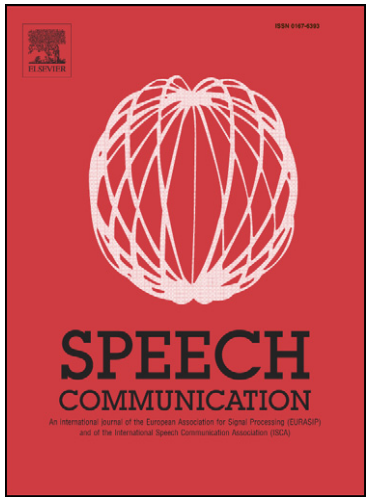
Towards a Neurocomputational Model of Speech Production and Perception

Bernd J. Kröger, Jim Kannampuzha, Christiane Neuschaefer-Rube



Towards a Neurocomputational Model of Speech Production and Perception

Bernd J. Kröger, Jim Kannampuzha, Christiane Neuschaefer-Rube

PII: S0167-6393(08)00130-1
DOI: 10.1016/j.specom.2008.08.002
Reference: SPECOM 1746

To appear in: *Speech Communication*

Received Date: 5 March 2008
Revised Date: 29 July 2008
Accepted Date: 27 August 2008

Please cite this article as: Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C., Towards a Neurocomputational Model of Speech Production and Perception, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.08.002

Towards a Neurocomputational Model of Speech Production and Perception

Bernd J. Kröger, Jim Kannampuzha, and Christiane Neuschaefer-Rube

Department of Phoniatrics, Pedaudiology, and Communication Disorders,

University Hospital Aachen and Aachen University, Aachen, Germany

bkroeger@ukaachen.de , jkannampuzha@ukaachen.de, cneuschaefer@ukaachen.de

Submitted to

Speech Communication

Special Issue NOLISP 2007 (editor: Prof. Dr. Kuldip Palival)

Abstract

The limitation in performance of current speech synthesis and speech recognition systems may result from the fact that these systems are not designed with respect to the human neural processes of speech production and perception. A neurocomputational model of speech production and perception is introduced which is organized with respect to human neural processes of speech production and perception. The production-perception model comprises an artificial computer-implemented vocal tract as a front-end module, which is capable of generating articulatory speech movements and acoustic speech signals. The structure of the production-perception model comprises motor and sensory processing pathways. Speech knowledge is collected during training stages which imitate early stages of speech acquisition. This knowledge is stored in artificial self-organizing maps. The current neurocomputational model is capable of producing and perceiving vowels, VC-, and CV-syllables (V = vowels and C = voiced plosives). Basic features of natural speech production and perception are predicted from this model in a straight forward way: Production of speech items is feedforward and feedback controlled and phoneme realizations vary within perceptually defined regions. Perception is less categorical in the case of vowels in comparison to consonants. Due to its human-like production-perception processing the model should be discussed as a basic module for more technical relevant approaches for high quality speech synthesis and for high performance speech recognition.

Keywords: speech, speech production, speech perception, neurocomputational model, artificial neural networks, self-organizing networks

1 Introduction

Current speech recognition systems are easily outperformed in the case of (i) non-restricted vocabulary, (ii) if the speaker in not well-known by the system and (iii) if noise reduces the speech signal quality (e.g. Benzeghiba et al. 2007, Scharenborg 2007). Current corpus-based speech synthesis systems are limited as well, especially concerning (i) flexibility in modeling different speaker and voice characteristics and concerning (ii) segmental as well as prosodic naturalness (e.g. Clark et al. 2007, Latorre et al. 2006). These limitations may be attributed to the fact that speech recognition as well as speech synthesis systems currently are not modeled with respect to the basic human neural processes of speech production and speech perception.

A variety of brain imaging studies clarify the role of different subcortical and cortical brain *regions* for speech production (e.g. Murphy et al. 1997, Kuriki et al. 1999, Wise et al. 1999, Bookheimer et al. 2000, Rosen et al. 2000, Scott et al. 2000, Benson et al. 2001, Huang et al. 2001, Blank et al. 2002, Vanlancker-Sidtis et al. 2003, Ackerman and Riecker 2003, Hillis et al. 2004, Shuster and Lemieux 2005, Kemeny et al. 2005, Riecker et al. 2006, Sörös et al. 2006) as well as for speech perception (e.g. Binder et al. 2000, Hickok and Poeppel 2000, Fadiga et al. 2002, Wilson et al. 2004, Boatman 2004, Poeppel et al. 2004, Rimol et al. 2005, Liebenthal et al. 2005, Uppenkamp et al. 2006, Zekveld et al. 2006, Obleser et al. 2006 and 2007). Other studies focus on the interplay of speech production and perception (Heim et al. 2003, Okada and Hickok 2006, Callan et al. 2006, and Jardri et al. 2007) but only few among them introduce *functional* neural models which explain and emulate (i) the complex neural sensorimotor processes of speech production (Bailly 1997, Guenther 1994, 1995, 2006, Guenther et al. 2006) and (ii) the complex neural processes of speech perception including comprehension (McClelland and Elman 1986, Gaskell and Marslen-Wilson 1997, Luce et al. 2000, Grossberg 2003, Norris et al. 2006, Hickok and Poeppel 2004 and 2007).

It is the aim of this paper to introduce a biologically motivated approach for speech recognition and synthesis, i.e. a computer-implemented neural model using artificial neural networks, capable of imitating human processes of speech production and speech perception. This production-perception model is based on neurophysiological and neuropsychological knowledge of speech processing (Kröger et al. 2008). The structure of the model and the process of collecting speech knowledge during speech acquisition training stages are described in detail in this paper. Furthermore it is described how the model is capable of producing vowels and CV-syllables and why the model is capable of perceiving vowels and consonants categorically.

2 The Structure of the Neurocomputational Model

While the *structure* of this neurocomputational model is based on neurophysiological and neuropsychological facts (Kröger et al. 2008), the speech *knowledge* itself is gathered by training artificial neural networks which are part of this model (Kröger et al. 2006a and 2006b). The organization of the model is given in Fig. 1. It comprises a cortical and a subcortical-peripheral part. The cortical part is subdivided with respect to neural processing within the frontal, the temporal, and the parietal cortical lobe. Functionally the model comprises a production and a perception part. In its current state the model excludes linguistic processing (mental grammar, mental lexicon, comprehension, conceptualization) but focuses on sensorimotor processes of speech production and on sublexical speech perception, i.e. sound and syllable identification and discrimination.

The *production part* is divided into feedforward and feedback control (see also Guenther 2006). It starts with the phonemic representation of a speech item (speech sound, syllable, word, or utterance) and generates the appropriate time course of articulatory movements and the appropriate acoustic speech signal. The phonemic representation of a speech item is generated by higher level linguistic modules (Levelt et al. 1999, Dell et al. 1999, Indefrey and Levelt 2004) subsumed as widely distributed frontal-temporal procedural and declarative neural processing modules (Ullman 2001, Indefrey and Levelt 2004) which are not specified in detail in this model. Subsequently each phonologically specified syllable (i.e. a phonemic state; a neural activation pattern on the level of the phonemic map) is processed by the *feedforward control* module. In the case of a *frequent syllable*, the sensory states (auditory and somatosensory state) and the motor plan state of the syllable (which are already learned or trained during speech acquisition; see below) are activated via the phonetic map. The *phonetic map* (Fig. 1) can be interpreted as the central neural map constituting the mental syllabary (for the concept of mental syllabary see Levelt and Wheeldon 1994 and Levelt et al 1999). For each frequent syllable a phonemic state initiates the neural activation of a specific neuron within the phonetic map, which subsequently leads to activation patterns of the appropriate sensory states and the appropriate motor plan state. In the case of *infrequent syllables* the motor plan state is assembled within the motor planning module on the level of sub-syllabic units, e.g. syllable constituents like syllable onset and syllable rhyme or single speech sounds (Varley and Whiteside 2001). This path is not implemented in our model at present. On the level of the *motor plan map* a high level motor state (motor plan) is activated for each speech item under production (current speech item). This high level motor state defines the temporal coordination of *speech gestures* or *vocal tract action units* (Goldstein et

al. 2006, Saltzman and Munhall 1989; for a general description of goal-directed action units see Sober and Sabes 2003, Todorov 2004, Fadiga and Craighero 2004). The motor plan of a speech item is processed by the motor execution module in order to define the spatio-temporal trajectories of articulator movements. Thus the motor execution module calculates the concrete specification of each speech gesture on the level of the *primary motor map* (cf. Ito et al. 2004, Sanguineti et al. 1997, Saltzman 1979, Saltzman and Munhall 1989, Saltzman and Byrd 2000). For example, a labial closing gesture involves coordinated movement of at least the lower jaw, the lower and upper lips. Thus each of these articulators must be controlled synergetically for the realization of a speech gesture. Subsequently the movement of an articulator is executed by activating the motor units controlling this articulator via the neuromuscular processing module.

-- insert Figure 1 about here --

The (lower level) primary motor map comprises 10 articulatory parameters (Kröger et al. 2006b). Each articulatory parameter value is coded by two neurons with complementary activation (see below) leading to 20 neurons to encoding the primary motor commands for each point in time. The conversion of physical parameter values (e.g. displacement of an articulator) into neuromotor activation patterns is done (i) by mapping the physical displacement range for each parameter onto a neural activation range [0, 1] (i.e. no activation to full activation of a neuron) and (ii) by defining two neurons for each parameter with complementary activation ($a_2 = 1 - a_1$) in order to hold the overall activation a ($a = a_1 + a_2$) constant (= 1) for each parameter value. The size of the (higher level) motor plan map depends on the length of the utterance under production. In the case of V-, CV-, and VC-items three vocalic higher level parameters (high-low, front-back, rounded-unrounded) and four higher level consonantal parameters (labial, apical, dorsal, exact closing position) are controlled. These vocalic parameters and the consonantal parameter closing position are encoded using 2 neurons with complementary activation each, while the three remaining consonantal parameters are encoded by one neuron each in order to reflect the activation of a specific vocal tract organ. Thus the motor plan map for V-, CV-, and VC-items consists of 11 neurons. Since a motor plan encodes a motor or sensory V-, CV-, or VC-item of a transition for C (encoded by 4 time labels) and a steady state portion for V (encoded by one time label) the (lower level) primary motor state of these items is encoded by five consecutive time labels. Thus the appropriate number of primary motor map neurons for a whole speech item is

5 x 20 = 100 neurons plus 10 neurons for coding 5 time intervals describing the temporal distance from label to label.

A computer-implemented numerical *articulatory vocal tract model* generates the time course of vocal tract geometries and subsequently the *acoustic vocal tract model* generates the acoustic speech signal. A three-dimensional articulatory-acoustic model is used here which is capable of generating high-quality articulatory and acoustic speech signals (Birkholz and Jackèl 2004, Birkholz and Kröger 2006 and 2007, Birkholz et al. 2006 and 2007, and Kröger and Birkholz 2007). These articulatory and acoustic signals are used for feedback control.

The articulatory and acoustic signals generated by feedforward control are continuously monitored or controlled. For this *feedback control* the articulatory and acoustic signals are converted into neural signals by auditory and somatosensory (i.e. tactile and proprioceptive) receptors. Somatosensory feedback signals (relative positions of articulators to each other and position and degree of vocal tract constrictions, see Saltzman and Munhall 1989, Shadmehr and Mussa-Ivaldi 1994, Tremblay et al. 2003, Nasir and Ostry 2006) are used for controlling motor execution. In addition sensory (i.e. somatosensory and auditory) signals are converted into *higher level cortical sensory states,* which represent the current speech item. These auditory and somatosensory (feedback) states of a currently produced speech item are processed by comparing them with the appropriate prelearned auditory and somatosensory state, activated by feedforward control before the current speech item is produced. This comparison is done on the level of the somatosensory and auditory processing modules. If the prestored (or feedforward) sensory state and the feedback sensory states indicate a reasonable difference an error signal is activated for correcting the motor plan during the ongoing feedforward control.

