Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer

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, RWTH Aachen, Germany
bkroeger@ukaachen.de, jim.kannampuzha@rwth-aachen.de, cneuschaefer@ukaachen.de

[2] Department of Computer Science, University of Rostock, Germany
piet@informatik.uni-rostock.de

ABSTRACT

A comprehensive neural model of speech motor control including a three dimensional articulatory speech synthesizer as a front-end device is described in detail in this paper. The training of the sensory-to-motor mappings – which can be interpreted as the prelinguistic phase of speech acquisition – is described in detail for quasi-static as well as for dynamic articulation.

Index Terms: speech production, neural model, articulatory model, articulatory speech synthesis, speech acquisition

1. INTRODUCTION

Modeling sensorimotor control of speech production is rare (cp. [1] and [2]). This may result from the complexity of cortical and subcortical representations as well as from the complexity of cortical and subcortical processing occurring in speech production. A comprehensive overview of the topology of the speech motor control network is given by [1]. Three cortical maps – i.e. a *speech sound map*, a *sensory map* and a *motor map* (articulatory velocity and position map in terms of [1]) – provide the linguistic, sensory, and motor representations of the speech sound, syllable, or word currently produced. The sensory map is subdivided in an auditory and a somatosensory state and an auditory and a somatosensory error map. Neurons of the speech sound map represent already learned sounds, syllables, or words. These neurons activate the motor plan of the appertaining sound, syllable, or word (feedforward control). Simultaniously the auditory and somatosensory representation of this item is coactivated and compared with the sensory representation currently produced by the articulatory-acoustic vocal tract model. If both representations differ, a sensory error signal is generated in order to correct the current motor plan (feedback control).

Two basic training phases can be separated for the speech motor control network. (i) During the *babbling phase* the sensory-to-motor mappings are trained on the basis of motor-to-sensory data generated by the front-end articulatory-acoustic model. (ii) During the *imitation phase* the sound-to-sensory mappings as well as the feedforward sound-to-motor mapping are trained by perceiving and reproducing sounds, syllables and words.

A crucial point within modeling sensorimotor control of speech production is the quality of the feedback signals produced by the articulatory-acoustic vocal tract model. Our three dimensional articulatory speech synthesizer [3] is capable of generating high quality acoustic and articulatory signals, which serve as a basis for auditory and somatosensory feedback signals. The modeling of *sensory-to-motor mappings* on the basis of this articulatory-acoustic vocal tract model is described in detail in this paper for static as well as for dynamic articulation.

2. THE FEEDBACK CONTROL NETWORK

The feedback control network comprises a sensory and a motor representation. The sensory representation comprises an *auditory* and a *somatosensory map* (Fig. 1). The motor map is subdivided in a *spatial coordinate* and a *joint coordinate motor map* in our approach. The mappings between the sensory maps and the spatial coordinate motor map – i.e. the *sensory-to-motor mappings* (Fig. 1) – are trained during the babbling phase. Within this phase random motor states are generated on the joint coordinate level, transferred into articulatory representations, and inputted to the articulatory vocal tract model. The appertaining acoustic and articulatory signals are forwarded to the subcortical auditory and somatosensory processing units and subsequently forwarded to the sensory maps (Fig. 1).

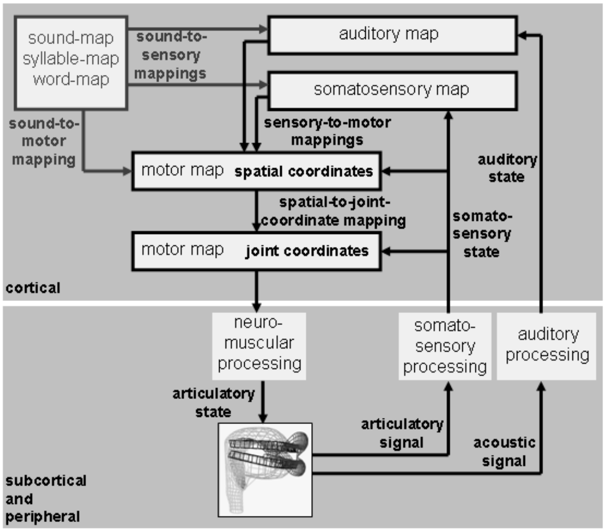
sound-map
syllable-map
word-map
sound-to-sensory mappings
auditory map
somatosensory map
sound-to-motor mapping
sensory-to-motor mappings
motor map spatial coordinates
auditory state
spatial-to-joint-coordinate mapping
somato-sensory state
motor map joint coordinates
cortical
neuro-muscular processing
somato-sensory processing
auditory processing
articulatory state
articulatory signal
acoustic signal
subcortical and peripheral

Figure 1: The neural model of speech production.

