# 3

# Phonetotopy within a neurocomputational model of speech production and speech acquisition

BERND J. KRÖGER
JIM KANNAMPUZHA
ANJA LOWIT
CHRISTIANE NEUSCHAEFER-RUBE

**Abstract:** A neurocomputational model of speech production is introduced in this paper. The model comprises neural maps and mappings, i.e. structure and knowledge. The structure is postulated on the basis of neurophysiological and neuropsychological facts while the knowledge is acquired during training or learning phases. The structure of the model and the training phases are described in detail in this paper. The training phases can be attributed to prelinguistic and early linguistic phases of speech acquisition. A phonetic neural map is postulated to be a central part of this model. After prelinguistic and early language-specific training phases this phonetic map exhibits an ordering of phonetic states with respect to phonetic features like the vocalic dimensions 'high-low' and 'front-back' or the consonantal place of articulation ('labial', 'apical', and 'dorsal'). This feature is labelled as *phonetotopy* in our approach.

## 1. INTRODUCTION

*Neurologically based computational models of speech production and speech acquisition* are rare. Only few approaches exist that focus on learning to produce speech and especially on learning to produce *articulatory speech movements* during speech acquisition (cf. Bailly, 1997; Guenther, 1994; 1995; 2006; and Guenther et al., 2006). These approaches complement more linguistically oriented neural- or cognitive-based computational models of speech production which focus mainly on semantic, syntactic, and lexical processes of generating the phonological description of a word or an utterance (e.g. Levelt et al., 1999; Dell et al., 1999). These latter approaches describe *linguistic processes* but not their *concrete phonetic or sensorimotor implementation*. The organization of the neurocomputational model introduced here is based on general neurophysiological and neuropsychological principles of movement control and speech production (Kröger et al., 2008). The organization of the model (i.e. its structure) and the knowledge incorporated in the model, which results from neural learning, is described in detail in this paper.

In parallel to *tonotopy*, i.e. the fact that cells within the primary auditory cortex which are sensitive to different pitches are ordered with respect to pitch (cf. Kandel et al., 2000, p. 609), and in parallel to *somatotopy*, i.e. the fact that cells within the primary somatosensory cortex which are sensitive to somatosensory signals from different parts of the body are ordered with respect to the anatomical location of these body parts (cf. Kandel et al., 2000, p. 460), here the concept of *phonetotopy* is introduced. Our learning results using this neurocomputational production model indicate that phonetic states like sensory or motor representations of vowels are ordered with respect to phonetic dimensions like low-high and front-back and they indicate that sensory and motor states of consonantal closing gestures are ordered with respect to the place of articulation.

The structure of the model and the computer-implementation of neural maps and mappings are described in section 2. Five training experiments for silent articulation, for proto-vocalic and vocalic, and for proto-consonantal and consonantal articulation are described in sections 3 to 7. The overall results are discussed in section 8.

## 2. THE STRUCTURE OF THE MODEL

The organization of our neurocomputational model (Figure 1) is based on general neurophysiological and neuropsychological knowledge (Kröger et al., 2008). A main feature of the model is its subdivision into feedforward and feedback control (cf. Guenther et al., 2006; and Guenther 2006).

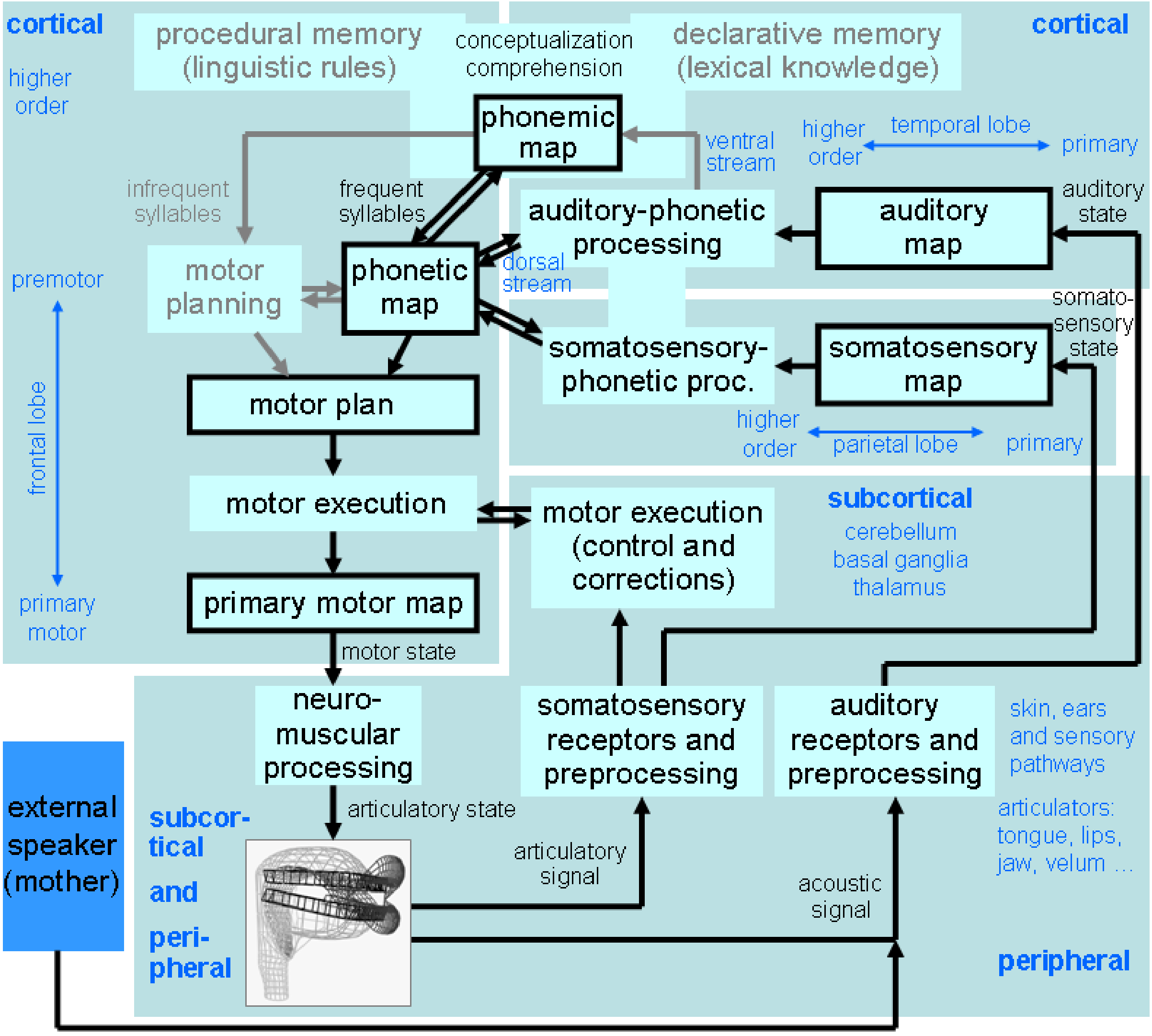

**Figure 1:** The neural model of speech production. Boxes with a black outline: neural maps; other boxes: neural processing units comprising neural maps and mappings not specified in detail. Arrows between neural maps indicate neural mappings or associations; arrows towards or from processing units indicate projections. The associations of the phonemic map, sensory maps, and motor plan map via the phonetic map are bidirectional, leading to a co-activation of phonemic, sensory, and motor states. In addition a bilateral connection occurs between the phonetic map and the motor planning module as well as between the cortical and the subcortical motor execution modules. Cortical, subcortical, and peripheral regions are separated. Additionally the cortical regions are separated with respect to the frontal, temporal, and parietal lobes. Modules given in grey letters and grey lines: conceptual modules, mappings and projections which are not yet been implemented.

On the *linguistic level* phonemic word forms are selected from the mental lexicon. They pass linguistic rule modules including the syllabification module (Levelt, 1992; Levelt et al., 1999; Indefrey & Levelt, 2004) and subsequently build up a chain of phonemic items (syllables) ready for sensorimotor production. This linguistic module can be subdivided into a procedural and a declarative linguistic module (Ullman, 2001) but is conceptual and not implemented in the current version of our model.

Sensorimotor *feedforward control* in our model starts with a chain of syllables specified on the level of the *phonemic map* (Kröger, Birkholz, Kannampuzha et al., 2007). If the syllable under production is a frequent syllable within the speaker's language – i.e. an already well-practised or "overlearned" syllable – the phonemic state leads to a co-activation of the appropriate auditory, somatosensory, and motor plan state via the *phonetic map*. The prelearned associations of motor and sensory states for frequent syllables or sounds are stored by the synaptic link weights of the phonemic-phonetic, phonetic-sensory and phonetic-motor mappings. Phonetic maps arise and are trained during speech acquisition for different types of sounds and different types of syllables. Sounds are ordered within these maps with respect to phonetic features like the vocalic features low-high and front-back or the consonantal feature place of articulation. This ordering is directly reflected by the neurons of the phonetic map (see our experimental results given below). The phonetic map is a self-organizing map (SOM; see Kohonen, 2001) representing the associations between the phonemic, motor, and sensory representations for all types of sounds and for all frequent syllables within the target language. Thus the phonetic map links the *phonemic map* with the *motor plan map* and with the *sensory maps (auditory and somatosensory map)*. From the viewpoint of self-organization, the phonetic map is a part of the mapping between phonemic, motor, and sensory maps. This SOM is not introduced in the approach of Guenther (2006) and Guenther et al., (2006), but in our opinion it is advantageous to introduce this neural map explicitly in a sensorimotor model of speech production since its neurons can be interpreted as *hyper-* or *supramodal state neurons* connecting the phonemic, motor and sensory states of a speech item. We hypothesize that this level is an explicit level of speech-relevant mirror neurons (cf. Fadiga et al., 2002; Fadiga and Craighero, 2004; Rizzolatti and Craighero, 2004). All maps and mappings described thus far form the mental syllabary as postulated by Levelt and Wheeldon (1994).