The conversion of physical or psychophysical sensory parameter values (e.g. bark scaled formant values) into neural activation patterns is done (i) by mapping the whole physical parameter range onto the "neural" range [0, 1] (i.e. no activation to full activation of a neuron) and (ii) by defining two neurons per parameter with complementary activation (see above for the primary motor map). Since auditory states are processed as whole patterns, parameter values for our V-, CV-, and VC-items (see above) are obtained at 5 positions (labels) in the acoustic signal. Three formants were processed leading to 3 x 5 = 15 parameter values and thus to 30 neurons per item for the auditory state map. 10 proprioceptive and 9 tactile parameters were processed (Kröger et al. 2006b) leading to 19 parameter values and thus 28 neurons for each item. Only one tactile and proprioceptive state is coded for the whole speech item representing the gestural target region of the vocalic part in the case of a V-item

and representing the gestural target regions of the vocalic and the consonantal part in the case of VC- or VC-items (overlay of tactile and proprioceptive patterns).

The *perception part* of the neurocomputational model starts from an acoustic speech signal, generated by an external speaker (Fig. 1). This signal is converted into neural signals by auditory receptors and is further processed into a cortical higher level auditory signal via the same auditory pathway that is used for the feedback control of speech production (self-productions). Speech perception comprises two pathways (cf. Hickock and Poeppel 2000, 2004, and 2007). The *auditory-to-meaning pathway (ventral stream)* directly activates neural states within the mental lexicon by the high level cortical auditory state for a speech item (e.g. a word). This pathway is not included in our model, since high level mental lexical representations are out of the scope of this study. The *auditory-to-motor pathway (dorsal stream)* activates the phonetic state of the current speech item (e.g. sound or syllable) within the cortical frontal motor regions. This pathway is included in our model and it will be shown below that this pathway is capable of modeling categorical perception of speech sounds and is capable of modeling differences in categorical perception of vowels and consonants.

The structure of the neurocomputational model differentiates neural maps and neural mappings. *Neural maps* are ensembles of neurons which represent the phonemic, phonetic, sensory or motor speech states. These maps are capable of carrying states of different speech items by different neural activation patterns. These activations change from speech item to speech item under production or perception. *Neural mappings* represent the neural connections between the neurons of neural maps (Fig. 2). These connections can be excitatory or inhibitory. The degree of excitatory or inhibitory connection is described by link weight values. These values w_{ij} characterize the neural connection between each pair of neurons. They define the degree of activation of a connected neuron b_j within a neural map 1 (comprising M neurons $j = 1, \ldots, M$) resulting from the degree of activation of all connecting neurons a_i within a neural map 2 (comprising N neurons $I = 1, \ldots, N$) (see Eq. 1 and Fig. 2).

$$b_j = actfunc\left(\sum_{i=1}^{N} a_i w_{ij}\right) \quad for \;\; j = 1,..., M \;, \qquad \text{(eq. 1)}$$

Here *actfunc* is the activation function (a sigmoid function in the case of our modelling; see Zell 2003) which represents the total activation of neuron b_j in map 1 as function of the sum of activations from all neurons i within map 2. The link weight values w_{ij} are limited to the interval [-1, +1] (i.e. maximal inhibitory to maximal excitatory link weight value).

-- insert Figure 2 about here --

The link weight values reflect the whole knowledge inherent in the training data and thus the knowledge gathered during the training procedures. Link weight values are adjusted during training stages, i.e. during speech acquisition stages (see below). They are allowed to be modified continuously in order to reflect new knowledge gained over life time.

One-layer feedforward networks (Fig. 2) are of limited power and are used in our model exclusively for calculating articulatory joint-coordinate parameters from articulatory tract-variable parameters (cf. Kröger et al. 2006c). In this paper we will focus on the central phonetic map and the multilateral co-activation of phonemic states, sensory states, and motor plan states via the phonetic map. This multilateral co-activation is achieved by using self-organizing maps or networks (Kohonen 2001 and Fig. 3). Each neuron of the central self-organizing map (i.e. the phonetic map) represents a speech item. Different phonetic submaps (i.e. different parts within the phonetic map) are defined for each class of speech items, i.e. for vowels, for CV-, and for VC-syllables. Multilateral co-activation of phonemic, sensory, and motor plan states for a speech item via the phonetic map means that an activated neuron of the phonetic map (representing a currently perceived or produced speech item) leads to a co-activation of neural activation patterns within the phonemic, motor plan, or sensory side layer maps representing this current speech item. The set of link weight values of the connections between all neurons of the phonemic, motor plan, or sensory side layer map and a neuron within the central phonetic map characterize the phonemic, motor plan, or sensory state of the speech item represented by this neuron within the phonetic map. Activation patterns of neurons within the side layer maps induced by an activation pattern of the phonetic map as well as activation patterns of the phonetic map induced by an activation pattern of one of the side layer maps are calculated in the same way as it is described above for simple one-layer feedforward networks (eq. 1)

-- insert Figure 3 about here --

The structure of the neurocomputational production-perception model introduced here is based on the structure of the DIVA model introduced by Guenther (2006) and by Guenther et al. (2006). The approach described in this paper as well as the Guenther approach comprise a feedforward and a feedback control path. Both approaches comprise self-organizing networks for processing neural states and comprise neural maps for storing phonemic, motor, and sensory states representing speech items. Both approaches introduce pre-linguistic and early linguistic language-specific training (i.e. babbling and imitation training, see below) in

order to shape the neural mappings within the computational models and both approaches include the concept of a mental syllabary (Levelt and Wheeldon 1994, Levelt et al 1999) and basic ideas of the mirror neuron concept (Fadiga and Craighero 2004, Rizzolatti and Craighero 2004) since both approaches claim a simultaneously occurring activation of sensory and motor states for speech items.

But there are three major differences between both approaches. Firstly, the DIVA approach does not *separate motor planning and motor execution* as is introduced here. This separation results from the fact that for all types of voluntary movements (actions) just the goal of an action (e.g. grasping a definite object or pressing a sequence of buttons) and the temporal overlap or temporal sequencing of actions are determined on the planning level while the details of movement execution are determined on lower neural levels (Kandel et al. 2000, Kröger et al. 2008). In the case of speech production *vocal tract action units* or *speech gestures* are well established as basic units of speech production (Browman and Goldstein 1989 and 1992, Goldstein et al. 2006 and 2007) separating motor speech planning – i.e. the level of action scores (Goldstein et al. 2006) – and motor speech execution (Saltzman and Munhall 1989, Goldstein et al. 2006) – i.e. the detailed determination of all articulator movements. The practical importance of dynamically defined speech action units becomes apparent if modelling of segmental reduction effects resulting from high speech rate (Kröger 1993) or if modelling of speech errors (Goldstein et al. 2007) is attempted. Secondly, the DIVA model does not explicitly introduce a *phonetic map* or at least a map, reflecting the self-organization of speech items between sensory, motor, and phonemic representation; and the DIVA model does not explicitly claim *bidirectional mappings* between phonemic, sensory, and motor representations. But the assumption of bidirectional associations is essential in our production-perception model. Production is modelled in our approach using neural connections from the phonemic map directed towards the motor and sensory maps via the phonetic map and perception is modelled in our approach using the neural connections from sensory maps directed toward phonemic map via the phonetic map. Furthermore the phonetic map itself is a central concept in our approach. On the one hand, the phonetic map introduces a *hypermodal description* of speech items which connects the sensory and motor representations of a speech item as is claimed in the mirror neuron theory. Our simulation results indicate that it is very feasible to introduce this level of neural self-organization (phonetic map) since it elucidates the ordering of speech items with respect to phonetic features (*phonetotopy*, see below). Furthermore the notion of the phonetic map is important for modelling speech perception since perceptual discrimination is defined in our approach as a

distance between activated states on this neural level (see below). Thirdly, the DIVA model is a production model not aiming for modelling *speech perception*. But according to the arguments given above the modelling of speech production and speech perception as two closely related processes is of great importance. This is achieved in our approach.

3 Gaining Speech Knowledge: Training Stages for Speech Acquisition

Speech knowledge is gained during training stages which model *basic stages of human speech acquisition.* This knowledge is stored within the mappings of the model, i.e. by the link weight values connecting the neurons within the production-perception model. Link weight values are adjusted during training stages. Two basic training stages can be differentiated, i.e. the babbling and the imitation stage (Oller et al. 1999).

For *babbling training* the training sets comprise pre-linguistic speech items, i.e. proto-vocalic and proto-syllabic speech items. The model randomly produces proto-vocalic and proto-syllabic speech items and listens to its own productions using the auditory feedback path (Fig. 1). The link weights between the sensory maps and the motor plan map are adjusted via the phonetic map during this training stage. No phonemic activation occurs since these training items are pre-linguistic. The knowledge which is gained during babbling training is language independent *general phonetic knowledge*. During this training stage the neuro-computational model learns the *sensorimotor interrelationship* of the vocal tract apparatus and its neural control, i.e. the interrelationship between various motor plan states and their resulting somatosensory and auditory states. The babbling training can be subdivided into training stages for training of proto-vocalic states and for proto-syllabic CV- and VC-states.

The *proto-vocalic babbling training set* comprises a set of proto-vocalic states which exhibit a quasi-continuous variation of the vocalic motor plan parameters low-high and front-back. The training set used here comprises 1076 proto-vocalic states which cover the language independent articulatory vowel space between the cardinal vowel qualities [i], [a], and [u] (Fig. 4). Each proto-vocalic motor plan state is defined by the vocalic parameters back-front and low-high. Thus the proto-vocalic training stimuli form a two-dimensional plane within the F1-F2-F3 acoustic vowel space (Fig. 4). Other motor parameters like tongue body height, tongue body horizontal and vertical position, and like lip opening are functions of these two motor plan parameters (Kröger et al. 2006c). Even lip-protrusion is a function of the parameter front-back in the case of this preliminary training set since the training set does not include rounded front proto-vowels (e.g. [y]-like vowel qualities).

-- insert Figure 4 about here --

The *proto-syllabic CV and VC babbling training sets* are based on a set of 31 proto-vocalic motor plan states which are characterized by the motor plan parameters back-front and low-high and which are covering the whole articulatory vowel space. Labial, apical, and dorsal opening and closing gestures (proto-CV- or proto-VC-gestures) starting or ending with a full closure are superimposed on these proto-vocalic states. Three proto-consonantal places of articulation (front-mid-back) are defined per gesture. This leads to a total amount of 279 training items for each of the two proto-syllabic training sets (proto-CV- and proto-VC-training set). The articulatory velocity of the gesture-executing articulator for all closing and opening gestures is proportional to the distance between actual articulator position and gestural target position. This leads to an exponential time function for the displacement of this articulator. A gesture is ending if the articulator-target distance is below 10% in the case of opening or proto-CV-gestures or if a full closure is reached in the case of closing or proto-VC-gestures (target of closing gestures is beyond the full closure position). In summary the motor plan state for these proto-CV- and proto-VC-gestures is defined by (i) two vocalic parameters (back-front and low-high), by (ii) the gesture-performing articulator (labial, apical, or dorsal) and by (iii) the exact proto-consonantal closing position (front-mid-back). The lower level (or primary) motor parameters and their time courses are calculated from these motor plan parameters by the motor execution module. The appropriate auditory state of these opening and closing gestures (proto-CV- and proto-VC-syllables) is the time courses of the first three formants F1, F2, and F3 (Fig. 5 and see section 2) and the appropriate somato-sensory state for each proto-syllabic motor plan state comprises tactile information of the proto-consonantal closure and proprioceptive information of the proto-vocalic opening.

