Learning to associate speech-like sensory and motor states during babbling

Bernd J. Kröger[1], Peter Birkholz[2], Jim Kannampuzha[1], Christiane Neuschaefer-Rube[1]

[1]Department of Phoniatrics, Pedaudiology and Communication Disorders
University Hospital Aachen (UKA) and Aachen University (RWTH)
Pauwelsstr. 30, 52074 Aachen, Germany

[2]Institute for Computer Science
University of Rostock, Albert-Einstein-Str. 21, 18059 Rostock

bkroeger@ukaachen.de, piet@informatik.uni-rostock.de,
jim.kannampuzha@rwth-aachen.de, cneuschaefer@ukaachen.de

Abstract. *Background: Development of the feedback loop of speech production starts during the babbling phase of speech acquisition. Within the first year of lifetime toddlers acquire the ability of imitating auditory stimuli, i.e. they acquire the ability of associating speech-like sensory and motor states. Method: Self-organizing maps and one-layer feed-forward networks were used for modeling this learning behavior within a neural model of speech production. The model includes auditory, proprioceptive, tactile, and motor representations for static as well as for dynamic speech-like events, i.e. proto-vocalic and proto-consonantal states. A three-dimensional articulatory speech synthesizer serves as a front-end device for generating high quality sensory signals. Results: Self-organizing maps are useful for modeling auditory-to-somatosensory as well as for auditory-to-motor mappings. The somatosensory-to-motor mapping is modeled successfully using a one-layer feed-forward net. Conclusion: The neural model of speech production introduced here is capable of describing learning during the babbling phase of speech acquisition. The model is now ready for building up the mental syllabary, i.e. for processing sounds, syllables and words of a specific language.*

1. Introduction

Young children use their articulatory organs for sound production from the first day of their lifetime. During the first year of life children produce non-speech acoustic signals like crying, laughter, moaning and vegetative sounds (e.g. couching, sneezing, burping). But they also produce speech-like sounds – i.e. speech-like phonation, vowel-like articulation, and primitive speech-like articulatory movements (Oller et al. 1999). Thus from the viewpoint of speech acquisition the first year of lifetime is very important. It is called *babbling phase* or *prelinguistic phase of speech acquisition* since the toddler learns to produce first static and dynamic articulatory primitives – here called "proto-vocalic", "proto-consonantal", or "proto-gestural" articulation. Within this year the toddler already learns to associate auditory patterns (i.e. formant patterns) with articulatory or motor states.

This paper presents a neural model of speech production capable of reflecting these early processes of speech acquisition. The paper focuses on the development of associations

between auditory and motor representations, i.e. on the auditory-to-motor mappings for proto-vocalic and proto-consonantal articulatory gestures. Two different setups for the motor and sensory maps of the feedback subsystem were tested.

2. Brief outline of the neural model

Neural models of speech production comprise a *feedback subsystem* and a *feed-forward subsystem* (Guenther et al. 2006). The feed-forward subsystem directly associates states of the speech sound map – i.e. neural representations of sounds, syllables or words – with their motor states. The feedback system is very complex. Here, the auditory and somatosensory state activated by the speech sound map are compared with the online sensory signals produced by the current state of the speech organs. The occurring difference signal is used for updating or correcting the current motor state.

Our neural model of speech production comprises auditory, somatosensory, and motor maps, the appropriate mappings, subcortical and peripheral processing of signals, and a high-quality 3D-articulatory-acoustic speech synthesizer as a front-end system for producing high quality tactile, proprioceptive, and auditory feedback signals (Kröger et al. 2006a and Fig. 1). The sensory-to-motor mappings evolve during training using a set of static or dynamic articulatory states. First results of training the neural model using one-layer feed-forward artificial networks were collected (Kröger et al. 2006a). During the babbling phase of speech acquisition mainly the feedback control subsystem evolves. During the imitation phase of speech acquisition the feed-forward control subsystem becomes more and more relevant.