Infrequent syllables are not processed by the mental syllabary but by a separate motor planning module, generating the motor plan of a syllable on the basis of subsyllabic units (e.g. the sound chain; cf. Levelt and Wheeldon, 1994; Levelt et al., 1999; Varley and Whiteside, 2001). This motor planning module is linked with the phonetic map since it profits from the phonetic knowledge on the production of frequent syllables stored within the phonetic map. The motor planning unit for infrequent syllables is conceptual but not implemented in the current version of the model.

While feedforward control as described above is the main control mechanism within normal (adult) speech production and is implemented in our model for frequent syllables, online *feedback control* for supervising the ongoing flow of speech production (cf. Guenther et al., 2006) is also just conceptualised and not yet implemented in our model. However, the *feedback control* paths as given in Figure 1 are activated during the training procedures of speech acquisition in the current version of the model. Feedback control starts with the auditory and somatosensory processing of the articulatory and acoustic signals produced by the speaker's vocal tract (Figure 1). Lower level somatosensory signals (i.e. proprioceptive articulator-related joint-coordinate parameters or articulatory parameters in Kröger et al., 2006b) are directly projected to and processed by the motor execution modules (Figure 1; cf. Shadmehr and Mussa-Ivaldi, 1994; Tremblay et al., 2003; Nasir and Ostry, 2006). Higher level somatosensory and auditory information (i.e. proprioceptive tract-variable related and tactile parameters, Kröger et al., 2006b) is projected to the auditory-phonetic and somatosensory-phonetic processing module via the auditory and somatosensory map and can be compared there with the stored sensory state of a speech item (speech sound, syllable or word). In the case of differences between stored and current perceptual states for a speech item, an error signal can be generated (cf. Guenther et al., 2006).

A further feature of this model is the separation of a higher and a lower level of motor representations, i.e. the separation of a *motor plan level* and a *primary motor level* (cf. the organization of movement control in action theory, Fadiga and Craighero, 2004). In contrast, Guenther et al., (2006) directly connect phonemic, sensory and motor representations without a separation of different levels of motor control. In our approach the motor plan level is introduced in parallel to the sensory and phonemic levels. Motor plans as well as phonemic and sensory representations of syllables are processed as a whole in our approach.

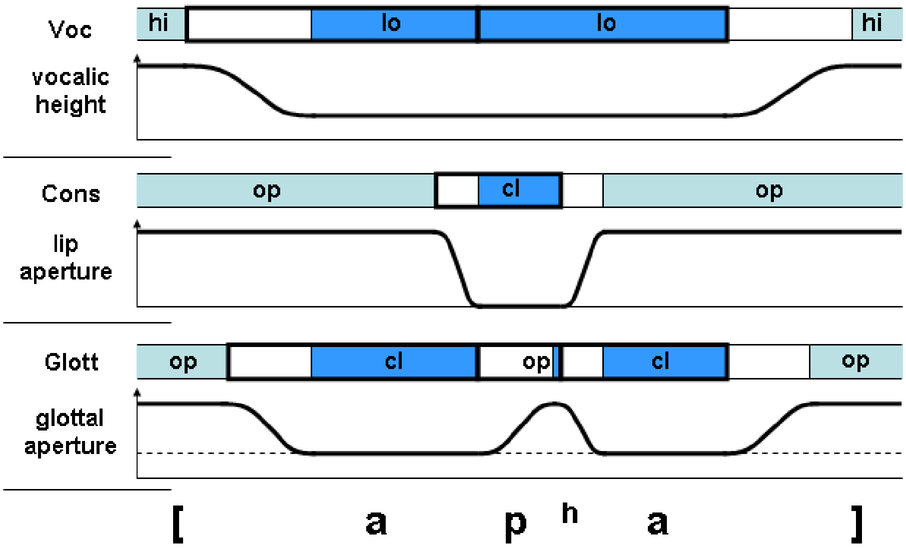

**Figure 2:** Temporal coordination of speech gestures for [ap$^h$a]. Time intervals of gestural activation are marked by bold-framed boxes: two vocalic [a]-forming gestures (lo for low) on the vocalic row, one labial closing gesture (cl) on the consonantal row, and two glottal closing (cl) and three glottal opening (op) gestures on the glottal row. The time interval of gestural activation can be subdivided into transition or movement portion (white) and target portion (dark grey). Movement transitions of vocal organs are given for tongue height, for lip aperture and for glottal aperture below the tier of activation intervals for each gesture. Low and high positions of the tongue are indicated by the naming of the appropriate gesture (lo and hi). In addition target time intervals of all "neutral gestures" are indicated by light grey boxes. Neutral gestures are always activated if no other specific gesture occurs on the vocalic, consonantal or glottal row. Neutral gestures represent the pre- or post-speech articulation, i.e. high tongue position, no consonantal closure and glottal opening for breathing. For simplification, velopharyngeal gestures are omitted in this figure; for a more complete description of our gestural concept, see Kröger and Birkholz (2007).

The *motor plan* of a syllable comprises a score of executable articulatory actions. These are goal-directed speech gestures describing higher level motor features such as "produce a vocal tract closure using lips for [b]", “produce a glottal opening for [p]”, or “produce a velopharyngeal opening for [m]” as well as their temporal coordination. The motor plan of the speech item [apa] is given in Figure 2.

One labial closing gesture and one glottal opening gesture for the production of [p], two coordinated gestures for tongue, lower jaw, and lips for producing the [a]-articulation, and two glottal closing gestures for producing phonation have to be coordinated in time on the motor plan level. Subsequently, motor execution leads to a concrete specification of each gesture on the level of the *primary motor map*. For example, a labial closing gesture involves coordinated movement of at least three articulators, i.e. the lower jaw and the lower and the upper lips, and each of these articulators is controlled by an ensemble of different motor units. Thus the concrete realization of a gesture is not specified on the higher motor plan level but spelled out during motor execution. Concepts for speech gesture coordination are suggested by different authors (cf. Ito et al., 2004; Sanguineti et al., 1997; Saltzman, 1979; Saltzman and Munhall, 1989; Saltzman and Byrd, 2000).

*One-layer feedforward networks* (Figure 3) are currently used for modelling the motor plan to primary motor mapping, and *self-organizing networks* are used in the case of the phonemic to sensory and phonetic to motor mapping (Figure 4), where the phonetic map represents the self-organizing map (SOM, i.e. the central map of the self-organizing network; cf. Kröger, Birkholz, Kannampuzha et al., 2007; Kröger, Birkholz, Neuschaefer-Rube, 2007).

Furthermore our model is capable of processing acoustic signals of external speakers (Figure 1). This information is processed via the same auditory signal processing pathway as is used for feedback control during training. An external auditory signal is also able to activate states of the phonetic map via the dorsal stream of the auditory pathway (cf. Hickok and Poeppel, 2007) and can co-activate pre-stored sound or syllable information of the auditory, phonetic, and phonemic parts of the mental syllabary (Figure 1). The ventral stream of the auditory pathway (cf. Hickok and Poeppel, 2007), which directly activates lexical items (e.g. words), is conceptualised but not implemented in the current version of the model.

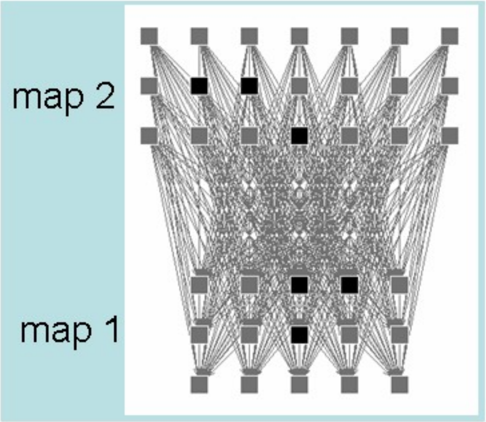

**Figure 3:** Example of a one-layer feedforward network (cf. Zell 2003). Grey lines indicate the heap of neural connections between two maps (motor plan and primary motor maps). The two heaps of squares indicate neuron collectives (i.e. neural maps). Black squares indicate activated neurons. The neural activation pattern of each neural map determines a distinct motor plan state (map 2) or primary motor state (map 1).