-- insert Figure 5 about here --

Proto-vocalic, proto-CV-syllabic, and proto-VC-syllabic babbling training is performed independently from each other. Three self-organizing maps (size: M = 15 x 15 = 225 neurons) form three phonetic submaps and are trained by using the three training sets described above. Training leads to an adjustment of link weight values w_{ij} between the N side layer neurons a_i and the M central layer neurons b_j. The side layers consist of the motor plan map (i = 1, … , K) and the sensory (auditory and somatosensory) maps (i = K+1, … , N) while

the central layer represents the phonetic map ($j = 1, \ldots, M$). Link weight values are initialized by random values within the interval [0, 1] (i.e. no activation to full activation). The link weights $w_{ij}(t_{init})$ are initialized using random values between 0 and 1 (Eq. 2). This adjustment of the link weights is done incrementally, i.e. step by step, using Hebbian learning (Eq. 3). When a new stimulus I with $I = (x_0, \ldots, x_N)$ is presented, the winner neuron b_{winner} is identified in the central layer by calculating the minimum of Euclidian norm between I and W_j, $j = 1, \ldots, M$; i.e. $winner = \arg\min_j(\|I - W_j\|)$ where W_j is a vector containing the link weights of all links from the central layer neuron b_j to the side layer neurons a_i, i.e. $W_j = (w_{1j}, \ldots, w_{Nj})$. Once the winner neuron b_{winner} is identified the link weights for a step t with $t_{init} < t < t_{max}$ are updated as

$$w_{ij}(t_{init}) = rand(0,1) \qquad \text{(eq. 2)}$$

$$w_{ij}(t+1) = w_{ij}(t) + N_{winner,j}(t) \cdot L(t) \cdot (I_i - w_{ij}(t)), \qquad \text{(eq. 3)}$$

where $0 < L(t) < 1$ is a constantly decreasing learning factor defined as

$$L(t) = 0.00001 + (L_{init} - 0.00001)\left(1 - \frac{t}{t_{\max}}\right) \qquad \text{(eq. 4)}$$

and $N_{winner,j}(t)$ is a neighborhood kernel (see Eq. 5). Only the link weights of the neurons in the neighborhood around the winner neuron are updated. A 1-neighborhood is defined as all 8 neurons around the winner neuron, if they exist. A (n+1)-neighborhood contains all neurons of a n-neighborhood and their 1-neighbors, if they exist. Thus a neighborhood kernel $N_{winner,j}(t)$ is defined as

$$N_{winner,j}(t) = \begin{cases} 1 & if\ b_j \in r(t) - neighborho\ od \\ 0 & if\ b_j \notin r(t) - neighborho\ od \end{cases} \qquad \text{(eq. 5)}$$

with neighborhood radius of b_{winner}. The additional step dependent function $r(t)$ is introduced to get a constantly decreasing neighborhood radius (see Eq. 6).

$$r(t) = 1.0 + (r_{init} - 1.0)\left(1 - \frac{t}{t_{\max}}\right) \qquad \text{(eq. 6)}$$

For the babbling training an initial neighborhood radius $r_{init} = 12$ and an initial learning rate $L_{init} = 0.8$ are chosen.

Proto-vocalic and proto-syllabic test sets were defined for testing the proto-vocalic and proto-syllabic training results. The proto-vocalic test set comprises 270 proto-vocalic states which cover the language independent articulatory vowel space between the cardinal vowel qualities [i], [a], and [u]. This proto-vocalic test set is organized in the same way as the proto-vocalic training set but the test set exhibits a much lower density within the articulatory or auditory vowel space. This also results in different training and test items. Both proto-syllabic test sets are based on a set of 22 quasi-vocalic motor plan states covering the whole language independent articulatory vowel space. Both proto-syllabic test sets are organized in the same way as the proto-syllabic training sets but the test sets exhibit a lower density within the articulatory or auditory vowel space for the proto-vocalic starting or ending positions of the VC- or CV-proto-syllables. Both proto-syllabic test sets comprise 198 items. The test items were different from the training items defined above.

An estimation of the quality of the proto-vocalic and the proto-syllabic training results is done by calculating a mean error over all test set items for estimating an articulatory state of a test set item from its auditory state. The calculation of the error value for each test item comprises six steps: In a first step the motor plan state of a test item is applied to the motor execution module for calculating the appropriate articulatory patterns (i.e. the time course of articulatory parameters for a speech item) by using the feedforward part of the model. This calculated articulatory pattern is called *initial articulatory pattern*. In a second step the appropriate auditory state pattern is calculated by using the output of the three-dimensional articulatory-acoustic model for the initial articulatory pattern and by applying this output to the auditory feedback pathway of the model. In a third step the motor plan state is recalculated from the auditory state pattern calculated in the second step. Note that the trained self-organizing network is used for this step. This step leads to an *estimated motor plan state* which results from the sensorimotor knowledge stored within the self-organizing network, i.e. which results from the learning or training procedure. In a fourth step the *estimated articulatory pattern* is calculated for the estimated motor plan states by reusing the feedforward part of the model. In a fifth step the estimated and initial articulatory patterns are compared. An error value is calculated for each test item which is the difference between estimated and initial articulatory pattern. This difference is normalized with respect to the initial articulatory pattern. In a sixth step the mean error over all test set items is calculated for the trained network.

500.000 training steps are sufficient for predicting associated articulatory states from the auditory states of the test items with a precision below 2% error rate on the primary motor level in the case of the proto-vocalic training (using the proto-vocalic training set) and 280.000 training steps are sufficient for predicting the articulatory states from the auditory states with a precision below 5% error rate in the case of both proto-syllabic trainings (using both proto-syllabic training sets). Thus the complete babbling training requires less than five minutes on standard PC's.

The resulting link weight values for the neurons connecting the self-organizing phonetic maps with the motor plan and auditory map are graphically displayed for the proto-vocalic training in Fig. 6 and for the proto-CV-syllabic training in Fig. 7. It appears that motor plan states are organized with respect to phonetic categories. In the case of the vocalic phonetic submap vocalic states are ordered continuously with respect to the motor plan parameters high-low and front-back. Experimental evidence for this kind of ordering is given by Obleser et al. (2006). In the case of the syllabic submap three regions occur which represent the gesture-performing articulator (labial, apical, and dorsal), i.e. an ordering occurs with respect to the motor-plan parameter gesture-performing articulator. This neural behavior resulting from self-organization of vocalic and consonantal or syllabic states with respect to phonetic categories (high-low, front-back, gesture-performing articulator) can be labeled as *phonetotopy* in parallel to tonotopy for the cortical ordering of auditory states with respect to their fundamental frequency (Kandel 2000, p. 609) or in parallel to somatotopy for the ordering of somatosensory states with respect to their location on the body surface (Kandel 2000, p. 460f).

It should be kept in mind at this point that the general phonetic sensorimotor knowledge stored in these phonetic maps is knowledge of sensorimotor relations exclusively generated by the three-dimensional articulatory and acoustic vocal tract model. Thus it is important for the performance or quality of neurocomputational models of speech production and perception that these models comprise realistic articulatory and acoustic vocal tract models as front-end modules which are capable of generating high quality articulatory and acoustic signals, since the signals generated by the articulatory-acoustic model are the basis for the calculation of all sensory signals.

-- insert Figure 6 and 7 about here --

After babbling training the neurocomputational model is capable of reproducing (or imitating) the motor plan state (i.e. the articulation) of any pre-linguistic speech item – in our case of any proto-vowel, proto-CV-syllable and proto-VC-syllable (with C = proto-consonantal closing gestures) – from their acoustic (or auditory) state patterns. Thus the neurocomputational model is now ready for language-specific *imitation training*. For imitation training the training sets comprise *language-specific* speech items; in our case vocalic and syllabic speech items. Beside the adjustment of link weights of the mapping between the phonetic map and the sensory maps and of the mapping between the phonetic map and the motor plan map, which is mainly done during babbling training, now in addition the link weights of the mapping between the phonetic map and the phonemic map are adjusted. Language-specific imitation training results in (i) specifying *regions of typical phoneme realizations (phone regions)* within the phonetic map, i.e. in specifying regions of neurons within the phonetic map, which represent typical realizations of a phoneme or of a syllable phoneme chain (see Fig. 6 and Fig. 7) and in (ii) fine-tuning of the sensorimotor link weights already trained during babbling. This fine-tuning mainly occurs at the phone regions. Thus the knowledge which is gained during imitation is language dependent. In other words during this training stage the neurocomputational model mainly learns to link neurons which represent different phonemes or phonemic descriptions of syllables with the motor plan states and with the sensory states of their appropriate typical realizations. In parallel to babbling training also imitation training can be subdivided into training procedures for vowels, CV- and for VC-syllables.

The *vowel imitation training set* comprises a set of 100 acoustic vowel realizations per phoneme for a typical five vowel phoneme system /i/, /e/, /a/, /o/, and /u/ (e.g. Bradlow 1995 and Cervera et al. 2001). A three-dimensional Gaussian distribution was chosen for each phoneme for distributing the 100 realizations per phoneme over the F1-F2-F3-space (Fig. 8 for the F1-F2-space). The distribution of the phoneme realizations in the acoustic vowel space (F1-F2-F3-space) is chosen as realistically as possible. The acoustic vowel realizations within the acoustic vowel space slightly overlap. These 500 vowel realizations are supposed to be realizations given by different external speakers, but matched with respect to the models babbling vowel space. It should be noted that vowel phonemes normally are learned in the context of words during speech acquisition. This is replaced in this model by training of isolated vowels by reason of simplicity. More complex training scenarios are beyond the scope of this paper.

-- insert Figure 8 about here --

During vowel imitation training each external acoustic (or auditory) vowel item is processed by the proto-vocalic babbling network in order to estimate the appropriate motor plan parameters. Thus the model is capable of re-articulating (imitating) these externally produced vowels and the model is capable of generating the appropriate internal auditory feedback states. In natural human speech acquisition scenarios the imitated vowel item is then judged as right or wrong (i.e. is accepted or not accepted) by an external listener; i.e. the produced item is awarded or not by the external listener, e.g. by communication between carer and toddler. If the item is accepted as a proper realization of the intended phoneme, its motor and sensory states can be linked to the neuron representing this phoneme in the phonemic map. In the case of our model all realizations (re-articulations or imitations) can be accepted and thus can be added to the imitation training data set since the acoustic realizations of all re-articulations (or imitations) occur within the phoneme realization clouds (Fig. 8). Thus the vocalic imitation training set comprises 500 items of appropriate phonemic, motor plan, and sensory states. These data are the basis for imitation training.

The *syllable CV and VC imitation training sets* are based on a set of a labial, apical, and dorsal closing and opening gesture ending or starting at 31 different vowel realizations per vowel phoneme. That leads to 31 acoustic realizations for each of the phonemic CV- or VC- syllables (i.e. /bi/, /di/, /gi/, /be/, /de/, /ge/, /ba/, /da, /ga/, /bo/, /do/, /go/, /bu/, /du/, and /gu/) and results in 465 training items. Each of these externally produced acoustic items are imitated in the same way as described above for the vowel items. Thus 465 training items of appropriate phonemic, motor plan, and sensory states for CV- or VC-stimuli are generated.

Only 5.000 training steps for vowels and only 5.000 training steps for CV- and VC-syllables had to be added to the proto-vocalic and proto-syllabic babbling training for obtaining clear *phoneme realization regions* (*phone regions*) within the phonetic maps (see the outlined neuron boxes in Fig. 6 and Fig. 7). A neuron of the phonetic map is defined to be a part of a phone region if the phonemic link weight value for this neuron of the phonetic map and the appropriate neuron of the phonemic map is above the level of 0.95. Thus for the neurons which form a phone region within the phonetic map, strong excitatory connections exist towards the neuron representing the appropriate phoneme within the phonemic map.

For imitation training the imitation training sets are used in addition to the ongoing applied babbling training set. The network is not reset; the already trained babbling network is used as a basis for further training. The algorithms for adjusting the network link weight are

identical for babbling and for imitation training. Thus the succession from babbling to imitation training needs not to be abrupt. Imitation training can start in parallel to babbling training if some sensorimotor knowledge, i.e. if some knowledge how to estimate motor plan states from auditory states, is already available from early babbling training. The complete imitation training requires less than one minute on a standard PC for the training sets used here.

4 Producing and perceiving Vowels and CV-Syllables

It should be emphasized that babbling and imitation training is not only the basis for learning to *produce* speech items of a target language. Since the sensory states of all self-productions are perceived by the feedback loop during babbling training and since external acoustic speech items as well as self-productions are perceived during imitation training it can be hypothesized that babbling and imitation training are also important for learning to *perceive* speech items of a target language.

