Chapter 2

Phonemic, sensory, and motor representations in an action-based neurocomputational model of speech production (ACT)

Bernd J. Kröger, Peter Birkholz, and Anja Lowit

Abstract

Neural models of speech production are highly valuable for our understanding of normal as well as of disordered speech production processes. However, only few quantitative neural models of speech production are available. This paper introduces an action-based neurocomputational model of speech production (called ACT) comprising a high-quality articulatory-acoustic speech synthesizer as a front-end module. Some basic features of ACT are (i) the use of sensory information in motor control of speech production including feed-forward and feedback processing, (ii) a separation of motor planning and motor execution, which can be attained from using the concept of ACTion (concept of vocal tract action units or articulatory gestures), and (iii) the inclusion of principles of self-organization for the buildup of neural mappings for speech production during speech acquisition. The organization of ACT including sensory (i.e. somatosensory and auditory), motor, phonetic, and phonemic neural representations and the gathering of knowledge during learning or training stages of speech acquisition is described in this paper. Currently ACT is capable of producing speech items of a "model language" comprising a set of 65 CV- and CCV-syllables where C stands for voiced or voiceless plosives, nasals, or the apical lateral.

2.1. Introduction

Neural models of speech production describe (i) the *neurolinguistic processes* from conceptualization to phonological encoding and (ii) the *neurophonetic processes*, i.e. phonetic encoding, sensorimotor control, and the articulatory-acoustic realization of speech items. Within neurolinguistic models of speech production (e.g. Dell 1999; Levelt 1999), many neurophonetic aspects of speech production are not addressed in detail. A promising comprehensive quantitative neurophonetic model of speech production has been proposed by Guenther (DIVA model, see Guenther 2006; Guenther et al. 2006). In DIVA, feed-forward and feedback control is separated.

Feed-forward control starts with the activation of a phonemic state within the speech sound map for a speech item (i.e. sound, syllable, word, or utterance), subsequently activating virtual motor commands at the level of the articulatory velocity and position map. These motor commands directly control an articulatory-acoustic model which is capable of producing simulations of articulatory speech movements and acoustic speech signals. In DIVA, *feedback control* starts with the calculation of sensory (i.e. auditory and somatosensory) feedback signals on the basis of the articulatory-acoustic output of the articulatory-acoustic model. These sensory feedback signals are compared with stored sensory states for the speech item under production and in the case of noticeable differences between stored and current sensory signals appropriate error signals are calculated and corrective motor commands are issued to the articulators (Guenther et al. 2006).

The organization of the speech production model ACT introduced in this chapter is based on DIVA. However, it emphasizes three aspects: (i) The basic unit of sensorimotor control of speech articulation is the *vocal tract action unit* or *articulatory gesture* (Goldstein et al. 2006, 2007). Consequently a planning and execution stage can be separated for vocal tract movements, leading to two motor representations: on the one hand a *primary motor representation* directly defines activations of functional groups of muscles for each time instant and directly controls the execution of articulation, and on the other hand a *motor plan representation* defines motor plans for speech items (here: speech sounds, syllables, words, or short utterances) by specifying vocal tract action scores or gestural scores (Goldstein et al. 2006; Kröger and Birkholz 2007) on the planning level of the model. (ii) The phonemic to motor state and the phonemic to sensory state mappings introduced in the DIVA model are implemented in our model by using *self-organizing maps* (Kohonen 2001). Thus the phonemic to motor as well as the phonemic to sensory mappings are implemented by including a self-organizing map as central neural layer of these mappings. This central neural layer is called *phonetic map* in our approach. (iii) Since the phonemic to motor state mapping including the phonetic map implies the storage of motor states for each speech item, this neural processing path is assumed to represent the processing of *frequent* speech items (cp. concept of mental syllablary, Levelt and Wheeldon 1994; Levelt et al. 1999). In addition, infrequent speech items can be generated in ACT by using an alternative neural pathway via a *motor planning module*.

2.2. **Neural representations**

ACT differentiates neural maps, neural mappings (also called neural associations), and neural processing modules. *Neural processing modules* (see boxes without black outline in Fig. 2.1) are complex units that themselves comprise maps and mappings which currently are not specified in detail. It is the task of processing modules to transform *neural states* or *neural activation patterns* of one neural representation into states or activation patterns of other neural representations. *Neural maps* (indicated by black outlined boxes in Fig. 2.1) are ensembles of virtual neurons coding phonemic, phonetic, sensory, or motor states of speech items. *Neural mappings* (indicated by arrows in Fig. 2.1) represent ensembles of synaptic links between neurons in different neural maps. Within these mappings each neuron of one map is connected with each neuron of the other map (Kröger et al. 2006a, 2006b). The degree of connectivity per link is called *link weight* or *synaptic weight* and ranges from fully excitatory connection (synaptic link weight value = 1) to fully inhibitory connection (synaptic link weight value = −1). Link weights or synaptic weights are adjusted during the learning or training stages, i.e. during the babbling and the imitation stage of speech acquisition in the current model. ACT comprises self-organizing maps and mappings and makes use of non-supervised learning (Kohonen 2001) which models early phases of speech acquisition, i.e. babbling and imitation training (see Section 2.2.3).