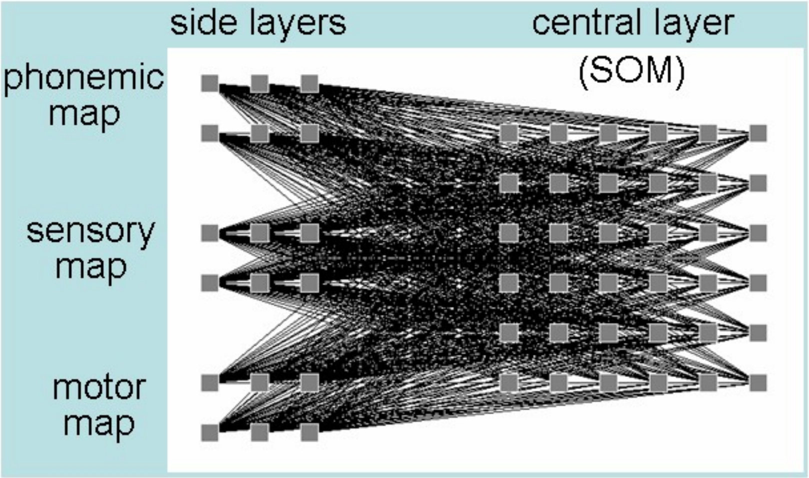

**Figure 4:** Example of a self-organizing network (cf. Kohonen, 2001). The heaps of grey squares indicate neuron collectives (i.e. neural maps). Black lines indicate the neural connections between side layers (i.e. neural maps for input and output representations) and the central layer (i.e. self-organizing map, SOM; the phonetic map in our model). The central layer (SOM) and all neural connections represent the self-organizing network.

The validity of the training results for our neurophonetic model, i.e. the quality of the neural mappings, strongly depends on the quality of the acoustic and articulatory signals generated by the included *vocal tract* *model*. These signals are the basis for calculating somatosensory and auditory feedback information and are needed for training the neural networks of the model during the speech acquisition phases. Our vocal tract model comprises a highly elaborate three-dimensional articulatory model (Birkholz, 2005; Birkholz et al., 2006; Birkholz and Kröger, 2006; 2007) and a highly elaborate articulatory-acoustic model (Birkholz and Jackel, 2004; Birkholz, 2005; Birkholz et al., 2007) capable of producing all speech-relevant vocal tract states (vocalic openings, critical closures for fricatives, complete closures, velopharyngeal coupling, etc.). The articulatory model generates vocal tract geometries for each time point on the basis of the lower level motor commands (articulatory commands; primary motor level). A vocal tract area function is calculated on the basis of the resulting geometrical information, which stipulates the acoustically relevant information of the vocal tract cavities (pharyngeal, mouth, and nasal cavity). The acoustic speech signal is then calculated using a multi-tube approximation of the vocal tract by applying the transmission line circuit approach (Birkholz and Jackel, 2004; Birkholz, 2005; Birkholz et al., 2007).

The gross or broad anatomical location of each map and mapping considered within this model can be assigned with respect to brain imaging literature (Figure 1; cf. Huang et al., 2001; Blank et al., 2002; Hillis et al., 2004; Guenther et al., 2006; Sörös et al., 2006). It is important that most of these maps, mappings, and modules are located *bilaterally* since they model the general phonetic sensorimotor processes of speech production. Lateralization occurs mainly for the *higher level language processing units* (Blank et al., 2002; Indefrey and Levelt, 2004; Liebenthal et al., 2005; Rimol et al., 2005).

In order to validate the functional aspects of the model, a number of experiments were performed to test whether it could simulate normal speech acquisition observed in infants. Central to this approach are the neural link weights of the mappings, i.e. the strength of the synaptic connection of neurons of two associated neural maps, which arise during the learning or training phase of the model (Guenther et al., 2006; Kröger et al., 2006b; Kröger, Birkholz & Neuschaefer-Rube, 2007).

Normal *speech acquisition* is divided in our model into three stages, i.e. (i) silent articulations, (ii) proto-vocalic and proto-consonantal articulation, and (iii) language-specific training of vowels and consonants (Kröger et al., 2006b). Whilst the first two phases are part of the *prelinguistic speech acquisition phase* (also called *babbling phase*), the third phase is part of *language-specific learning or training* (also called *imitation phase*; cf. Oller et al., 1999). The proto-vocalic training phase defined in our model is comparable to the phonation stage defined by Oller et al. (1999). The proto-consonantal training phase of the model can be associated mainly with the primitive articulation stage and in part with the expansion and the canonical stage defined by Oller et al. (1999). During babbling the toddler tries to collect "sensorimotor experience" from playing with his/her own speech apparatus and thus produces all kinds of possible speech (as well as non-speech) motor events and perceives the resulting sensory consequences. This contrasts with the language-specific stage, where the toddler learns to imitate external speech from carers (mother, father, etc.). Each stage or phase in the learning process depends at least in part on the previous phase, i.e. information acquired at an early stage is further modified and built on at the next stage.

All three model training phases defined above have been processed in our approach so far. Table 1 outlines how the training phases were attributed to mappings of our model trained during these phases. Both vowels and consonants were used in this training.

**Table 1:** The assignment of training phases defined within the neurophonetic model with the mappings trained during these phases. Mappings are indicated in Figure 1 by arrows.

| *Training phase* | *Mappings trained during each training phase* |
|---|---|
| silent articulations (prelinguistic babbling) | motor plan → primary motor |
| proto-vocalic or proto-consonantal articulation (prelinguistic babbling) | phonetic → motor plan → primary motor<br>somatosensory ←→ phonetic, auditory ←→ phonetic |
| language-specific training: vocalic and consonantal (development of the mental syllabary by imitation of external speech) | phonetic → motor plan → primary motor<br>somatosensory ←→ phonetic, auditory ←→ phonetic<br>phonemic ←→ phonetic |

The following four sections describe the experimental set-up and the results of the training phases for the acquisition and the later production of vowels and consonants within VC-syllables using the neurocomputational model.

## 3. EXPERIMENT I: SILENT ARTICULATION TRAINING

### 3.1. Method

The training phase for *silent articulation* is based on a training set consisting of 4608 items of lower level and related higher level motor states (i.e. primary motor and motor plan states). The 4608 items cover the whole range of static articulatory states which can be produced by the vocal tract model (cf. Kröger et al., 2006a; 2006c). Within this training phase the mapping of higher level to lower level motor parameters, i.e. motor plan to primary motor mapping (see Figure 1), is learned or trained. This mapping is also known as the spatial-to-joint coordinate mapping (cf. Saltzman and Munhall, 1989).

### 3.2. Results

Training this static part of the motor plan to primary motor mapping was successful using a one-layer feedforward network (Figure 3) and using standard training algorithms (Zell, 2003). Within a one-layer feedforward network all neurons representing the higher level spatial motor parameters are connected with all neurons representing the lower level joint motor parameters (Figure 3; a detailed discussion of neural representations for motor parameters is given in Kröger et al., submitted). Approximately 100,000 training steps were needed to obtain acceptable training results, i.e. to obtain a mean error for predicting a lower level motor state by higher level parameters below 10%. The prediction error was measured using a test set of 1152 test items which also cover the whole range of articulatory states but use combinations of parameter values different from those used in the training set. An example of training results is given in Figure 5 for the production of three types of vocal tract closures, a labial, an apical, and a dorsal closure. Vocal tract closure is controlled here by higher level motor parameters. The closures were stable, even with strong distortions on the level of the primary motor parameters, i.e. by fixing of the lower jaw in a very high or very low position. This result is a typical illustration of motor equivalence (for the concept of motor equivalence in speech, see Perkell et al., 1993; Guenther 1995).

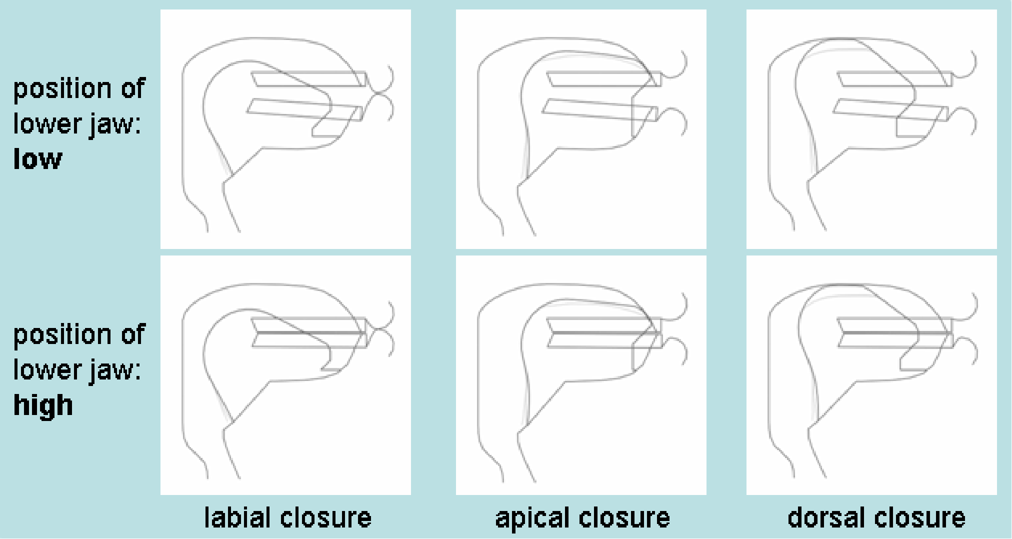

**Figure 5:** Mediosagittal contours generated by our three-dimensional vocal tract model for the production of a labial, an apical, and a dorsal closure defined by higher level motor parameters (the light grey lines indicate lateral articulator contours). Higher level motor parameters are constant for each column defining the labial, apical, or dorsal closure. The jaw parameter is fixed by lower level motor parameters as low or high (see both rows). It can be seen that the production of these three closures is stable despite the distortion introduced by different jaw positions.

