Developing a Model of Speech Production using the Neural Engineering Framework and the Semantic Pointer Architecture

Bernd J. Kröger[1], Trevor Bekolay[2], Peter Blouw[2,3] & Terence C. Stewart[4]

[1]*Neurophonetics Research Group, Department for Phoniatrics, Pedaudiology and Communication Disorders, Medical Faculty, RWTH Aachen University, Germany*
[2]*Applied Brain Research, Waterloo, ON, Canada*
[3]*Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON, Canada*
[4]*National Research Council of Canada, University of Waterloo Collaboration Centre, Waterloo, ON, Canada*

bernd.kroeger@rwth-aachen.de, trevor.bekolay@appliedbrainresearch.com, pblouw@uwaterloo.ca, terry.stewart@gmail.com

Abstract

Only a few approaches exist that are capable of combining the symbolic linguistic part of speech production with the phonetic or sensorimotor part. The latter is especially difficult to model due to the need for a concrete articulatory-acoustic model that is controlled by a motor module activating a temporally well synchronized set of speech articulator movement units (SAMUs) based on the output of the phonological-linguistic module. In this conference contribution we will explain how a comprehensive symbolic-linguistic and sensorimotor model of speech production can be shaped using the well-established Neural Engineering Framework (NEF) for designing large scale neural models. We also employ the Semantic Pointer Architecture, an extension of the NEF that enables us to define and process neural signals representing symbolic cognitive linguistic units like words, lemmas, and phonological forms of syllables as well as sensorimotor units like SAMUs.

Keywords: speech production, articulatory speech synthesis, speech gestures, speech articulator movement units, coproduction, coarticulation, speech dynamics, Neural Engineering Framework, Semantic Pointer Architecture

1. Introduction

The input level of most biologically inspired computational models of speech production is the phonological specification of a syllable or word (see, e.g., Civier et al. 2013, Guenther et al. 2006, Guenther & Vladusich 2012, Hickok 2012, Kröger et al. 2009, Kröger & Cao 2015, Kröger et al. 2020). One issue that all these speech production models share is that they do not incorporate *speaking rate* as an explicit control parameter even though it is well known that speaking rate can change the articulation of a word or utterance in a nonlinear way. For example, speaking rate can change the amount of temporal overlap of speech articulation movement units (SAMUs or speech gestures), which can lead to assimilation as well as to segmental reduction effects (see, e.g., Browman & Goldstein 1992, Kröger 1993).

In this paper we introduce a more complete model that starts at the lexical concept level and that is based on the idea of building up syllables or words by gestures or SAMUs and by controlling the temporal coordination of SAMUs as suggested in the gestural framework by Browman & Goldstein (1992) and more concretely by Saltzman & Byrd (2010). This concept was adapted and modified for Standard German by Kröger (1993) and by Kröger & Birkholz (2007). Furthermore, our current implementation of this approach is neurobiologically underpinned and uses the Neural Engineering Framework (NEF, see Eliasmith & Anderson 2004, Eliasmith 2013) and the Semantic Pointer Architecture (SPA, see Stewart & Eliasmith 2014). This combined NEF-SPA approach is capable of modeling large scale brain models, i.e., capable of performing cognitive processes, modeling short-term and long-term memories, and processing different types of sensory input and motor output (Eliasmith et al. 2012).

Our current model of speech production and speech perception is capable of modelling aspects of normal and disordered speech (Kröger et al. 2020). The production part of the model comprises a cognitive module initiating the production of a word, a lexical component comprising a semantic, lemma, and phonological level as well as a mental syllabary that stores the temporal coordination of SAMUs for the most frequent syllables of the target language (Kröger & Bekolay 2019). At the level of the mental syllabary a phonological specification of a syllable is transformed into a motor plan, i.e. to a score of SAMUs, defining the types of SAMUs or speech gestures and their temporal coordination (ibid.). This neural speech production model has already been applied to medical research questions (Senft et al. 2016, Senft et al. 2018, Stille et al. 2019, Stille et al. 2020) and to basic linguistic research questions concerning the feedback mechanisms involved in the production and repair of word production errors (Kröger et al. 2020).

