The Organization of a Neurocomputational Control Model for Articulatory Speech Synthesis

Bernd J. Kröger[1], Anja Lowit[2], and Ralph Schnitker[3]

[1] Department of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and Aachen University, Aachen, Germany
bkroeger@ukaachen.de
[2] Speech and Language Therapy Division,
Department of Educational and Professional Studies, University of Stratclyde, UK
a.lowit@strath.ac.uk
[3] Central Service Facility "Functional Imaging" at the ICCR-BioMat,
University Hospital Aachen and Aachen University, Aachen, Germany
ralph@izkf.rwth-aachen.de

Abstract. The organization of a computational control model of articulatory speech synthesis is outlined in this paper. The model is based on general principles of neurophysiology and cognitive psychology. Thus it is based on such neural control circuits, neural maps and mappings as are hypothesized to exist in the human brain, and the model is based on learning or training mechanisms similar to those occurring during the human process of speech acquisition. The task of the control module is to generate articulatory data for controlling an articulatory-acoustic speech synthesizer. Thus a complete "BIONIC" (i.e. BIOlogically motivated and techNICally realized) speech synthesizer is described, capable of generating linguistic, sensory, and motor neural representations of sounds, syllables, and words, capable of generating articulatory speech movements from neuromuscular activation, and subsequently capable of generating acoustic speech signals by controlling an articulatory-acoustic vocal tract model. The module developed thus far is capable of producing single sounds (vowels and consonants), simple CV- and VC-syllables, and first sample words. In addition, processes of human-human interaction occurring during speech acquisition (mother-child or carer-child interactions) are briefly discussed in this paper.

Keywords: Computational model, neural model, speech production, articulatory speech synthesis, speech acquisition, self-organizing maps.

1 Introduction

While a lot of knowledge has been collected over the last decades concerning articulatory-acoustic models, i.e. front-end components for articulatory speech synthesis, which convert articulatory information into acoustic speech signals using an artificial vocal tract (Badin et al. 2002, Beautemps et al. 2001, Birkholz et al. 2006, Birkholz et al. 2007a, Engwall 2003, Kröger and Birkholz 2007), a remaining goal is the

A. Esposito et al. (Eds.): HH and HM Interaction 2007, LNAI 5042, pp. 121–135, 2008.

development of a human-like control component to generate articulatory, including phonatory, speech information. A comprehensive gestural control concept for articulatory speech synthesis has been introduced by Kröger and Birkholz (2007), but the best way of collecting realistic gestural speech data for generating high-quality segmental and prosodic control information for controlling the articulatory-acoustic front-end system, i.e. the artificial vocal tract model, remains an open question.

One possible solution is to employ *articulatory data* from a test subject with the objective of reproducing or mimicking the articulatory strategies of this speaker (e.g. Birkholz et al. 2007b). However, it is not trivial to extract the underlying data for controlling the movements of model articulators from articulatory flesh-point movement data in the form in which they are currently available, for instance from widely used tracking devices like EMA systems (Perkell et al. 1992, Stone 1997). A different approach is the direct use of *acoustic data*, since toddlers are extremely successful at learning to speak just by listening to persons without knowing their implicit articulatory strategies.

In this paper a neuro-computational control concept for articulatory speech synthesis is introduced, which is capable of mimicking important aspects of the natural process of speech acquisition like babbling and imitation. During the natural process of speech acquisition the toddler starts to explore his own vocal tract just by producing speech-like phonation while the vocal tract is not constricted. This enables the toddler to learn the acoustic-auditory consequences of different articulatory vowel-like states. Later on the toddler produces random vocal tract closing and opening movements and in addition learns the acoustic-auditory consequences of different articulatory consonant-like states. Both processes take place within the *babbling phase of speech acquisition* (Oller et al. 1999). Then the toddler is ready for imitating syllables or words produced by external speakers. This second phase is called the *imitation phase of speech acquisition* (Oller et al. 1999). During this phase the basis for the development of the mental syllabary and the mental lexicon is established for the target language (Levelt et al. 1999, Levelt and Wheeldon 1994, Indefrey and Levelt 2004).

A consideration of these natural processes of speech acquisition suggests that a successful strategy to develop high-quality articulatory speech synthesis could be to mimic these processes, i.e. to allow the model to explore its sensorimotor mappings during an artificial babbling phase, and to explore the production of syllables, words and utterances by imitating given acoustic speech items. Using this approach the organization of the computational control model should be designed in a similar way to the presumed *organization of speech production paths in the human brain* and the *learning strategies for acquiring phonetic and linguistic knowledge* should be similar to those processes which take place during natural speech acquisition. The organization of our computational control module with respect to neurophysiologic and neuropsychological knowledge as well as an outline of the learning processes for acquiring general phonetic and language specific speech knowledge are described in this paper.