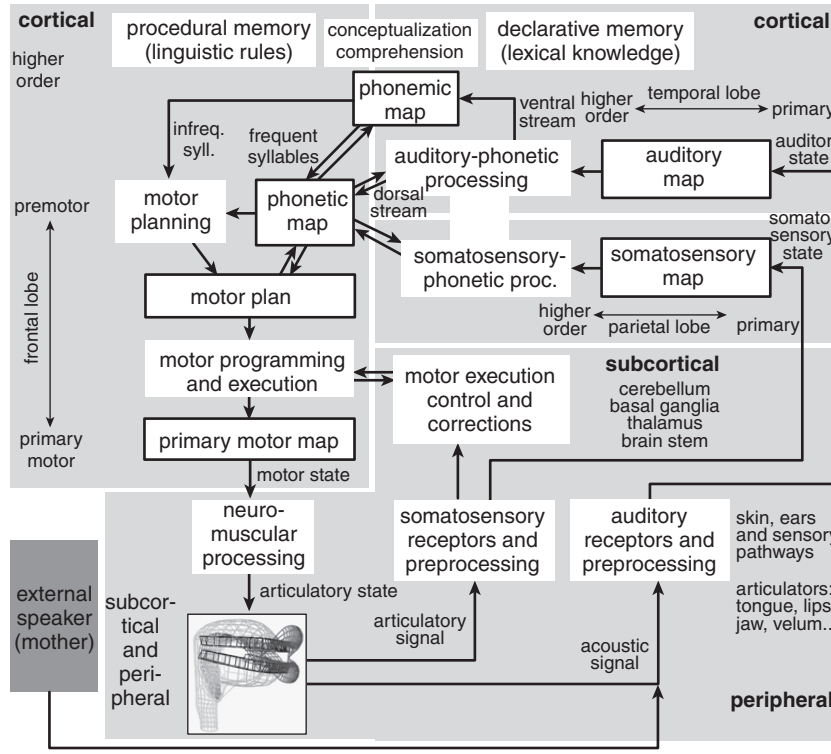

Fig. 2.1 Organization of the quantitative computer-implemented neural model of speech production. Boxes with black outline: neural maps; boxes without black outline: neural processing units comprising neural maps and mappings not specified in detail; arrows: neural mappings, associations, or projections. The associations of the phonemic map, sensory maps, and motor programme map via the phonetic map are bidirectional, leading to a multidirectional activation of phonemic, sensory, and motor plan states.

2.2.1. **Primary motor state representation and articulatory state representation**

The three-dimensional articulatory-acoustic vocal tract model (VTM) (Fig. 2.2, Birkholz et al. 2006; Birkholz and Kröger 2006, 2007) is controlled for each time instant by a set of *articulatory parameters*. These parameters define the location of articulators relative to other articulators, i.e. the location of lower lips relative to jaw, the location of tongue body relative to jaw, and the location of tongue tip relative to tongue body (Kröger et al. 2006b, 2006c). The set of model articulators in our three-dimensional VTM are lower jaw, tongue body, tongue tip, velum, lips, hyoid, and glottis. Hyoid position directly defines the vertical position of the larynx and the vertical and horizontal position of the tongue root. Phonatory parameters (i.e. arytenoid lateral position for controlling glottal aperture and lung volume for controlling subglottal pressure) are part of this set of articulatory parameters. While the position of velum and lower jaw are controlled by one parameter (one degree of freedom: vertical displacement), all other articulators are controlled by two parameters reflecting two degrees of freedom for displacement of the articulator relative to other articulators. The articulators are controlled in our model by a set of *primary motor parameters*. These primary motor parameters transform articulatory parameters into velocity and

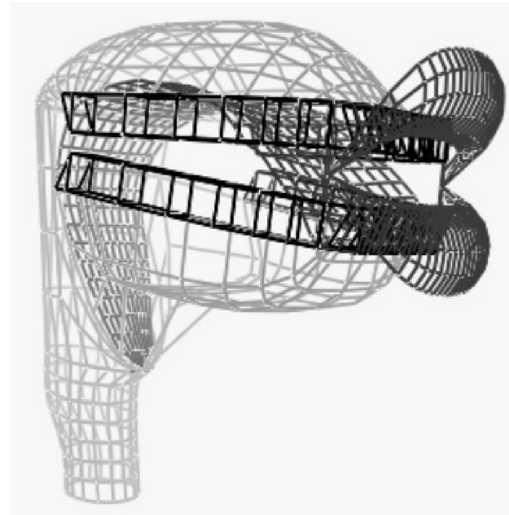

Fig. 2.2 The geometrical grid-representation of the three-dimensional articulatory-acoustic vocal tract model (VTM).

position specifications for each degree of freedom for each articulator by specifying a neural primary motor activation pattern for each point-in-time. With respect to the task dynamics concept (Saltzman and Munhall 1989) the primary motor parameters as well as the articulatory parameters are related to the task dynamics body articulator space while task-space specific parameters are represented at the motor plan level in ACT (see below).