In this paper we will introduce our approach for triggering and executing SAMUs based on a neural oscillator approach (Kröger et al. 2016). This approach allows us to elegantly generate the set of temporally synchronized SAMUs needed for the production of a syllable or word. A main benefit of this control approach is that the speaking rate can be controlled by changing one parameter, i.e., the frequency of the neural syllable generation oscillators. The temporal coordination of all SAMUs is controlled by phasing rules that define the points in time for starting and ending the activation of lower-level neural oscillators that control single SAMUs.

2. The model

2.1. Basic architecture of the model: the main modules

The architecture of the model is displayed in Fig. 1 and described in detail in Kröger et al. (2020). The cognitive processing module together with the control module initiate the production of one or more words, i.e., activate a semantic idea for an utterance. The sequence of words is projected downwards within the production pathway module by activating the concept, lemma and phonological form of each word, which is already stored in the mental lexicon, and furthermore by activating the motor plan and all gestures (or speech articulation movement units, SAMUs) of the appropriate syllables which

are already stored in the mental syllabary. The SAMUs activate motor units of muscle groups controlling the position and thus the movements of all articulators in the vocal tract model (see peripheral system in Fig. 1). The articulatory-acoustic (vocal tract) model generates a temporal sequence of model articulator movements and an acoustic speech signal. The organization of the motor plan, also called the score of SAMUs (score of vocal tract actions or gesture score) is discussed in detailed in section 2.2.

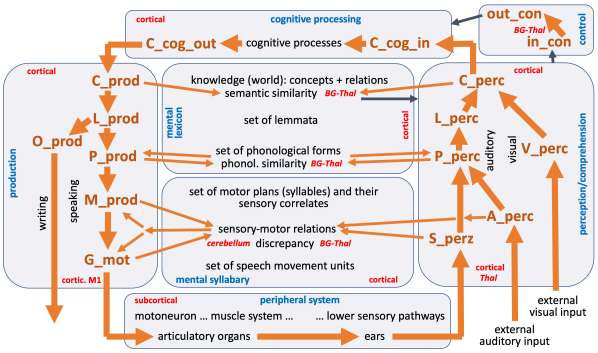

Figure 1: *The architecture of the speech production model (adapted from Kröger et al. 2020). Names starting with C_ denote concepts, L_ lemmas, P_ phonological forms, M_ motor plans, G_ gestures, O_ orthography, S_ somatosensory, A_ auditory, and V_ visual. BG-Thal is the basal ganglia and thalamus.*

Tactile and proprioceptive signals, which result from the temporal sequence of articulator positions within the vocal tract model as well as auditory signals from the acoustic speech signal are processed within the somatosensory and auditory pathways, which are subsumed as the perception-comprehension module in our model (Fig. 1). Both sensory pathways enable different feedback mechanisms to take place at different levels of the model (see horizontal arrows between production and perception-comprehension pathway in Fig. 1). All orange names refer to *neural state buffers* representing neural states of activated words or syllables (for details see Kröger et al. 2020). The orange arrows refer to *neural transformations* and to a projection mechanism for neural activation patterns realized by associative memories and their neural connections with the neural state buffers (ibid.).

2.2. The organization of syllabic motor plans: SAMUs and their temporal organization

The *motor plan* of a syllable is activated in the motor plan buffer, M_prod, of our production model (Fig. 1). The neural state activation of a motor plan is triggered from the phonological form of a syllable (the lowest level of the mental lexicon, see Fig. 1). The motor plan starts with a sequence of well-timed go-signals (start signals) activating a sequence of syllable oscillators representing all syllables of a word (for details of syllable oscillator activation see Kröger et al. 2016). The organization of the M_prod (motor plan) level is displayed in Fig. 2. Within the syll_init level, the temporal sequence of go-signals for the syllable oscillators is generated. The syll_osz level comprises all syllable oscillators representing all frequent syllables in the target language (note that only three syllable oscillators are displayed in Fig. 2). These oscillators allow the triggering of SAMU activation, i.e. they start the SAMU oscillators on the gesture level (G_mot, Fig.1).