# 4. EXPERIMENT II: PROTO-VOCALIC ARTICULATION TRAINING

## 4.1. Method

The training of *proto-vocalic articulation* is based on a training set comprising 1078 items, i.e. proto-vocalic higher level motor states, which cover the entire vowel space bordered by a high-front [i]-like, a high-back [u]-like and a low [a]-like articulation (Figure 6a, and see Kröger et al., 2006a and 2006b). For this training set the primary motor parameters are constrained with respect to an articulatory variation in two vocalic motor plan parameter dimensions: high-low and front-back. This set of training items is used for training the sensory to motor mapping via the phonetic map (Figure 1). The main goal of this training phase is to predict proto-vocalic higher level motor states from sensory states.

## 4.2. Results

Training was successful using a self-organizing neural network or self-organizing map (SOM) comprising 15x15 neurons and using standard training algorithms (Kohonen, 2001). The 15x15 neurons represent the central SOM within this mapping (see Figure 4) and this SOM represents the vocalic part of the phonetic map (Figure 1). Preliminary experiments indicated that the size of the SOM can be varied between 10x10 and 20x20 without changing the results substantially. In order to keep the computational time practicable (below one hour for a complete training phase), a 15x15-sized SOM was chosen. Approximately 500,000 training steps were necessary to obtain good training results, i.e. a mean error below 2% for predicting a proto-vocalic motor state from a sensory state. The prediction error was measured using a test set of 270 test items which also cover the same range of vocalic states as defined by the training set, which use different proto-vocalic parameter values with respect to the training items. The synaptic link weights for the neural connections between self-organizing phonetic map and auditory map are displayed by the intersection points (i.e. net nodes) of the grid network in Figure 6b. Each intersection point within the grid network (Figure 6b) represents one neuron of the SOM and its link weights with the auditory map. The grid network itself displays the neighbourhood relations of these neurons.

Firstly a comparison of Figure 6a and Figure 6b indicates that the grid network for proto-vowels (Figure 6b) not only unfolds in a regular way but also covers the whole (stimulus) range of the vowel space (Figure 6a).

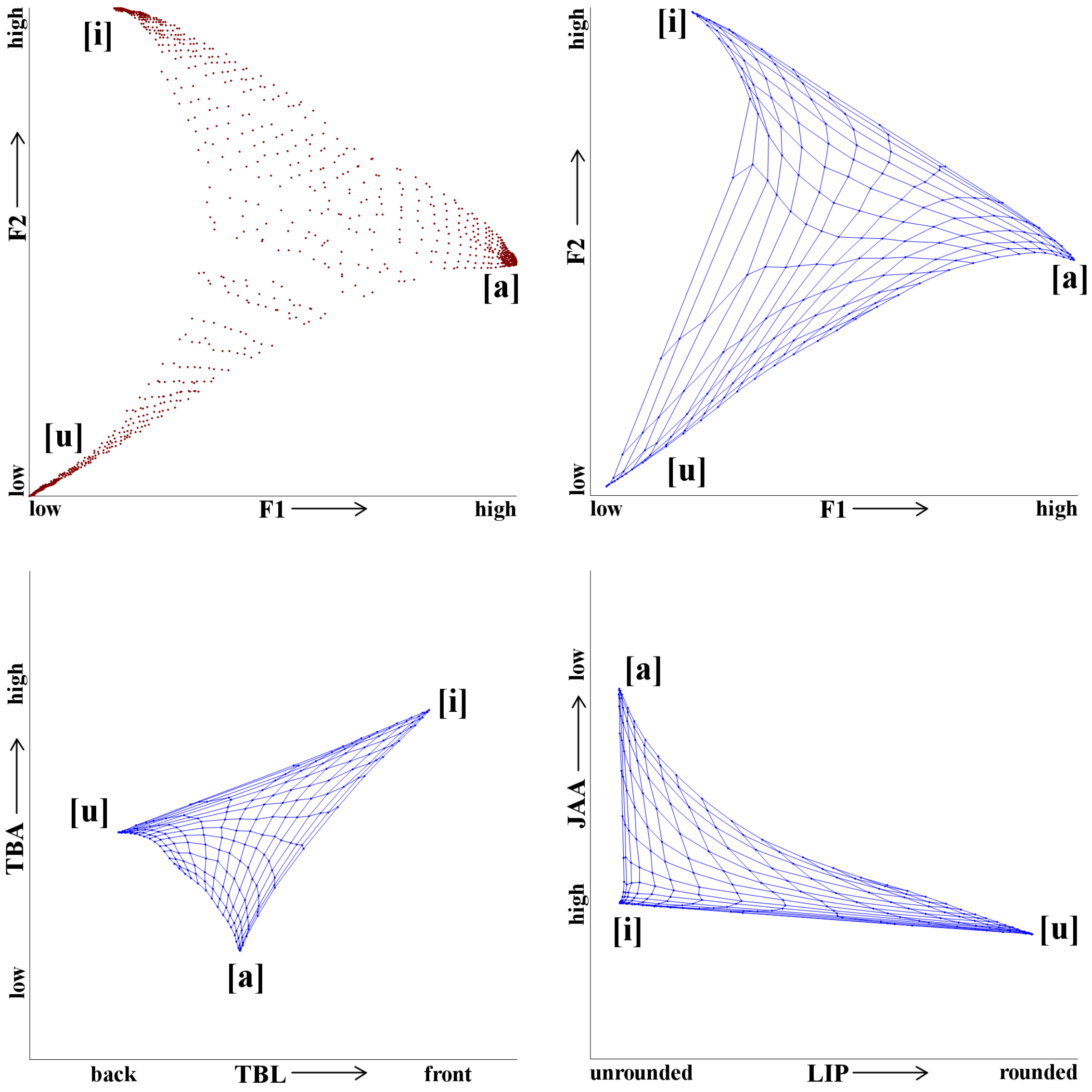

**Figure 6:** (a) Top, left : Distribution of the proto-vocalic training items in the auditory domain (bark-scaled F1 and F2 values). (b) Top, right: Typical training result for the self-organizing map displayed as a grid network. The nodes of the grid network represent the auditory F1-F2-link weights of the self-organizing map after proto-vocalic training. Data and training results are displayed in the acoustic-auditory F1-F2-plane (F3-link weights were also trained but are not displayed here). Both formant frequency axes are bark scaled. (c) Bottom, left: Training result for the same SOM for the motor parameter link weights tongue body horizontal location (TBL) and tongue body angle (TBA). (d) Bottom, right: Training result for the same SOM for the motor parameter link weights lip protrusion (LIP) and lower jaw angle (JAA).

Secondly the regular unfolding grid network (Figure 6b, 6c, and 6d) indicates that the differences of synaptic link weight values of neighbouring SOM neurons (represented by the distance of the neurons within the grid network) are small, i.e. link weight values from neuron to neuron vary continuously with respect to the neural ordering within the SOM. Thirdly this continuous variation of link weight values displayed in Figure 6b for the auditory link weights and in Figure 6c and Figure 6d for the motor plan link weights indicates in addition an ordering of proto-vocalic phonetic states within the grid network. Since the motor parameter dimensions of tongue body parameters (tongue body horizontal location, TBL, and tongue body angle, TBA) and the motor parameter dimensions of lip and lower jaw parameters (lip protrusion, LIP, and jaw angle, JAA) as well as the auditory (bark-scaled) F1-F2-dimensions can also be interpreted as vocalic low-high and front-back dimensions, it appears that proto-vocalic states represented by the net nodes of the grid network are ordered with respect to these dimensions. Thus the SOM displayed in Figure 6b, 6c, and 6d for the auditory and the motor plan dimensions is capable of organizing or arranging vowel states in an ordered and phonetic manner with respect to the vocalic attributes high-low and front-back. This feature of our neurocomputational network is called *phonetotopic ordering* of vocalic states or just *phonetotopy*. This SOM and its link weight distribution displayed in Figure 6b, 6c, and 6d form the *vocalic part of the phonetic map*.