2.2.2. **Sensory state representations**

Somatosensory information is gained from the articulator positions and from the contact areas quantified by the VTM. The somatosensory neural activation pattern (or somatosensory state) on the level of the *somatosensory map* (Fig. 2.1) represents the proprioceptive and tactile feedback state at the current point-in-time. A proprioceptive state is characterized by the specification of all proprioceptive model parameters, i.e. by the neural specification of the absolute position of all articulators with respect to the craniofacial reference system (i.e. with respect to the scull or hard palate, Kröger et al. 2006b). The absolute positioning of articulators is comparable to tract variable or task space information in terms of the task dynamics approach (Saltzman and Muhnall 1989) and can be calculated in our approach from the set of articulatory parameters for each point-in-time. A tactile state is characterized by the specification of a set of tactile model parameters, e.g. contact area of articulators (e.g. lips, tongue tip, tongue body) and contact area of vocal tract walls (e.g. alveolar ridge, hard palate, soft palate, Kröger et al. 2006a, 2006b). This contact area can be specified for each point-in-time directly within the VTM.

The auditory neural activation pattern is calculated from the acoustic transfer function of the VTM. The resulting frequency values of the first three formants F1, F2, and F3 are extracted from the transfer function by using a peak picking algorithm. The resulting formant values are bark-scaled and forwarded in steps of 10 ms towards the *auditory map* (Fig. 2.1).

The sensory (somatosensory and auditory) states represent the feedback sensation (or the auditory sensation of external speech signals) at the current point-in-time at the level of the somatosensory and auditory maps (Fig. 2.1). These point-in-time sensory states are then stored in

a short-term memory in syllable-sized chunks at the level of the sensory-phonetic processing modules (Fig. 2.1). Thus point-in-time and period-in-time sensory states are differentiated in ACT.

2.2.3. **Phonemic state representation**

At the level of the *phonemic map* each neuron represents a phoneme (single sound neurons) or a frequent syllable (syllable neurons) of the *model language*, i.e. the language which is trained by ACT. The model language comprises 5 vowel phonemes (/i/, /e/, /a/, /o/, and /u/), 45 CV-syllables (with C = /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, and /l/), and 20 CCV-syllables (with C1 = /b/, /g/, /p/, /k/ and C2 = /l/, i.e. with consonant clusters /bl/, /gl/, /pl/, and /kl/). Thus the model language comprises 65 phonemic representations for syllables in total. Two of these 65 syllables (i.e. /lo/ and /ble/) are defined to be infrequent in our model language and thus are not trained during the imitation training phase of speech acquisition. During the production of a speech item, the neural activation occurring within the phonemic map causes an activation of motor plans via the phonetic map for frequent syllables (neural pathway for frequent syllables). Thus, in the case of frequent syllables, the activation of a syllable neuron within the phonemic map initiates the activation of the corresponding motor plan. In the case of infrequent syllables, the activation pattern within the phonemic map (activation of one to three single sound neurons) leads to the co-activation of neurons within the phonetic map that represent phonetically similar syllables. For example, in the case of /le/, all neurons within the phonetic map are activated which represent /lV/-syllables. In addition the co-activation of gesture neurons within the gesture map (a co-activation which is initiated by the single sound neurons of the phonemic map) directly specifies the set of gestures which is needed for assembling the motor plan of the syllable on the level of the motor planning module. The detailed specification of inter-gestural timing is then gained from the motor plans of the phonetically similar syllables which are activated at the level of the phonetic map (see arrow between phonetic map and motor planning module in Fig. 2.1).

2.2.4. **Motor plan state representation**

The motor plan state for each speech item needs (i) the specification of a set of vocalic, consonantal, velopharyngeal, and glottal vocal tract action units, assembling the speech item and (ii) the specification of all spatial and temporal characteristics for each action unit, i.e. end-articulator, target, gestural rapidity, duration of gesture, and inter-gestural temporal coordination (see Kröger and Birkholz 2007). ACT uses vocal tract action units in a bivalent way (i) as sensorimotor control units and (ii) as phonological units. Browman and Goldstein (1989, 1992) described gestures as phonological units and as units of dynamically defined control regimes. The concept of actions as sensorimotor control units is based on Bernstein (1967) and adapted for speech movements by Saltzman and Kelso (1983), Saltzman (1985), Kelso et al. (1986), and Saltzman and Munhall (1989). It is very important to note that the goal of actions in general is *functional* (Bernstein 1967). A functional goal in the case of speech gestures is to produce perceptual relevant distinct cues for discriminating and identifying phonological units. This elucidates the phonological character of vocal tract actions and is thus in agreement with Browman and Goldstein's notion of gestures. For a current characterization of vocal tract action units or articulatory gestures see also Goldstein et al. (2006, 2007).

