MODELING OF SPEECH PRODUCTION FROM THE PERSPECTIVE OF NEUROSCIENCE

Bernd J. Kröger[1,2]

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

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

bernd.kroeger@rwth-aachen.de

Abstract: Recent neural models are capable of generating quantitative patterns of speech articulation and speech acoustics. Five models are discussed here: the DIVA model, the task dynamics model, the ACT model, the Warlaumont model and the Hickok model. These models have a more or less strong background in neuroscience. Directions are identified in this paper for a further development of quantitative production models in order to bring models more in line with recent research outcomes from neuroscience.

1 Introduction

The perhaps best-known model of speech production is the Levelt model [1]. This approach describes the whole process of speech production from intention to articulation. The Levelt model includes the whole linguistic processing from utterance planning towards the generation of a phonological representation and subsequently towards the generation of articulatory speech patterns. In contrast to the Levelt model we will focus in this paper on models which particularly describe the *sensorimotor processes of speech production*, starting with a phonological representation of a speech item, and subsequently generating articulatory movement trajectories and an acoustic speech signal [2, 3, 4, 5, and 6]. These models are *quantitative production models* because they generate measurable articulatory movement patterns and subsequently measurable acoustic speech signals (or at least are prepared to generate these signals in future versions: Hickok model [7]). These models already include some knowledge gained from brain imaging as well as from behavioral experiments and thus are at least particularly *neuroscience based models*.

It is the goal of this paper to identify directions how a quantitative model should be organized in order to be in line with neuroscience related knowledge. Furthermore it should be stated here that it is not the goal of this paper to give a complete survey on all existing models of speech production. This paper mainly reflects models which are closely related to our own model [4] and which to our opinion are important for the future development of this type of speech production models.

2 Existing Production Models

2.1 DIVA Model

The structure of the DIVA model (directions into velocities of articulators model [2]) comprises a feedforward and a feedback control subsystem. The starting level of this model is a *speech sound map*, where a set of model neurons represent language specific speech items

(i.e. phonemes, syllables, or short sequences of syllables, with the syllable being the most typical unit represented by a single model neuron [8]). Starting with the activation of a specific model neuron at the level of the speech sound map (i.e. activation of a specific speech item), a *feedforward command* (also labeled as motor command) is activated. At the same time neural representations of an auditory and a somatosensory target region (sensory expectation for that speech item) are co-activated. The forward command generates an articulatory movement pattern via a subsequent co-activation of neural patterns at the level of articulator velocity and position maps (level of primary motor map) and subsequently generates an acoustic speech signal by using an articulatory-acoustic model (speech synthesizer). If the auditory and somatosensory feedback signals derived from these articulatory and acoustic signals (generated by the articulatory-acoustic model) are within the expected auditory and somatosensory target regions mentioned above, no additional motor command (i.e. no additional feedback command) is generated. But if one or both sensory feedback signals exceed the limits defined by sensory target regions at least for a short time interval within the time interval representing the whole speech item, *sensory error signals* are generated for that speech item at the level of the auditory and/or somatosensory error map, and a corrective motor command, also called *feedback command* is activated at the level of the feedback control map within the feedback control loop.

In the DIVA model, the activation of feedforward commands, the co-activated associated sensory target regions for each language specific word or syllable, as well as the sensory-to-motor mapping is part of the feedback control map. This map and its mapping towards other neural maps are trained during a *babbling* and during an *imitation learning process*. The synaptic projections between sensory error maps and motor cortex are tuned during babbling (by using prelinguistic proto-speech items) and build up the feedback control map. All other synaptic projections are trained during imitation. Firstly, an auditory target is learned for each word or syllable of a language. If the model then attempts to produce that speech item, corrective motor commands (feedback commands) are activated for updating and storing the current forward command for that speech item. Normally, more than one attempt is needed in order to train the association of speech items and appropriate forward commands. Secondly, during these production (or imitation) attempts in addition somatosensory target regions are learned from the somatosensory states which were activated for each production attempt.

Auditory error signals mainly occur during early phases of speech acquisition and thus are mainly used for the adjustment and storage of feedforward commands during imitation. Auditory error signals in addition occur, if speech production is perturbed externally, e.g. by shifting the frequencies of F1 and/or of F2 for a defined time interval. DIVA produces fast *compensation* via feedback motor commands (added to the already learned feedforward motor commands) starting approximately 75-150 msec after the perturbation onset. In addition, if the auditory perturbation lasts over a longer time period (e.g. for 25-50 epochs, including approximately 30 word productions within each epoch), the auditory error cell activation leads to a further tuning or *adaptation* of feedforward commands (i.e. to an additional learning effect) which subsequently in addition results in a significant after effect, i.e. which results in occurrence of altered feedforward commands, even after removal (switch off) of the auditory feedback perturbation [9].

