vocal tract actions within a syllable and between syllables. Thus, the intrinsic timing specifies the temporal coordination of vocal tract actions within and between syllables. It is the aim of this paper to introduce a comparable approach for modeling the temporal coordination of vocal tract actions in a biologically based and quantitative manner using the NEF. Simulation results from a spiking neuron model of speech production using intrinsic timing are presented in subsequent sections. Key features of this model will also be discussed.

2 Method

2.1 The model

The neural model (Fig. 1) includes cortical and subcortical components. The initiation of syllable production is triggered by *visual input* (written syllables). The input is encoded in a *visual input neuron ensemble* (labeled as "vision" in Fig. 1) and then processed by model components corresponding to the *basal ganglia* and *thalamus*. The neural output from thalamus activates a premotor representation for each visually initiated syllable within the model components labeled *premotor syllable buffer* and *premotor syllable associative memory*, which subsequently activates a set of recurrently connected neuron ensembles (i.e., neural oscillators). Each neural oscillator represents a specific syllable at the *premotor syllable level* (three syllable oscillators are shown in Fig. 1). Basal ganglia and thalamus implement an action selection system that controls the sequencing of syllables and the initiation of each syllable oscillator [15].

The *neural syllable oscillators* occurring at the premotor syllable level activate an "internal clock" for syllable production and subsequently define the time points at which each vocal tract action (also labeled as "speech action" or "gesture") must be activated (for a review of the concept of vocal tract actions see [16]). The frequency of these syllable oscillators (*syllable oscillator frequency*) is dependent on the rate of speech and syllable stress level. An increase in speaking rate is realized by an increase in syllable oscillator frequency, which shortens the duration of each syllable. A higher syllable stress level is realized by lowering the syllable oscillator frequency, because stressed syllables are voiced for longer durations.

All *vocal tract actions* are represented as neural oscillators as well (see vocal tract action level in Fig. 1). Thus, at the level of each *vocal tract action oscillator*, a further intrinsic temporal scale is defined which mainly specifies the duration of the articulator movements controlled by this vocal tract action from the time point at which the action starts to the time point at which the articulatory target (e.g., a consonantal constriction or closure, a vocalic tract shape, a velopharyngeal closure as needed for obstruents or a velopharyngeal opening as needed for nasals, a glottal configuration for phonation, or a glottal opening as needed for voiceless sounds) is reached. This temporal phase is called the *movement phase* of a speech action, while the following time period until the speech action ends is called the *target phase* (the movement phase is called the "transition portion" in [16]). During the target phase, the speech action has reached its articulatory goal. In the case of constriction forming

speech actions (consonantal speech actions), this phase often indicates saturation (ibid.) due to the contact of articulators with each other (e.g., the upper and lower lips) or the contact of articulators with vocal tract walls (e.g., the tongue tip or tongue dorsum with the palate).

Subsequently, each vocal tract action generates a time dependent activation of specific muscle groups which control the movement of the articulators involved in the realization of a specific vocal tract action. Each muscle group is represented by a specific neuron ensemble in our model. The twelve *muscle group neuron ensembles* build up the *muscle group activation level*.