The *motor plan map* is an ensemble of neurons that specifies the motor plan or vocal tract action score (also called gestural score) of a speech item. The *vocal tract action score* or *gestural score* not only specifies the set of vocal tract actions or gestures needed for the production of a speech item, but also specifies all *intra-gestural parameters* and the *parameters for inter-gestural temporal coordination* (*inter-gestural timing parameters*). A vocal tract action or gesture defines a

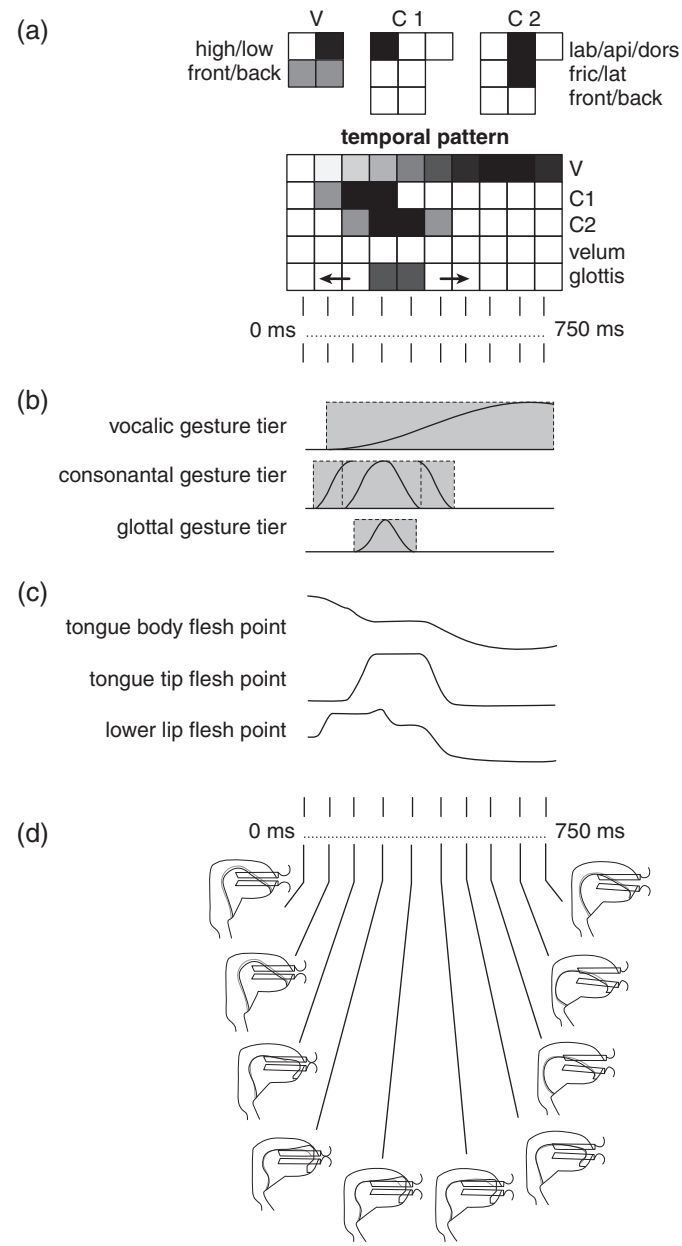

Fig. 2.3 Realization of the syllable /pla/: (a) neural activation pattern of motor plan, (b) plot of degree of realization for gestures ordered with respect to three gestural tiers, (c) plot of vertical displacement of model articulator flesh points, and (d) temporal sequence of 10 vocal tract states for a realization of the syllable /pla/. The *neural activation pattern* comprises four neuron clusters, i.e. a cluster representing the *temporal pattern* for the speech item (V-, CV-, or CCV-item) and three neuron clusters representing the vocalic and the two consonantal gestures (V, C1, C2). Each box within a neuron cluster represents a neuron. Each neuron can be activated on a scale from no activation (white) to full activation (black). The neuron cluster for the temporal pattern comprises