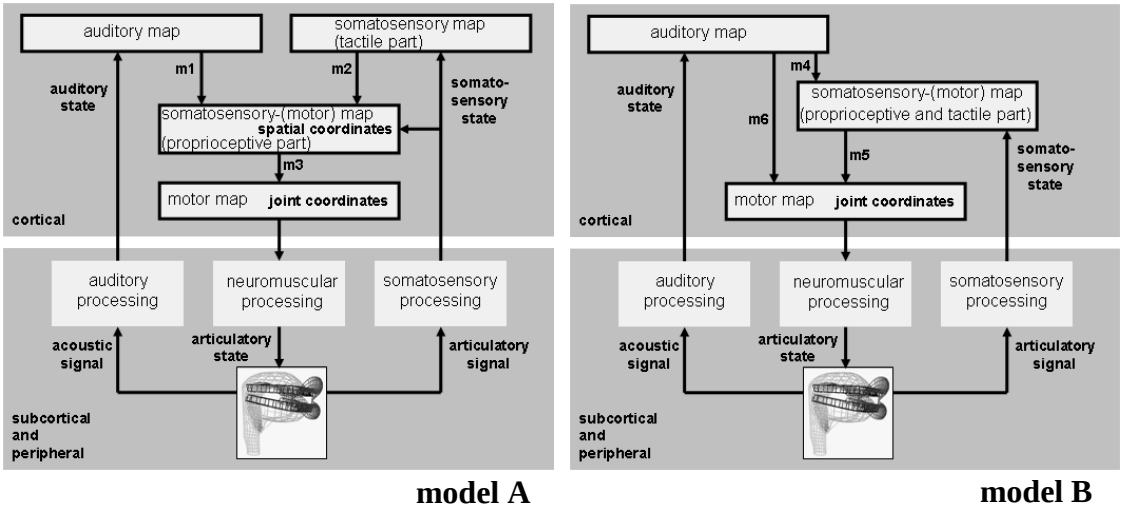

Figure 1. Two different models of the neural feedback subsystem (A and B) as a part of a neural control model of speech production.

In our earlier simulation study (Kröger et al. 2006a) each model parameter is represented by a single neuron. Thus the parameter value is represented simply by the degree of activation of this neuron. The neural maps of our current approach indicate two different kinds of vector representations. Within these representations each model parameter is encoded in a complex way by a set of two or more neurons. Furthermore the previously used simple one-layer feed-forward networks are replaced by *self-organizing maps* (SOM's, Kohonen 2001) for nearly all mappings. These SOM-networks are biologically more realistic in the case of modeling cortical sensory mappings.

Our neural feedback models (Fig. 1) comprise auditory, somatosensory, and motor maps and the related mappings which co-activate the appropriate sensory and motor states. The maps and mappings of the models are based on parameters mainly defined by the front-end 3D-

articulatory-acoustic speech synthesizer (Birkholz et al. 2006). The synthesizer is controlled by a set of 10 articulatory parameters (Fig. 2 and Tab. 1) and provides a set of 10 tract variable parameters, a set of 9 tactile parameters, and a set of 3 auditory parameters (Fig. 3 and Tab. 1). These parameters and representations are discussed in detail in Kröger et al. (2006a). In addition, the current model provides an aerodynamic state parameter describing the pressure built-up which occurs in the vocal tract in the case of an oral obstruction.

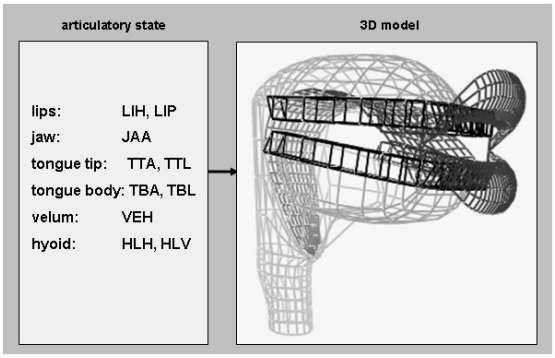

Figure 2. Articulatory parameters and geometrical grid-representation of the 3D model (see also Tab. 1).

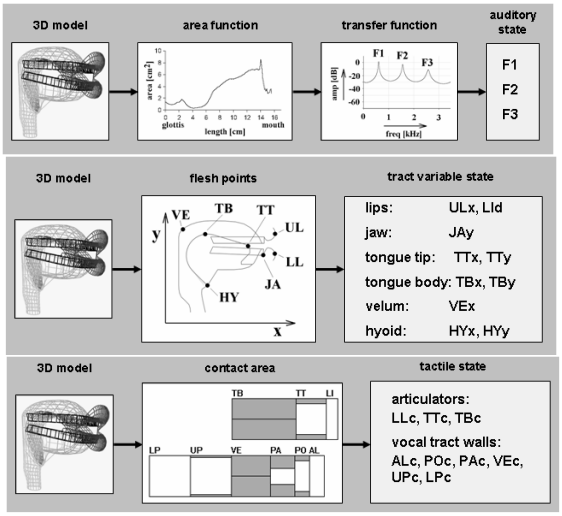

Figure 3. Tract-variable, tactile, and auditory parameter sets generated by the articulatory-acoustic model (see also Tab. 1).