2 The Computational Control Model

The control model is implemented quantitatively (computer-implemented) using artificial neural network algorithms (feed-forward networks and self organizing networks, cf. Kohonen 2001, Zell 2003) but the overall organization of the model (Fig. 1) mainly

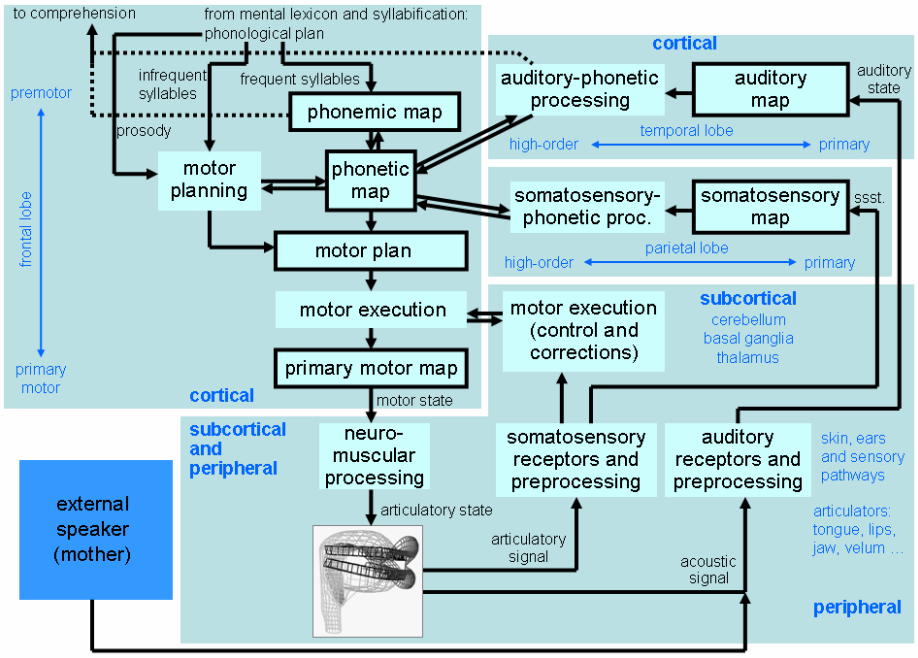

Fig. 1. Computational neural model of speech production. Boxes with black outlines represent neural maps, arrows indicate processing paths or neural mappings. Boxes without black outline indicate processing modules.

refers to basic neurophysiologic and neuropsychological principles (cf. Kandel et al. 2000, Frackowiak et al. 2004, Fadiga and Craighero 2004) given below.

Parallel processing: Primary somatosensory and primary motor cerebral cortical maps are topographically organized (somatotopy). The articulators (lips, tongue tip, tongue body, tongue root, velum, larynx) are controlled by parallel neural projections from the cerebral motor cortex (from the primary motor map, Fig. 1) to the peripheral motor units controlling different effectors (efferent pathways for lips, tongue, jaw, velum, larynx). *Motor information* is forwarded in parallel for different articulators. Furthermore motor information is forwarded in parallel by using two different main processing routes. A *direct route* forwards motor information from the primary cortical motor map to the motor units via the pyramidal tract. An *indirect route* is a mediated forwarding of motor information from the primary cortical motor map to motor units via basal ganglia and cerebellum. *Somatosensory information* is forwarded by parallel neural projections from the somatosensory receptor cells to the somatosensory cerebral cortex (afferent pathways for auditory and somatosensory information, Fig. 1). Somatosensory information from different articulators and vocal tract walls (e.g. hard and soft palate, pharynx wall) is also forwarded by parallel neural projections. *Auditory information* in different frequency regions (i.e. from different regions of the basilar membrane) is forwarded by parallel neural projections from the auditory

receptor cells within the cochlea to the primary auditory map of the cerebral cortex by tonotopical projections.

Serial or sequential processing: Afferent sensory and efferent motor pathways are organized hierarchically. *Sensory information* is processed at different stages of the afferent sensory pathway starting with the neural activation of peripheral receptor cells, passing through sequentially ordered processing stages (neural nuclei e.g. within the thalamus). This information leads to primary and higher-order neural cerebral cortical activations (somatosensory and auditory map in Fig. 1). *Motor information* is also processed at different levels within the efferent motor pathway (i.e. premotor cortex, motor cortex, motor units), starting with a higher-order *abstract motor representation of a goal-directed action,* i.e. an articulatory speech gesture such as lip closure for realization of the speech sound [b] (*motor plan level* in Fig. 1). This motor information leads to lower-order neural activations, providing a more and more detailed description of the realization of the current action. It ends with a detailed specification of the activation of motor units leading to coordinated articulator movements (e.g. coordinated movement of lower jaw, lower lips and upper lips in the case of a lip closure action).

Higher-order and lower-order motor and sensory representations: Higher-order and lower-order cortical maps contain motor and sensory neural representations of e.g. a sound or syllable. These are coded by the activation of a single neuron, or an ensemble of neurons within defined *neural maps* (represented by the boxes with black outlines in Fig. 1). Neural maps are found within the primary cortical sensory or motor areas, within unimodal association areas, and within hetero- or multimodal association areas of the cerebral cortex. Unimodal association areas process information coming from one single sense (e.g. auditory or somatosensory); hetero- or multimodal association areas process information coming from more than one sense. *Neural interactions, associations or mappings* connect neural maps (represented by the arrows between the boxes with black outlines in Fig. 1). These cortico-cortical mappings are complex: Parallel and sequential processing, known to operate in lower-order motor and sensory pathways, also occur for cortico-cortical sensory associations from the primary, via the uni-modal sensory to the multimodal sensory maps. In addition, complex mappings occur for cortico-cortical motor associations from the prefrontal cortex via the premotor cortex to the motor cortex. Furthermore, multimodal cortico-cortical associations (mirror neuron circuits, Fadiga et al. 2002 and Kohler et al. 2002) connect cortical maps and lead to a hypermodal co-activation of sensory, motor, and abstract linguistic (phonological) states for speech items. A hypermodal neural map or state is a neural map or state collecting information from all other maps. This map or state itself can not be attributed to a single (auditory or somatosensory) modality. It is beyond any modality. In the case of our approach, a hypermodal *phonetic map* is defined (see below).