Fig. 2.3 (*Cont.*) rows representing the mean degree of realization for the vocalic (V), consonantal (C1, C2), velopharyngeal, and glottal gestures within distinct time intervals. The columns represent a time series of 75 ms intervals (here 750 ms in total). Arrows within the temporal pattern indicate directions for possible variations of intergestural temporal coordination and for possible variation of gestural length (exemplified here for the glottal opening gesture). The neuron cluster representing the vocalic gesture comprises four neurons representing the vocalic target dimensions high/low and front/back. The remaining neuron clusters for the consonants represent the gesture performing vocal tract organ (labial, apical, dorsal), the type of closure (fricative or lateral), and in the case of the fricatives the exact place of articulation (front or back). The degree of realization for all gestures displayed in the temporal pattern (a) is identical with the curves for degree of realization for each gesture displayed in (b) but the temporal resolution is better in (b). The flesh-point displacements shown in (c) indicate the resulting gestural movement patterns generated for different model end-articulators. The temporal succession of 10 vocal tract states in (d) represents the model vocal tract configuration in the middle of time intervals represented by the ten columns of the temporal pat-tern (see vertical lines in (a) and (d)).

goal-directed movement of an ensemble of articulators (e.g. lower jaw, lower lips, and upper lips in the case of a labial closing gesture, cp. Saltzman and Munhall 1989; Browman and Goldstein 1989, 1992; Goldstein et al. 2006, 2007). This action goal is functionally defined leading to discrete categories such as 'produce a bilabial closure'. In addition, the functionally defined goal includes temporal aspects, e.g. 'produce a closure for a voiced plosive realization with appropriate rapidity (i.e. not too slowly in order to separate it from an approximant realization)' or 'produce the bilabial closure for a voiced plosive realization with appropriate duration of closure (i.e. not too long in order to separate it sufficiently from a voiceless plosive realization)'. (Intra-)gestural parameters or vocal tract action parameters which must be specified for each gesture or vocal tract action are the gesture performing vocal tract organ (or gesture performing end-articulator), gestural target, gestural rapidity, and gestural duration. Gestural rapidity together with current gestural duration specifies the current degree of realization of a gesture (Kröger and Birkholz 2007). In addition, parameters for inter-gestural temporal coordination specify the temporal start and ending of each gesture with respect to the temporal location of each other gesture. An example of a *neural motor plan activation pattern* is presented in Fig. 2.3. A neuron's activation level within the *temporal pattern* specifies the degree of realization for the gesture represented by that neuron at that particular time instant. Thus, the temporal pattern displays the points in time for the start and ending of gestures as well as the degree of realization in between.

Gestures are organized on *rows* with respect to the type of gesture, i.e. on the vocalic (tract shaping), consonantal, velopharyngeal, and glottal row (Goldstein and Fowler 2003). Intergestural coordination can be specified by phasing rules in articulatory phonology (Browman and Goldstein 1992) or by using coupling graphs (Goldstein et al. 2006). It should be noted that the temporal coordination is not specified by rules in ACT but results from neural learning or training during the imitation phase of speech acquisition. After neural learning or training the resulting neural motor plan activation patterns (such as that in Fig. 2.3) are stored for each frequent syllable by the link weights of the phonetic to motor plan mapping.

2.2.5. **Phonetic state representation**

The *phonetic map* (Fig. 2.1) is a neural self-organizing map as defined by Kohonen (2001). Each neuron of this map represents a phonetic state by linking a sensory and motor plan state. Some phonetic states are linked with phonemic states in order to tag phonetic realizations of phonemic states (i.e. vowels or frequent syllables). The detailed organization of the phonetic map emerges during babbling and imitation training (see below). The phonetic map is defined here

typically as a neural map comprising 25 × 25 (= 625) neurons. The size of the map has been varied between 10 × 10 to 30 × 30 maps without indicating substantial changes with respect to the organization and to the phonetic features of the map. Thus we used a map size which on the one hand is able to include sufficient phonetic detail but which on the other hand limits computational processing time in practicable limits (i.e. lower than 2 hours for processing a specific set of training data). The feature of phonetotopy occurring in phonetic maps is discussed in detail in Kröger et al. (2008).

2.3. **Training the model: modelling early stages of speech acquisition – babbling and imitation**

The main goal of training is to enable the model to produce speech items of a (simple) 'model language' comprising V(owel)-items and one- or two-syllabic words composed of CV- and CCV-syllables, with C(onsonant) as defined in Section 2.2.3. Training comprises *frequent syllables*, which are therefore also labelled as 'well practiced', or 'overlearned'. *Infrequent syllables* are defined to be produced without training. Other syllable structures such as CVC can be produced as well by using ACT but are not included in this chapter. While the structure of the maps is defined a priori in our model, training leads to a specification of all link weights quantifying a neural mapping. This specification of link weights allows the association of neural states of different maps connected by a mapping.

