Neural models of speech production and speech acquisition

Bernd J. Kröger

Department of Phoniatrics, Pedaudiologie and Comunication Disorders, University Hospital Aachen (UKA) and Aachen University (RWTH), Pauwelsstr. 30, 52074 Aachen, Germany
Email: bkroeger@ukaachen.de

Abstract

Neural models of sensorimotor control of speech production are rare. One current approach is the model given by Guenther (2006 and Guenther et al. 2006). Following the ideas of Guenther a comprehensive neural model of speech production using self-organizing neural networks can be given (Kröger et al. 2006a and 2006b). Both approaches simulate early processes of speech acquisition (babbling and imitation) as occur by toddlers during their first (and second) year of lifetime. The Kröger model is capable of generating acoustic speech signals and sensory feedback signals by using a high quality 3-dimensional articulatory-acoustic speech synthesizer as a front-end device. A mental syllabary forms the central layer within this model. The mental syllabary comprises a heap of neural SOM layers (phonetic map) which can be interpreted as a system of mirror neurons co-activating phonemic, sensory, and motor states of a syllable under production (feed-forward control). Feedback control is modeled by comparing the current sensory feedback state produced by the articulatory-acoustic model with the prestored sensory state, activated during feed-forward control via the mental syllabary. Both models can be integrated easily as a phonetic module within a more general linguistic model of speech production.

Introduction

A model of speech production can be subdivided into a control module – comprising feed-forward and feedback control – and an articulatory-acoustic module, i.e. into a controller and a controlled system or plant. While a lot of knowledge has been collected concerning the plant over the last decades – i.e. concerning static articulatory data and time-dependent kinematic articulatory data and concerning the articulatory-acoustic modeling – much less knowledge is availabe concerning the neural control of speech production (cp. Guenther 2006 and Guenther et al. 2006). The goal of our current work is to develop a comprehensive neural model of speech production which closely reflects as many aspects as possible of the natural human speech production process including self-organization of cortical structures occurring during speech acquisition.

The articulatory-acoustic module and motor and sensory parameters

Our three-dimensional speech synthesizer (i.e. articulatory-acoustic vocal tract model) (Birkholz et al. 2006) is controlled by a set of 10 articulatory parameters (Tab. 1). This set of parameters represents quasi-static articulatory states of all model articulators - i.e. lips, tongue, jaw, velum, and larynx (Fig. 1). 10 neurons of the joint coordinate motor map (Fig. 1) directly control the articulatory (or low-level motor) state. The acoustic model is driven by the vocal tract area function calculated from the geometrical data of the articulatory model for each articulatory state. The acoustic model is capable of generating vocal tract transfer functions as well as the acoustic speech signal.

Table 1: List of articulatory parameters, i.e. joint coordinate motor parameters

ABBR.	NAME OF ARTICULATORY PARAMETER
JAA	lower jaw angle
TBA	tongue body angle
TBL	tongue body horizontal location
TTA	tongue tip angle
TTL	tongue tip horizontal location
LIH	relative lip height
LIP	lip protrusion
VEH	velum height
HLH	hyoid horizontal location
HLV	hyoid vertical location

Relative lip height means: lip height relative to jaw.

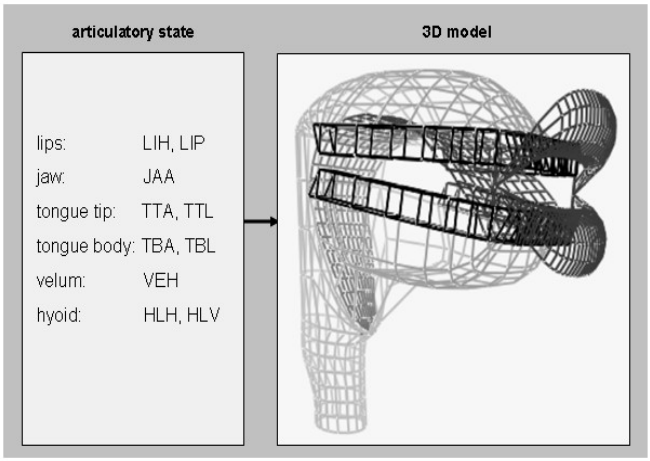
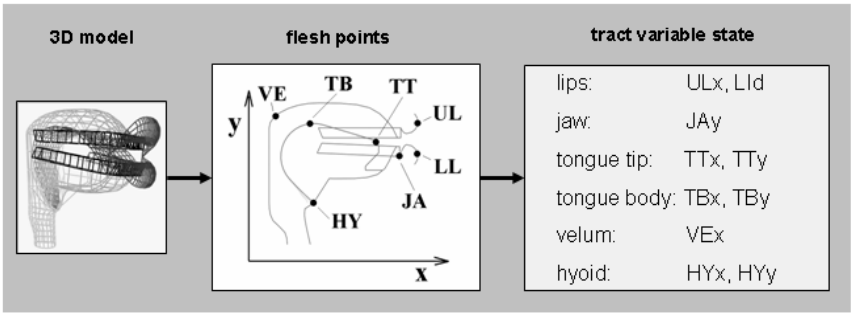
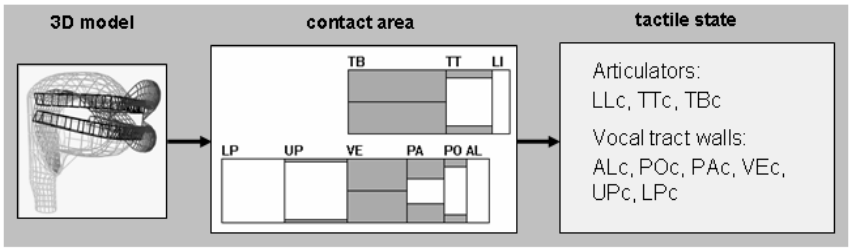