Sensory neural processing leads to *integration* of information: Speech items are handled in parallel by different sensory systems (auditory and somatosensory) and are then integrated into a *hetero-, multi-, or hypermodal abstract percept* (e.g. a percept of a syllable or word). This integration is represented by the *phonetic map* in our model (Fig. 1). Here, auditory, somatosensory, phonemic, and motor planning information is brought together in one map.

Motor neural processing: planning and execution. Motor neural processing can be subdivided into planning (selection and preparation) and execution (see Fig. 1). The *selection phase of planning* leads to a temporally coordinated plan of abstract internal representation of goal-directed actions (motor plan of speech gestures, Fig. 2). On this level gestures are defined just by their goal. An example is presented in Figure 2., where three gestures are shown, namely (i) labial closing (clla): closing the lips; (ii) vocal tract formation: shaping the vocal tract – mainly tongue and lower jaw – into a specific form, e.g. into the form of an [a]-vowel (aavt); (iii) glottal opening (opgl): opening the glottal slit with the goal of ceasing vocal fold vibration. In addition, at this level the duration of each gesture (boxes with black outlines in Fig. 2) and their temporal coordination are defined. The duration of a gesture is the sum of the time required for the movement quantified as gestural rapidity or gestural onset (first light portion within the outlined box, representing each gesture in Fig. 2) and the following target portion (following dark portion within the outlined box representing each gesture in Fig 2; A more complete survey on the concept of gesture and gestural organization of speech is given in Kröger and Birkholz 2007).

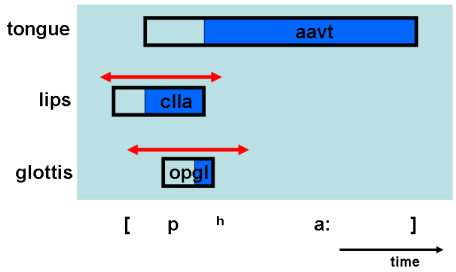

Fig. 2. Gestural organization (motor plan) of the syllable /pa/. Outlined boxes indicate gestural duration; aavt = [a]-forming vocal tract gesture, clla = labial closing gesture; opgl = glottal opening gesture. The arrows indicate the fact that temporal coordination of gestures (i.e. degree of temporal overlap of gestures) may vary.

The *preparation phase of planning* leads to a context-dependent higher-order specification of each gesture, e.g. specification of amplitude, temporal duration and temporal coordination of the gestures before they are performed. The planned (i.e. selected and prepared) gestures are then held in *working memory* for execution. *Execution* is the concrete lower-order realization of the gestures with respect to all of the articulators involved (i.e. lips and lower jaw in the case of a labial closing gesture). The earliest possible starting time for execution of an utterance is the completion of planning, at least for the first syllable of an utterance. The preparation and execution of a plan of speech gestures always takes place within a specific speaking context and is always modified within respect to this specific speaking context, e.g. whether the speaker is talking casually in a quiet environment, whether the speaker is speaking aloud in a noisy environment, or whether the speaker is presenting to a large audience.

Motor equivalence: A distinct gesture can be executed with a high degree of flexibility at the primary motor and at the articulatory level. If more than one articulator is

involved in gesture-execution, these articulators contribute to the execution with different strengths: in the case of a labial closing gesture, the jaw, the lower and upper lips are involved. The articulators produce larger movement amplitudes in [aba] than in [ibi] since overall jaw lowering is higher in the [a]- than the [i]-environment. Furthermore, depending on the context, one distinct gesture may start from different initial positions of the gesture executing vocal tract organ, and may exhibit different gestural duration and amplitude. Motor equivalence is mainly handled by the motor execution module (Fig. 1).

Control circuits: Speech motor control is based on the principles of *feed-forward (or anticipatory) control* and *feedback control* which underlie the performance of *skilled actions,* in a similar way to the motor control of other voluntary motor actions such as grasping. Speech movements are skilled actions and thus can be learned from "novel" or "untrained" to "automatic", "trained" or "overlearned". Learning of actions is described as *implicit learning*. Thus, at the beginning of training, a person uses feedback control circuits to monitor movements (see feedback loops for auditory and somatosensory signals in Fig. 1, starting with auditory and somatosensory preprocessing and leading to possible corrections of the motor plan via the auditory-phonetic and somatosensory-phonetic processing units, not described in detail in this paper), but at the end of successful training the action will be performed automatically "without thinking about it" by using mainly feed-forward control circuits (see the feed-forward pathway in Fig. 1, starting from phonological information and going down to neuromuscular processing). *Feedback or closed loop control* is accomplished by continuously monitoring the performance of an action via sensory feedback. If the desired state (reference sensory signal) does not correspond to the feedback signal, an error signal is produced, which leads to compensatory changes to the motor execution (error signals are calculated within the auditory-phonetic and somatosensory-phonetic processing units in our model, Fig. 1). *Feed-forward, or open loop control* is the direct motor execution of pre-learned skilled actions. The actual programming of these actions also needs sensory information about the state of the vocal organs of the speaker and about the environmental state (see above), but this is solely sensory information which is available *before* the action is performed. No sensory information controlling the accomplishment of the goal-directed movement, i.e. sensory information gained *during* the performance of the action is necessary. This is realized in our model by using the neural path starting with a phonological representation of a speech item (sound, syllable, word, or utterance) and generating its motor representation via the phonetic map (Fig. 1). In this case sensory states of the speech item may be co-activated as well via the phonetic-sensory mappings (Fig. 1), but that is not mandatory for the feed-forward execution of a speech item.