Each mapping within our network (see chapter 4, 5, and 6) is accomplished using a one-layer feed-forward network (cf. [4]). Logistic activation functions and identity output functions are used for modeling the output neurons. Additionally all output neurons are allowed to exhibit an activation threshold (bias) different from zero. Training was performed using the JAVA-version of the Stuttgart Neural Network Simulator SNNS [5].

3. THE SPEECH SYNTHESIZER AND THE FEEDBACK SIGNALS

Our three-dimensional speech synthesizer (i.e. articulatory-acoustic vocal tract model) [3] is controlled by a set of 10 articulatory parameters (Tab. 1). This set of parameters represents quasi-static articulatory states of all model articulators - i.e. lips, tongue, jaw, velum, and larynx (Fig. 2). 10 neurons of the joint coordinate motor map (Fig. 1) directly control the articulatory state. The acoustic model is driven by the vocal tract area function calculated from the geometrical data of the articulatory model for each articulatory state. The acoustic model is capable of generating vocal tract transfer functions as well as the acoustic speech signal.

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

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

Relative lip height means: lip height relative to jaw.

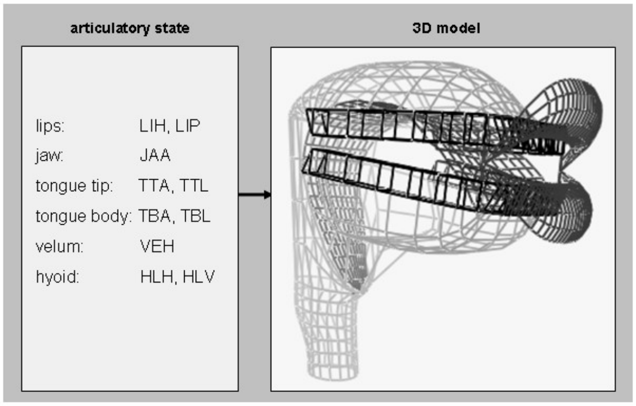

Figure 2: Articulatory parameters (for abbreviations see Tab. 1) and geometrical grid-representation of the 3D model.

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

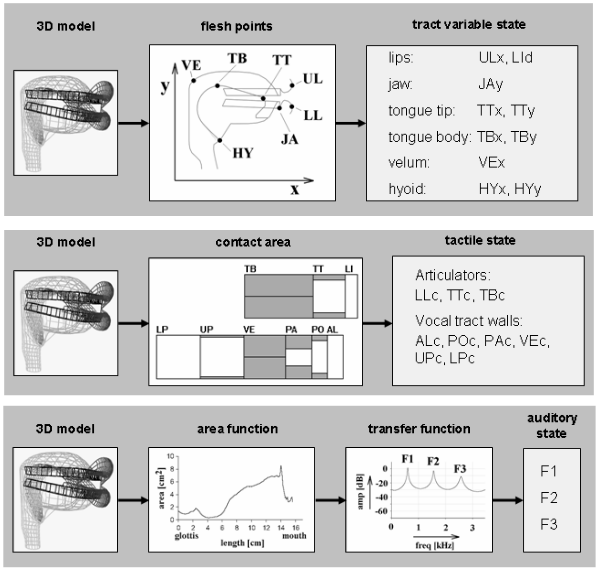

Figure 3: Generation of a tract variable, a tactile, and an auditory feedback signal (for abbreviations see Tab. 2 and Tab. 3)

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

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

Table 3: List of tactile parameters

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

4. THE SPATIAL-TO-JOINT-COORDNATE MAPPING

In the case of the *spatial-to-joint-coordinate mapping* a training set combining the minimum and maximum values of all 10 articulatory parameters was used [7]. 100000 cycles batch training were sufficient for obtaining a mean error of 9.1% for predicting the joint coordinate parameters of any articulatory state from the appertaining spatial coordinate pattern. This net is capable of modeling features of motor equivalence like maintaining labial, apical and dorsal vocal tract closure for different jaw positions (Fig. 6).

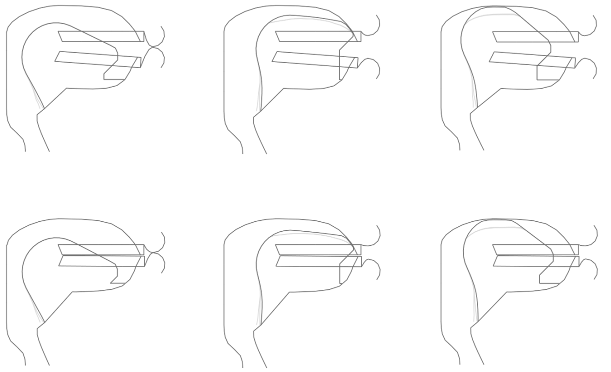

Figure 6: Production of labial, apical, and dorsal vocal tract closure using two different jaw positions (low and high). The same set of tract variables is used for each type of vocal tract closure with exception of the tract variable JAy. (light-gray lines: lateral tongue contours)