The *production pathway* (phonemic map → phonetic map → motor plan map → primary motor map → articulation) has been introduced in section 2. The speech items which were trained in this study can be labeled as frequent syllables. The description of the processing of infrequent syllables is beyond the scope of this paper. Our training results given above indicate strong neural connections from a phonemic state within the phonemic map to a set of neurons within the phonetic map. Each of these sets of neurons within the phonetic map represent a region of phoneme realizations (phone regions) and thus represent *production variability* since neighboring neurons within the phonetic map represent slightly different motor and sensory states (for natural variation in vowel realizations see Perkell et al. 1993). If a phonemic speech item is activated (phonemic map) this leads to an activation of *several* neurons within the phonetic map (see the outlined boxes or phone regions for example for the vocalic phonetic map; Fig. 6). Thus in our model the maximal activated neuron within the phonetic map can differ from realization to realization. Therefore the motor plan and the subsequent articulatory realization of a phonemic item are allowed to vary within a perceptually acceptable region. These regions for phonemic items are the phoneme realization regions or phone regions and they are language-specific and are defined during imitation training (see Fig. 6 and Fig. 7).

Furthermore *coarticulation* is introduced in our neurocomputational model. Two sources of coarticulation are implemented in our model. Firstly, coarticulation results from the fact that the exact coordination of articulators for executing a speech gesture is controlled by

the motor execution module and that a speech gesture is not encoded in all details on the motor plan level. That leads to variability in gesture execution with respect to context. For example the realization of /b/ in /ibi/ or /aba/ is different in our model. In /aba/ the lower jaw is more involved in the execution of the labial closing gesture than in /ibi/ because of the wide mouth opening occurring in /a/ in comparison to /i/. Because of this wide mouth opening in /a/ it would be ineffective to execute the closing gesture in /aba/ just by using the lips. It is more effective to add a synergetic elevation of the lower jaw. Thus, the lower jaw elevation and the lower lip elevation form a labial closing gesture in a *synergetic* way. Secondly, coarticulation results from the fact that gesture specifications can vary even on the level of the motor plan. For example lip protrusion is allowed to vary for a consonantal labial closing gesture since lip protrusion is a *non relevant phonemic feature* in the case of a labial closing gesture in our target language. Since the labial closing gesture within a CV-syllable temporarily overlaps with the following vocalic gesture (e.g. for a gesture for realizing an /i/- or /u/) our simulations show anticipatory lip protrusion on the motor execution level in /pu/ while lips are not protruded during the labial closure in /pi/.

In the case of language-specific perception of speech items it can easily be shown that the neurocomputational model trained thus far for vowels and simple CV- and VC-syllables is capable of producing *categorical perception* for vowels and in an even stronger way for consonants (i.e. voiced plosives in the case of our model). The *auditory pathway* for perception of external speech items (auditory receptors → auditory map → phonetic map → phonemic map) has already been introduced in section 2 (auditory-to-motor pathway, see Hickok and Poeppel 2000 and 2004). Thus the phonetic map is not only a central neural representation in speech production but also in speech perception at least for sublexical speech units like speech sounds and syllables. In order to show that the current neuroncomputational production-perception model perceives vowels (for the five vowel system /i/, /e/, /a/, /o/, and /u/) and consonants (for the voiced plosives /b/, /d/, and /g/) in a speech-like categorical way, speech identification and discrimination experiments were carried out using the model. In order to be able to perform these experiments using the model, 20 different instances of the model were trained using (i) different sets of training data due to different randomization procedures for determining the vocalic items within all training sets, using (ii) a different ordering of training stimuli during each training stage, and using (iii) different sets of randomly generated initial link weight values for each of the 20 instances. The resulting 20 instances of the model are called *virtual listeners*.

Identification of an external acoustic stimulus is performed in our model by a virtual listener by identifying the most excited neuron within the phonemic map. Discrimination of two external acoustic stimuli is performed in our model by calculating the most activated neuron on the level of the phonetic map for each acoustic stimulus and subsequently by calculating the city block distance between these both neurons for each virtual listener. The phonetotopic ordering of speech items on the level of the phonetic map (see above) is a first hint that distance between speech items (states) on the level of this map indicates phonetic similarity or dissimilarity. Moreover we assume that the sensory resolution of two states (i.e. the capability for discrimination between these states) is governed by the spatial distance of these two states on the level of the phonetic map. This assumption holds for tonotopic ordering and thus for F0-discrimination of auditory stimuli (see the discussion of tonotopic cortical maps, Kandel 2000, p. 609) and this assumption also holds for somatotopic ordering and thus for the spatial discrimination of tactile stimuli (see the discussion of somatotopic maps, Kandel et al. 2000, p. 460ff). Consequently it can be hypothesized that two stimuli can be discriminated if the distance of the activated neurons representing the stimuli on the level of the phonetic map exceeds a certain neuron distance within this map and it can be hypothesized that discrimination becomes stronger with increasing neuron distance.

Vocalic and consonantal identification and discrimination tests were performed on the basis of *quasi-continuous acoustic stimulus continua* (for an introduction to speech perception experiments see e.g. Raphael et al. 2007). The stimulus continua generated for these tests model an /i/-/e/-/a/-continuum for vowels and a /ba/-/da/-/ga/-continuum for CV-syllables (Fig. 9 and Fig. 10). The resulting identification and discrimination scores are given in Fig. 11 and Fig. 12. It can be seen that the measured identification scores (measured for the 20 virtual listeners by identifying the most excited neuron within the phonemic map via the phonetic map for each stimulus) indicate more abrupt phoneme boundaries in the case of consonants than in the case of the vowels. Additionally it can be seen that the measured discrimination scores (measured for the same 20 virtual listeners by estimating the distance for both stimuli on the level of the phonetic map; see above) indicate higher discrimination scores at least for consonant perception. Beside *measured discrimination* (naturally perceived discrimination) also *calculated discrimination* scores are shown in Fig. 11 and Fig. 12. Calculated discrimination scores are theoretical constructs (see Liberman et al. 1957). They are calculated from (measured) identification scores for each single (virtual) listener. Thus calculated discrimination is a discrimination of stimuli which merely results from differences in identification of these stimuli. The probability p_{discr} for a certain percentage of calculated

discrimination of two stimuli a and b is based just on the identification probabilities p_{id} of these two stimuli for each phonemic category i = 1, 2, or 3 (with 1 = /b/, 2 = /d/, and 3 = /g/ in case of consonants and with 1 = /i/, 2 = /e/, and 3 = /a/ in the case of vowels, see Eq. 7 and Liberman et al. 1957, p. 363).

$$p_{discr} = 0.5 + 0.5 \cdot \sum_{i=1}^{3} (p_{id}(a,i) - p_{id}(b,i))^2 \qquad \text{(eq. 7)}$$

Consequently calculated discrimination just indicates that part of discrimination of stimuli which results from the ability of subjects to classify stimuli to different categories. Calculated discrimination or *discrimination based on identification* (Liberman et al. 1957, Eimas 1963) and its difference to (naturally) measured discrimination is discussed as an important feature of categorical perception (Damper and Harnad 2000). Calculated discrimination indicates discrimination which is just based on discrete linguistic or phonemic categorical knowledge, while measured discrimination scores indicate the *complete* discrimination of two stimuli based on all available auditory information given by these stimuli; not just the linguistic, phonemic, or categorical information, needed for (categorical) identification. It can be seen from Fig. 11 and Fig. 12 that measured discrimination rates are always higher than calculated discrimination rates. That is in agreement with identification and discrimination scores extracted from identification and discrimination experiments carried out with humans and can be interpreted in the way that acoustic speech stimuli always convey categorical (linguistic) and non-categorical (para-linguistic or non-linguistic extra) information. While measured and calculated discrimination scores are nearly identical in the case of consonants, it comes out from our modeling data that measured discrimination is better than calculated discrimination especially in the case of vowels. This is in agreement with result of natural speech perception (Fry et al. 1962, Eimas 1963) and reflects the typical differences in categorical perception of consonants and vowels.

-- insert Figure 9, 10, 11, and 12 about here --

5 Discussion and Conclusions

The experimental results presented in this paper indicate that a model of speech production and perception which is shaped with respect to *basic neurophysiological facts* is capable of embedding important features of speech production and speech perception in a straight

forward way even if the neurocomputational modeling is relatively basic as is here by using simple standard self-organizing networks. Typical and therefore important features of speech production and perception like production variability of phoneme realizations and categorical speech perception and especially the fact of different degrees of categorical perception for consonants and vowels, occur in a straightforward way in this production-perception model. Since human speech production and perception easily outperforms speech synthesis and speech recognition systems at least in difficult conditions, it could be useful to include *human-like* speech processing routines into such technical speech processing systems. This may help to increase the quality and the level of performance of technical speech processing systems.

Furthermore this modeling study indicates the close relationship of speech *production* and speech *perception*. Speech perception theories such as the motor theory of speech perception (Liberman et al. 1967, Liberman and Mattingly 1985) or the direct-realist theory (Fowler 1986) have already postulated this close relationship. And recent experimental results provide support for this claim and suggest that the development of an integrative model on speech production and perception is highly desirable. For example perceptual feedback loops (also called self-monitoring processes) are known to activate parts of the speech perception mechanism during overt (external perceptual loop) as well as covert speech production (internal perceptual loop, cf. Indefrey and Levelt 2004, Postma 2000, Hartsuiker and Kolk 2001). In addition imaging studies focusing on speech perception have demonstrated that perception is capable of activating parts of the speech production cortical networks (Fadiga et al. 2002, Wilson et al. 2004, Hickok and Poeppel 2004 and 2007).

Bidirectional mappings between phonemic and phonetic and between sensory and phonetic maps are introduced in our neural model in order to illustrate the close relationship between production and perception. The introduction of these bidirectional mappings is the basis for important features of the model like categorical perception. Physiologically a bidirectional mapping comprises two related unidirectional mappings since neurons always forward their firing pulses in one direction (Kandel et al. 2000). Thus physiologically bidirectional mappings are represented by two neural paths connecting the maps in both directions (see the separate arrows in Fig. 1). The phonetic map – which forms the central map for all bidirectional mappings in our model (see Fig. 1) can be interpreted as the central part of the mental syllabary (Levelt and Wheeldon 1994 and Levelt et al 1999). Neural cortico-cortical connections exist in both directions between this part of the frontal cortex and

the sensory areas as well as between this part of the frontal cortex and those temporal regions which process phonemic information (Kandel et al. 2000).

Other computer implemented models of speech production (Bailly 1997, Guenther 1994, 1995, 2006, and Guenther et al. 2006) as well as the model introduced here reflect the relationship between perception and production by incorporating *perceptual feedback control loops* or by incorporating production-perception pathways for self-monitoring processes (Indefrey and Levelt 2004). *Dual stream models of speech perception* have recently been published which introduce a ventral stream for passive auditory processing and a dorsal stream activating auditory-motor networks (e.g. Hickok and Poeppel 2004 and 2007) but passive models of speech perception that do not refer to production processes can also be found (McClelland and Elman 1986, Gaskell and Marslen-Wilson 1997, Luce et al. 2000, Norris et al. 2006). The model introduced here reflects the close relationship between speech production and speech perception since on the one hand our model comprises basic features of speech production models (cf. Guenther et al. 2006) and since on the other hand our model is capable of incorporating in addition the dual stream idea (Hickok and Poeppel 2007) in a straight forward way (see the labels "ventral stream" and "dorsal stream" in Fig. 1).

Mirror neurons (visual and audio-visual mirror neuron system) appear to be one of the neural systems that are involved in the association of production and perception processes (Rizzolatti and Arbib 1998, Studdert-Kennedy 2002, Kohler et al. 2002, Fadiga and Craighero 2004, Rizzolatti and Craighero 2004, Wilson et al. 2004, Iacoboni 2005, Wilson and Knoblich 2005, Arbib 2005). Systems of mirror neurons have been detected which code the abstract meaning of goal-directed actions (e.g. grasping) and which are capable of co-activating motor *and* sensory (visual and audio-visual) representations of these actions by neural cortico-cortical associations. These visual and audio-visual mirror neuron systems also co-activate abstract concepts (preferably for action words) and thus are capable of associating higher order linguistic representations with goal-directed actions. A *speech* mirror neuron system ("mirror resonant system" after Fadiga and Craighero 2004, p. 167, "auditory mirror neuron system" or "echo neurons" after Rizzolatti and Craighero 2004, p. 185f) is postulated which is newer from the viewpoint of evolution in comparison to the mirror neuron system introduced above and which is directly linked with the capacity of humans to learn speech items by imitation. It can be assumed that this *speech mirror neuron system* in parallel co-activates motor representations, sensory representations, and phonemic representations of speech items. Given that from a phonetic viewpoint speech items also are built up by goal-directed actions (called *speech gestures*) which build up the motor plans for speech items in our model (see

section 2), it can be hypothesized that a mirror neuron layer also exists for the association of motor, sensory, and phonemic representations of speech gestures (see also Westerman and Miranda 2004).