## 5. EXPERIMENT III: LANGUAGE-SPECIFIC VOCALIC ARTICULATION TRAINING

### 5.1. Method

After proto-vocalic training the model is capable of learning vocalic phoneme realizations by using a *language-specific vocalic training set*. The aim of this phase is to train the model to produce correct phoneme realizations, i.e. to activate correct motor and sensory representations for each phonemic state. This training is based here on a training set representing a five-phoneme vowel system (/i/, /e/, /a/, /o/, and /u/) comprising 100 realizations of each phoneme. The distribution of these 500 phoneme realizations (training items) over the acoustic-auditory F1-F2 vowel space is shown in Figure 7a. The 100 realizations per phoneme form a “phoneme realization region” within the vowel space. They are generated by random spreading or scattering with respect to a Gaussian density distribution. The training is done after proto-vocalic training and leads to a further adjustment of the link weights of the phonetic-sensory and phonetic-motor mappings. In addition, however, the phonetic-phonemic mapping is trained at this point (Figure 1). The link weight values of the phonetic-phonemic mapping are initially zero during proto-vocalic training since that training does not by definition include a phonemic labelling of the stimuli. Thus during the proto-vocalic acquisition phase only those synaptic link weights within the SOM are trained which connect the vocalic phonetic map with the sensory and motor maps. However, during language-specific vocalic training the synaptic link weights connecting the SOM with the phonemic map (Figure 1 and Figure 4) are also adjusted.

## 5.2. Results

Using standard training SOM algorithms, approximately 5,000 training steps were necessary to obtain a mean error below 1% for predicting a motor state from a sensory state at the language-specific stage for the 5-vowel system. This training leads to a shift of neural net nodes within the sensory and motor link-weight dimensions. These net nodes then concentrate in the phoneme regions of the vowel space in all motor and sensory dimensions (see Figure 7b, 7c, and 7d). Each phoneme realization region is thus represented by a heap of neighbouring net nodes within the grid network.

Since the language-specific training also comprises training of the phonetic-phonemic mapping (Figure 1) in addition to the synaptic link weights for the phonetic-motor and phonetic-sensory mappings, the synaptic link weights of the phonetic-phonemic mapping (i.e. the phonemic link weights for the SOM neurons) are adjusted and can also be displayed (Figure 8). In Figure 8 the link weight distribution is not displayed using a *grid network display* but a *neuron box display*.

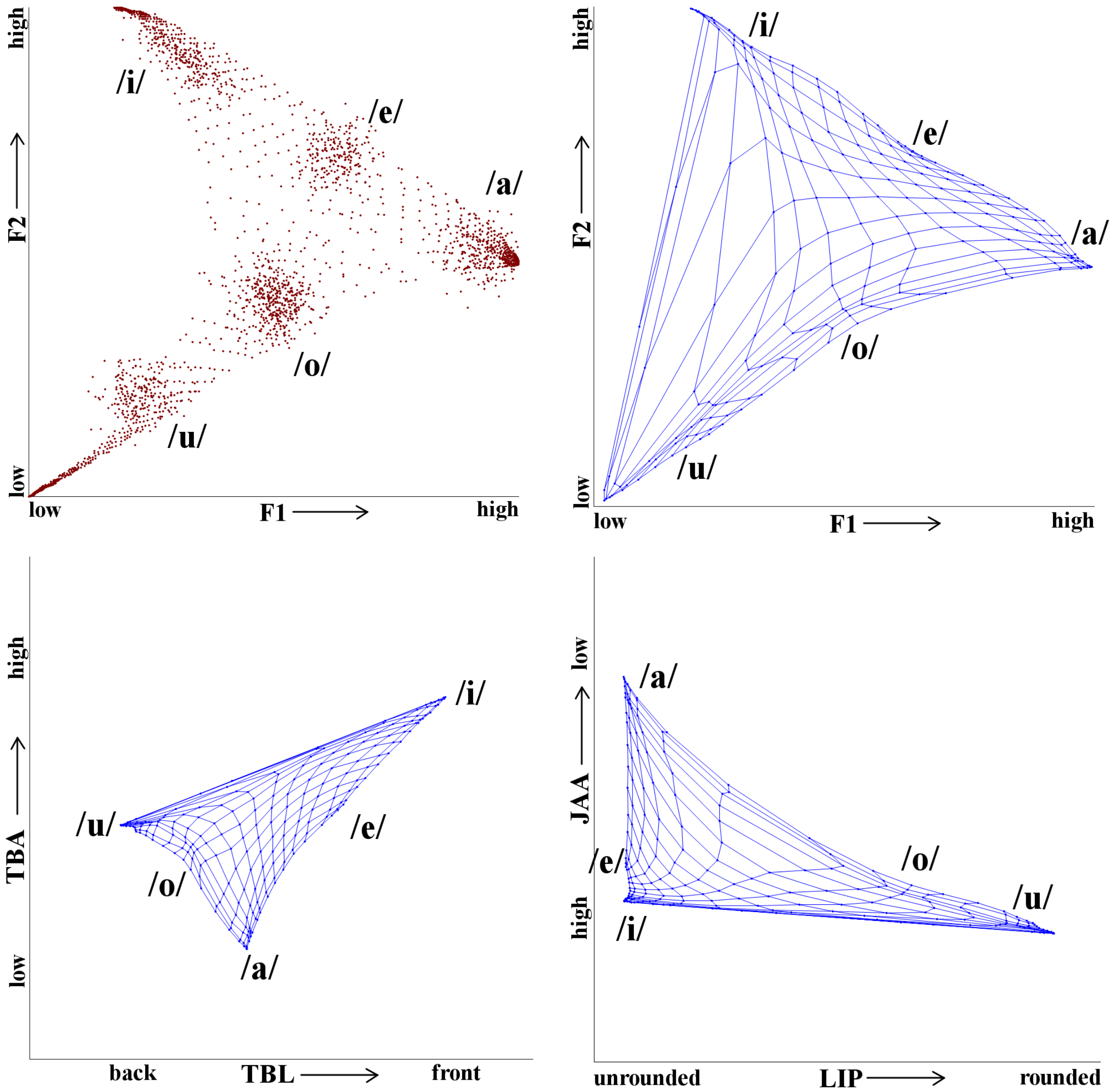

**Figure 7:** (a) Top, left: The distribution of 1078 proto-vocalic training items of the proto-vocalic training set (light grey dots) is overlaid by 500 language-specific training items (dark grey dots) representing a typical five-vowel phonemic system. (b) Top, right: Training results for the phonetic-auditory mapping of the self-organizing map displayed as a grid network. Both formant frequency axes are bark scaled. (c) Bottom, left: Training result for the same SOM for the motor parameter link weights tongue body horizontal location (TBL) and tongue body angle (TBA). (d) Bottom, right: Training result for the same SOM for the motor parameter link weights lip protrusion (LIP) and lower jaw angle (JAA).

Here (Figure 8) the neurons of the 15x15 SOM (i.e. vocalic part of the phonetic map) are displayed as boxes and the link weights (in this case the phonemic link weights) are displayed by a bar chart. The distribution of these phonemic link weights clearly indicates that bundles of neighbouring neurons within the phonetic map are strongly connected with one neuron of the phonemic map, i.e. that bundles of neighbouring neurons represent realizations of a definite phoneme on the level of the phonetic map. Furthermore the phonetotopic ordering of vowel states within the phonetic map is apparent. It should be noted that this phonetotopic ordering of phoneme realizations only appears when language-specific training is done in parallel or even after proto-vocalic training. Phonetotopic ordering always occurs in these cases, but the distribution of the phonetic vocalic dimensions low-high and front-back differs from training to training because of the random distribution of initial link weight values (see Figure 9). If proto-vocalic training is omitted, our training results indicate no phonetotopic ordering (Figure 10).

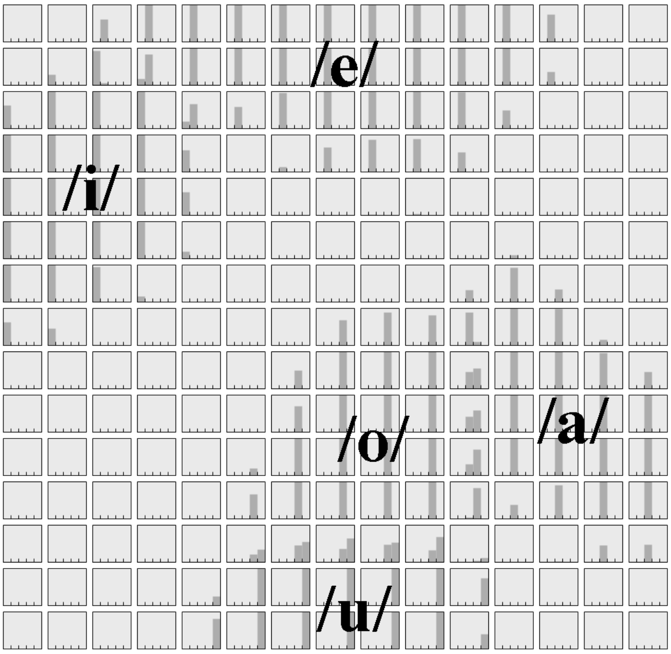

**Figure 8:** Training results for the phonemic-phonetic mapping of the self-organizing map given in the neuron box display. Each grey box represents one neuron of the SOM (as do the nodes in the grid network display in Figure 8). The five bars within each box represent the phonemic link weights of the SOM after language-specific vocalic training. Bars from left to right represent link weight values for /i/, /e/, /a/, /o/, and /u/. This bar plot indicates that some neurons can be clearly associated with distinct vowel phonemes while other neurons exhibit no association with any vowel phoneme.