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

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

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

Table 1. Model parameters for articulatory, tract-variable, and tactile states (see also Fig. 2). The model parameters for the auditory state are the bark-skaled formant values F1, F2 (, and F3).

3. Somatosensory-to-motor mappings

Within feedback model A the learning of proto-vocalic or proto-consonantal articulation is preceded by learning the spatial-to-joint coordinate mapping (m3, Fig. 1a). The spatial-to-joint mapping is also referred to as the tract variable to articulatory state transformation numerically solved in the task-dynamic approach (Saltzman et al. 1989). In our approach the *tract variable representation* is labeled synonymously as the *proprioceptive representation*, since it repre-

sents the location of an end-articulator within the cranial reference system (Fig. 1a). This representation is on the edge between motor and sensory representations.

This *proprioceptive representation* comprises 2 or 3 neurons per tract variable parameter depending on the number of states per parameter used in the current training set (Fig. 4). The training set is described in detail in Kröger et al. (2006b). Each state – i.e. each parameter value occurring during the training phase – is represented by one neuron. The neuron which represents a distinct parameter value is activated maximally, if the parameter indicates this value. This kind of neural coding is called "map representation" (Bullock et al. 1993).

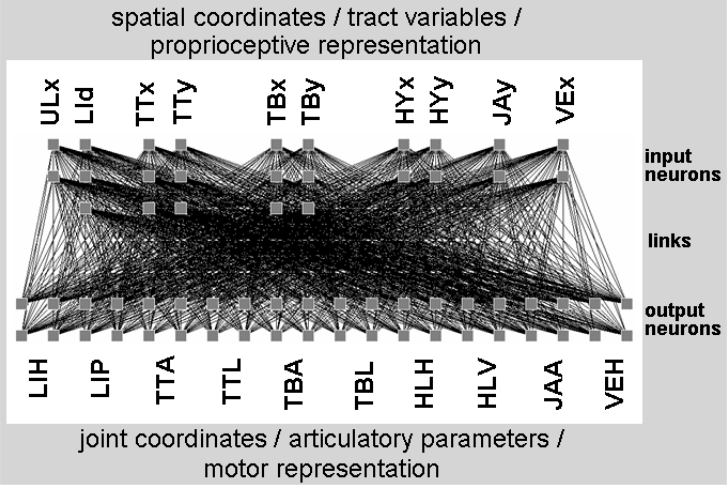

Figure 4. The proprioceptive-to-motor network.

The *articulatory representation* synonymously called *motor representation* or *joint coordinate representation* (Fig. 4) – comprises 4 neurons for each articulatory parameter. These four neurons can be interpreted as a pair of 2 neurons each. Each pair of neurons represents the agonist-antagonist muscular activation for each articulatory parameter. The two neurons of each pair of neurons are activated consecutively for parameter values from 0 to 0.5 and from 0.5 to 1, since only groups of neurons (in this case only 2 neurons are forming the group) are capable of modeling the large dynamic range occurring for muscular activation (Kandel et al. 2000).

The *tactile representation* (Fig. 3) comprises 2 neurons per tactile parameter. The neurons within each pair are activated consecutively for increasing parameter values (see above). The parameter value 1 represents the state of complete contact (full closure) while the parameter value 0 represents the state of no contact.

The *proprioceptive-to-motor mapping* (m3, Fig. 1a and Fig. 4) is implemented by a one-layer feed-forward network exhibiting 1000 link weights connecting 25 input neurons (proprioceptive map) with 40 output neurons (motor map). Training the proprioceptive-to-motor mapping – i.e. adjusting the link weights of this mapping – was performed using a min-max-combination training set comprising 4608 patterns (Kröger et al. 2006b). 5000 cycles of batch training were sufficient for reaching a mean error of 9.1% for predicting an articulatory state. The resulting mapping is capable of modeling motor equivalence (Kröger et al. 2006a and b).

The *somatosensory-to-motor mapping* (m5, Fig. 1b) is also implemented using a one-layer feed-forward network. The net exhibits 1800 link weights connecting 45 input neurons (proprioceptive and tactile map, Fig. 1b) with 40 output neurons (motor map). Training the proprioceptive-to-motor mapping is performed using the same min-max-combination training set (Kröger et al. 2006b). 5000 cycles of batch training were sufficient for reaching a mean error of 10.6% for predicting an articulatory state. The resulting mapping is also capable of modeling motor equivalence (Kröger et al. 2006a). Thus, both mappings (m3 and m5, Fig. 1a and Fig. 1b) behave similar. Despite the fact that the size of the input representation increases quantitatively by nearly a factor 2, the error of the somatosensory-to-motor mapping (m5) is only slightly higher than the error of the proprioceptive-to-motor mapping (m3).