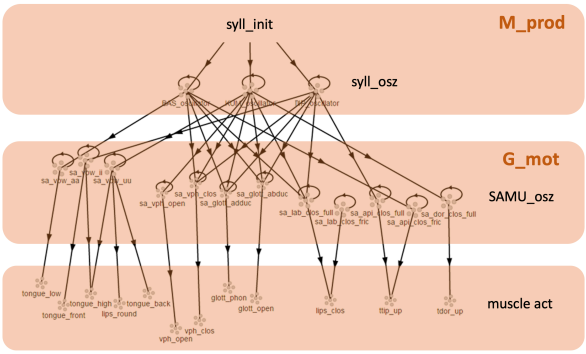

Figure 2: *Neural oscillators realizing the motor plan and SAMU levels of the production model. The neuron ensembles covering the muscle activation levels for controlling the model articulators are already part of the peripheral system module of our production model.*

The number of frequent syllables for Standard German is about 2000 (for about 95% of all syllables in standard texts, see Kröger et al. 2011, p. 290). Because each neural oscillator only requires about 100 neurons (located in the pre-motor area of the cortex) the storage of all motor plans of frequent syllables as represented in our model by these neural oscillators does only a small region of the premotor cortex is occupied for the storage of motor plans of frequent syllables.
The SAMU level comprises neural oscillators for all SAMUs existing in a target language (see Table 1 for a subset of SAMUs needed for simulating Standard German).

Table 1: *List of SAMUs, e.g. for Standard German (incomplete), see Kröger & Bekolay (2019), p. 20; phonetic symbols of sound realizations in SAMPA notation; abbrev. is abbreviation of name of SAMU.*

abbrev.	SMU for....	e.g., for realization of ...
srtb	strongly raising the tongue body (vocalic)	/i/, /u/
ratb	raising the tongue body (vocalic)	/e/, /o/
lotb	lowering of the tongue body	/E, O/
dltb	deep lowering of the tongue body	/a/
rttb	retraction of tongue body	/u/, /O/, /o/
fwtb	forwarding the tongue body	/i/, /e/
roli	rounding the lips	/u/
…		
vow_aa	equals **dltb**	/a:/ (German)
vow_ii	coproduction of **srtb** & **fwtb**	/i:/ (German)
vow_uu	coproduction of **srtb** & **rttb** & **roli**	/u:/ (German)
…		
clli	closing the vocal tract by lips	/b/, /p/, /m/
cltt	closing the vocal tract by tongue tip	/d/, /t/, /n/
cltb	closing the vocal tract by tongue body	/g/, /k/
actt	producing an alveolar constriction by tongue tip	/s/, /z/
pctb	producing a postalveolar constriction by tongue blade	/S/, /Z/
…		
opgl	glottal opening	voiceless sounds
phgl	producing a phonatory glottal closure	voiced sounds
…		
clvp	closing the velopharyngeal port (raising the velum)	obstruents (plosives and fricatives)
opvp	opening the velopharyngeal port (lowering the velum)	nasals