In a series of experiments, brain regions are identified to host specific maps of the DIVA model. These results are reported in detail in [10]. Interestingly the (language specific) speech sound map here is mainly associated with the left ventral premotor cortex while the feedback control map, which mainly develops during prelinguistic babbling training, and thus merely reflects general sensorimotor than language-specific behavior, is mainly associated with the right ventral premotor cortex. This is in agreement with studies, reporting a more bilateral activation of brain regions for more general (not langue specific) lower-level speech production mechanismus [11]. Also the locations of auditory and somatosensory feedback

processing are listed in detail here [10]. In addition, beside the cortex, the role of cerebellum hosting processing routines for motor commands, and the role of basal ganglia and thalamus, hosting processing routines for initiation of articulation is emphasized.

2.2 Task Dynamic Model

From a linguistic perspective, syllables are structured with respect to constituents like syllable onset, nucleus, and coda, where the syllable nucleus in most cases is represented by a vowel and where syllable onset as well as syllable coda (if occurring) are represented by one or more consonants (consonant clusters). Thus, between the level of phonemic representation (abstract symbolic level) and the primary motor level (i.e. level of representation of ongoing articulatory movement patterns) at least one level should exist, reflecting this organization of syllables. Such a *planning level* as well as the calculation of movement patterns on the basis of this planning is introduced in a quantitative form in the task dynamic approach [3, 13]. Here *vocal tract action units (gestures)* are assumed as basic units of speech production and phasing relations are assumed for quantifying the temporal coordination of these basic speech action units within a syllable.

The task dynamics approach separates two levels, i.e. an intergestural coordination level and an interarticulatory coordination level. At the *intergestural coordination level*, gesture activation is specified (gesture activation "…can be interpreted as the strength with which the associated gestures "attempts" to shape vocal tract movements …" [3, p. 335]) and activation intervals as function of time indicate the temporal organization of all gestures within a syllable, called "gestural score" [13]. At the *interarticulatory coordination level*, the movement pattern is calculated for each model articulator on the basis of articulator-related as well as on the basis of vocal-tract-shape-related (i.e. tract variable) coordinates. Tract variables are assumed in this approach to specify the goal of each gesture (i.e. location and aperture of the vocal tract constriction) in a context independent way, while model articulator variables show the resulting context dependent movement pattern for each model articulator during the articulation of a speech item.

Gestures are modeled quantitatively as *time-invariant dynamical systems* (more specifically *point-attractor systems* or critically damped *oscillator systems*) and thus each gesture defines a *class* of goal-directed movements. But the production of a speech item and thus the underlying gestures (i.e. group of dynamical systems) become time dependent with respect to the fact that gesture activation is time-dependent: gesture activation starts and ends at specific points in time. The contextual variation of articulation is simulated in this approach by the interplay of intergestural and interarticulatory coordination. In addition this approach is capable of modeling *compensatory articulation* with respect to mechanical perturbations at the level of the vocal tract (e.g. fixation of the lower jaw by bite-blocks [14, 15]) due to the interplay between the two basic levels introduced in the task dynamic model.

2.3 ACT Model

Our ACTion-based model of speech production, speech perception, and speech acquisition [4] is comparable to the DIVA model but augmented in a way that we not only assume a phonemic map, were one model neuron represents one syllable, but that in addition a *high-level motor representation (motor plan)* is assumed, where the temporal coordination and degree of activation of all vocal tract actions, building up a syllable, is represented in a comparable way as it is represented in a gesture score in the task dynamic approach. Model articulator movements are calculated on the basis of this motor plan for each syllable [16]. Subsequently an articulatory-acoustic model (articulatory synthesizer) generates articulatory movement patterns and an acoustic speech signal [17]. Auditory and somatosensory feedback signals are generated and these signals can be compared with auditory and somatosensory

expectations, co-activated with the activation of a specific speech item at the phonemic level (cf. DIVA model).