4. Learning proto-vocalic articulation

The *auditory representation* comprises 2 neurons per auditory parameter. The activation of one neuron directly represents the (relative) formant value F_i+ and $F_i- = 1 - F_i+$ (i = 1, 2, or 3 and range of F_i- and F_i+ is 0...1). This representation was chosen in order to guarantee a balanced neural activation for each formant parameter value (see also Guenther et al. 1996).

Within the neural feedback models A and B the auditory-to-motor mapping is subdivided into an auditory-to-proprioceptive mapping and a proprioceptive-to-motor mapping (m1 and m3, model A, Fig. 1a) and an auditory-to-somatosensory mapping and a somatosensory-to-motor mapping (m4 and m5, model B, Fig. 1b). In addition, the neural feedback model B proposes a direct auditory-to-motor mapping (m6, model B, Fig. 1b). All three paths were tested.

In order to get useful training results it was necessary to restrict the amount of possible articulatory vowel-like (or proto-vocalic) states with respect to a set of 3 basic lingual gestures (front-high, back-high, and low) and with respect to two labial gestures (rounded and spread). These basic gestures can be defined easily within the proprioceptive representation. Variation of lip rounding covaried with the lingual front-back dimension resulting in a two dimensional articulatory proto-vocalic space. 540 proto-vocalic training patterns were generated for covering the whole proto-vocalic space (Fig. 5).

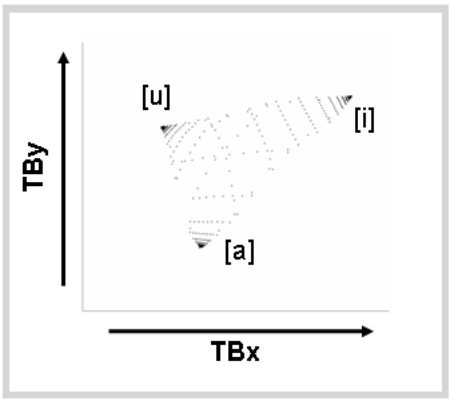

Figure 5. Space of proto-vocalic states given in proprioceptive dimensions (tongue body horizontal vs. vertical position TBx and TBy, see Tab. 1) and display of the set of 540 training patterns.

In all three cases *self-organizing maps* (SOM's, Kohonen 2001) were used for modeling the auditory-to-proprioceptive (m1, Fig. 1a), the auditory-to-somatosensory (m4, Fig. 1b), and the auditory-to-motor mapping (m6, Fig. 1b) respectively. The self-organizing map consists of 10x10 neurons in each case (Fig. 6). Only the input representation differs with respect to the actual mapping (m1, m4, or m6). It should be noted that auditory as well as somatosensory and motor neurons occur side by side as input of the SOM's in in terms of Kohonen (2001).

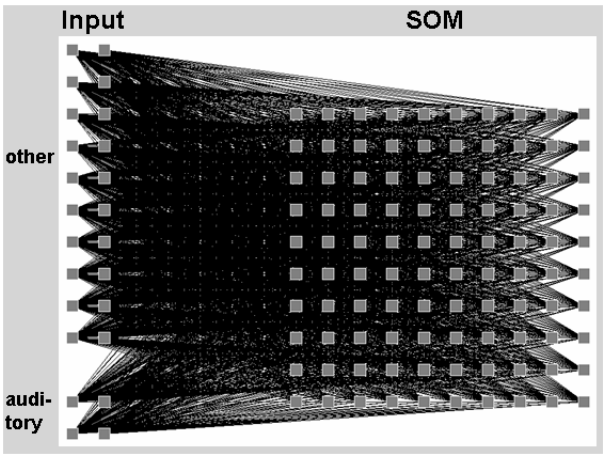

Figure 6. Self-organizing map (10x10 neurons) and the appropriate network for different neural input representations. The input representation is only specified for auditory. “Other” means “proprioceptive” in the case of m1 (Fig. 1a), “proprioceptive plus tactile” in the case of m4 (Fig. 1b) and “motor” in the case of m6 (Fig. 1b).