Neuron ensembles (i.e. small neuron buffers) are needed for representing the activation for each modeled muscle group. Thus, each muscle group controls the displacement of one model articulator in one direction (see Table 2). Specifically, each model articulator is controlled by one, two or four model muscle groups depending on (i) whether an articulator mainly moves in one dimension (e.g. the velum: up-down) or in two dimensions (e.g. tongue body: up-down, front-back) and (ii) whether the neutral or rest position of a model articulator movement dimension is in the middle or the end of the

displacement range in that dimension. When movement is possible in both the positive and negative directions of that dimension, positive and negative displacement values are reached by different model muscle groups, e.g., front and back positioning of the tongue body like in /i/ vs. /u/. When movement is possible in only one direction, only one muscle group is necessary (e.g., consonantal movement of the tongue tip: raising the tongue tip for a consonantal closure like in /t/). The set of articulator control parameters of our vocal tract model are described in detail in Kröger et al. (2014, p. 204)
Because our model is a simplification of reality, a model muscle group may represent more than one physiological muscle group controlling an articulator. Moreover, in our model the activation and thus the contraction of each model muscle group directly represents a specific degree of displacement of a model articulator from its rest position in the appropriate movement direction. The neuron ensembles coding the neural activation level of each muscle group are already part of the peripheral system module of our production model (Fig. 1).

Table 2: *List of names (abbreviations) of model muscle groups, and corresponding movement dimension and direction for each model articulator controlling the vocal tract model (Kröger et al. 2014). "Port" is velopharyngeal port; "stop" is glottal stop.*

abbrev.	movement direction & (dimension)		model articulator	
tb_high	towards high	(vertical)	tongue body	(e.g. /i/)
tb_low	towards low	(vertical)	tongue body	/e.g. /a/)
tb_front	towards front	(horizontal)	tongue body	(e.g. /i/
tb_back	towards back	(horizontal)	tongue body	(e.g. /u/)
…				
tt_up	consonantal raising	(vertical)	tongue tip	(e.g. /t/, /s/)
tb_up	consonantal raising	(vertical)	tongue body	(e.g. /k/)
tt_front	consonantal back	(horizontal)	tongue tip	(e.g. /S)
tb_back	consonantal raising	(vertical)	tongue body	
…				
li_round	rounding	(horizontal)	lips	(e.g. /u/)
li_spread	spreading	(horizontal)	lips	(e.g. /i/)
li_clos	consonantal closing	(vertical)	lips	(e.g. /p/)
…				
vph_close	closing the port	(vertical)	velum	(obstruents)
vph_open	opening the port	(vertical)	velum	(nasals)
…				
gl_phon	soft closing	(horizontal)	glottis	(voiced sounds)
gl_close	closing	(horizontal)	glottis	(e.g. /?/ stop)
gl_open	opening the port	(horizontal)	glottis	(voiceless sounds)

2.3. The concept or neural oscillators: How to model the temporal coordination of SAMUs in syllables

Each syllable oscillator (syll_osz in Fig. 2) triggers the execution of all SAMUs needed to produce a specific syllable. The syllable oscillator frequency defines the intrinsic time scale for each syllable and thus reflects the speaking rate. The triggering of SAMUs is realized by defining specific phase values at which each SAMU starts within one oscillation cycle of the syllable oscillator (Kröger et al. 2016).
Each SAMU is characterized by a neural oscillator as well (Fig. 2 and Kröger et al 2016). The oscillation frequencies of these oscillators define the speed at which the SAMU is executed, which allows for the high articulator velocities necessary for most consonantal SAMUs and for slow articulatory velocities as are seen in vocalic SAMUs (see Kröger and Bekolay 2019, p. 19).

3. Simulation experiment

The production of three CVC-syllables building up a three-syllabic psuedoword was simulated (cf. Kröger et al. 2016). It was shown in a previous study that the syllable frequency could be varied in a wide range from 1 Hz to 3 Hz, leading to mean syllable duration from 500 msec to 167 msec (ibid.). In this study we measured the maximum movement velocities of the main articulator for different types of SAMUs. The maximum velocity appears within the movement phase of a SAMU.

4. Results

The results for four different types of SAMUs (vocalic, consonantal, velopharyngeal and glottal) as well as for three different speaking rates (slow: f = 1.33 Hz, normal: f = 2.0 Hz and fast: f = 3.0 Hz) are presented in Table 3.