Learning and storing (acquiring) speech items: Two basic learning paradigms can be differentiated for the acquisition of speech gestures. *Babbling* is learning to associate the auditive and the somatosensory (tactile and proprioceptive) outcome of (random or semi-random) vocal tract motor activity with a (higher-order) representation of this motor activity. Learning during babbling mainly forms the sensorimotor mappings via the phonetic map (Fig. 1). *Imitation* means the active repeating and rehearsing of a perceived speech item stimulated by human-human interaction (i.e. between toddler and carer). Imitation needs basic sensorimotor knowledge already gained

during babbling. On the one hand, learning or training of motor skills can be described as *implicit learning* and thus leads to implicit knowledge stored in implicit memory (information on *how* a gesture is performed; involving the parietal heteromodal association cortex, cerebellum, and supplementary motor cortex, see Frackowiak et al. 2004, p. 25). Implicit training needs a high degree of repetition and the skill increases slowly until the task or gesture is overlearned and the action can be performed automatically and without attention. On the other hand, an action is defined on an abstract level mainly by the goal with which it is associated, and corresponds to sensory (visual, auditory, and/or somatosensory) cues. Thus a motor action can be linked to its sensory goal by *explicit learning,* leading to explicit knowledge (information on *what* gesture has to be executed; involving the prefrontal cortex and basal ganglia system, see Frackowiak et al. 2004, ibid.).

Integrative vs. cellular paradigm: Our control model in its current state is based on the neurophysiologic and neuropsychological principles given above and thus incorporates the *integrative neural paradigm*. However, it is also necessary to take into consideration the *detailed cellular paradigm* in order to understand the neurophysiology of speech production satisfactorily. The integrative approach only stipulates *where* neural processing occurs (i.e. which brain region is activated) and *what* is processed, but not *how* it is processed by an ensemble of or by single nerve cells (neurons). This detailed quantitative neural processing is achieved in the current model by using artificial neural networks. The architecture and function of these networks is based on general or basic neurophysiologic assumptions (e.g. Kohonen 2001, Zell 2003) as follows:

Neural maps (on a cortical level): Actions, percepts, or abstract representations of e.g. a speech item, also called *neural states* or *neural representations,* are represented on the detailed cellular neural level by *specific activation patterns* within distinct cellular collectives, i.e. in distinct neural maps (e.g. sensory maps, motor maps, phonetic, or phonological maps, Fig. 1).

Neural mappings are capable of linking two or more neural maps (arrows in Fig. 1). Two kinds of networks are used within our approach: *One-layer feed-forward networks* are capable of projecting states of one map (input map) on neural states of a second map (output map), e.g. from a lower-order auditory representation to a higher-order auditory representation or from a higher-order motor representation to a lower-order motor representation (Fig. 1). In addition, *self-organizing maps (SOMs)* and their mappings to related maps are capable of projecting states of one related map to states of one or more other related maps, e.g. from a higher-order sensory representation to a higher-order motor representation or to a symbolic linguistic (i.e. phonemic) representation. The central SOM itself is represented as *phonetic map* in our model and the related maps are the phonemic, auditory, somatosensory, and motor plan map (Fig. 1). The organization of these mappings may lead to co-activation (parallel activation) of phonemic, sensory and motor states (mirror neuron assumption, Fadiga et al. 2002, Kohler et al. 2002). In both types of networks (feed-forward and SOM), a vast number of neural connections (dendrites and axons) connect each neuron of one map with each neuron of another (associated) map. And in both types of neural networks, neurons fire if their activation threshold value is exceeded. Firing leads to positive or negative activation (exhibition or inhibition) of associated cells in

associated maps. The sign and degree of activation of the associated cells is determined by the *synaptic link weight* between two associated cells.

Neural learning or training: Synaptic link weights of all neural connections within a neural mapping are adjusted by exposing the network to a defined set of external stimuli. This procedure can be called training or learning. Standard learning or training algorithms (cf. Kohonen 2001, Zell 2003) are used for learning or training in both types of networks in our approach.

Overall organization of the control model: On the basis of the neurophysiologic and neuropsychological principles given above, a concrete organization of a neural model of speech production can be proposed (Fig. 1). The model comprises peripheral, subcortical, and cortical modules. Peripherally a 3-dimensional articulatory-acoustic model (also called "artificial vocal tract") provides the front-end device for producing articulatory and acoustic signals (cf. Kröger and Birkholz 2007). At the subcortical and peripheral level the motor signals control neuromuscular units and subsequently the vocal tract organs or articulators of the articulatory-acoustic model. The articulatory and acoustic signals subsequently generated by the articulatory-acoustic model are processed by somatosensory and auditory peripheral and subcortical preprocessing instances and are forwarded to the primary cortical auditory and a primary cortical somatosenory (tactile and proprioceptive) level (i.e. current sensory state related to a produced speech item). On the cortical level definite motor, sensory, phonemic, or phonetic states within neural maps (outlined boxes in Fig. 1) are activated via neural mappings (arrows in Fig. 1; neural processing units comprising maps and mappings are indicated by non-outlined boxes in Fig. 1). Cortico-cortical sensorimotor mappings are trained mainly during the babbling phase of speech acquisition using self-organizing maps (SOMs) for different types of motor states (see below). Self-organizing maps for different types of sounds (vowels, plosives, fricatives etc.) and for different types of syllables (CV, VC, CVC, …) are outlined as phonetic map. In addition a phonemic-phonetic mapping is trained during the imitation phase of speech acquisition, storing the relations between phonemic speech items and sensory states (i.e. how a speech sound, syllable, or word should sound and should feel like; i.e. the internal sensory representation of a speech item) and storing the relations between phonemic speech items and motor states (i.e. how a speech item has to be produced). Thus after imitation training the model is capable of (i) executing automated or highly overlearned speech items without extensively using feedback control (i.e. using previously stored or pre-learned motor plans as occurs for frequent syllables) directly by activating the appropriate stored motor state and (ii) of calculating and executing motor plans for infrequent syllables by using the motor planning module (via motor planning module, Fig. 1; the motor planning module and its mappings are not described in detail in this paper). The training of concrete phonetic submaps e.g. for proto-vocalic states during babbling and of vocalic states during the imitation phase of speech production is briefly described in Kröger et al. (2007).