Figure 1: Articulatory parameters (for abbreviations see Tab. 1) and geometrical grid-representation of the 3D model.

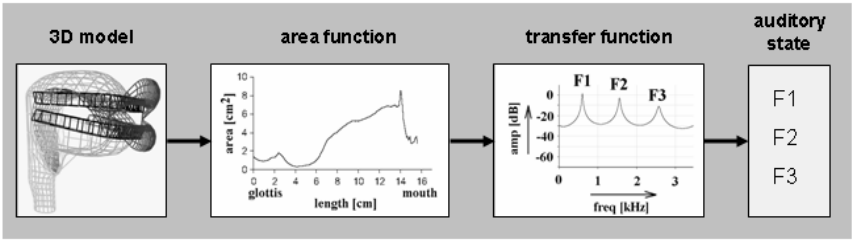
Somatosensory preprocessing comprises proprioceptive and tactile preprocessing. *Proprioceptive preprocessing* is accomplished by extracting the location of 7 flesh points relative to the cranial coordinate system (Fig. 2 and Tab. 2). This sensory information is directly used as a high-level motor representation, namely as the *spatial coordinate motor parameters* or *tract variables* within the spatial coordinate motor map. *Tactile preprocessing* is represented in our approach by extraction of the contact area between (i) lower and upper lip and between (ii) tongue and hard palate, soft palate, and pharyngeal wall (Fig. 2 and Tab. 3). The first 6 tactile parameters in Tab. 2 indicate the contact area at vocal tract walls while the last 3 parameters indicate the contact area at the movable articulators. *Auditory preprocessing* is represented in our approach by extraction of bark-scaled formant values F1, F2, and F3 from the vocal tract transfer function (Fig. 2).



(a)



(b)



(c)

Figure 2: Generation of (a) tract variable, (b) tactile, and (c) auditory feedback signal (for abbreviations see Tab. 2 and Tab. 3)

Table 2: List of tract variables, i.e. spatial coordinate motor parameters

ABBR.	NAME OF TRACT VARIABLE
ULx	upper lip horizontal position
JAy	lower jaw vertical position
TTx	tongue tip horizontal position
TTy	tongue tip vertical position
TBx	tongue body horizontal position
TBy	tongue body vertical position
VEx	velum horizontal position
HYx	hyoid horizontal position
HYy	hyoid vertical position
LId	lips vertical distance

Table 3: List of tactile parameters

ABBR.	NAME OF TACTILE PARAMETER
ALc	contact area of alveolar ridge
POc	contact area of postalveolar region
PAc	contact area of palatal region
VEc	contact area of velar region
UPc	contact area of upper pharyngeal region
LPc	contact area of lower pharyngeal region
LIc	contact area of lips
TTc	contact area of tongue tip
TBc	contact area of tongue body

One-layer feed-forward mappings, self-organizing maps (SOM's) and multidirectionality

Two different kinds of neural networks have been used in our modeling thus far, i.e. *unidirectional one-layer feed-forward neural networks* (Kröger et al. 2006a) and *multidirectional self-organizing neural networks* (Kröger et al. 2006b). While one-layer feed-forward networks can be used successfully for modeling the somatosensory-to-motor mapping, self-organizing neural networks are feasible for modeling many aspects of sensory-to-motor mappings for different tasks, e.g. for modeling the auditory-to-motor mapping for vowels and simple syllables like vowel-plosive combinations.