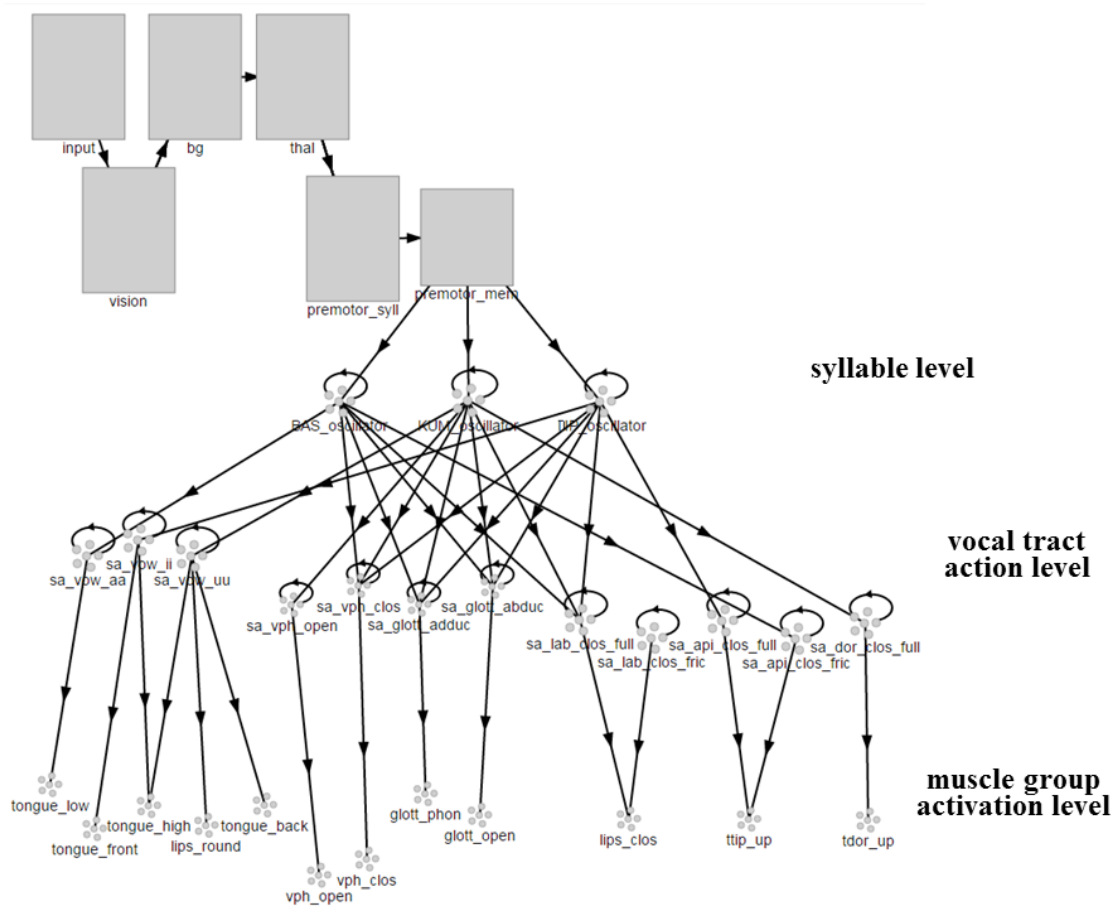

Figure 1 – Structure of the neural model for the mental syllabary (see also text): bg = basal ganglia, thal = thalamus, syll = syllable buffer, mem = memory; oscillators are defined here for three syllables only: /bas/, /kum/, and /dip/; types of vocal tract actions (also called sa = speech actions): vow = vocalic actions, vph = velopharyngeal actions, glott = glottal actions, lab = labial, api = apical, dors = dorsal actions, clos_full = full closing action, clos_fric = near closing actions for fricatives; muscle groups are defined for reaching low, fronted, or high tongue position (tongue_low, tongue_front, tongue_high) , rounded lips (lips_round), opened or closed velopharyngeal port (vph_open, vph_clos), opened glottis (glott_open), closed glottis for phonation (glott_phon), closed lips (lips_clos), consonantal upward position of tongue tip or tongue dorsum (ttip_up, tdors_up).

Our model postulates four cortical layers that organize the preparation and execution of a syllable (Fig. 1): (i) At the premotor buffer and premotor associative memory, the sequence of go-signals for a syllable sequence is stored. (ii) At the premotor syllable level, the overall time interval for the execution of a syllable and the time points for the temporal coordination of all vocal tract actions within a specific syllable are determined. (iii) At the vocal tract action level, the execution of each specific vocal tract action as part of a specific syllable is prepared. (iv) At the muscle group activation level (assumed to be located in primary motor cortex), the

neuromuscular activation patterns for controlling the set of speech articulators over time are generated.

It can be seen from Fig. 1 that each neural oscillator within the premotor syllable layer (representing a specific learned syllable of the target language) is connected only with those speech action oscillators which are needed for the realization of that syllable. Further, the neural connections between the syllable oscillators and the vocal tract action oscillators indicate which vocal tract actions are needed for the articulatory realization of which syllable. In a comparable way, the vocal tract action oscillators are connected only with those muscle group neuron ensembles that are needed for the realization of that vocal tract action.