Important features of the neural speech production model outlined above are (i) the differentiation of feed-forward and feedback control (cf. Guenther et al. 2006 and Guenther 2006), (ii) the separation of production paths for frequent syllables (via phonetic map) and infrequent syllables (via motor planning module) and (iii) the differentiation between a higher level motor representation of speech items (motor

plan level, Fig. 1) and a lower level motor representation (primary motor representation). While on the motor plan level all speech gestures forming the speech item, their intragestural parameters (gestural target values, gestural rapidity and gestural duration), and the temporal coordination between speech these gestures is defined (Fig. 2), the concrete execution of the gestural plan by concrete articulator movements is defined at the primary motor level (see also neurophysiologic and neuron-psychological principles given above).

3 Babbling and Imitation Phase of Speech Acquisition

Following Oller et al. (1999) the prelinguistic phase of speech acquisition or *babbling phase* can be subdivided into a phonation phase, a primitive articulation phase, an expansion phase, and a canonical phase. In the phonation phase speech-like proto-vocalic articulation occurs while during the later phases the toddler starts to produce sounds which include primitive vocal tract closing and opening movements for building up and releasing vocal tract constrictions or vocal tract closures. In our approach two cases of proto-articulation (or prelinguistic articulation) are modelled, i.e. the production of proto-vocalic and simple proto-consonantal closing gestures (cf. Kröger et al. 2007). During this phase of speech acquisition the model learns to relate sensory states to motor states for these proto-speech gestures. Thus with increasing babbling training, the model is capable of predicting motor states from sensory states. For example the model is then capable of predicting the articulatory target positions of a proto-vocalic gesture from a definite static F1-F2-F3 formant pattern or is capable of predicting the place of articulation of a proto-consonantal gesture from the temporal F1-F2-F3 transition pattern. In terms of the gestural concept (Kröger und Birkholz 2007) the model is capable of predicting *intra*gestural parameters from auditory states, i.e. gestural parameters defining the gestural target, the speed at which the target is reached (rapidity), and the overall duration of a gesture (cf. Fig. 2). Since all types of gesture are learned separately, they constitute different phonetic (sub-) maps for describing the sensorimotor relations for proto-vocalic and different proto-consonantal speech gestures. Motor plan and auditory link weight values are visualized in Fig. 3 for the vocalic phonetic map after babbling training. The training items comprise 540 proto-vocalic self-productions, covering the whole proto-vocalic articulator space between a cardinal [i], [a], and [u] (cf. Kröger et al. 2007). Standard training procedures were used (Kohonen 2001). Due to the knowledge stored within this vocalic map, proto-vocalic motor states can be predicted with high accuracy from their auditory states (prediction error below 1%). Furthermore it can be seen, that the vocalic states are ordered within this vocalic map with respect to the motor and auditory link weight values; link weight values change smoothly from neuron to neuron.

Since babbling training enables the model to predict motor states from sensory states, the *sensorimotor knowledge* stored within the phonetic map serves as a helpful basis for imitation of acoustic speech items produced by external speakers (mother or carer) during the imitation phase of speech acquisition.