Table 3: *Maximum articulator movement velocities (max vel.) for different types of SAMUs and different speaking rates (slow, normal, fast)*

abbrev.	movement direction & (dimension)		max vel. (percentage)		
SAMU			slow	normal	fast
aa_vow	lowering tongue body	(vertical)	100	100	100
li_clos	closing the lips	(vertical)	72	88	100
vph_open	lowering the velum.	(vertical)	76	88	100
gl_open	opening the glottis.	(horizontal)	70	94	100

The results indicate that despite the fact that the speaking rate increases by 50% from slow to normal and by a further 50% from normal to fast, maximum articulator velocities remain stable for vocalic SAMUs and increase relatively slowly (i.e., from about 70% to 100%) for all other types of SAMUs.
The stability of the movement portions of the vocalic gestures indicates that even if speaking rate increases and thus the time interval of activation of vocalic SAMUs decreases, no increase in effort occurs in order to reach a vocalic target earlier. Thus, the maximum vocalic articulator displacement decreases with increase in speaking rate (reduction of vocalic gestures). In the case of consonantal SAMUs (here, the lip closing action) the maximum articulator velocity increases slightly because even in the case of a high speaking rate and shorter time interval of SAMU activation, a certain degree of articulator displacement (here, lip closure) needs to be reached. The same holds for all other consonantal SAMUs acting on the tongue tip (vocal tract closure at the alveolar ridge or at the hard palate) as well as for the tongue body (vocal tract closure between tongue body and hard or soft palate). Moreover, the same holds for SAMUs controlling the aperture of the velopharyngeal port. Articulator velocity increases here slightly if speaking rate increases in order to guarantee a sufficient lowering of the velum (sufficient opening of the velopharyngeal port) to produce nasals even at high speaking rates as well as to guarantee a sufficient elevation of the velum to guarantee a tight closure of the velopharyngeal port when producing obstruents (plosives and fricatives). The same holds for SAMUs controlling the glottal aperture. Articulator velocity (here of the arytenoids) increases slightly from slow to fast speaking rate in order to guarantee a sufficient opening of the glottis during the production of voiceless sounds in the case of all speaking rates as well as to guarantee a sufficient glottal closure for phonation and a sufficient tight closure for a glottal stop for all speaking rates.

5. Discussion and conclusion

Our implementation was motivated by prior gesture or SAMU timing models using task dynamics and coupled oscillators (Goldstein et al. 2006, Saltzman & Byrd 2010). In these models, vocal tract actions (speech gestures) are assumed to be intrinsically timed and modelled by harmonic oscillators. This idea can explain inter-gesture timing within and between syllables by relative timing or "phasing" values. While the approach of Goldstein et al. (2006) and Saltzman & Byrd (2010) is grounded in cognition basically, our approach can easily be interpreted in a neurobiological way as well. Our model is

embedded in the comprehensive NEF-SPA framework, i.e. in a neurobiologically inspired framework for implementing large-scale neural models capable of describing cognitive and sensorimotor aspects of speech production.
Our simulations indicate that a wide range of speaking rates can be modelled by varying syllable oscillator frequencies. Moreover, the idea of variable syllable oscillator frequencies in combination with stable SAMU oscillator frequencies is in agreement with results of experimental measurements. Our simulations indicate that the movement phase of gestures (SAMUs) remains stable (regarding maximum articulator velocity) while the temporal overlap of gestures increases with speaking rate. Thus, while the kinematic shape of gestures remains relatively stable, their temporal coordination varies (see the iceberg-concept, Fujimura 1992). This nonlinear articulatory behavior leads to the typical assimilation and reduction phenomena occurring in many spoken languages as speaking rate increases (for German see Kröger 1993).

6. Acknowledgements

We would like to thank Prof. Dr. Chris Eliasmith, head of the Centre for Theoretical Neuroscience at the University of Waterloo, Canada, for his kind support of this work on speech production modeling.

7. References

Browman, C.P., Goldstein, L. (1992) Articulatory phonology: an overview. *Phonetica, 49*, 155-180.