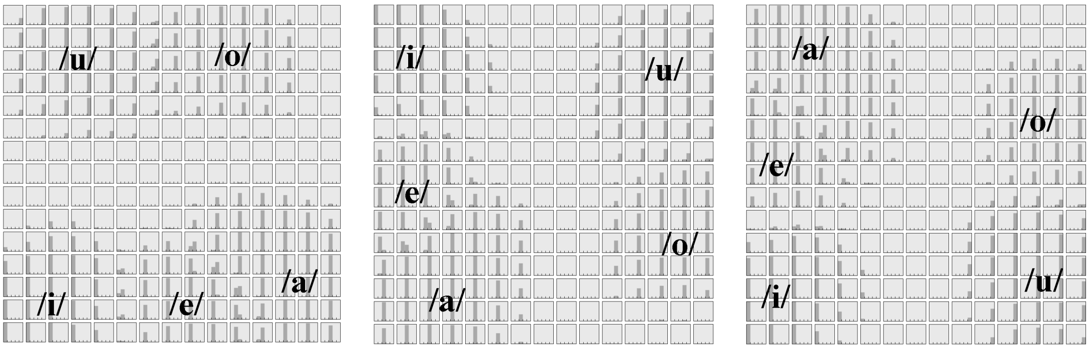

**Figure 9:** Training results for the phonemic-phonetic mapping of the self-organizing map given in the neuron box display for three other instances (training procedures). Each grey box represents one neuron of the SOM (cf. Figure 8). In all three cases a phonetotopic ordering occurs (as occurs in Figure 8), but the ordering differs with respect to the x- and y-axes of the neural map.

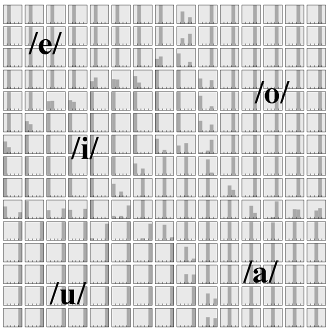

**Figure 10:** Training results for the phonemic-phonetic mapping of the self-organizing map given in the neuron box display for a further instance of the model (cf. Figure 8 and Figure 9). Not babbling training, but imitation training was executed and no phonetotopic ordering occurred.

## 6. EXPERIMENT IV: PROTO-CONSONANTAL ARTICULATION TRAINING

Up to this point, only steady state motor events, i.e. vowels, had been trained. As a next step, the model underwent prelinguistic and linguistic training phases for consonants, more specifically for labial, apical and dorsal full closing gestures.

## 6.1. Method

The training of *proto-consonantal articulation* is based on a training set comprising 225 closing gestures. The training items start from 25 different proto-vocalic states, covering the whole vowel space. Each proto-vocalic state is combined with nine different closing positions, i.e. three positions for labial, apical, and dorsal closures. In natural babbling, the exact closing position for a vocal tract organ (lips, tongue tip, tongue body) is in part determined by the vocalic starting position of this articulator (henceforth "natural closing position"). This is reflected in this training set by the association of rounded vowels with rounded labial closures, unrounded vowels with unrounded labial closures, back vowels with a slightly more back velar dorsal closure and front vowels with a slightly more front velar closure. In addition, a forward- or backward-shift from these natural closing positions was implemented in the training set in order to have a training set comprising all physiologically possible consonantal closure positions. This results in three closing positions (front, mid, and back) for each of the three articulatory gestures (labial, apical, and dorsal), i.e. it results in nine different closing positions per proto-vocalic starting position.

The *motor plan representation* of each proto-consonantal closing gesture comprises the following motor plan parameters: (i) The parameters 'front-back' and 'high-low' define the proto-vocalic state; (ii) the parameter 'gesture-executing vocal organ' (lips, tongue tip, or tongue body), and (iii) a label for the 'exact proto-consonantal place of articulation' (front, mid, or back) define the closing gesture. The exact movement of the gesture-executing vocal tract organ on the level of the primary motor map is defined by calculating the distance between the current and target position of the closure-producing vocal tract organ and then by specifying a movement velocity which is proportional to this distance.

The *auditory representation* of each proto-consonantal closing gesture comprises the first three (bark-scaled) formant transitions of the closing gesture starting from the proto-vocalic state and ending at the beginning of vocal tract closure (Figure 11). Formant values were extracted at 5 equidistant time points covering the entire transition interval. In addition, somatosensory information are added to the training items (i) representing the tactile pattern (contact pattern) of the tongue tip or tongue body with the hard and soft palate during vocal tract closure and (ii) representing the proprioceptive state of  tongue and lips for the initial proto-vowel and for the final proto-consonantal closure.

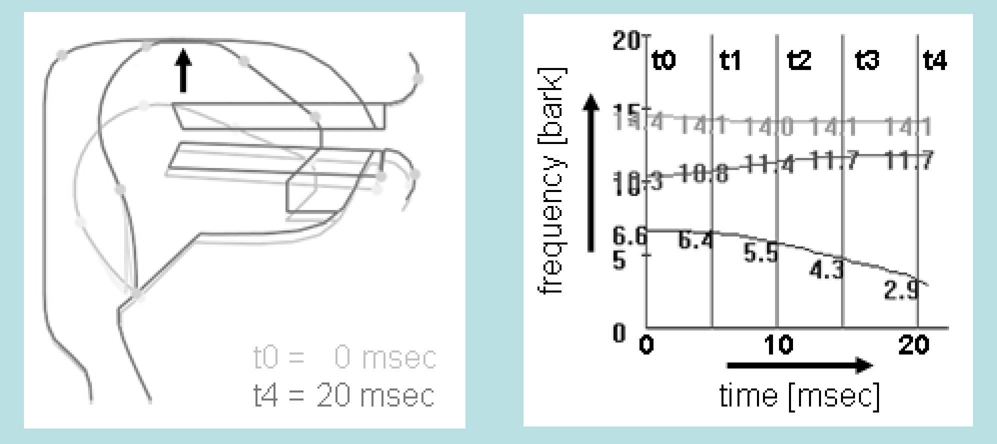

**Figure 11:** Auditory state (right side) for a dorsal closing gesture (left).

## 6.2. Results

The aim of the *proto-consonantal training phase* is to predict the higher level motor representation of closing gestures from the auditory representation, i.e. from the formant transitions. Training was successful using a 10x10 self-organizing map (voiced plosive VC part of the phonetic map) for training the phonetic-sensory and phonetic-motor mappings (simulations using 15x15 maps led to similar results; 20x20 maps have not been tested yet). Standard training algorithms were used (Kohonen, 2001). Approximately 150,000 training steps were sufficient to obtain good training results, i.e. a mean error below 5% for predicting all motor plan parameters of a proto-consonantal closing gesture from its auditory state (i.e. its formant transition). The prediction of just the gesture-executing vocal tract organ (lips, tongue tip or tongue body) leads to a prediction error of lower than 1%. The prediction errors were measured using a test set of 198 test items which cover the whole range of proto-vocalic states and proto-consonantal closures as defined by the training set (Figure 6a). Different motor plan parameter values were used here in comparison to the training set items. The synaptic link weights after training are given in a neuron box display for the neurons of the SOM (Figure 12). This SOM represents the *VC-proto-consonantal part* of the phonetic map. The first three bars indicate the motor link weight values for identifying the gesture-executing vocal tract organ (labial, apical, or dorsal). The map shows a clear separation of three regions for labial, apical, and dorsal closures indicating that these three types of closing gestures can be clearly separated by our production model on the level of this SOM (Figure 12b). It should be emphasized that the formant transitions displayed within each box in Figure 12a do *not* show training *data* but training *results*. They display the auditory *link weight values* of this SOM. Thus this display delivers *100 typical formant transitions*, learned from all combinations of proto-vocalic starting positions and closure positions given in the training set. This is the *formant-transition knowledge for closure gestures* gained from the proto-consonantal training. In addition, on the basis of this 10x10 system of formant transitions, a clear auditory separation of labial, apical, and dorsal closing gestures is given. This means that the network is capable of predicting motor plan representations for closing gestures or at least the closure-performing articulator and thus the place of articulation on the basis of these 100 auditory states.

The SOM displayed in Figure 12 indicates not only an ordering of phonetic states with respect to the type of closing gesture (labial, apical, and dorsal; Fig 12b) but also with respect to the proto-vocalic starting position (Figure 12c). Distinct SOM regions can not only be detected for the closure-forming vocal tract organ but also for the proto-vocalic starting position, at least for the front-high and back-high feature combinations (Figure 12c) across different types of closing gestures (labial, apical, and dorsal). This illustrates *phonetotopy* for the type of closing gesture. Strict phonetotopy cannot occur here in the sense that a strict ordering occurs for the proto-vocalic starting positions of the gestures with respect to the phonetic dimensions ‘front-back’ and ‘high-low’, as it is impossible from a topological viewpoint to order all closing gesture states with respect to three phonetic dimensions (i.e. consonantal closing position ‘labial-apical-dorsal’ and two vocalic dimensions ‘front-back’ and ‘high-low’), since it is a physiological fact that cortical neural maps are two-dimensional.