Self-organizing neural networks comprise one central layer (self-organizing map, SOM) and one or more side layers (phonemic, sensory, and motor layers, Fig 3). The central layer can be interpreted as a *cortical layer of mirror neurons* (cp. Kohler et al. 2002) leading to a multidirectional co-activation of all side layers – i.e. of the phonemic, the sensory, and the motor layers. Thus an activation of a syllable within the phonemic map leads to a co-activation of the appropriate auditory, somatosensory and motor states (production). Or an activation of the sensory state of a syllable leads to co-activation of the appropriate motor and phonemic state (perception). At least also the activation of the high-level motor state of a syllable (i.e. the motor plan of a syllable; covert speech) – for example induced by visual stimulation – leads to co-activation of the sensory and phonemic state.

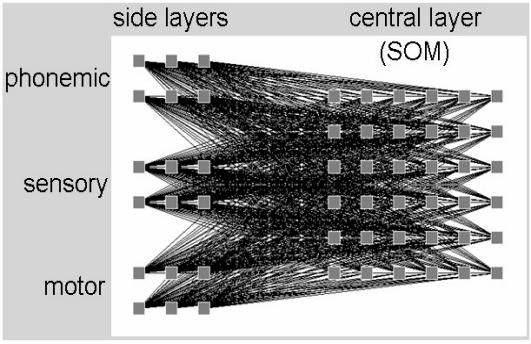


Figure 3: The organization of a multidirectional self-organizing network.

Towards a phonetic control model including multidirectional mappings and a mental syllabary

Our modeling results collected thus far for early phases of speech acquisition (i.e. babbling phase and imitation phase) can be subsumed in a comprehensive production model (Fig. 4). Within this control model the SOM layers of all mappings described above are called *phonetic map*. The *phonemic map* comprises the phonological codes of all sounds and frequent syllables of a language. Infrequent syllables are processed via a motor planning module. Multidirectional co-activation of phonemic, sensory, and motor states occur as described above via the phonetic map. The phonemic map and parts of the phonetic map form the *mental syllabary* (cp. Levelt and Wheeldon 1994). Feed-forward control startes with activation of a phonemic state for a currently activated syllable, followed by co-activation of the appertaining sensory and (high-level) motor state (i.e. motor plan) of this syllable via the mental syllabary or motor planning module. Feedback control is activated if the (prestored) co-activated sensory state deviates from the current feedback sensory state. A separation of motor planning and motor execution is also included in our approach (Fig. 4)