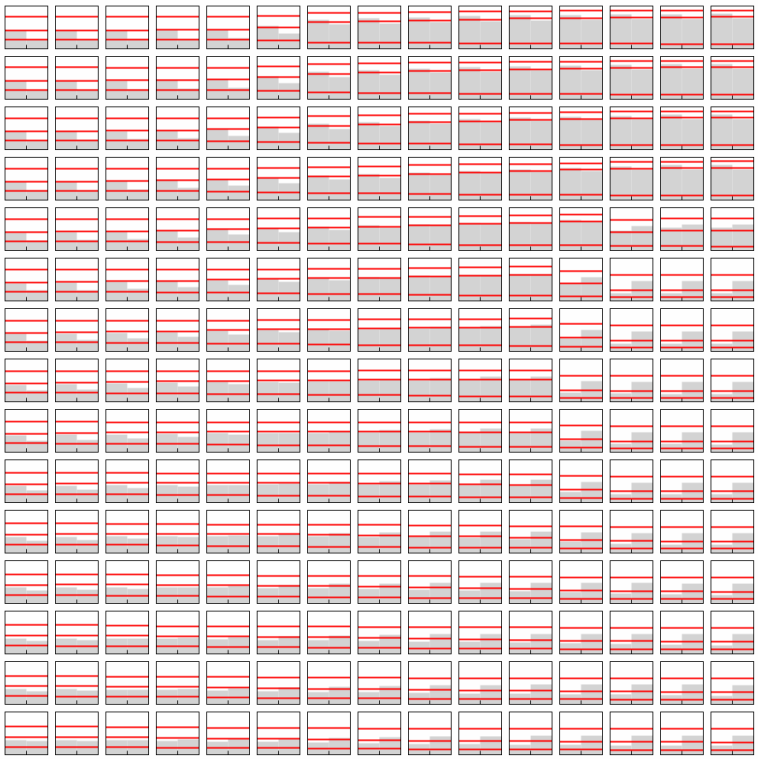

Fig. 3. Diagram of motor plan and auditory link weight values after proto-vocalic babbling training for each neuron within the vocalic phonetic map (15x15 neurons). Link weight values are given for two motor plan parameters within each neuron box: back-front (left bar) and low-high (right bar). Link weight values are given for three auditory parameters: bark scaled F1, F2, and F3 (horizontal lines within each neuron box).

In contrast to with babbling, during the *imitation training* of language specific vowels, consonants, syllables, or words, knowledge concerning phonological categories of sounds and concerning meaning of words is needed. For this imitation training a training set is used in our approach that comprises knowledge concerning the phonological category together with the knowledge concerning the auditory state of a speech item. The model starts training the phonological-auditory mapping via the phonetic map, i.e. enforces the neural connections between the phonological and auditory representation of this speech item (cf. Fig. 1). But how does a model get the knowledge concerning phonological categories and concerning the meaning of a word? This knowledge results from complex processes of human-human interaction: The toddler develops the neural connections between the acoustic form of a spoken object word and its meaning for example by pointing at an object (table, chair, door etc.) and by looking simultaneously at the carer, enforcing her/him to name the object. In the case of abstract words the learning process for combining auditory form and word meaning is even more complex. Currently this complex human-human interaction is not included in our modeling process. In our model this knowledge is directly

incorporated in the training sets for imitation training. The training set is organized as a heap of N acoustic realizations for each phonemic speech item (N=100 for different vowel phonemes and for consonant phonemes in different vocalic contexts, cf. Kröger et al. 2007) for training the phonological-auditory mapping followed by one imitation of each realization for training the phonological-motor mapping. For the vocalic phonetic map the motor plan and auditory link weight values are visualized in Fig. 4 after an imitation training of 500 vocalic productions which represent a 5 phoneme system /i/, /e/, /a/, /o/, and /u/ (cf. Kröger et al. 2007). This imitation training was performed after the vocalic babbling training described above. It is obvious that the vocalic states are ordered within this map in the same way as after babbling training. In addition five regions can be found within this vocalic phonetic map, which label high phonemic link weights and therefore potential realiations of the five vocalic phonemes defined in the training set.

Furthermore motor plan and auditory link weight values are visualized in Fig. 5 for the VC phonetic map for consonantal closing gestures after babbling training of three labial, apical, and dorsal closing gestures starting from 25 different proto-vocalic

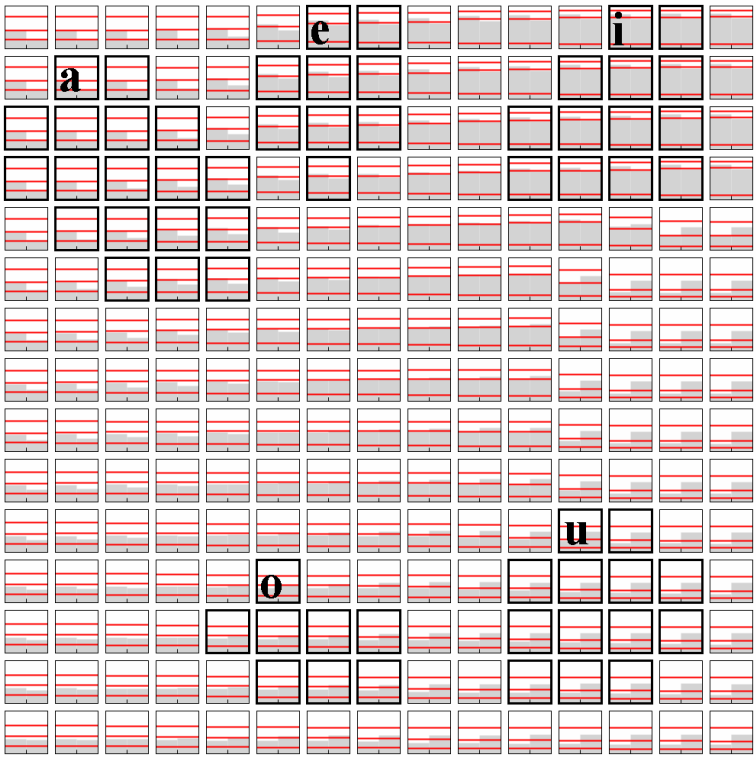

Fig. 4. Diagram of motor plan and auditory link weight values after vocalic imitation training for each neuron within the vocalic phonetic map (15x15 neurons; cf. Fig. 3). In addition the outlined boxes mark neurons within the vocalic phonetic map, which represent phoneme realizations.