2.2 Simulation of speech production

The sequencing of three CVC syllables is simulated at four different rates of speech. These CVC syllables are composed from three *vowels* and different types of consonants. For vowels, we use a high front vowel /i/, a high back vowel /u/, and a low vowel /a/ (see Fig. 2c and Fig. 2d). For consonants, we use (i) *voiced plosives,* which comprise a full closing action (labial, apical, dorsal), a velopharyngeal closing action, and a glottal phonation action (see /b/ and /d/ in Fig. 2c and Fig. 2d). We use (ii) *nasals,* which differ from voiced plosives by replacing the velopharyngeal closing action with a velopharyngeal opening action (see /m/ in Fig. 2c and Fig. 2d). We use (iii) *voiceless plosives*, which differ from voiced plosives by replacing the glottal closing action (for phonation) with a glottal opening action (see /k/ and /p/ in Fig. 2c and Fig. 2d). Finally, we use (iv) *voiceless fricatives*, which differ from voiceless plosives by replacing the full closing action (labial, apical, dorsal) with a fricative near closing action (see /s/ in Fig. 2c; both full closing and near closing actions are labeled as “up” movements in Fig. 2d).

Different *speaking rates* were simulated by altering the *syllable oscillator frequency* in four steps from 1 Hz (very slow speaking rate) to 3 Hz (fast speaking rate) with the intermediate steps 1.5 Hz (slow speaking rate) and 2 Hz (normal speaking rate; note that because the speech sounds of the syllable are realized in 50% of the duration of a syllable oscillator cycle at the acoustic level, the voiced syllable durations range from 500 msec (for 1 Hz) to 167 msec for 3 Hz). The time steps for visual input are adapted to speaking rate (faster time steps with increasing speaking rate). The resulting neural activations for different muscle groups can be seen in Fig. 2d and in Fig. 3a-c for different speaking rates. Visual input representation, neural activity at the premotor buffer, as well as neural activity of the syllable oscillators is shown in Fig. 2a-c for very slow speaking rate.

3 Results

The model is capable of generating neural activation patterns at the syllable level as well as at the vocal tract action and muscle group activation level. These activations can be generated for a wide range of speaking rates from very slow (1 Hz) to fast (3 Hz). Vocal tract actions are coordinated with each other in the temporal domain using a relative time scale. For example, for these CVC syllables, the consonantal constriction action at syllable onset starts at 0.2 and stops at 0.5, while the consonantal action at syllable offset starts at 0.6 and stops at 0.9. These time values are relative; the value 0 represents the start of the syllable and the value 1 represents the end of the syllable oscillation cycle. In order to have reached the vocalic target at the time point at which the consonantal constriction of syllable onset releases, vocalic actions need to start at 0.2 as well, but vocalic actions exhibit a longer movement (transition) phase so that the vocalic target is reached not earlier than about 0.4 to 0.5 on the relative syllable time scale. The time interval of the target portion of consonantal, vocalic, as well as of velopharyngeal and glottal closing actions can be seen in Fig. 3. The dashed horizontal

lines indicate that the vocal tract targets have been reached in the case of closing/constriction actions (i.e., saturation, see above).