In the case of the *auditory-to-proprioceptive mapping* (m1, Fig. 1a) 25 proprioceptive neurons plus 4 auditory neurons (F1 and F2) lead to 29 input neurons and thus to 2900 link weights

which are adjusted during training. 200 cycles per 540 training patterns give 108.000 training steps. Standard SOM learning algorithms were used (random initialization; update radius = 5 neurons; rectangular neighborhood function; update radius decay factor = 0.999; learning rate factor = 0.1; learning rate decay factor = 0.99). The mapping was trained 10 times. In all cases the mapping was *not* capable of producing stable proto-vocalic states. Especially in the case of proto-vocalic corner states ([i], [a], and [u]) oral closures were not avoidable.

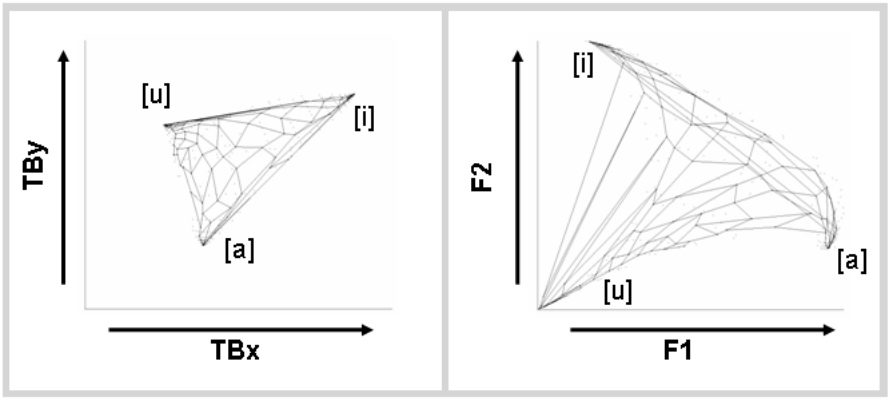

Figure 7. Display of the SOM-weights for each of the 10 x 10 neurons of the self-organizing map (a) within the TBx-TBy-space, i.e. proprioceptive space, and (b) within the F1-F2-space, ie. auditory space (case: mapping m4)

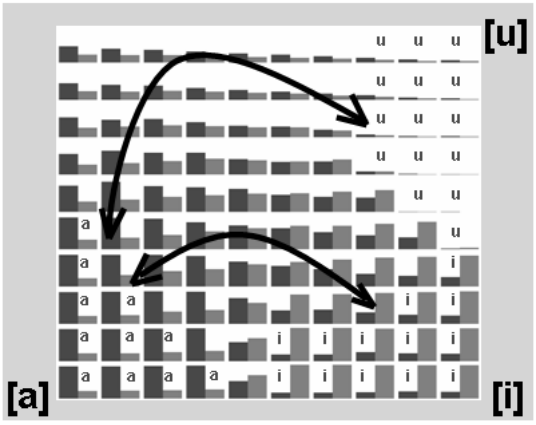

Figure 8. Display of the F1-F2-SOM-weights for each of the 10x10 neurons of the self-organizing map using bar charts. The arrows are described in the text (case: mapping m4).

In the case of the *auditory-to-somatosensory mapping* (m4, Fig. 1b) 45 somatosensory neurons plus 4 auditory neurons lead to 49 input neurons and thus to 4900 link weights which are adjusted during training. The training patterns and the number of training cycles are as described above for m1. Standard SOM learning algorithms were used with the parameter settings given above. This mapping does not exhibit the shortcomings stated above for m1. Stable proto-vocalic articulatory states were generated on the basis of the auditory representations (i.e. F1 and F2). Moreover the resulting mapping indicates typical features of SOM's: The topology of the two-dimensional self-organizing map reflects the topology of the training items within each two dimensional subspace of the input parameters (Fig. 7). The SOM-F1-F2-weights are also displayed – ordered with respect to the SOM-neurons – using bar charts (Fig. 8). It can be seen, that the corner-states (front-high, back-high and low or [i], [u], and [a]) indicated by the F1-F2-vectors (0, 1), (0, 0), and (1, 0.3) respectively are well represented within the 10x10-SOM. Additionally smooth transitions occur from [u] to [a] as well as from [i] to [a] (arrows in Fig. 8).

In the case of the direct *auditory-to-motor mapping* (m6, Fig. 1b) 40 motor neurons plus 4 auditory neurons lead to 44 input neurons and thus to 4400 link weights which are adjusted during training. The training patterns and the number of training cycles remain as stated above for m4. Standard SOM learning algorithms were used with the parameter settings given above. Stable proto-vocalic articulatory states were predicted very precisely on the basis of the auditory representation using this net (prediction error smaller than 1%). Also the

topology of the two-dimensional self-organizing map reflects the topology of the training items within each two dimensional subspace of the input parameters as given above for m4.