It must be emphasized that proto-consonantal training can start in parallel with proto-vocalic training. A finalization of proto-vocalic training is not needed for starting the proto-consonantal training.

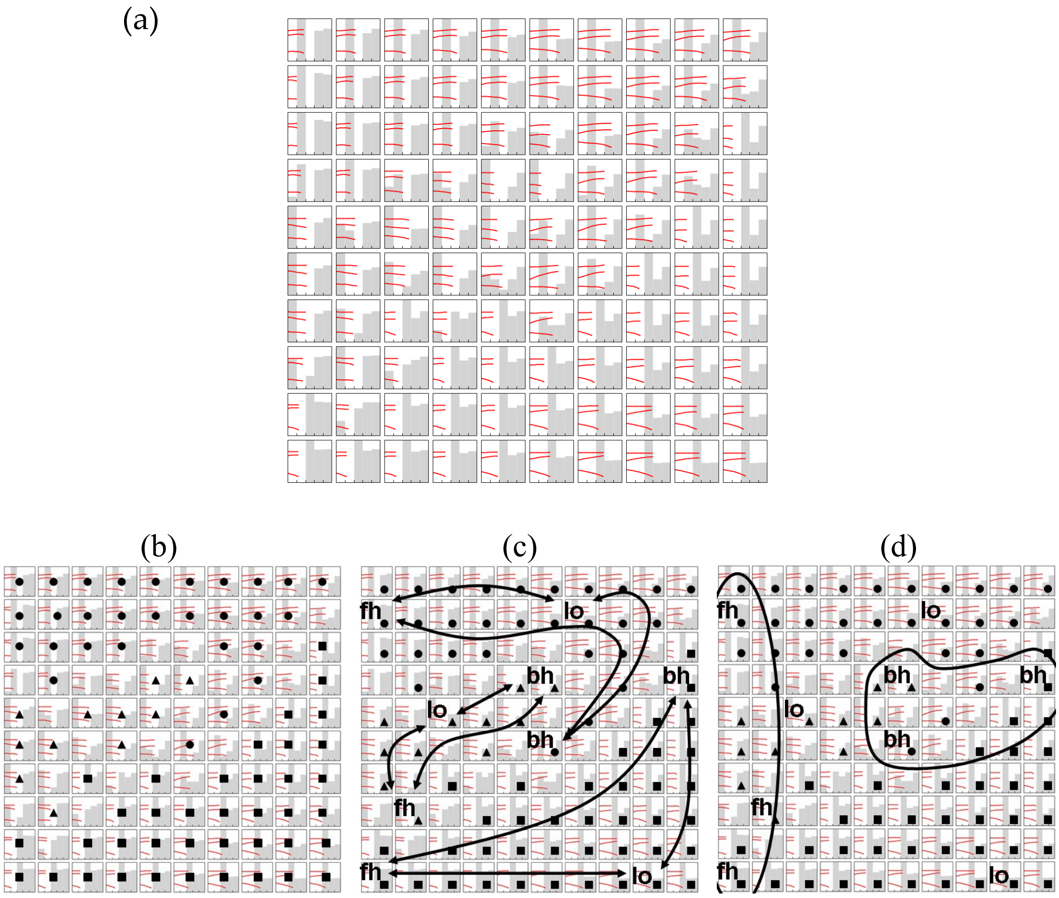

**Figure 12:** (a) Neuron box display of motor plan parameter link weights for the SOM representing the VC-proto-consonantal part of the phonetic map. Each box represents one neuron of the 10x10 SOM. Grey bars within each box from left to right: the first three bars represent link weights for the gesture-executing vocal organ (lips, tongue tip, tongue body); bars 4 and 5 represent two phonetic parameters of the proto-vocalic starting vowel: back-front (bar 4) and low-high (bar 5). The F1-F2-F3-formant trajectories (lines) represent the auditory link weights. (b) Same SOM as in Figure 10; in addition gesture-executing vocal organs are marked if the link weight value is above 98% for this organ: labial (triangle), apical (dot), dorsal (square). (c) Same SOM; in addition feature combinations for the proto-vocalic starting position are marked: front-high or [i]-like (fh), back-high or [u]-like (bh), and low or [a]-like (lo); also transitions from front-high to back-high and from front-high or back-high to low are marked by arrows for each gesture-executing vocal tract organ. (d) Same SOM; in addition connected regions can be found for front-high ([i]-like) and back-high ([u]-like) feature combinations for the proto-vocalic starting positions across labial-apical-dorsal boundaries.

# 7. EXPERIMENT V: LANGUAGE-SPECIFIC CONSONANTAL ARTICULATION TRAINING

## 7.1. Method

After proto-consonantal training a typical *language-specific consonantal phoneme system* is trained using a VC-syllable training set, i.e. the *VC voiced plosive training set*. This training set comprises all combinations of 50 vowel phoneme realizations (10 realizations per vowel /i/, /e/, /a/, /o/, /u/ were chosen from the language-specific vocalic training set) and three consonantal closures (labial, apical, and dorsal) using the natural closing position (see section 6.1). Thus the training set comprises 150 items. Similar to the process described in the training experiments for vowels, training items are now in addition labelled as /Vb/, /Vd/, or /Vg/ (V = vowel phoneme label).

## 7.2. Results

Training was successful using a 10x10 SOM. In comparison with the proto-consonantal SOM, now in addition the phonemic-phonetic mapping was trained. 100 training cycles (approximately 5,000 training steps) were sufficient to obtain good training results, i.e. to predict the phonemic state (/b/, /d/, or /g/) from an auditory state with a prediction error below 1% for all VC-combinations. The resulting map of phonetic-sensory and phonetic-motor link weights is nearly identical with those trained earlier during the proto-consonantal training. In addition phonemic link weights for /b/, /d/, and /g/ were trained now, leading to a distribution similar to the motor link weight distribution for the gesture-executing vocal tract organ (lips, tongue tip, tongue body). Thus in the case of proto-consonantal and language-specific consonantal training, the fact of three discrete closure-forming articulators (lips, tongue tip, and tongue body) already anticipates the phonemic separation in the case of a three-plosive phoneme system /b/, /d/, and /g/ during the proto-consonantal (prelinguistic) training phase.

It must be emphasized that language-specific consonantal training can also start in parallel with language-specific vocalic training as well as in parallel with proto-vocalic and proto-consonantal training. A finalization of proto-vocalic and vocalic training as well as a finalization of proto-consonantal training is not needed to start the language-specific consonantal or syllabic training.

## 8. MAIN RESULTS AND DISCUSSION

A neurocomputational model of speech production is introduced which defines the structure of neural maps and mappings. Training or learning phases are described for acquiring sensorimotor, phonetic and phonemic knowledge for the mappings which build up this model.

In the structural part the model differentiates feedforward and feedback control (cf. Guenther, 2006). In the case of feedforward control, motor plans and sensory states of sounds, syllables or words are activated from phonological (i.e. linguistic) descriptions of these speech items via a phonetic map. The phonetic map introduced in our approach exhibits bilateral mappings with phonemic, sensory, and motor maps and can be interpreted as a computer-implemented spell-out of the mental syllabary (cf. Levelt and Wheeldon, 1994). In the case of feedback control, already learned and stored sensory states of speech items can be compared with auditory signal states currently produced by the model (or speaker). The neural feedback pathway is active during all training phases described above but cannot be activated in the current implementation of our model as an online control system during speech production.

The phonetic map introduced in our approach results from the fact that the neural mapping of sensory, motor plan, and phonemic states is implemented as a self-organizing map (SOM; Kohonen, 2001). This SOM is the central part of the sensory-motor-phonemic mapping. The cells within the SOM representing these connections can be labelled as *hyper-* or *supramodal* since they comprise direct neural connections towards all sensory maps and towards the motor plan map. The SOM in addition can be labelled as *phonetic* since it comprises the whole range of phonetic realizations of a phonemic item. Different SOMs were trained for different phonemic items (e.g. V and VC in this paper). Our training results indicate that the ordering of phonetic states within these SOMs reflects phonetic dimensions like the vocalic dimensions 'high-low' and 'front-back' or classes of plosives like 'labial', 'apical', and 'dorsal' plosives. This ordering is called *phonetotopic ordering* and has recently been verified by brain imaging experiments for vowels (Obleser et al., 2006). Furthermore, regarding the phonetic map, i.e. the vocalic and consonantal phonetic submaps trained during this study, it has been found that the neurons within the phonetic submaps exhibit an ordering of phonetic states with respect to phonetic features like high-low and front-back in the case of the vocalic submap or high-low, front-back and consonantal place of articulation (labial, apical, dorsal) in the case of VC-syllables. The same results gained for VC-syllables can also be reproduced for CV-syllables (not shown in this paper).