Self-organization is a central principle of learning and self-organizing maps are used for modeling cortical networks (Kohonen 2001). Within our neurocomputational model artificial self-organizing neural networks are implemented since self-organizing neural networks are biologically plausible and have been used successfully for modeling semantic lexical networks (Ritter und Kohonen 1989), for (i) modeling semantic and phonological aspects during early lexical development (Li et al. 2004), and for (ii) modeling the generation and recognition of goal-directed movements (Bullock et al. 1993, Tani et al. 2004). A further argument for using self-organizing maps is their success in modeling the mapping between phonemic and phonetic aspects of speech production as demonstrated by the learning experiments for vowels and syllables described in this study.

In our current model different submaps are used for different classes of speech items (V, CV, VC) and separate training procedures were introduced for training these classes of speech items. This separation of the phonetic map in submaps as well as the separation of training procedures for different speech items was done in order to simplify the modeling of the speech acquisition procedure for these three classes of speech items from the computational viewpoint. But in principle all types of speech items (i.e. all types of syllables and words or word components) can be trained simultaneously by introducing just one comprehensive learning task and by using *one single phonetic map*. Recent preliminary experiments indicate that a comprehensive single phonetic map shapes different subregions representing different classes of speech items. The ordering of speech items within these subregions is similar to the phonetotopic ordering presented in this paper for the different submaps discussed here.

It is unclear whether the training sets used here constitute a representative natural model of babbling and imitation training during early states of human speech acquisition. Our training sets comprise a widespread set of vocalic vocal tract positions and a widespread set of opening and closing movements. At least these sets comprise all vocal tract positions and all opening and closing movements which are physiologically possible. But it is conceivable that toddlers very quickly reduce their set of training items from all physiological possible positions and movements towards a subset of positions and movements which are especially important for speech.

It should be noted that our neural modeling approach does not include modeling of *temporal aspects* of neural functioning. Rather the temporal aspects of production and perception are included in the speech items and thus in the sensory, motor, phonetic, and phonemic states. In our production-perception model sensory and motor states of vowels and syllables are processed as a whole. Our modeling approach thus is sufficient as long as only a description of the training and processing of syllables is wanted. In contrast a detailed temporal organization becomes important if speech items comprise more than one syllable. In this case processing delays must be introduced for all pathways postulated in the model (cf. Guenther et al. 2006) and temporal aspects of neural activity need to be considered (cf. Maass and Schmitt 1999).

The two training stages identified by our modeling study distinguish between *babbling* (i.e. the build-up stage for sensorimotor representations of pre-linguistic proto-vocalic and proto-consonantal speech gestures) and *imitation* (i.e. the build-up stage for language-specific perceptual, motor, phonetic, and phonemic representations of speech items). A closer modeling of early stages of speech acquisition (Oller et al. 1999) is beyond the scope of this paper. Furthermore in reality the two training stages introduced here overlap in time. This is partially realized in our approach, since babbling and imitation training items are applied in parallel during the imitation training stage after a short babbling training stage.

The next important step would be to introduce processes for building up the *mental lexicon* and for modeling the process of word segmentation and identification (cf. Batchelder 2002, Werker and Yeung 2005, Jusczyk 1999, Brent 1999). The representation of the mental lexicon of the target language is very important for including top-down processes of speech perception and thus for speech recognition. However consideration of these processes currently goes beyond the scope of the current implementation of our model. But the model in generally is open for integrating a mental lexicon.

Last but not least it has to be stated that the neurocomputational production-perception model developed thus far by no means is an alternative solution for high-performance speech recognition or speech synthesis systems. At present the model described here is capable of producing and perceiving simple CV- and VC-syllables under ideal conditions. Concerning a further development of the model introduced here two different strategies are imaginable. On the one hand, this model can be further developed in order to handle more complex classes of speech items (words, sentences, or a whole discourse) under ideal and non ideal conditions (e.g. different speakers, different emotional states, external noise). On the other hand, the organization of the neurocomputational model outlined in this paper could be integrated at

least partially into the architecture of current or new speech recognition and speech synthesis systems.

Acknowledgments

This work was supported in part by the German Research Council Grant Nr KR 1439/13-1.

References

Ackermann H, Riecker A (2003) The contribution of the insula to motor aspects of speech production: a review and a hypothesis. Brain and Language 89: 320-328

Arbib MA (2005) From monkey-like action recognition to human language: an evolutionary framework for neurolinguists. Behavioral and Brain Sciences 28: 105-167

Bailly G (1997) Learning to speak: sensory-motor control of speech movements. Speech Communication 22: 251-267

Batchelder EO (2002) Bootstrapping the lexicon: a computational model of infant speech segmentation. Cognition 83: 167-206

Benson RR, Whalen DH, Richardson M, Swainson B, Clark VP, Lai S, Liberman AM (2001) Parametrically dissociating speech and nonspeech perception in the brain using fMRI. Brain and Language 78: 364-396

Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C, Rose R, Tyagi V, Wellekens C (2007) Automatic speech recognition and speech variability: A review. Speech Communication 49: 763-786

Binder JR, Frost JA, Hammeke, TA, Bellgowan PSF, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex 10: 512-528

Birkholz P, Jackèl D (2004) Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. Proceedings of the International Conference on Speech and Language Processing (Interspeech 2004, Jeju, Korea) pp. 1125-1128

Birkholz P, Kröger BJ (2006) Vocal tract model adaptation using magnetic resonance imaging. Proceedings of the 7th International Seminar on Speech Production (Belo Horizonte, Brazil) pp. 493-500

Birkholz P, Kröger BJ (2007) Simulation of vocal tract growth for articulatory speech synthesis. Proceedings of the 16th International Congress of Phonetic Sciences (Saarbrücken, Germany) pp. 377-380

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

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

Blank SC, Scott SK, Murphy K, Warburton E, Wise RJS (2002): Speech production: Wernike, Broca and beyond. Brain 125: 1829-1838

Boatman D (2004) Cortical bases of speech perception: evidence from functional lesion studies. Cognition 92: 47-65

Bookheimer SY, Zeffiro TA, Blaxton TA, Gaillard W, Theodore WH (2000) Activation of langauge cortex with audtomatoc speech tasks. Neurology 55: 1151-1157

Bradlow AR (1995) A comparative acoustic study of English and Spanish vowels. Journal of the Acoustical Society of America 97: 1916-1924

Brent MB (1999) Speech segmentation and word discovery: a computational perspective. Trends in Cognitive Sciences 3: 294-301

Browman C, Goldstein L (1989) Articulatory gestures as phonological units. Phonology 6: 201-251

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

Bullock D, Grossberg S, Guenther F (1993) A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. Journal of Cognitive Neuroscience 5: 408-435

Callan DE, Tsytsarev V, Hanakawa T, Callan AK, Katsuhara M, Fukuyama H, Turner R (2006) Song and speech: brain regions involved with perception and covert production. Neuroimage 31: 1327-1342

Cervera T, Miralles JL, Gonzales-Alvarez J (2001) Acoustical analysis of Spanish vowels produced by laryngectomized subjects. Journal of Speech, Language, and Hearing Research 44: 988-996

Clark RAJ, Richmond K, King S (2007) Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication 49: 317-330

Damper RI, Harnad SR (2000) Neural network models of categorical perception. Perception and Psychophysics 62: 843-867

Dell GS, Chang F, Griffin ZM (1999) Connectionist models of language production: lexical access and grammatical encoding. Cognitive Science 23: 517-541

Diehl R, Lotto AJ, Holt LL (2004) Speech perception. Annual Reviev of Psychology 55: 149-179

Eimas PD (1963) The relation between identification and discrimination along speech and non-speech continua. Language and Speech 6: 206-217

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

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

Fowler CA (1986) An event approach to the study of speech perception from a direct-realist perspective, Journal of Phonetics 14: 3-28

Fry DB, Abramson AS, Eimas PD, Liberman AM (1962) The identification and discrimination of synthetic vowels. Language and Speech 5: 171-189

Gaskell MG, Marslen-Wilson WD (1997) Integrating form and meaning: A distributed model of speech perception. Language and Cognitive Processes 12: 613-656

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

Goldstein L, Marianne Pouplier, Larissa Chen, Elliot Saltzman, Dani Byrd (2007) Dynamic action units slip in speech production errors. Cognition 103 : 386-412

Grossberg S (2003) Resonant neural dynamics of speech perception. Journal of Phonetics 31: 423-445

Guenther FH (1994) A neural network model of speech acquisition and motor equivalent speech production. Biological Cybernetics 72: 43-53

Guenther FH (1995) Speech sound acquisition, coarticulation, and rate effects in a neural model of speech production. Psychological Review 102: 594-621

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

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

Hartsuiker RJ, Kolk HHJ (2001) Error monitoring in speech production: a computational test of the perceptual loop theory. Cognitive Psychology 42: 113-157

Heim S, Opitz B, Müller K, Friederici AD (2003) Phonological processing during language production: fMRI evidence for a shared production-comprehension network. Cognitive Brain Research 16: 285-296

Hickok G, Poeppel D (2000) Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences 4: 131-138

Hickok G, Poeppel D (2004) Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92: 67-99

Hickok G, Poeppel D (2007) Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences 4: 131-138

Hillis AE, Work M, Barker PB, Jacobs MA, Breese EL, Maurer K (2004) Re-examining the brain regions crucial for orchestrating speech articulation. Brain 127: 1479-1487

Huang J, Carr TH, Cao Y (2001) Comparing cortical activations for silent and overt speech using event-related fMRI. Human Brain Mapping 15: 39-53

Iacoboni M (2005) Neural mechanisms of imitation. Current Opinion in Neurobiology 15: 632-637

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

Ito T, Gomi H, Honda M (2004) Dynamical simulation of speech cooperative articulation by muscle linkages. Biological Cybernetics 91: 275-282

Jardri R, Pins D, Bubrovszky M, Despretz P, Pruvo JP, Steinling M, Thomas P (2007) Self awareness and speech processing: an fMRI study. Neuroimage 35: 1645-1653

Jusczyk PW (1999) How infants begin to extract words from speech. Trends in Cognitive Sciences 3: 323-328

Kandel ER, Schwartz JH, Jessell TM (2000) Principles of Neural Science (McGraw-Hill, New York, 4th Edition)

Kemeny S, Ye FQ, Birn R, Braun AR (2005) Comparison of continuous overt speech fMRI using BOLD and arerial spin labeling. Human Brain Mapping 24: 173-183

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

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

Kröger BJ (1993) A gestural production model and its application to reduction in German. Phonetica 50: 213-233

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

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

Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2006c) Spatial-to-joint mapping in a neural model of speech production. DAGA-Proceedings of the 32th Annual Meeting of the German Acoustical Society (Braunschweig, Germany) pp. 561-562 (see http://www.speechtrainer.eu)

Kröger BJ, Lowit A, Schnitker R (2008) The organization of a neurocomputational control model for articulatory speech synthesis. In: Esposito A, Bourbakis N, Avouris N, Hatzilygeroudis I (eds.) Verbal and Nonverbal Features of Human-Human and Human-

Machine Interaction. Selected papers from COST Action 2102 International Workshop (Springer Verlag) pp. 121-135

Kuriki S, Mori T, Hirata Y (1999) Motor planning center for speech articulation in the normal human brain. Neuroreport 10: 765-769

Latorre J, Iwano K, Furui S (2006) New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer. Speech Communication 48: 1227-1242

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

Li P, Farkas I, MacWhinney B (2004) Early lexical development in a self-organizing neural network. Neural Networks 17: 1345-1362

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychological Review 74: 431-461

Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54: 358-368

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21: 1-36

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA (2005) Neural substrates of phonemic perception. Cerebral Cortex 15: 1621-1631

Luce PA, Goldinger SD, Auer ET, Vitevitch MS (2000) Phonetic priming, neighborhood activation, and PARSYN. Perception and Psychophysics 62: 615-625