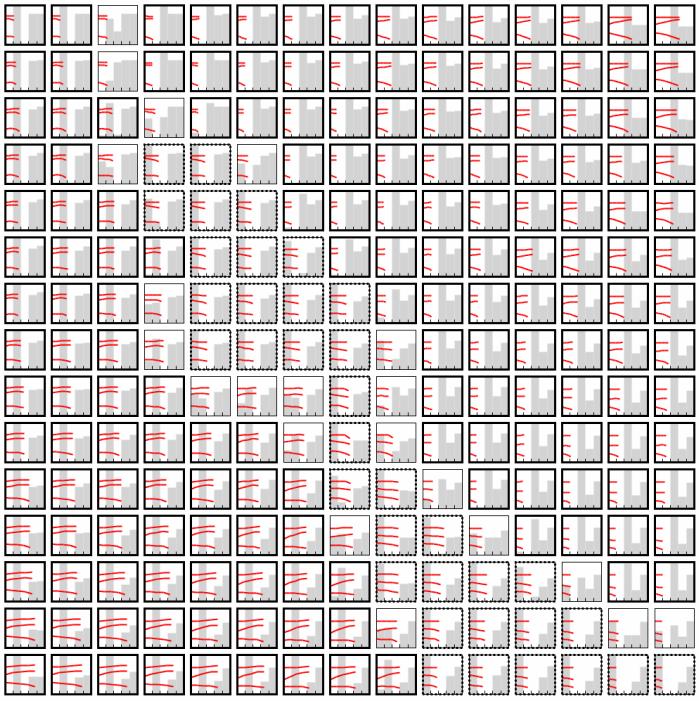

Fig. 5. Diagram of motor plan and auditory link weight values after proto-consonantal babbling training for each neuron within the vocalic phonetic map (15x15 neurons). Link weight values are given for five motor plan parameters within each neuron box. First three columns: vocal tract organ which performs the closing gesture (labial, apical, dorsal); two last columns: back-front value (forth column) and low-high value (fifth column) of the starting vowel within the VC-sequence. Link weight values are given for three auditory parameters: bark scaled F1, F2, and F3 (formant transitions within each neuron box). Outlined boxes: see text.

states each (i.e. set of 225 training items). Phonemic identification as /b/, /d/, or /g/ is obtained here by identifying all neurons by means of a dominant motor parameter "labial" as /b/, of a dominant motor parameter "apical" as /d/ and of a dominant motor parameters "dorsal" as /g/ ("dominant" means neural activation above .8 or 80% for the articulator motor neuron; see outlined boxes in Fig. 5: full outline bottom left: /d/, full outline top right: /g/; dashed outline: /b/). Also it can be seen that the VC states in this consonantal phonetic map are ordered within this map with respect to the motor and auditory link weight values. Three distinct regions can be detected within this consonantal phonetic map, which represent potential realizations of the three consonant phonemes. Standard training algorithms (Kohonen 2001) are used for the training of this VC-SOM resulting in a rate below 5% for the correct detection of a VC-syllable from the auditory state.

On the level of the mental lexicon (not displayed in our model), single neurons are defined for each word, leading to sub-activations of the phonological forms of the syllables and sounds, which make up the word on the level of the phonological map (Fig. 1). Also on the level of the phonological map, one single neuron represents a

definite syllable or speech sound. Phonological-auditory neural connections are built syllable by syllable between the phonological form of syllables and the appropriate auditory form via the phonetic map by (passive) listening. The following step of imitation training is necessary to learn the appropriate production, i.e. the appropriate motor plan for each phonological speech item. Starting with auditory states of simple gestures realizing vocalic sounds or simple syllables (CV and VC) composed of closing or opening gestures, the appropriate intragestural parameter values can be estimated by using the sensorimotor knowledge gained during the babbling phase (see above). In addition these intragestural parameter values are fine-tuned with respect to the language-specific auditory sound representations just learned for different phonological items. This results in language-specific gesture representations for all phonemes and simple syllables of a language.

In a further step this training is expanded to more complex speech items, i.e. to complex syllables and to words. Since the intragestural parameters are already trained for the language specific gestures, only learning of *inter*gestural timing of gestures within syllables and words remains. This training or learning procedure is performed by generating different motor plans of a syllable or word with differing intergestural timing. During the learning procedure the optimal intergestural timing values are estimated by minimizing the distance between the auditory representation for each motor plan and the already stored target auditory representation for the speech item. While successful training or fine-tuning of vocalic and consonantal gestures is obtained using self-organizing maps, our first modeling results suggest that this later training of intergestural timing can be performed successfully using simple one-layer feed-forward networks, connecting the phonological representation of the speech item directly with a quantitative description of gestural timing on the motor plan level.

4 Results and Discussion

The problem of developing high-quality articulatory speech synthesis by using "natural" control rules has not yet been solved, but this paper gives an outline for a "BIONIC" (biologically motivated technical) control concept for articulatory speech synthesis. The organization of this neurocomputational control concept for speech synthesis, or more generally for speech production, is described in this paper and is discussed with respect to general and basic principles of neurophysiology and neuropsychology. Based on this approach, structure and knowledge can easily be separated within the neural speech production system. The structure is given by the organization of the control model into neural maps (cf. Fig. 1), while the knowledge of speech production is stored within the neural link weights of the mappings connecting different maps. This knowledge is gathered during learning or training procedures (speech acquisition). In its current state the neurocomputational model was trained to produce static vowels and simple CV- and VC-syllables, i.e. simple opening and closing gestures producing voiced plosives.

An important feature of this neurocomputational control concept is the *separation of motor planning and motor execution* by using the gestural concept for describing articulatory speech movements and their control. First experiments for learning complex speech items composed of more than one gesture and especially for training the intergestural timing have been carried out.