2.3.1. **Babbling training**

Method: It is assumed that a set of pre-linguistic proto-speech items (proto-vowels and proto-syllables) can be produced by ACT directly at the beginning of speech acquisition training. This set of proto-gestures used for babbling training comprises (i) proto vocalic (labial and tongue body) vocal tract forming gestures, e.g. a labial rounding or spreading gesture, or a tongue body front raising, back raising or lowering gesture; (ii) proto consonantal (labial, tongue body, or tongue tip) vocal tract closing gestures; (iii) proto glottal opening and closing gestures for devoicing and voicing; (iv) proto velopharyngeal opening and closing gestures for production of nasality or no nasality; and (v) proto pulmonary gestures for providing the vocal tract system with sufficient subglottal pressure and airflow for speech-like phonation. Length and temporal coordination of proto-gestures is varied, leading to a training set comprising round about 1300 babbling training items (see Table 2.1). Training is performed by taking training items (motor plan states) from this training set in random order, calculating the appropriate sensory states. If the succession of motor plan state and appropriate sensory states are activated at the level of the motor plan map and sensory maps, the link weights of the appropriate mappings connecting the motor plan map and the sensory map with the phonetic map, i.e. the link weights of the sensory to phonetic and motor plan to phonetic mapping are continuously adjusted. The mathematical procedure for adjusting link weights for these mapping in each training step (i.e. with each training item) and thus the self-organization of the phonetic map is described in Kröger et al. (in press) for a smaller model language comprising only CV-syllables where C are voiced plosives.

Results: About 2,600,000 training steps were sufficient in order to reproduce motor plan states from a specific sensory state with a mean error below 2%. This performance value was calculated by comparing the neural motor plan state calculated from the sensory to motor plan mapping with the initial motor plan state chosen for testing the performance of the network. On the level of the phonetic map the self-organization of phonetic states, resulting from unsupervised learning, indicates that all proto speech items defining the babbling training set are ordered with respect

Table 2.1 Organization of the babbling training set.

Subset	variation #1	variation #2	sum
proto-V (1)	Target of vocalic gesture within [i]-[a]-[u]-continuum (150 × 1 = **150**)		**150**
proto-CV with C = voiced proto-plosives (lab, api, dors) or lateral (4)	Target of vocalic gesture within [i]-[a]-[u]-continuum (50 × 4 = **200**)	Length of consonantal closing gesture and temporal coordination of vocalic and consonantal gesture (15 × 4 = **60**)	**260**
proto-CV with C = proto- nasals (lab, api) (2)	Target of vocalic gesture within [i]-[a]-[u]-continuum (50 × 2 = **100**)	Length of consonantal closing gesture and temporal coordination of consonantal closing and velopharyngeal gesture (15 × 2 = **30**)	**130**
proto-CV with C = voiceless proto-plosives (lab, api, dors) (3)	Target of vocalic gesture within [i]-[a]-[u]-continuum (50 × 3 = **150**)	Length of consonantal closing gesture and temporal coordination of consonantal closing and glottal gesture (15 × 3 = **45**)	**195**
proto-CCV with C1 = voiced proto-plosives (lab, dors), C2 = /l/ (2)	Target of vocalic gesture within [i]-[a]-[u]-continuum (50 × 2 = **100**)	Length of both consonantal closing gesture and temporal coordination of both consonantal closing gestures (81 × 2 = **162**)	**262**
proto-CCV with C1 = voiceless proto-plosives (lab, dors), C2 = /l/ (2)	Target of vocalic gesture within [i]-[a]-[u]-continuum (50 × 2 = **100**)	Length of both consonantal closing gesture and temporal coordination of both consonantal closing and glottal gesture (90 × 2 = **180**)	**280**
Total	**800**	**477**	**1277**

This training set comprises five subsets which each are organized with respect to up to two criteria of variation. The number in brackets indicates the number of subset items (column 1) or the number of training items per variation (bold) = (number of variations) × (number or subset times). The sum of all training items per subset (covering all variations) is listed in the last column. Abbreviations: lab = labial; api = apical; dors = dorsal.

to three main groups: proto V items, proto CV items, and proto CCV items. Moreover, proto CV and proto CCV items form subgroups with respect to manner, place, and voicing of initial consonant. The phonetic map at this stage of babbling comprises phonetic states with gestural targets representing the whole [i]-[a]-[u]-continuum in the case of vocalic gestures, a continuum of closure duration in the case of consonantal gestures, and a continuum of inter-gestural timing relations for all proto [V], proto [CV], and proto [CCV] items. This 'grossness', 'roughness', or 'non-distinctiveness' in spatial and temporal specification during babbling is the basis for training language-specific gestures and language-specific speech items. After babbling, the intra-gestural parameters of proto gestures and the inter-gestural temporal coordination of proto speech items become 'fine-tuned' during imitation training due to the demands of a specific language.