Speech items like syllables are processed as a whole in this model. The motor plan and its sensory consequences are stored for each frequent syllable *as a whole pattern* by the phonetic to sensory and by the phonetic to motor mappings. Thus the temporal succession of a motor and its appropriate sensory states is represented as one chunk on the level of the sensory and motor maps including the time succession for this speech item. Therefore time occurs implicitly in our model. As a result this model in its current implementation does not process temporal aspects of speech production in the same way or in such a detailed way as is done in the Guenther et al. (2006) approach.

Moreover in contrast to the gestural control approach introduced by Browman and Goldstein (1992), which is a mainly articulatory-based and not acoustically or perceptually based concept, the goals of gestures are coded within a hypermodal phonetic domain in our approach. For example, targets of speech gestures, which are defined on the level of the motor plan in our approach (and which are defined in parallel within the tract-variable space in the Browman and Goldstein approach), have in addition closely related sensory (i.e. auditory and somatosensory) correlates in our approach. This results from the bidirectional mappings of motor plan and sensory states (Figure 1). Thus gestural targets are coded in parallel in motor plan and sensory domains in our approach.

In order to gather speech knowledge, the mappings are trained in five basic training or learning phases. These phases are described in detail in this paper (i.e. silent articulation training, proto-vocalic and vocalic articulation training, proto-consonantal and consonantal articulation training). Silent articulation training results in the ability to accomplish a certain local or overall vocal tube geometry (i.e. formation of a certain consonantal closure or near-closure or formation of a certain vocalic tube geometry) using different relative articulatory positioning patterns (e.g. accomplishment of a certain vocal tube geometry using different articulatory positions of the lower jaw). Proto-vocalic and proto-consonantal articulation training are prelinguistic phases and lead to the gathering of general (language-independent) phonetic knowledge. These training phases result in the ability to predict proto-speech (or speech-like) articulation from sensory, mainly auditory speech-like patterns. This prelinguistic sensorimotor babbling training is the basis for language-specific imitation training. In the case of imitation training, vocalic and consonantal articulation training are language-specific and are exemplified in this paper for (i) a five-vowel phoneme system (/i/, /e/, /a/, /o/, and /u/) and (ii) for VC-syllables with a phoneme system of three voiced plosives (/b/, /d/, and /g/). The vocalic phoneme realizations form realization clouds within the articulatory and acoustic vowel space as is illustrated above (Figure 7).

It may be a shortcoming of our study that the (external) acoustic phoneme realization clouds for imitation training are generated in an artificial way representing a "hypothetical" 5-phoneme vowel system, but preliminary tests which vary the location and the degree of overlap of the phoneme realization clouds indicate that the resulting self-organizing phonetic map is stable with respect to the features mentioned in this paper, e.g. the feature of phonetotopy.

It may be a further limitation of our model that the babbling training sets are associated with random production of motor events, although this is an oversimplification (cf. McNeilage et al., 1997; McNeilage et al., 2000). Better training sets should be found, but currently there exist no complete phonetic data sets of toddlers' productions during the first year of life.

As a result of the phonetotopic ordering of phonetic states within the phonetic SOM for *all* motor and sensory dimensions, this SOM network is capable of predicting motor and somatosensory states for each speech item directly from the auditory state (auditory information). However, in addition the network is also capable of predicting, for example, sensory (auditory and/or somatosensory) states from motor states with comparable accuracy. This *multidirectionality* – i.e. multidirectional prediction of motor and sensory states for a speech item via the phonetic map – is represented by the bidirectional arrow bundles in Figure 1. Furthermore this multidirectionality means *co-activation* of motor or sensory states: If a specific state within one map in the side layer of the self-organizing network (motor, auditory, or somatosensory) is initially stimulated, a co-activation of all other side-layer maps immediately occurs via the central layer (phonetic map).

Thus assuming the proposed structure or organization of this neurocomputational model as introduced here – which is based on existing neurophysiological and neuropsychological knowledge – our simulation experiments for acquiring prelinguistic and basic linguistic or language-dependent knowledge lead to three main results: (i) The model is capable of predicting effects like ordering of vocalic and consonantal states with respect to phonetic features, which is labelled as phonetotopy. (ii) The model is capable of predicting motor states from auditory states after prelinguistic babbling training, here exemplified for proto-vocalic and proto-consonantal articulation training. This ability is the basis for imitation training. (iii) During language-specific imitation training the model is capable of learning sets of sound or syllable realizations for each phonemic representation (here exemplified for V and VC with C = voiced plosives). These sets of sound or syllable realizations are clustered within the phonetic map.

Future work has to be done in order to train other sound types (nasals, fricatives, approximants, etc.) and other and even more complex syllable types (like CV, CVC, CCV, CCVC, etc.). This would lead to an exemplification of the mapping between the phonetic map and the motor planning module (Figure 1) and would lead to the concretisation of language-specific timing rules for speech gestures. Furthermore, corrective processes due to feedback signals occurring during the sensorimotor process of speech production (cf. Guenther et al., 2006) should be implemented in our approach immediately.

## ACKNOWLEDGEMENTS

This work has been supported in part by the German Research Council DFG research grant No. KR 1439/13-1. I deeply thank the editors of this book and all people who gave fruitful comments on this paper and who contributed in correcting this manuscript.

## REFERENCES

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

Birkholz, P. (2005). *3D articulatory speech synthesis*. University of Rostock: Unpublished PhD thesis. [In German: 3D-Artikulatorische Sprachsynthese].

Birkholz, P. & Jackel, 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, 1125-1128.

Birkholz, P. & Kröger, B.J. (2006) Vocal tract model adaptation using magnetic resonance imaging. *Proceedings of the 7th International Seminar on Speech Production*, Ubatuba, Brazil, 493-500.

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

Birkholz, P., Jackel, D., and Kröger, B.J. (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, 873-876.

Birkholz, P., Jackel, D., and Kröger, B.J. (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, S.C., Scott, S.K., Murphy, K., Warburton, E., and Wise, R.J.S. (2002). Speech production: Wernike, Broca and beyond. *Brain* 125, 1829-1838.

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

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

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

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

Guenther, F.H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics* 72, 43-53.

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

Guenther, F.H. (2006). Cortical interaction underlying the production of speech sounds. *Journal of Communication Disorders* 39, 350-365.

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

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

Hillis, A.E., Work, M., Barker, P.B., Jacobs, M.A., Breese, E.L., and Maurer, K. (2004). Re-examing the brain regions crucial for orchestrating speech articulation. *Brain* 127, 1479-1487.

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

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

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

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

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

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

Kröger, B.J., Birkholz, P., Kannampuzha, J., and 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),* Pittsburgh, USA, 565-568.

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

Kröger, B.J., Birkholz, P., Kannampuzha, J., and 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, 561-562. (for download see also http://www.speechtrainer.eu)

Kröger, B.J., Birkholz, P., Kannampuzha, J., and Neuschaefer-Rube, C. (2007). Multidirectional mappings and the concept of a mental syllabary in a neural model of speech production. *DAGA-Proceedings of the 33th Annual Meeting of the German Acoustical Society,* Stuttgart, Germany), 91-92. (for download see also http://www.speechtrainer.eu)

Kröger, B.J., Birkholz, P., and Neuschaefer-Rube, C. (2007). Modeling developmental aspects of sensorimotor control os speech production. *Laryngo-Rhino-Otologie* 86, 365-370. [In German: Ein neuronales Modell zur sensomotorischen Entwicklung des Sprechens]

Kröger, B.J., Schnitker, R., and Lowit, A. (2008). The organization of a neurocomputational control model for articulatory speech synthesis. In: A. Esposito, N. Bourbakis, N. Avouris, and I. Hatzilygeroudis (eds.) *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction,* Berlin: Springer, 126-141.

Kröger, B.J., Birkholz, P., and Lowit, A. (submitted). Phonemic, sensory, and motor representations in a neural model of speech production.

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

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

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

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., and Medler, D.A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex* 15, 1621-1631.

McNeilage, P.F., Davis, B.L., and Matyear, C.L. (1997). Babbling and first words: phonetic similarities and differences. *Speech Communication* 22, 269-277

McNeilage, P.F., Davis, B.L., Kinney, A., and Matyear, C.L. (2000). A motor core of speech: a comparison of serial organization patterns in infants and languages. *Child Development* 71, 153-163.

Nasir, S.M. & Ostry, D.J. (2006) Somatosensory precision in speech production. *Current Biology* 16, 1918-1923.

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

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

Perkell, J.S., Matthies, M.L., Svirsky, M.A., and Jordan, M.I. (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.

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

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

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, K.G. (1989). A dynamic approach to gestural patterning in speech production. *Ecological Psychology* 1, 333-382.

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

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

Sörös, P., Sokoloff, L.G., Bose, A., McIntosh, A.R., Graham, S.J., and Stuss, D.T. (2006). Clustered functional MRI of ouvert speech production. *NeuroImage* 32, 376-387.

Tremblay, S., Shiller, D.M., and Ostry, D.J. (2003). Somatosensory basis of speech production. *Nature* 423, 866-869.

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

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

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