Maass W, Schmitt M (1999) On the complexity of learning for spiking neurons with temporal coding. Information and Computation 153: 26-46

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cognitive Psychology 18: 1-86

Moore RK (2007) Spoken language processing: piecing together the puzzle. Speech Communication 49: 418-435

Murphy K, Corfield DR, Guz A, Fink GR, Wise RJS, Harrison J, Adams L (1997) Cerebral areas associated with motor control of speech in humans. Journal of Applied Physiology 83: 1438-1447

Nasir SM, Ostry DJ (2006) Somatosensory precision in speech production. Current Biology 16: 1918-1923

Norris D, Cutler A, McQueen JM, Butterfield S (2006) Phonological and conceptual activation in speech comprehension. Cognitive Psychology 53: 146-193

Obleser J, Boecker H, Drzezga A, Haslinger B, Hennenlotter A, Roettinger M, Eulitz C, Rauschecker JP (2006) Vowel sound extraction in anterior superior temporal cortex. Human Brain Mapping 27: 562-571

Obleser J, Wise RJS, Dresner MA, Scott SK (2007) Functional integration across brain regions improves speech perception under adverse listening conditions. Journal of Neuroscience 27: 2283-2289

Okada K, Hickok G (2006) Left posterior auditory-related cortices participate both in speech perception and speech production: neural overlap revealed by fMRI. Brain and Language 98: 112-117

Oller DK, Eilers RE, Neal AR, Schwartz HK (1999) Precursors to speech in infancy: the prediction of speech and language disorders. Journal of Communication Disorders 32: 223-245

Perkell JS, Matthies ML, Svirsky MA, Jordan MI (1993) Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: a pilot "motor equivalence" study. Journal of the Acoustical Society of America 93: 2948-2961

Poeppel D, Guillemin A, Thompson J, Fritz J, Bavelier D, Braun AR (2004) Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex. Neurophysiologia 42: 183-200

Postma A (2000) Detection of errors during speech production: a review of speech monitoring models. Cognition 77: 97-131

Raphael LJ, Borden GJ, Harris KS (2007) Speech Science Primer: Physiology, Acoustics, and Perception of Speech. Lippincott Williams & Wilkins, 5th edition.

Riecker A, Kassubek J, Gröschel K, Grodd W, Ackermann H (2006) The cerebral control of speech tempo: opposite relationship between speaking rate and BOLD signal change at stratal and cerebellar structures. Neuroimage 29: 46-53

Rimol LM, Specht K, Weis S, Savoy R, Hugdahl K (2005) Processing of sub-syllabic speech units in the posterior temporal lobe: an fMRI study. Neuroimage 26: 1059-1067

Ritter H, Kohonen T (1989) Self-organizing semantic maps. Biological Cybernetics 61: 241-254

Rizzolatti G, Arbib MA (1998) Language within our grasp. Trends in Neuroscience 21: 188-194.

Rizzolatti G, Craighero L (2004) The mirror neuron system. Annual Review of Neuroscience 27: 169-192

Rosen HJ, Ojemann JG, Ollinger JM, Petersen SE (2000) Comparison of brain activation during word retrieval done silently and aloud using fMRI. Brain and Cognition 42: 201-217

Saltzman E (1979) Levels of sensorimotor representation. Journal of Mathematical Psychology 20: 91-163

Saltzman E, Byrd D (2000) Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. Human Movement Science 19: 499-526

Saltzman E, Munhall KG (1989) A dynamic approach to gestural patterning in speech production. Ecological Psychology 1: 333-382

Sanguineti V, Laboissiere R, Payan Y (1997) A control model of human tongue movements in speech. Biological Cybernetics 77: 11-22

Scharenborg O (2007) Reaching over the gap: A review of efforts to link human and automatic speech recognition research. Speech Communication 49: 336-347

Scott SK, Blank CC, Rosen S, Wise RJS (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123: 2400-2406

Shadmehr R, Mussa-Ivaldi A (1994) Adaptive representation of dynamics during learning of a motor task. Journal of Neuroscience 14: 3208-3224

Shuster LI, Lemieux SK (2005) An fMRI investigation of covertly and overtly produced mono- and multisyllabic words. Brain and Language 93: 20-31

Sober SJ, Sabes PN (2003) Multisensory integration during motor planning. Journal of Neuroscience 23: 6982-6992

Sörös R, Guttman Sakoloff L, Bose A, McIntosh AR, Graham SJ, Stuss DT (2006) Clustered functional MRI of overt speech production. Neuroimage 32: 376-387

Studdert-Kennedy M (2002) Mirror neurons, vocal imitation, and the evolution of particulate speech. In: Stamenov MI, Gallese V (eds.) Mirror Neurons and the Evolution of Brain and Language (Benjamins, Philadelphia) pp. 207-227

Tani J, Masato I, Sugita Y (2004) Self-organization of distributed represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB. Neural Networks 17: 1273-1289

Todorov E (2004) Optimality principles in sensorimotor control. Nature Neuroscience 7: 907-915

Tremblay S, Shiller DM, Ostry DJ (2003) Somatosensory basis of speech production. Nature 423: 866-869

Ullman MT (2001) A neurocognitive perspective on language: the declarative/procedural model. Nature Reviews Neuroscience 2: 717-726

Uppenkamp S, Johnsrude IS, Norris D, Marslen-Wilson W, Patterson RD (2006) Locating the initial stages of speech-sound processing in human temporal cortex. Neuroimage 31: 1284-1296

Vanlancker-Sidtis D, McIntosh AR, Grafton S (2003) PET activation studies comparing two speech tasks widely used in surgical mapping. Brain and Language 85: 245-261

Varley R, Whiteside S (2001) What is the underlying impairment in acquired apraxia of speech. Aphasiology 15: 39-49

Werker JF, Yeung HH (2005) Infant speech perception bootstraps word learning. Trends in Cognitive Sciences 9: 519-527

Westerman G, Miranda ER (2004) A new model of sensorimotor coupling in the development of speech. Brain and Language 89: 393-400

Wilson M, Knoblich G (2005) The case for motor involvement in perceiving conspecifics. Psychological Bulletin 131: 460-473

Wilson SM, Saygin AP, Sereno MI, Iacoboni M (2004) Listening to speech activates motor areas involved in speech production. Nature Neurosciences 7: 701-702

Wise RJS, Greene J, Büchel C, Scott SK (1999) Brain regions involved in articulation. The Lancet 353: 1057-1061

Zekveld A, Heslenfeld DJ, Festen JM, Schoonhoven R (2006) Top-down and bottom-up processes in speech comprehension. Neuroimage 32: 1826-1836

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

Figures

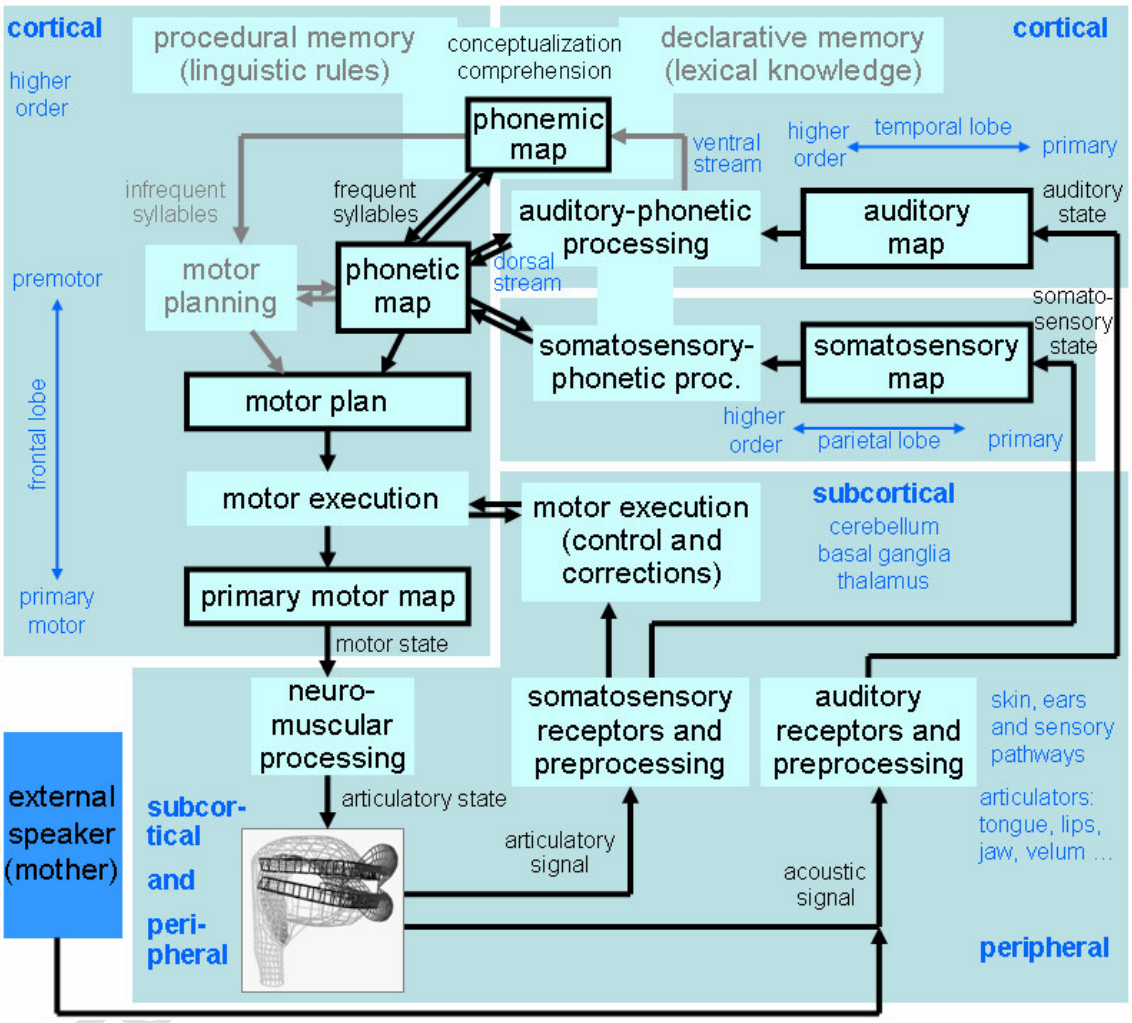


Figure 1: Organization of the neurocomputational model. Boxes with black outline represent neural maps. Arrows indicate processing paths or neural mappings. Boxes without outline indicate processing modules. Grey letters and grey arrows indicate processing modules and neural mappings which are not computer implemented in the current version of the model.

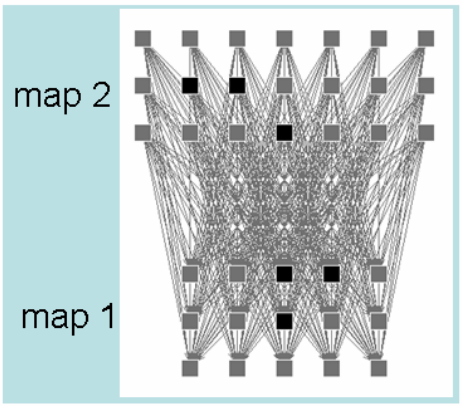


Figure 2: One-layer feedforward network connecting two neural maps 1 and 2. Grey lines indicate the neural connections connecting each neuron of map 1 with each neuron of map 2.