5. THE SENSORY-TO-MOTOR MAPPINGS FOR QUASI-STATIC ARTICULATION

In the case of the *tactile-to-motor mapping* training did not lead to feasible results using the min-max-combination training set described above. A constriction forming training set was shaped by using 5 labial, 5 apical, and 5 dorsal constriction forming articulatory states exhibiting an increasing degree of constriction up to full closure. Each set of these 15 constriction forming articulatory states is based on a set of 25 underlying vocalic articulations leading to a complete amount of 375 training patterns. The 25 underlying vocalic articulations form a subset of the vocalic training set, introduced below. 20000 cycles batch training were sufficient for obtaining a mean error of 8.4% for predicting the spatial coordinate parameters of any articulatory state from the appertaining tactile contact pattern. The mapping is capable of generating for example labial, apical, and dorsal constrictions or closures on the basis of tactile parameter settings.

In the case of the *auditory-to-motor mapping* the first goal was to learn *quasi-static vocalic articulation*. According to the problem of acoustic-to-articulatory inversion (e.g. [8]) the variety or vocalic articulatory states was constrained to the set of linear interpolations between three vocalic states within the tract variable space. For this purpose a *critical palatal, velar, and pharyngeal vocalic constriction forming state* was generated in advance, fulfilling the acoustic criterion of maximization of F2 for the palatal, minimization of F2 for the velar, and maximization of F1 for the pharyngeal vocalic constriction forming state. These three critical vocalic states can be interpreted as language-independent cardinal [i]-, [a]-, and [u]-articulations defined on the level of tract variables. The interpolation of vocalic states is parameterized by two tract variable parameters (TBx and Tby), representing the vocalic dimensions high-low and front-back. A vocalic training set was generated comprising 540 articulatory states concentrated at the edges of the vowel space defined by the three critical vocalic constriction forming states. The articulatory and acoustic-auditory states constituting this vocalic training set are given in Fig. 7. 20000 cycles batch training were sufficient for obtaining a mean error of 5.1% for predicting the spatial coordinate parameters (i.e. the tract variables) of any vocalic state from the appertaining formant pattern. The net is capable of generating a variety of vowel qualities (e.g. [i], [e], [ɛ], [a], [ɔ], [o], and [u]).

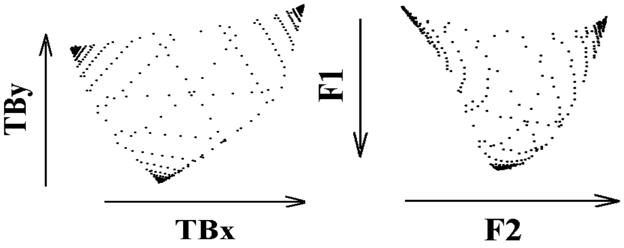

Figure 7: Articulatory and acoustic-auditory states forming the vocalic training set. Left side: articulatory vowel space (tract variables Tby vs. Tbx, relative values). Right side: acoustic-auditory vowel space (F1 vs. F2, relative values)

Thus the babbling phase of our neural model of speech production can be divided into (i) a *silent mouthing phase* for training the spatial-to-joint coordinate mapping and the tactile-to-motor mapping – since no auditory information is used here – followed by (ii) a *vocalic training phase* for training the auditory-to-motor mapping using quasi-static vocalic articulatory states. (iii) The next step is modeling a training phase for dynamic articulation, i.e. training the auditory-to-motor mapping using articulatory gestures (see below). It is important to realize, that this training phase is also part of the prelinguistic babbling phase. Therefore articulatory gestures used for prelinguistic training phases are labeled as *protogestures* or *raw gestures* in order to differentiate these gestures from phonologic or linguistic relevant gestures as defined in [9] and which are shaped later on during the imitation phase.

6. THE AUDITORY-TO-MOTOR-MAPPING FOR DYNAMIC ARTICULATION

So far the sensory-to-motor mappings describe quasi-static articulation. In the case of dynamic articulation auditory representations of protogestures are mapped to high-level motor representations of these gestures. This level of motor representation describes the planning of articulatory gestures in terms of spatial motor coordinate and temporal parameters (see below).

The *gestural auditory-to-motor mapping* comprises mappings for closing and opening protogestures (i.e. VC- and CV-syllables) as well as mappings for combinations of opening- and closing protogestures (i.e. VCV-sequences). The training sets comprise labial, apical, and dorsal protogestures based on different underlying vocalic states leading to simple (meaningless) CV-, VC-, or VCV-sequences.