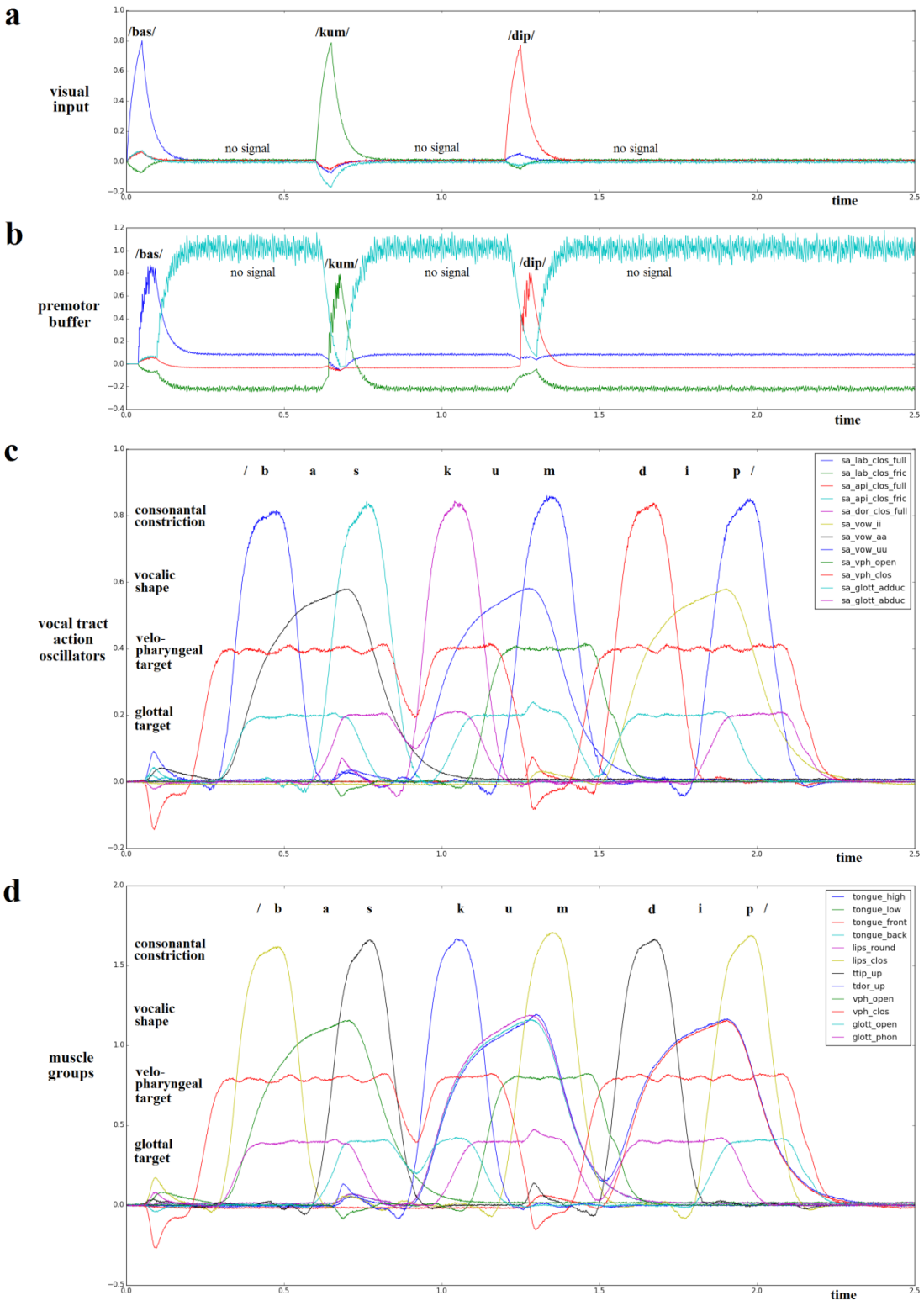

Figure 2 – Simulation results for the sequence of the three syllables /bas/, /kum/, and /dip/ uttered with very slow speaking rate. From top to bottom: Neural activation levels within (a) the visual input ensemble, (b) the premotor buffer for syllable representations (including "no signal" activation, i.e. if no visual input signal occurred), (c) the neural oscillators for vocal tract actions, and (d) the neuron ensembles representing muscle groups.