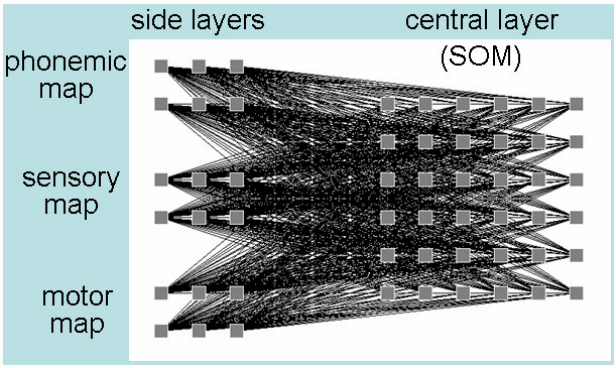
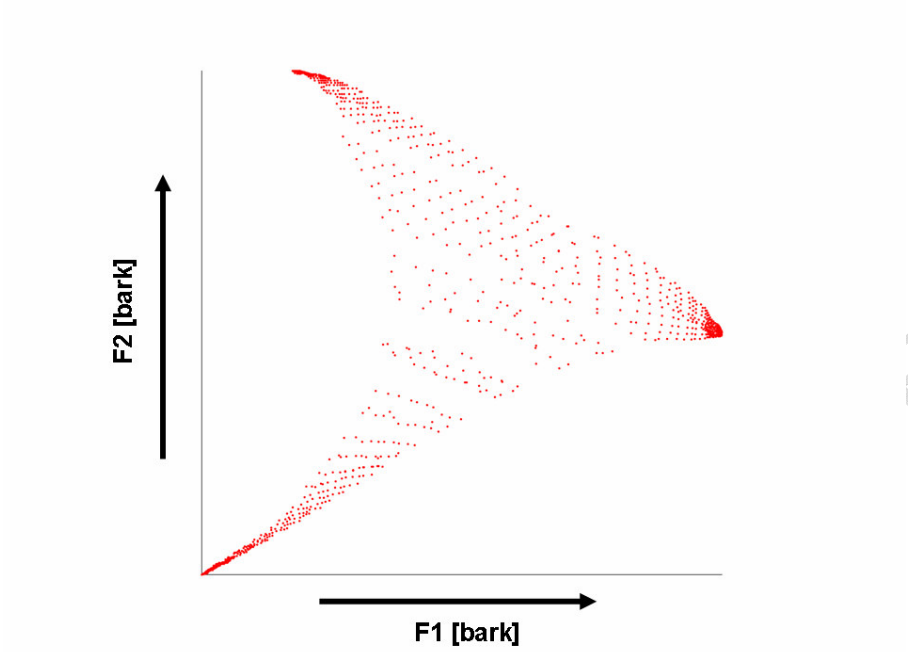


Figure 3: Self-organizing network connecting three neural maps (side layer maps) by a central self-organizing map (SOM or central layer map). Black lines indicate the neural connections, connecting each neuron of each side layer map with each neuron of the central map.

(a)



(b)

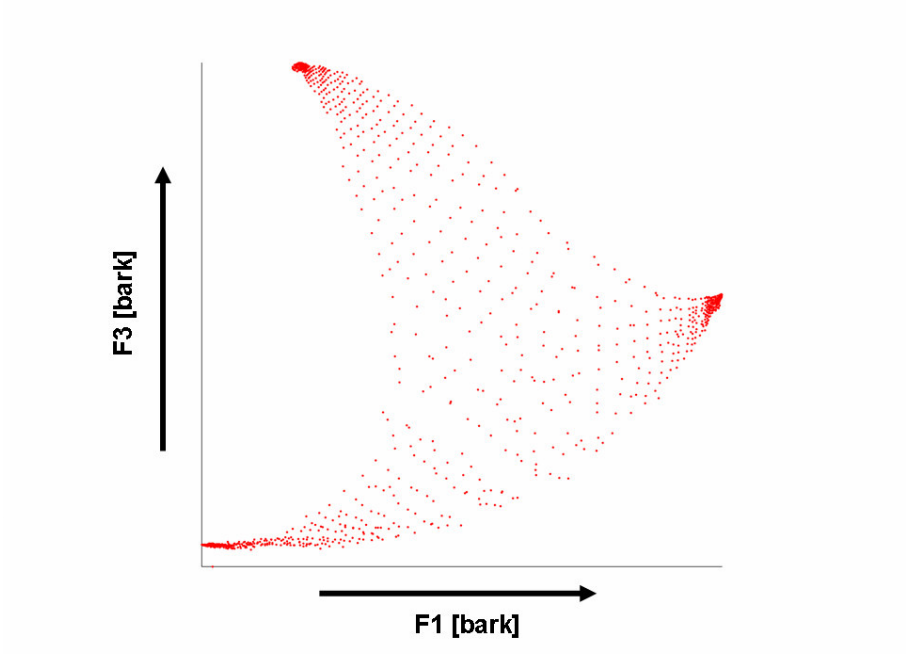


Figure 4: Position of all auditory patterns of the proto-vocalic training stimuli (grey points) in the normalized and bark-scaled (a) F1-F2 and (b) F1-F3 vowel space.

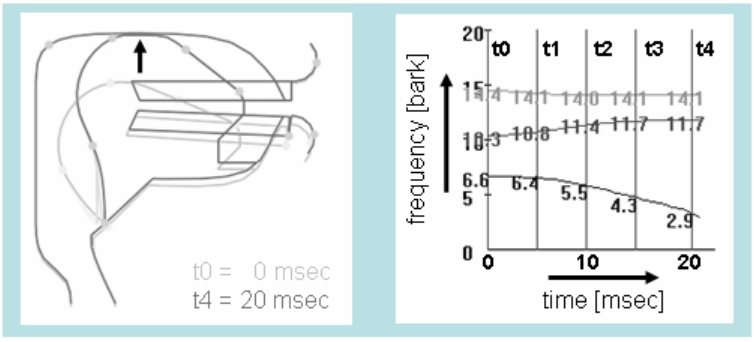


Figure 5: Auditory state (right side) for a dorsal closing gesture (left side).

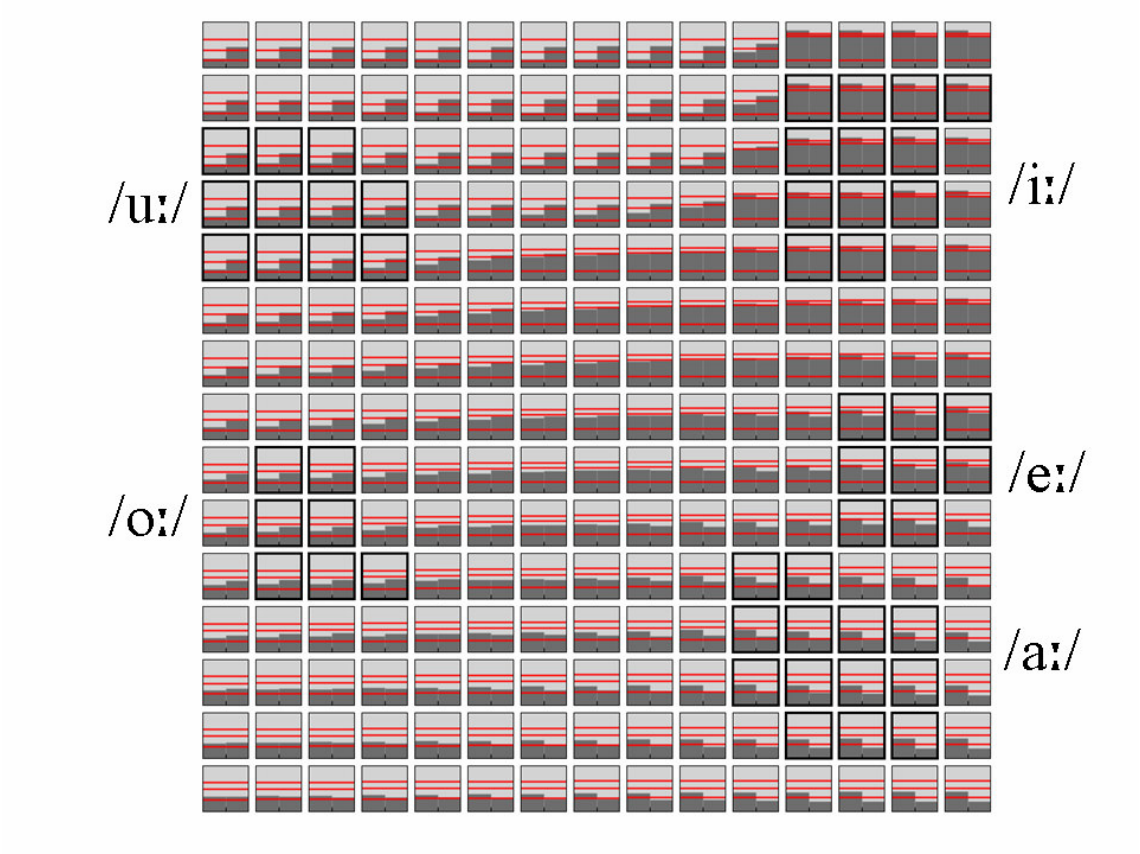


Figure 6: Motor plan and auditory link weight values after vocalic babbling and imitation training for each neuron within the vocalic phonetic map (15x15 neurons). Link weight values are given for two motor plan parameters within each neuron box: back-front (left bar) and low-high (right bar). Link weight values are given for three auditory parameters: bark scaled F1, F2, and F3 (horizontal lines within each neuron box). The outlined boxes indicate the association of neurons with vowel phoneme categories. These associations are established during imitation training (see text).

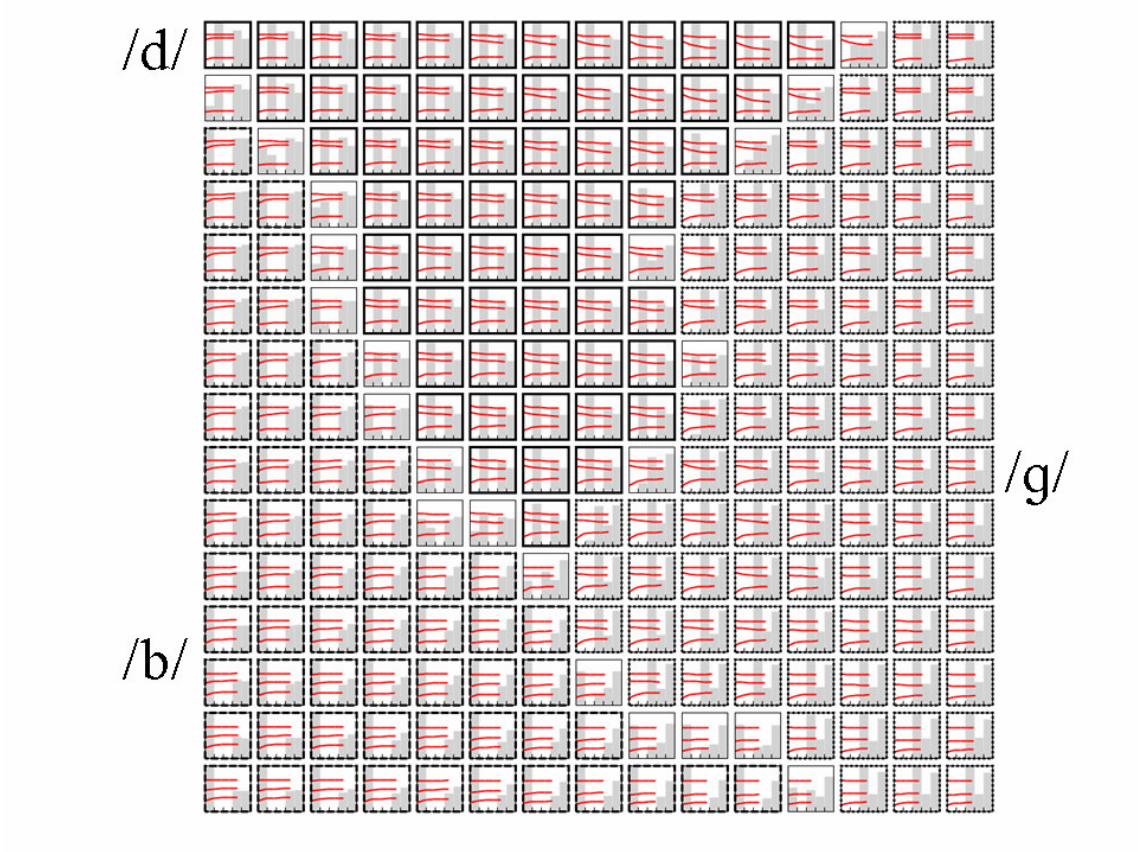
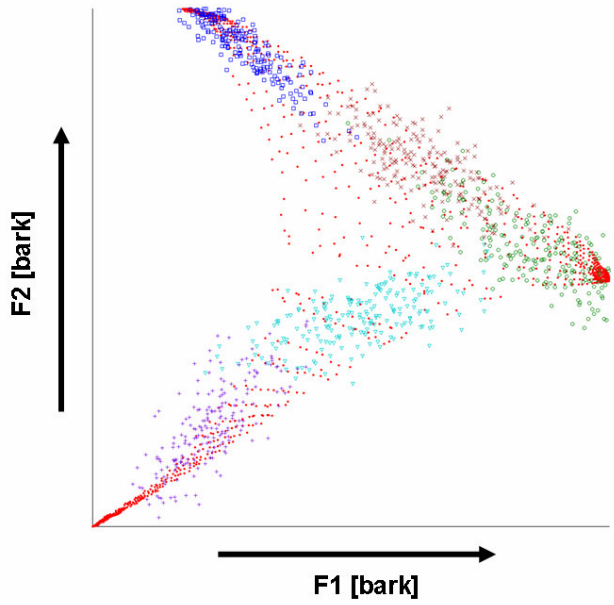


Figure 7: Motor plan and auditory link weight values after CV-syllabic babbling and imitation training for each neuron within the CV-phonetic map (15x15 neurons). Link weight values are given for five motor plan parameters within each neuron box. First three columns: vocal tract organ which performs the closing gesture (labial, apical, dorsal); two last columns: back-front value (forth column) and low-high value (fifth column) of the vowel within the CV-sequence. Link weight values are given for three auditory parameters: bark scaled F1, F2, and F3 (formant transitions within each neuron box). The outlined boxes indicate the association of neurons with consonant phoneme categories /b/, /d/, and /g/; each of these three regions comprises the appropriate consonant in all vocalic contexts. These associations are established during imitation training (see text).