2.3.2. **Imitation training**

The driving force for babbling training is to play around with one's own vocal tract organs, to perceive the occurring articulatory and acoustic results produced by this vocal tract and to make associations between articulation (motor plan states) and sensation (auditory and somatosensory states). The driving force of imitation training is to reproduce language-specific external speech items (e.g. words produced by an external speaker) perceived by the model. During babbling

training the model already developed knowledge for sensorimotor relations and is thus capable of using its knowledge for activating motor plans for imitating an externally produced auditory state. The model then fine-tunes this motor plan during a series of trial and error productions of the speech item in order to enable the reproduction of the external speech item satisfactorily.

Method: The set of imitation training speech items comprised 3 realizations for each of the 68 frequent speech items, i.e. for the 5 vowels, for the 44 CV syllables (/lo/ is defined to be infrequent, see Section 2.2.3), and for the 19 CCV syllables (/ble/ is defined to be infrequent, see Section 2.2.3) introduced above. The imitation training set thus comprised 204 training items, produced by a male speaker of Standard German. It should be noted that imitation training also included the notion of the phonemic state of the external speech items since all speech items trained during imitation were language-specific.

Results: In total 10 cycles of training leading to 2.040 training steps were sufficient in order to obtain link weights for each neuron of the phonemic map to at least one neuron within the phonetic map indicating a strong neural association (i.e. a link weight values > 0.8). Each training item comprised a motor plan, a sensory, and a phonemic state. Imitation training started with the already trained phonetic map resulting from babbling training. The ordering of phonetic states within the (self-organizing) phonetic map already established during babbling training remains stable during imitation training. Thus the model is now capable of producing each frequent V-item, CV-item, and CCV-item of the model language.

2.4. **The functional processes and the organization of ACT**

Following the DIVA model (Guenther 2006; Guenther et al. 2006), ACT comprises feed-forward and feedback control (Fig. 2.1). *Feed-forward control* in ACT starts with the phonological representation of a speech item under production (e.g. vowel, syllable, word, or utterance) activating a single or a succession of specific phonemic states at the level of the *phonemic map*. Each syllable within the model language, which is defined as frequent (see Section 2.3.3), is represented by a single 'neuron' at the level of the phonemic map. The activation of a syllable on the level of the phonemic map is called phonemic state or phonemic activation pattern within the phonemic map and leads to a co-activation of an appropriate motor plan state, auditory state, and somatosensory state (see Section 2.2). The motor plan and sensory states for frequent syllables are trained or learned during babbling and imitation training (see Section 2.3) and are stored by an ensemble of link weights quantifying the phonetic to phonemic, phonetic to motor plan, and the phonetic to sensory neural associations (and vice versa), i.e. the mappings between the *phonemic map*, the *motor plan map*, and the *sensory maps* (*auditory and somatosensory map*) via the *phonetic map* (Fig. 2.1). The motor plan of infrequent syllables is assembled via the *motor planning module* (Fig. 2.1) by adapting specifications for gestures and for the inter-gestural temporal coordination from phonetically similar frequent syllables.

The execution (or programming and execution) of each vocal tract action unit by articulators is defined by the motor programming and motor execution module. Programming and execution depends on the context of vocal tract actions, i.e. on the specification of preceding, following, and temporal overlapping actions, since actions share common articulators (e.g. the jaw in the case of overlapping vocalic and consonantal labial, apical, or dorsal actions) and thus often influence each other (Kröger et al. 2006c). Programming and execution of the motor plan of a speech item leads to an activation of a time series of *primary motor states* and subsequently to a time series of *articulatory* (*including phonatory*) *states* – i.e. a specification of articulatory movements for all model

articulators carrying out the appropriate speech item. Thus feed-forward control starts with the discrete phonemic state and proceeds via motor planning to motor programming and motor execution.

A time series (or temporal succession) of *articulatory states* is input to the VTM which generates an articulatory and acoustic speech signal. The articulatory signal comprises the three-dimensional surface of all model articulators (lips, lower jaw, tongue, velum, Birkholz et al. 2006). The acoustic signal results from simulation of the vocal tract acoustics in the time-domain by using a transmission line circuit analog (Birkholz et al. 2007). Both signals are used for calculating somatosensory and auditory feedback signals in the *somatosensory and auditory preprocessing modules* (Fig. 2.1).

Feedback control in ACT starts with somatosensory (proprioceptive and tactile) and auditory signals. Proprioceptive muscle-related signals are processed by the *motor execution control* and *correction module* (Fig. 2.1). Furthermore somatosensory (proprioceptive and tactile) and auditory information is processed by the *somatosensory-phonetic* and *auditory-phonetic processing modules* (Fig. 2.1). For the duration of the syllable under production the time course of somatosensory information is stored in short-term memory (part of the somatosensory-phonetic processing module) and subsequently this information can be compared with the learned somatosensory state of the syllable. In parallel, the acoustic signal produced by the VTM is transformed into an auditory activation pattern for each time instant at the level of the *auditory map* by the *auditory preprocessing module* (Fig. 2.1). Furthermore/moreover, this time course of the auditory activation pattern (in the auditory map) is stored in short-term memory on the level of the auditory-phonetic processing module. If current sensory (i.e. auditory and somatosensory) states of the speech item (feedback sensory states) deviate from the sensory state for this speech item activated via the phonetic to sensory mappings (i.e. sensory states activated during feed-forward control), an error signal is generated within the auditory-phonetic or somatosensory-phonetic processing module. This error signal is projected to the phonetic map for generating corrective motor plan information (cf. Guenther 2006; Guenther et al. 2006).