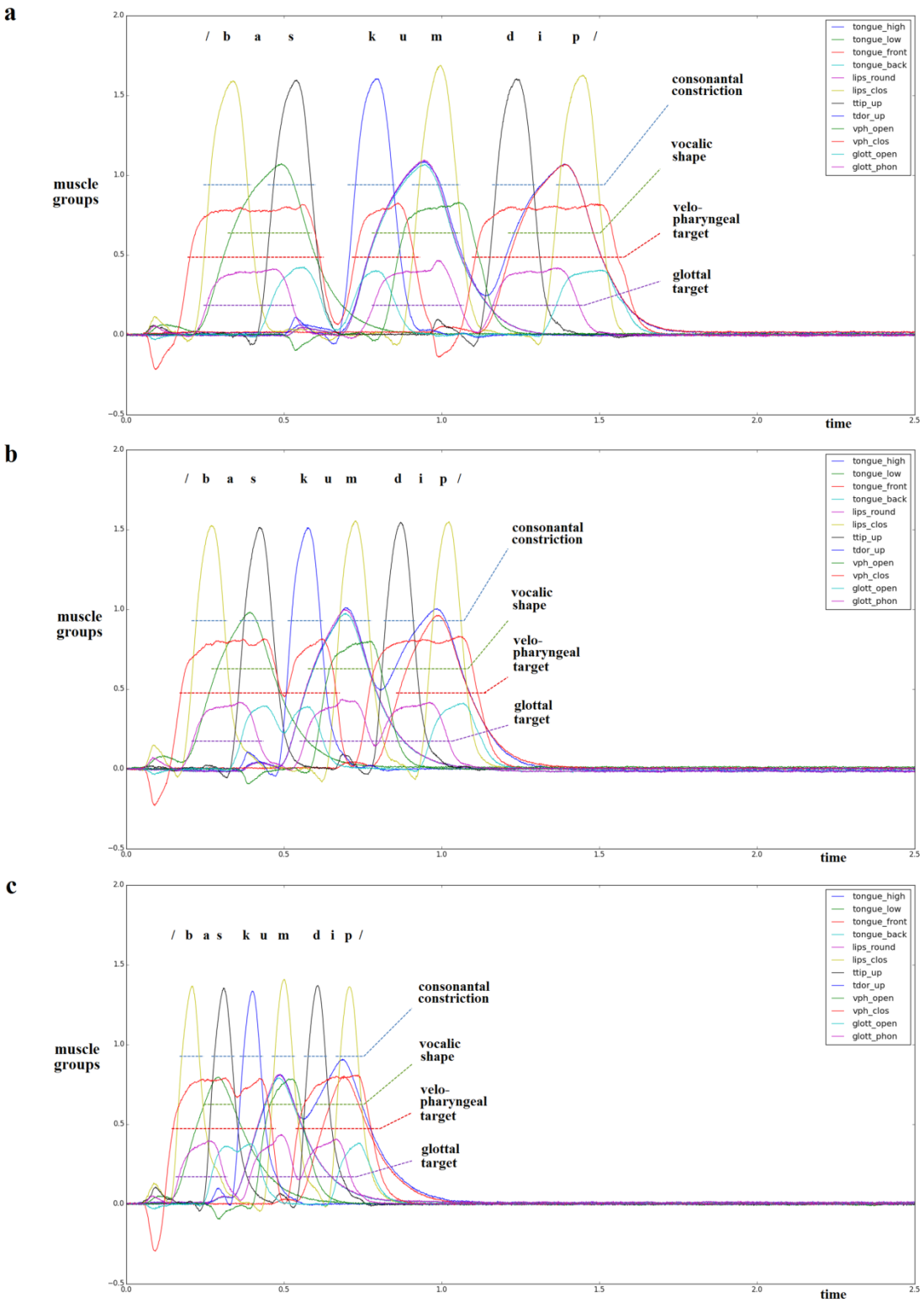

Figure 3 – Simulation results for the sequence of the three syllables /bas/, /kum/, and /dip/ uttered with (a) slow, (b) normal, and (c) fast speaking rate. Only the neural activation levels within muscle group neuron ensembles are shown. Horizontal dashed lines indicate saturation (see text).

It can be seen from Fig. 3 that the phasing of actions leads to stable relations in the temporal coordination of vocal tract actions. Thus, over a wide range of speaking rates, the following relations (*timing rules*) are always kept: (i) the vowel target region is reached before the constriction of the preceding consonant is released; (ii) the vowel target is held until the target region (constriction region) of the following consonant is reached; (iii) the velopharyngeal

closure is held during consonantal closures (except for nasals) and during the target phases of vowels; (iv) a veloparyngeal opening occurs during the consonantal closure of nasals; (v) the glottal closure for phonation is held during consonantal closures for voiced consonants and during target phases of vocalic actions (vowels are always voiced sounds); and (vi) a glottal opening occurs during the closure and at the beginning of the following vowel for voiceless consonants. These timing rules guarantee correct articulation of the sounds occurring within each syllable.

4 Discussion and Conclusions

A preliminary approach for modelling speech production and the intrinsic timing of vocal tract actions using spiking neurons is introduced here. By using neural oscillators, intrinsic time scales can be defined at the syllable level, and speaking rate can be varied over a wide range simply by altering one parameter, the syllable oscillator frequency. Because the temporal organization of vocal tract actions is regulated via constant relative timing (or phasing) values for starting and ending of vocal tract actions, the phase relations of vocal tract actions within syllables remain stable. This results in correct production of all speech sounds occurring within all syllables at different speaking rates (note that language-specific fine tuning (i.e., alteration) of phasing values at different speaking rates is possible in our model).