A main difference to DIVA can be seen in the fact that a supramodal self-organizing map, called *phonetic map* is introduced in ACT, which is associated with the higher-level motor map (containing the motor plan of a syllable), with the high-level auditory and somatosensory map (containing neural activation patterns of the sensory expectations for each syllable), and with the phonemic map (where each phonological representation of a syllable is represented by one model neuron). During imitation training (see below) the model neurons within the phonetic map represent phonetic realizations of syllables. The neural connections between a model neuron of the phonetic map and the neurons of the high-level motor and sensory map store the motor plan and sensory representation for a specific realization of a syllable.

Thus, our model in parallel to Levelt and Wheeldon [18] and to Levelt et al. [1] emphasizes the importance of knowledge and *sensorimotor skill repositories*. A higher level cognitive repository is the *mental lexicon*, comprising concepts, lemmas, and word forms (all symbolic), while a lower level sensorimotor repository, i.e. a *mental syllablary* is assumed, comprising complete gesture scores and sensory expectations of at least high frequent syllables for the spoken language [1, 18]. The existence of this repository reduces the computational load (i.e. the load for generation of the motor plan) during syllable articulation. After a syllabification process [1], the motor program for syllables need not to be assembled (or generated) on the basis of a segment chain, but can be activated as a whole at the level of the mental syllabary.

Babbling training is performed in ACT in order to supply the model with first auditory-to-motor associations. This enables first *language specific imitation* trials since due to babbling the model already has available some auditory-to-motor associative knowledge and thus is capable of producing first motor plans. If the resulting speech item is not awarded by the caretaker, more imitation trials are performed (cf. DIVA, but target "regions" are replaced here by a perceptual "acceptance range" defined by a caretaker). Thus in contrast to DIVA the *communicative interaction process* between model (or toddler) and teacher (or caretaker) is emphasized in our model. Sensorimotor babbling knowledge (sensory-to-motor associations for proto-syllables) as well as language specific knowledge after imitation training (i.e. motor plans and sensory expectations for syllables) is both stored by (i) the organization of the phonetic map and (ii) within the synaptic link weights between phonetic map and motor plan map, between phonetic map and sensory maps, and between phonetic map and phonemic map.

In contrast to the task dynamics approach, *no rules* are predefined in ACT for the relative timing of vocal tract actions within a syllable. This timing of vocal tract actions is (intuitively) learned by the model during imitation training. But due to the resulting self-organized phonetic map, language specific phasing rules can be derived from the occurring motor plans, which are already stored within the neural associations between phonetic map and motor plan map. Typically self-organizing maps allow generalization and thus the extraction of rules, if the number of training items is much larger than the number of model neurons within the self-organizing map.

Last but not least it should be mentioned that ACT includes a model of *speech perception* as well. Since DIVA as well as our approach include auditory feedback, it is obvious to widen the production model in order to become a production-perception model. In the case of ACT we are capable to show, that a stronger *categorical perception* occurs for consonants in comparison to vowels, which is related to the spatial (self-)organization of syllables within the phonetic map [4].

2.4 Warlaumont Model: Emphasizing reinforcement

This model concentrates on prespeech motor learning (mainly babbling) but beside babbling also includes the emergence of phoneme learning by using *reinforcement-gated self-organized learning* [5]. Thus, not imitation of caretakers' productions of speech items is focused on in this approach. But reinforcement by caregivers (i.e. whether a speech items sounds like a phoneme realization in the target language; extrinsic reinforcement) as well as self-reinforcement (intrinsic reinforcement) is introduced. A *self-organizing map* is postulated here at the motor neuron level. Learning results indicate that reinforcement learning leads to an emergence of muscle activation patterns for stable phonations and to an emergence of muscle activation patterns for phoneme realizations.

2.5 Hickok Model: Emphasizing Hierarchy and Neuroanatomy

Hickok [6] argues for a neuroanatomically grounded, hierarchical state feedback control model of speech production. The hierarchy comprises four levels, (i) a *conceptual level*, (ii) a *lemma level*, (iii) an *auditory level* and (iv) a *somatosensory level*. Hickok assumes that "the auditory system (is) driving higher-level control of the (syllabic opening-closing) cycles or half-cycles" [ibid., p. 139], while "the somatosensory system (is) driving lower-level online control that target the end point of a vocalic opening or closing" [ibid., p.139], i.e. consonantal and vocalic target points. This hierarchy is motivated by the fact that especially consonants like plosives show different acoustic (and thus auditory) patterns in different syllable contexts and thus cannot be associated with simple invariant auditory targets (as is the case for vowels and consonants, which can be produced in isolation), while clear articulatory and somatosensory targets or target regions exist.