One feature, which is beyond the scope of the actual model, but whose inclusion is an important aspect of the future development of this neurocomputational model is the modeling of *human-human interaction*. As has been discussed above, the toddler uses human-human interaction (toddler-carer interaction) for gaining the phonetics-phonology relations (i.e. for being able to associate meaning and phonetic word forms) and furthermore the toddler needs toddler-carer interactions in order to judge the quality of his productions during the imitation phase of speech acquisition: E.g. on the one hand no or negative reaction of the caretaker if a word is produced falsely, and on the other hand strong or positive reaction of the caretaker if a word is produced correctly the first time gives a strong support for learning word motor plans. Currently, the learning results of these human-human interactions, i.e. this knowledge is directly included in the training data used. It is beyond the scope of our current model to include strategies for gaining this knowledge.

Finally it should be mentioned that the control concept developed thus far focuses on the production path for *frequent syllables* via phonemic map, phonetic map to motor plan. In this pathway of the production model a phonemic-sensory and phonemic-motor mapping exists for each frequent syllable. An important question is how the toddler is able to generalize phonetic knowledge for sounds, syllabic subunits like onset and rhyme and for different types of syllables like CV, VC, CCV, CVC, etc. from the phonetic knowledge stored for each (frequent) syllable. This generalized phonological-phonetic knowledge could be used as basic knowledge for processing infrequent syllables, i.e. for computing motor plans for infrequent syllables from the appropriate phonological plans.

In summary the suggested modeling framework can serve as a basis for future models incorporating more explicit solutions for human-human interaction during speech acquisition. It should be used for modeling the vocabulary spurt, i.e. for modeling the development of the mental lexicon, not explicitly focused on in this paper. Furthermore the neural control model introduced here can be seen as a concrete neurophysiological and neuropsychological concept for the generation of speech movements starting from linguistic or phonological brain activations, and thus can be seen as a useful extension for more linguistically oriented word or utterance production models (cf. Levelt et al. 1999, Dell et al. 1999) generating the phonological form from a concrete communicative intention.

Acknowledgments. This work was supported in part by the German Research Council DFG grant Kr 1439/13-1. Acknowledgments go to the anonymous reviewers for giving helpful corrections and suggestions and to Jane F. Utting PhD for correcting the English.

References

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Savariaux, C.: Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. Journal of Phonetics 30, 533–553 (2002)

Beautemps, D., Badin, P., Bailly, G.: Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. Journal of the Acoustical Society of America 109, 2165–2180 (2001)

Birkholz, P., Jackèl, D., Kröger, B.J.: Construction and control of a three-dimensional vocal tract model. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), Toulouse, France, pp. 873–876 (2006)

Birkholz, P., Jackèl, D., Kröger, B.J.: Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech, and Language Processing 15, 1218–1225 (2007a)

Birkholz, P., Steiner, I., Breuer, S.: Control Concepts for Articulatory Speech Synthesis. In: Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, pp. 5–10 (2007b)

Dell, G.S., Chang, F., Griffin, Z.M.: Connectionist models of language production: lexical access and grammatical encoding. Cognitive Science 23, 517–541 (1999)

Engwall, O.: Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. Speech Communication 41, 303–329 (2003)

Fadiga, L., Craighero, L.: Electrophysiology of action representation. Journal of Clinical Neurophysiology 21, 157–168 (2004)

Fadiga, L., Craighero, L., Buccino, G., Rizzolatti, G.: Speech listening specifically modulates the excitability of tongue muscles: a TMS study. European Journal of Neuroscience 15, 399–402 (2002)

Frackowiak, R.S.J., Friston, K.J., Frith, C.D., Dolan, R.J., Price, C.J., Zeki, S., Ashburner, J., Penny, W.: Human Brain Function, 2nd edn. Elsevier Academic Press, Amsterdam (2004)

Guenther, F.H.: Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39, 350–365 (2006)

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

Indefrey, P., Levelt, W.J.M.: The spatial and temporal signatures of word production components. Cognition 92, 101–144 (2004)

Kandel, E.R., Schwartz, J.H., Jessell, T.M.: Principles of Neural Science, 4th edn. MacGraw-Hill, New York (2000)

Kohonen, T.: Self-organizing maps. Springer, Berlin (2001)

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

Kröger, B.J., Birkholz, P.: A gesture-based concept for speech movement control in articulatory speech synthesis. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 174–189. Springer, Heidelberg (2007)

Kröger, B.J., Birkholz, P., Kannampuzha, J., Neuschaefer-Rube, C.: Modeling the perceptual magnet effect and categorical perception using self-organizing neural networks. In: Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, pp. 789–792 (2007)

Levelt, W.J.M., Wheeldon, L.: Do speakers have access to a mental syllabary? Cognition 50, 239–269 (1994)

Levelt, W.J.M., Roelofs, A., Meyer, A.: A theory of lexical access in speech production. Behavioral and Brain Sciences 22, 1–75 (1999)

Oller, D.K., Eilers, R.E., Neal, A.R., Schwartz, H.K.: Precursors to speech in infancy: the prediction of speech and language disorders. Journal of Communication Disorders 32, 223–245 (1999)

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., Jackson, M.: Electronetic midsaggital articulometer (EMMA) systems for transducing speech articulatory movements. Journal of the Acoustical Society of America 92, 3078–3096 (1992)

Stone, M.: Laboratory techniques for investigating speech articulation. In: Hardcastle, J., Laver, J. (eds.) The Handbook of Phonetic Sciences, pp. 11–32. Blackwell, Oxford (1997)

Zell, A.: Simulation neuronaler Netze. Oldenbourg Verlag, München Wien (2003)