It is an important feature of this approach that an increase in speaking rate does not lead to an increase in muscle group activation for a vocal tract action, only to a change in duration and temporal overlap of muscle activation for different speech actions. Consequently, articulator velocities are not increased in the case of an increased speaking rate, while the temporal succession of time points representing the start of a speech action decreases in absolute value (increase in temporal overlap of speech actions). Thus the articulatory "behaviour" is highly nonlinear if speaking rate increases, and this nonlinearity can be modelled by altering a single parameter in our approach: the syllable oscillator frequency.

It is debatable whether we need to instantiate a neural oscillator for each frequent syllable (2000 syllable oscillators in Standard German, for example). It may be more feasible to have fewer (perhaps ten) neural syllable oscillators which represent the syllables under production. But this approach increases the number of neural connections between syllable oscillators and speech action oscillators, because information concerning the relative timing of speech actions for *all* frequent (i.e. already learned) syllables needs to be stored in these connections. In the model introduced here, only the timing information for one single syllable needs to be stored between a syllable oscillator and vocal tract action oscillators. In both cases, the number of neuron ensembles needed remains small enough that the syllable and vocal tract action levels can be stored in a few mm^2 of cortex.

Furthermore, it should be noted that our representation of the mental syllabary is comparable with a representation of the mental lexicon (cf. [17]) that introduces different levels for words and phonemes. Within the lexical model of Dell these levels are interconnected in a way that is comparable to how the syllable and vocal tract action levels are connected in our model.

In future work, we hope to include auditory and somatosensory representations of syllables and to model the neural connections between the mental syllabary and the mental lexicon, as is already outlined in our connectionist approach [6]. Moreover, a vocal tract model capable of realizing the model articulator movements controlled by the muscle group activation levels should be included.

Literature

[1] CIVIER O, BULLOCK D, MAX L, GUENTHER FH (2013) Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. Brain and Language 126: 263-278

[2] GUENTHER FH, GHOSH SS, TOURVILLE JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language 96: 280-301

[3] GUENTHER FH, VLADUSICH T (2012) A neural theory of speech acquisition and production. Journal of Neurolinguistics 25: 408-422

[4] KRÖGER BJ, KANNAMPUZHA J, NEUSCHAEFER-RUBE C (2009) Towards a neuro-computational model of speech production and perception. Speech Communication 51: 793-809

[5] KRÖGER BJ, KANNAMPUZHA J, KAUFMANN E (2014) Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. EPJ Nonlinear Biomedical Physics 2:2 (Springer)

[6] KRÖGER BJ, CAO M (2015) The emergence of phonetic-phonological features in a biologically inspired model of speech processing. Journal of Phonetics 53: 88-100

[7] LEVELT WJM, WHEELDON L (1994) Do speakers have access to a mental syllabary? Cognition 50: 239-269

[8] CHOLIN J, SCHILLER NO, LEVELT WJM (2004) The preparation of syllables in speech production. Journal of Memory and Language 50: 47-61

[9] CHOLIN J (2008) The mental syllabary in speech production: an integration of different approaches and domains. Aphasiology 22: 1127-1141

[10] ELIASMITH C, ANDERSON CH (2004) Neural engineering: Computation, representation, and dynamics in neurobiological systems. MIT press.

[11] 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

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

[13] GOLDSTEIN L, BYRD D, SALTZMAN E (2006). The role of vocal tract action units in understanding the evolution of phonology. In: Arbib MA (Ed.) Action to Language via the Mirror Neuron System. (Cambridge University Press, Cambridge), pp. 215-249

[14] SALTZMAN E, BYRD D (2010) Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. Human Movement Science 19: 499-526

[15] SENFT V, STEWART TC, BEKOLAY T, ELIASMITH C, KRÖGER BJ (2016) Reduction of dopamine in basal ganglia and its effects on syllable sequencing in speech: A computer simulation study. Basal Ganglia 6: 7-17

[16] KRÖGER BJ, BIRKHOLZ P (2007) A gesture-based concept for speech movement control in articulatory speech synthesis. In: Esposito A, Faundez-Zanuy M, Keller E, Marinaro M (eds.) Verbal and Nonverbal Communication Behaviours, LNAI 4775 (Springer Verlag, Berlin, Heidelberg) pp. 174-189

[17] DELL GS (1988) The retrieval of phonological forms in production: Tests of predictions from a connectionist model. Journal of Memory and Language 27: 124-142