As an example our preliminary modeling of the auditory-to-motor mapping for closing protogestures (VC-sequences) is described here. 10 underlying vocalic states distributed over the whole articulatory vowel space were selected from the vocalic training set. Each of these vocalic states serve as an articulatory basis for the execution of a proto-labial, -apical, and -dorsal closing gesture thus leading to a training set of 30 training patterns. 28 auditory and 6 motor neurons constitute the auditory and motor map in this case of dynamic articulation (Tab. 4 and Tab. 5).

The motor state description of a protogesture comprises (i) the spatial coordinates (TBx- and TBy-values) of the underlying vocalic state, (ii) the spatial coordinates of vocal tract closure (TBx and TBy-values for a dorsal, TTx- and TTy-values for apical, and ULx- and LId-values for labial protogestures), and (iii) the articulatory velocity for reaching the gestural target (i.e. closure). Articulatory velocity is defined by two factors: (a) distance of instantaneous articulator position from target and (b) by a factor GOpg < 1, describing the portion of the articulator-target distance, which the articulator covers during a definite time interval (i.e. the value of the volitional GO signal, cp. [4]).

Table 4: List of auditory parameters for closing gestures

ABBR.	NAME OF GESTURAL AUDITORY PARAMETERS
F1(ti), F2(ti), F3(ti)	formant values at 5 equidistant time instants ti during gestural transition
F1'(ti), F2'(ti), F3(ti)'	time differences of the same formant values for the 4 time intervals defined by ti
t_trans	time interval of formant transition

Table 5: List of motor parameters for closing gestures

ABBR.	NAME OF GESTURAL MOTOR PARAMETERS
TBx, TBy	static vocalic tract variables (see Tab. 2)
LAc, APc, DOc	type of closure: labial, apical, dorsal (value of neurons is 0 or 1)
GOpg	articulatory velocity of protogesture

The auditory state description of a protogesture is a direct pick up of the formant transition of F1, F2, and F3 by taking the bark-scaled formant values of F1, F2, F3 and their time derivatives at 5 equidistant time instants t0 to t4 covering the whole time interval from the begin of the closing gesture (t0, i.e. the vocalic state) up to vocal tract closure (t4, Fig. 8).

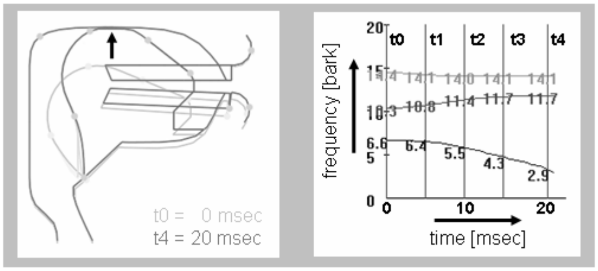

Figure 8: Spatial representation (left side) and auditory representation (right side) of a dorsal closing gesture.

The task of the gestural auditory-to-motor mapping is to predict a protogestural high-level motor description from its formant transitions. In the case of closing gestures 150000 cycles of batch training were sufficient for obtaining a mean error of 4.8% for predicting the type of closing gesture (i.e. dorsal, apical, or labial) from the formant pattern (Training and test set differed by number and position of underlying vocalic states).

7. DISCUSSION

A crucial problem in modeling speech motor control is modeling the sensory-to-motor mappings. Learning these mappings is unproblematic in our approach at least for two reasons: (i) High quality sensory signals are used based on the high quality of the articulatory and acoustic signals produced by our three dimensional articulatory speech synthesizer; (ii) a high-level motor representation, i.e. the spatial coordinate motor map is introduced. The importance of this level of representation within a model of speech motor control is underlined by the fact, that all our trials of training the auditory-to-motor or the tactile-to-motor net failed if directly the joint coordinate motor representation is used.

8. FURTHER WORK

The next step in enhancing our approach for modeling speech motor control is modeling the imitation phase, i.e. to train the sound-to-sensory as well as the sound-to-motor mapping. This is the the starting point for language-dependent training of sounds, syllables, and words.

9. ACKNOWLEDGEMENTS

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

10. REFERENCES

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

[2] Bailly G (1997) Learning to speak. Sensori-motor control of speech movements. *Speech Communication* 22, 251-267

[3] 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

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

[5] Stuttgart Neural Network Simulator. *http://www-ra.informatik.uni-tuebingen.de/SNNS/*

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

[7] Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2006) Spatial-to-joint coordinate mapping in a neural model of speech production. *Proceedings of DAGA 2006*, Braunschweig, Germany.

[8] Potard B, Laprie Y (2005) Using phonetic constraints in acoustic-to-articulatory inversion. *Proceedings of Interspeech, 9th European Conference on Speech Communication and Technology,* pp. 3217-3220

[9] Browman CP, Goldstein L (1992) Articulatory Phonology: An Overview. *Phonetica* 49, 155-180