Civier, O., Bullock, D., Max, L., Guenther, F.H. (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.

Eliasmith, C., Anderson, C.H. (2004). Neural engineering: Computation, representation, and dynamics in neurobiological systems. Cambridge, MA: MIT press.

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

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

Fujimura, O. (1992). Phonology and phonetics – a syllable-based model of articulatory organization. *Journal of the Acoustical Society of Japan, 13*, 39-48.

Goldstein, L., Byrd, D., Saltzman, E. (2006). The role of vocal tract action units in understanding the evolution of phonology. In M.A.Arbib (Ed.), Action to Language via the Mirror Neuron System, Cambridge, MA: Cambridge University Press, pp. 215-249.

Guenther, F.H., Ghosh, S.S., Tourville, J.A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*: 280-301.

Guenther, F.H., Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics, 25*, 408-422.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*, 135-145.

Kröger, B.J. (1993). A gestural production model and its application to reduction in German, *Phonetica, 50*, 213-233.

Kröger, B.J., Birkholz, P. (2007). A gesture-based concept for speech movement control in articulatory speech synthesis, In A. Esposito, M. Faundez-Zanuy, E. Keller, & M. Marinaro (Eds.). Verbal and Nonverbal Communication Behaviours, LNAI 4775, Berlin, Heidelberg: Springer Verlag, pp. 174-189.

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

Kröger, B.J., 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 A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud & A. Nijholt (Eds.), Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues. LNCS 6800, Berlin: Springer, Verlag, pp. 287-293.

Kröger, B.J., Bekolay, T., Eliasmith, C. (2014). Modeling speech production using the Neural Engineering Framework. *Proceedings of CogInfoCom 2014* pp. 203-208, IEEE Xplore Digital Library DOI=10.1109/CogInfoCom.2014.7020446

Kröger, B.J., Cao, M. (2015). The emergence of phonetic-phonological features in a biologically inspired model of speech processing. *Journal of Phonetics, 53*, 88-100.

Kröger, B.J., Bekolay, T., Blouw, P. (2016). Modeling motor planning in speech processing using the Neural Engineering Framework. In: Jokisch O (Ed.) Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2016 (TUDpress, Dresden, Germany), pp. 15-22.

Kröger, B.J., Bekolay, T. (2019). Neural Modeling of Speech Processing and Speech Learning. Cham: Springer Verlag.

Kröger, B.J., Stille, C.M., Blouw, P., Bekolay, T., Stewart, T.C. (2020). Hierarchical sequencing and feedforward and feedback control mechanisms in speech production: a preliminary approach for modeling normal and disordered speech. *Frontiers in Computational Neuroscience, 14*, 573554, doi: 10.3389/fncom.2020. 573554

Saltzman, E., Byrd, D. (2010). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science, 19*, 499-526.

Senft, V., Stewart, T.C., Bekolay, T., Eliasmith, C., Kröger, B.J. (2016). Reduction of dopamine in basal ganglia and its effects on syllable sequencing in speech: A computer simulation study. *Basal Ganglia, 6*, 7-17.

Senft, V., Stewart, T.C., Bekolay, T., Eliasmith, C., Kröger, B.J. (2018). Inhibiting Basal Ganglia Regions Reduces Syllable Sequencing Errors in Parkinson's Disease: A Computer Simulation Study. *Frontiers in Computational Neuroscience, 12*, 41.

Stewart, T. C., Eliasmith, C. (2014). Large-scale synthesis of functional spiking neural circuits. *Proceedings of the IEEE, 102*, 881-898.

Stille, C., Bekolay, T., Blouw, P., Kröger, B.J. (2019). Natural language processing in large-scale neural models for medical screenings. *Frontiers in Robotics and AI, 6*, 62, doi: 10.3389/frobt.2019.00062

Stille, C., Bekolay, T., Blouw, P., Kröger, B.J. (2020). 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*, doi: 10.3389/fpsyg.2020.01594