Neuroanatomical locations are specified explicitly for the higher-level auditory and lower-level somatosensory control loop. The higher-level auditory control loop comprises auditory processing regions (superior temporal gyrus STG and superior temporal sulcus STS), the Spt-region (Sylviain fissure at the pariotemporal boundary) for auditory-motor-association, and the Brodman area 44 for the activation of higher-level (syllable sized) motor programs. The lower-level somatosensory control loop comprises somatosensory processing regions (anterior supramarginal gyrus aSMG and primary somatosensory cortex S1), the cerebellum for somatosensory-motor-association, and the ventral Brodman area 6 as well as primary motor cortex for the activation of lower-level (speech sound sized) motor programs.

Furthermore Hickok [6] emphasizes the importance of *internal forward models* (cf. [19]) because "sensory feedback alone cannot support … a (high) correction effiency" [6, p. 136] and thus an "internal forward-looking mechanism" [ibid., p. 136], is needed, capable of "mak(ing) predictions regarding the current (articulatory) state" [ibid.; see also 19]. But it is emphasized that this "internal forward-looking mechanism is particularly useful for online movement control" [6, p.136], while feedback is crucial for three purposes: (i) learning sensorimotor relationships, (2) update of the internal model in case of persistent mismatches and (iii) to detect and correct for sudden perturbations [ibid., p. 136]. It is stated that we have to separate *external feedback control* and *internal feedback control* [ibid., p. 136], both occurring within the higher-level as well as within the lower-level feedback control circuits.

3 Directions for Developing Future Quantiative Neural Models

Models of speech production should clearly separate *structure* and *knowledge*. Structure should be hierarchical and should include top-down (e.g. motor commands) as well as with bottom-up interaction (e.g. generation and processing of feedback signals). We argue for a separation of a cognitive phonological level (phonemic map, processing abstract symbolic linguistic units), a multi- or supramodal phonetic level (phonetic map), unimodal high-level motor levels [3, 4] and sensory levels [2], as well as lower-level motor as well as sensory

levels. With respect to Hickock [6], a lower-level somatosensory level should directly provide feedback towards motor representations via the cerebellum, while the auditory level provides feedback to at least syllable-sized motor plans. Thus in our model ACT, lower-level sensory representations cover time intervals of 10-50msec, while higher-level sensory representations cover syllable sized time intervals from 50-500msec. Moreover lower-level somatosensory signals are assumed to be tactile and articulator-related proprioceptive information (relative coordinates) while higher-level somatosensory signals are assumed to be tactile as well as vocal-tract-related articulator positions (absolute coordinates), directly defining the vocal tract shape.

A further important concept is the introduction of a *sensorimotor repository*. A mental syllabary [1, 18] allows that 80% of all spoken syllables is based on just 500 (frequent) syllables (in the case of Standard German [20]). Thus the storage of motor plans of just 500 syllables discharged computational effort for motor plans dramatically. Thus in ACT a sensorimotor syllable repository is assumed. Lower frequent syllables can be assembled by co-activating phonetically similar higher frequent syllables and thus by taking motor plan parameters from these already learned syllables [21].

Concerning *learning* it is very interesting to state that the Warlaumont model [5] is capable of learning phoneme realizations without giving reference targets, i.e. without using imitation training [cf. 4]). The targets (phoneme regions) learned here, are derived exclusively from reinforcing specific babbling items. This allows training a speech production model without using imitation and thus the model (learner, toddler) has no to deal with the *vocal tract normalization problem* [22], because the learning only takes into account the speech items produced by the model itself.

The Warlaumont model [5] as well as the ACT model [4] use *self-organizing neural networks* (Kohonen networks [23]) for simulating neural learning and processing processes. Kohonen networks as well as the network approach used in DIVA [2] can be summarized as *rate models* (in contrast to *spiking neuron models*). Rate models integrate neuron activity over specific time intervals (e.g. 20 … 50 msec) and "model neurons" (i.e. groups on real neurons located close together and having similar functions) are defined in order to integrate over space as well. Rate models process neuron spike *rates* in contrast to real (but complex) *spike time patterns* [24]. Currently no comprehensive spiking neuron model is known, capable to describe main aspects of speech production and speech acquisition (e.g. as in ACT [4]), so that currently rate models seem to be the appropriate choice.