(a)



(b)

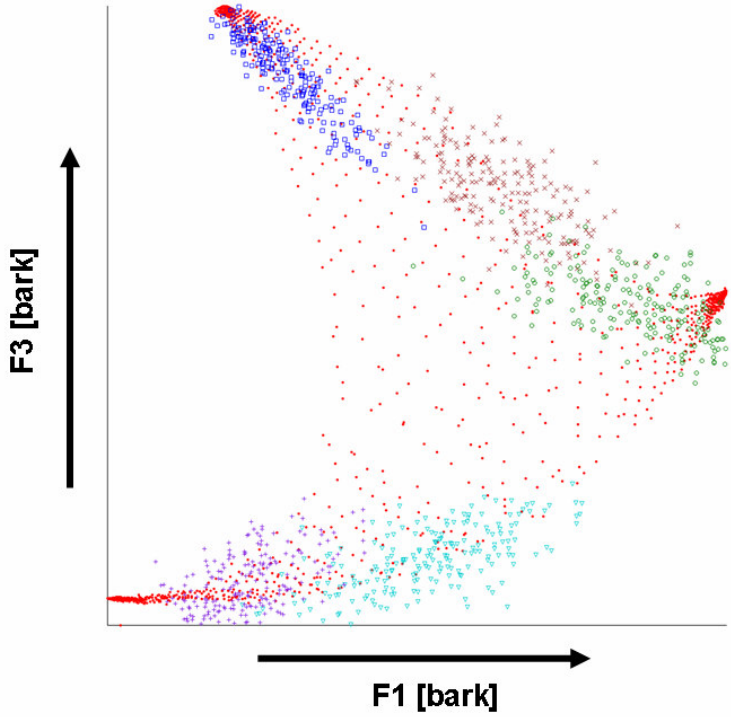


Figure 8: Positions of all auditory patterns of the language-specific vocalic training stimuli (phone clouds: 100 realizations per phoneme /i/ (square), /e/ (cross), /a/ (circle), /o/ (triangle), and /u/ (plus)) in the normalized and bark-scaled (a) F1-F2 and (b) F1-F3 vowel space. The patterns (or phone clouds) are added to the proto-vocalic training stimuli (points).

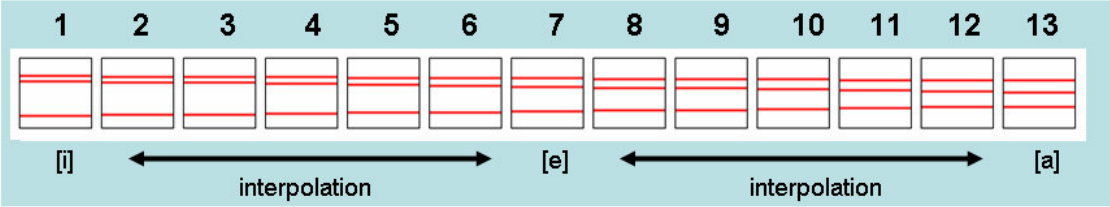


Figure 9: Bark-scaled formant pattern for 13 vocalic stimuli (/i/-/e/-/a/-continuum) for the vocalic perceptual identification and discrimination tests.

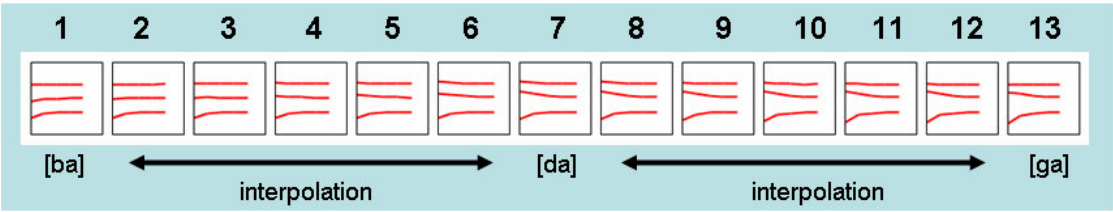


Figure 10: Bark-scaled formant pattern for 13 CV-stimuli (/ba/-/da/-/ga/-continuum) for the consonantal perceptual identification and discrimination tests.

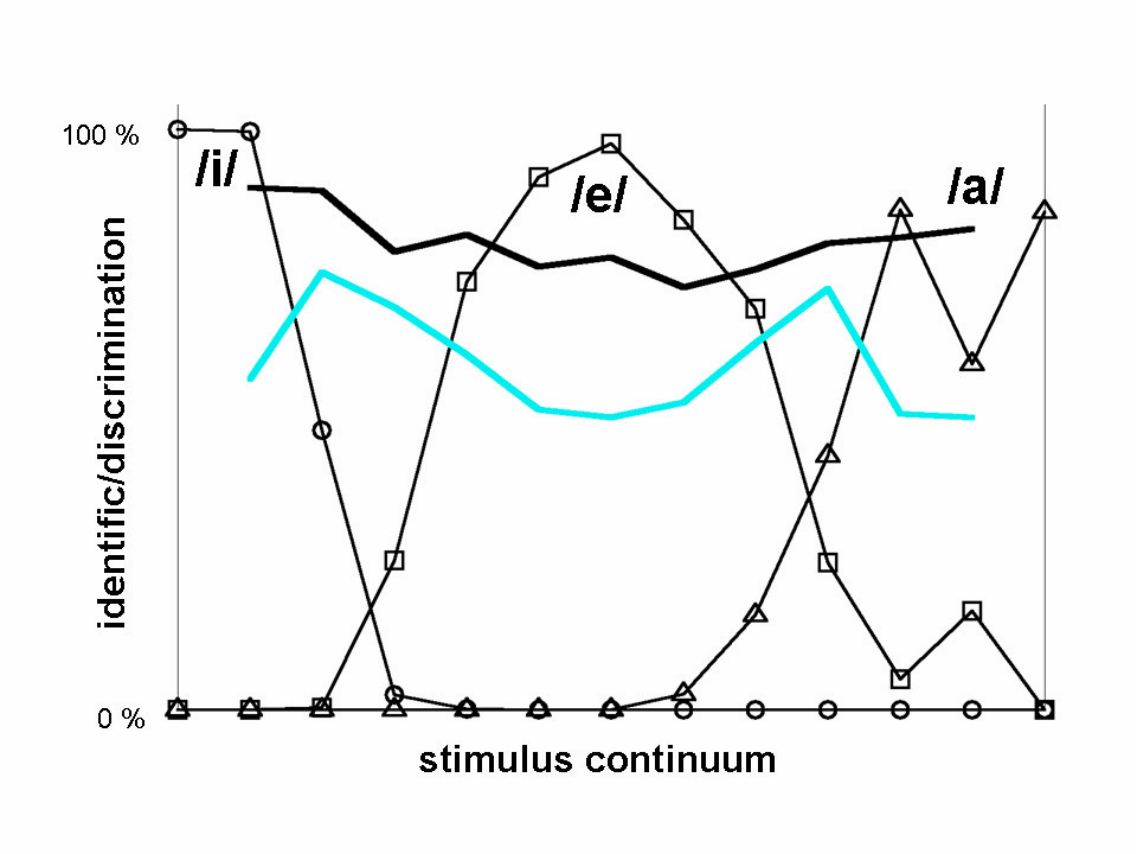


Figure 11: Measured identification scores (non-bold black lines) and measured (bold black line) and calculated (bold grey line) discrimination score for the vocalic /i/-/e/-/a/ stimulus continuum for 20 virtual instances of the model.

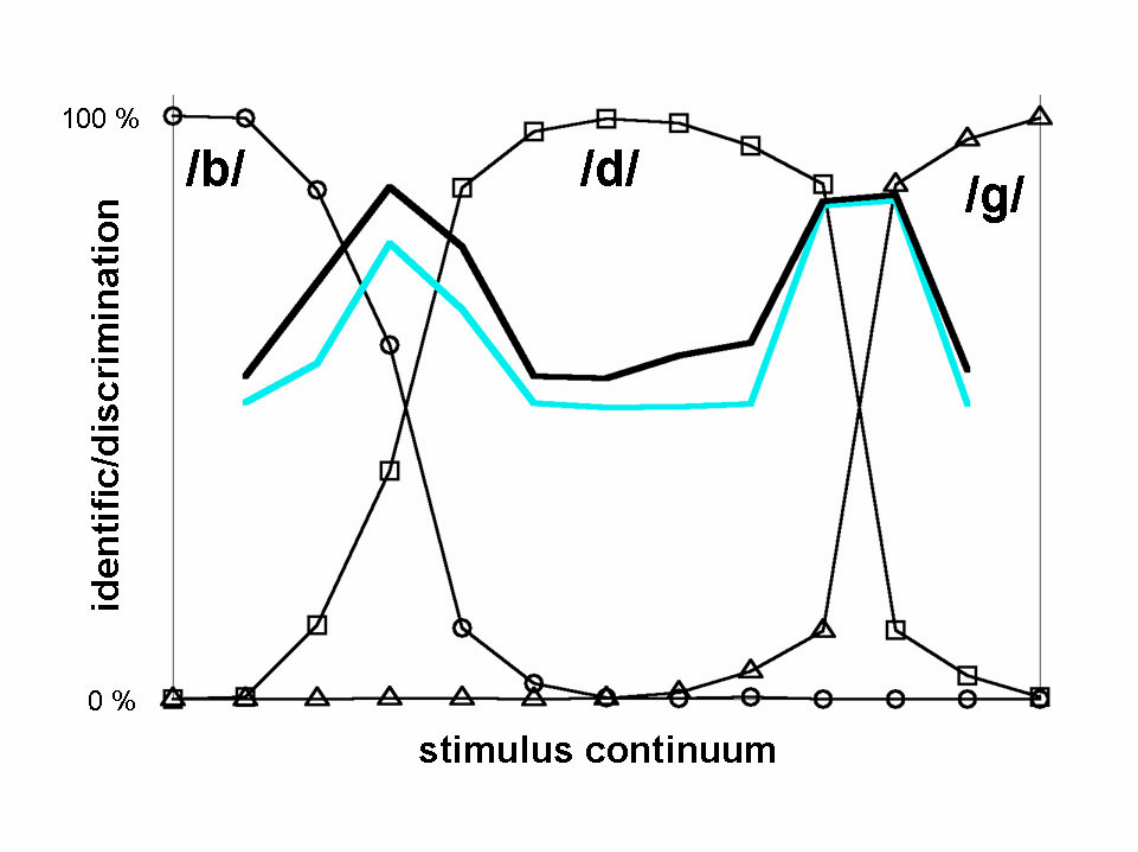


Figure 12: Measured identification scores (non-bold black lines) and measured (bold black line) and calculated (bold grey line) discrimination scores for the consonantal /ba/-/da/-/ga/ stimulus continuum for 20 virtual instances of the model.