5. Learning proto-consonantal articulation

Proto-consonantal articulation is defined here as toddlers first efforts in producing articulatory movements during the babbling phase. Proto-consonantal articulation comprises closing as well as opening proto-gestures, i.e. first VC- and CV-like articulations. The *auditory representation* is given by the formant pattern – i.e. by the formant transitions of the first 3 formants occurring during the closing or opening gesture (Kröger et al. 2006a). The *motor representation* of the closing or opening gestures comprises three groups of parameters: (i) The closure forming or gesture executing end-articulator (i.e. labial, apical, or dorsal), (ii) the somatosensory (or high-level motor) representation of the proto-vowel target and of the proto-closure target defining the begin and end of the gesture, and (iii) parameters defining the duration and the degree of realization of the gesture. For the definition of the motor state of a gesture it is reasonable to use somatosensory representations and not low level motor representations – as done in the case of proto-vocalic articulation – since the motor control of dynamically defined gestures is less ambiguous on the level of tract variables than on the articulatory or lower motor level.

Figure 9. (i) Chart bar display of SOM link weights for the parameter "gesture executing end-articulator" of the motor representation (from left to right: "labial", "apical", "dorsal", see light gray bars) and (ii) display of formant patterns in Bark (see black trajectories) for the auditory representation after training of a 10x10 SOM for opening proto-gestures. The absolute duration of the gestures varies since articulator velocity was kept constant.

A 10x10 SOM was used for training the *auditory-to-motor mapping* in the case of opening proto-gestures. The training set for opening proto-gestures consists of 6x17 patterns combining 1 labial, 2 apical, and 3 dorsal proto-consonantal target configurations with 17 proto-vocalic starting configurations covering the whole proto-vocalic space (Fig. 5). Training was performed using standard SOM training parameter adjustments. The SOM weights of the auditory representation (i.e. the gestural formant transitions of F1, F2, and F3) and 3 SOM weights of the high-level motor representation – indicating the motor parameter "gesture executing end-articulator" – are displayed in Fig. 9. It can be seen, that the SOM is capable of separating the 3 basic types of gestures i.e. labial, apical, and dorsal gestures. Moreover the SOM is capable of predicting the motor representation of an opening gesture on the basis of its auditory representation.

6. Results and conclusions

The neural feedback loop of speech production can be modeled successfully using relatively simple neural representations of sensory and motor states. *One-layer feed-forward networks* are

used for modeling the somatosensory-to-motor mapping. *Self-organizing maps* are used for modeling the auditory-to-somatosensory as well as the auditory-to-motor mapping. Two results of this work are important: (i) The prediction of proto-vocalic motor states from auditory states is much more precise by using control model B (mapping m6, Fig. 1) since in model A mapping m1 is succeeded by mapping m3 indicating an overall mean error of around 10% for predicting a motor state. Thus a direct auditory-to-motor mapping is advantageous for modeling proto-vocalic articulation. (ii) Integrating tactile and proprioceptive information into one single somatosensory map (model B) is advantageous in the case of vocalic as well as consonantal articulation. In the case of proto-vocalic articulation vocal tract closures can be avoided within the auditory-to-somatosensory mapping. In the case of proto-gestural VC-articulation tactile information – i.e. labial, apical, dorsal closure – can be predicted easily from auditory information – i.e. formant transitions. All in all the model is now ready for building up a mental syllabary, i.e. for processing sounds, syllables and words of a specific language.

7. Acknowledgments

This work was supported in part by the German Research Council Grant KR1439/10-1 and JA1476/1-1.

This work is dedicated to *Professor Dr. Georg Heike*, Head of the former Department of Phonetics at the University of Cologne (Germany) until 1999. He directed my attention to natural speech synthesis and modeling of coarticulation in German. He always gave precious comments on my work and he always supported my work of developing articulatory speech synthesis during the past decades.

8. References

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

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

Guenther FH, Gjaja MN (1996) The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America* 100: 1111-1121

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

Kandel ER, Schwartz JH, Jessell TM (2000) *Principles of neural science.* MacGraw-Hill, New York

Kohonen T (2001) *Self-organizing maps.* Springer, Berlin, 3rd edition

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, Pittsburgh, Pennsylvania) pp. 565-568

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

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

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