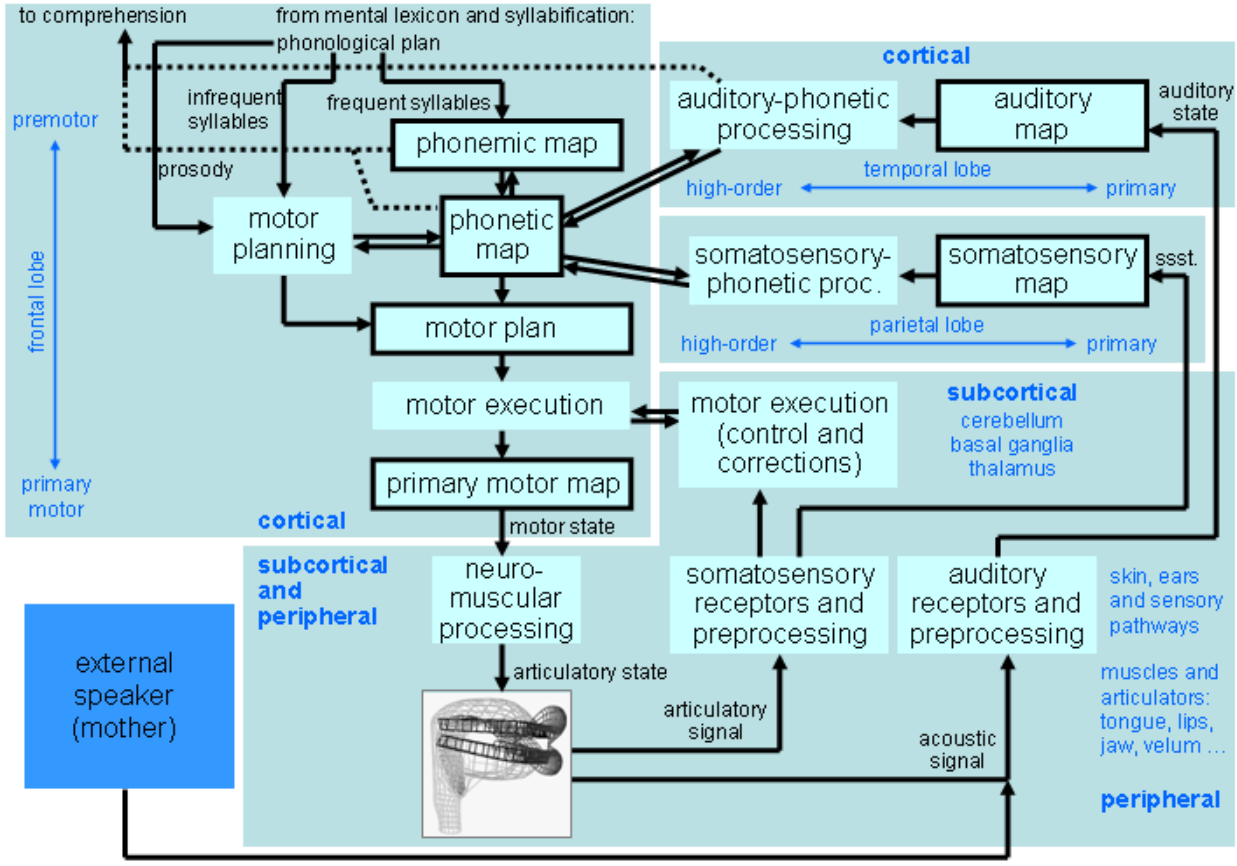


Figure 4: The organization of the multidirectional model of speech production (phonetic part).

A complete model of speech production

Our neurophonetic model of spech production can be seen as a part of the more general speech production model introduced by Levelt (1992), Levelt et al. (1999) and Indefreq and Levelt (2004). Two repositories – i.e. the *mental lexicon* and the *mental syllabary* are central components within this approach. Concepts, lemmas, and the phonological codes of lexical items are stored within the mental lexicon. The phonological codes and the motor plans of frequent syllables are stored within the mental syllabary. In our approach the mental syllabary in addition comprises the sensory representations of these syllables (efference copies) and is thus closely related to a *phonetic module*, i.e. to a heap of phonetic self-organizing maps, which are responsible for language specific production and perception effects (Fig. 5).

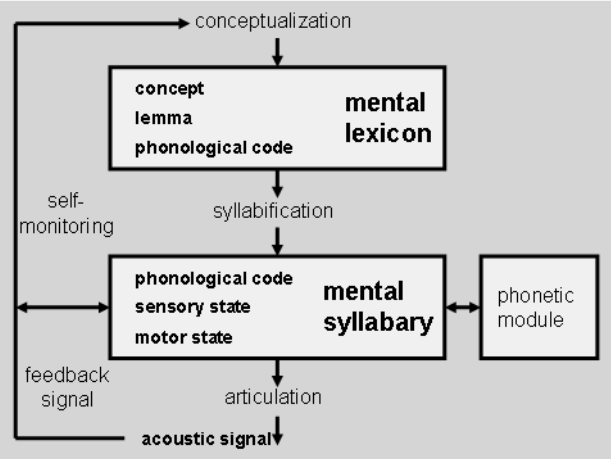


Figure 5: The organization of the complete linguistic and phonetic model of speech production.

Discussion

A comprehensive neural model of speech production is introduced. While this concept is mainly compatible with the approach of Guenther (2006), it strongly emphasizes the existence of multidirectional mappings on the level of the mental syllabary, leading to a co-activation of the phonemic state of a syllable with all co-occurring phonetic states, i.e. with sensory and motor states. A phonetic module – also occurring on the level of the mental syllabary – is introduced here as a central layer of mirror neurons, which are responsable for the co-activation of phonemic, sensory, and motor states. In addition a separation of motor planning and execution is introduced here. This model can be integrated as a phonetic part into a more general concept of speech production as introduced by Levelt et al. (1999).

References

Birkholz P, Jackel D, Kröger BJ (2006) Development and control of a 3D vocal tract model. *Proceedings of the IEEE International conference on Acoustics, Speech, and Signal Processing ICASSP 2006*, Toulouse, France, pp. 873-876

Guenther FH, Ghosh SS, Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96: 280-301

Guenther FH (2006) Cortical interaction underlying the production of speech sounds. *Journal of Communication disorders* 39: 350-365

Indefrey P, Levelt WJM (2004) The spatial and temporal signatures of word production components. *Cognition* 92: 101-144

Kohler E, Keysers C, Umilta MA, Fogassi L, Gallese V, Rizzolatti G (2002) Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297: 846-848

Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2006a) Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer. *Proceedings of the 9th International Conference on Spoken Language Processing* (Interspeech 2006 – ICSLP) pp. 565-568

Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2006b) Learning to associate speech-like sensory and motor states during babbling. *Proceedings of the 7th International Seminar on Speech Production* (Belo Horizonte, Brazil) pp. 67-74

Levelt WJM (1992) Accessing words in speech production: stages, processes and representations. *Cognition* 42: 1-22

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

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