A very important feature, which has been highlighted by Hickok [6] is, that two different types of feedback need to be differentiated (see above). *External feedback* allows adaptation (see also DIVA model [2]) while *internal feedback (inner models)* allow online corrections during motor execution of a motor plan (see also the *state feedback control model* of Houde and Nagarajan [19]). In concepts like ACT [4], computational or programming processes are assumed as not thus important. Moreover, neural processes are mainly forwarding neural activation by using already adjusted synaptic link weights (adjusted during learning phases). Correction processes as proposed by state feedback models are mainly motivated from models for arm and hand movement control, where the human with his arms and hands always needs to interact with *different environments* (i.e. different locations, different rooms etc.). But especially in the case of speech the target directed articulator movements always occur in the same "room", i.e. within speakers vocal tract. Thus the importance of state feedback control should remain a matter of debate.

4 Concluding Remark

We are at the very beginning concerning the development of neuroscience based models of speech production. Neuroscience based architectures (i.e. architectures based on imaging

experiments) are already suggested and used in some of these models. Also behavior based learning concepts have been used in simulation experiments for speech acquisition by using these models. But there is still a long way in order to model (or imitate) speech production from brain to articulation (including feedback) in a natural way, so that realistic neural function processes like spiking neuron approaches are used and so that a lot of "macroscopic" behavioral data (e.g. learning of articulatory skills, adaptation, compensation etc.) can be explained by realistic "microscopic" neural processes.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61233009)

Literature

[1] Levelt WJM, Roelofs A, Meyer AS (1999) A theory of lexical access in speech production. Behavioral and Brain Sciences 22, 1-75

[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] Saltzman EL, Munhall KG (1989) A dynamical approach to gestural patterning in speech production. Ecological Psychology 1, 333-382

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

[5] Warlaumont AS, Westermann G, Buder EH, Oller DK (2013) Prespeech motor learning in a neural network using reinforcement. Neural Networks 38, 64-75

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

[7] Walker G, Rong F, Hickok G (2012) An artificial neural network (ANN) model of sensorimotor development for speech. Abstracts of Neurobiology of Language Conference 2012 (San Sebastian, Spain), pp. 58-59

[8] Guenther FH, Vladusich T (2012) A neural theory of speech acquisition and production. Journal of Neurolinguistics 25, 408-422

[9] Guenther FH (2006) Cortical interaction underlying the production of speech sounds. Journal of Communication Disorders 39, 350-365

[10] Golfinopoulos E, Tourville JA, Guenther FH (2010) The integration of large-scale neural network modeling and functional brain imaging in speech motor control. NeuroImage 52, 862-874

[11] Riecker A, Mathiak K, Wildgruber D, Erb M, Hertrich I, Grodd W, Ackermann H (2005) fMRI reveals two distinct cerebral networks subserving speech motor control. Neurology 64, 700-706

[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] Fowler CA, Turvey MT (1980) Immediate compensation in bite-block speech. Phonetica 37, 306-326

[15] McFarland DH, Baum SR (1995) Incomplete compensation to articulatory perturbation. Journal of the Acoustical Society of America 97, 1865-1873

[16] Kröger BJ, Birkholz P, Kannampuzha J, Eckers C, Kaufmann E, Neuschaefer-Rube C (2011) Neurobiological interpretation of a quantitative target approximation model for speech actions. In: Kröger BJ, Birkholz P (eds.) Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011 (TUDpress, Dresden, Germany), pp. 184-194

[17] 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

[18] Levelt WJM, Wheeldon L (1994) Do speakers have access to a mental syllabary? Cognition 50, 239-269

[19] Houde JF, Nagarajan SS (2011) Speech production as state feedback control. Frontiers in Human Neuroscience 5, 82: doi: 10.3389/fnhum.2011.00082

[20] 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

[21] Kröger BJ, Miller N, Lowit A, Neuschaefer-Rube C. (2011) Defective neural motor speech mappings as a source for apraxia of speech: Evidence from a quantitative neural model of speech processing. In: Lowit A, Kent R (eds.) Assessment of Motor Speech Disorders. (Plural Publishing, San Diego, CA), pp. 325-346

[22] Johnson K (2008). Speaker normalization in speech perception. In: Pisoni DB, Remez RE (Eds.) The Handbook of Speech Perception. (Oxford, UK: Blackwell), pp. 363-389

[23] Kohonen T (2001) Self-Organizing Maps. (Springer, Berlin, Germany, 3rd edition)

[24] Kasabov N (2010) To spike or not to spike: A probabilistic spiking neuron model. Neural Networks 23, 16-19