Acoustic information of external speakers is processed via the same auditory signal pathway as is used for feedback control (Fig. 2.1). It is also possible to compare auditory activation patterns produced by external signals with learned auditory states of syllables, e.g. in order to tune motor plans of these syllables during imitation training (see Section 2.3.2).

2.5. **Discussion**

This chapter outlines a quantitative computer-implemented action-based neural model of speech production, called ACT. The chapter in addition describes training the model for a model language comprising a specific set of vowels CV-, and CCV-syllables. This training represents early stages of model-based speech acquisition (babbling and imitation stage). The model mainly focuses on the sensorimotor control part of speech production starting from a phonological description of the speech item under production. Following Guenther (2006) and Guenther et al. (2006) the model described here comprises feed-forward and feedback control subsystems. In addition a separation of motor planning and motor execution occurs in ACT as a result of introducing the *concept of action* in speech production which has been suggested as a control concept for speech production (e.g. Kelso et al. 1986; Saltzman and Munhall 1989; Goldstein et al. 2006, 2007). Furthermore, a dual route concept for separating the planning of frequent and infrequent syllables as was suggested, e.g. by Levelt et al. (1999) is part of this model.

During the babbling stage (Oller et al. 1999), the sensory to phonetic and the phonetic to motor plan mappings are trained on a variety of proto-speech items, comprising a continuum of

proto-vocalic and proto-syllabic states. Each neuron within the phonetic map represents a realization of these proto-speech items. Self-organization (Kohonen 2001) causes these proto-speech items to be ordered with respect to phonetic features such as low-high, front-back in the case of proto-vowels, and with respect to phonetic features such as place and manner of articulation in the case of proto-CV-syllables and with respect to different types of proto-syllables (CV vs. CCV). After babbling training, ACT is capable of generating motor plans for these proto-speech items from auditory and/or somatosensory information.

While babbling training is based on self-productions, imitation training is driven by externally produced speech items (speech items produced by mother or caregiver). The goal of babbling training was to associate sensory and motor plan states from self-productions. The goal of imitation training is to imitate external meaningful speech items (such as words) or parts of these items (such as syllables) as good as possible in order to associate phonemic, sensory, and motor plan states. Babbling training is an important precursor of imitation training since the model speaker after babbling training already is capable of activating motor plans from sensory states since sensory-to-motor-inversion knowledge and thus auditory-to-articulation-inversion knowledge is available. This capability is required during imitation training since only the auditory states of external speech items are available. During imitation training, in addition, neural associations arise between the phonetic and the phonemic map. Thus after imitation training, phonetic realizations of frequent language specific speech items (e.g. syllables) are each represented by a subgroup of neurons within the phonetic map.

In the case of this modelling study 1,300 babbling training items and 204 imitation training items representing 68 frequent speech items (V, CV, and CCV), forming the basis of a so-called model language were trained. The high amount of 2,600,000 babbling training steps results from the fact that all temporal and spatial combinations of proto-gesture were trained. It is more realistic to introduce imitation training already in parallel to babbling training (cf. Oller et al. 1999) in order to reduce the amount of babbling training items and to focus on important subsets or sub-dimensions of temporal coordination of actions occurring in the model language under acquisition. This on the one hand leads to a reduction of babbling training steps but on the other hand leads to an increase in imitation training trials since imitation may be imperfect if babbling knowledge (i.e. auditory-to-articulatory inversion knowledge) is incomplete.

The separation of motor planning on the one hand and motor programming and execution on the other hand results from the concept of vocal tract actions, which have to be arranged in vocal tract action scores. This separation is very advantageous for the modelling of motor speech disorders such as apraxia of speech (AOS) or dysarthria, as well as fluency disorders (Van Lieshout et al. 1996). While motor planning is a central concept in ACT, motor programming and motor execution are not differentiated in our model.

A dual neural pathway approach (Levelt et al. 1999; Varley and Whiteside 2001) is assumed for generating motor plans for frequent and infrequent syllables in ACT. The pathway for frequent syllables accesses the *mental syllabary* in Levelt's approach (Levelt and Wheeldon 1994; Levelt et al. 1999; Indefrey and Levelt 2004) or the phonetic map in ACT, respectively. In parallel to the Levelt approach, in ACT motor plans of frequent syllables are stored as a whole in the phonetic to motor plan mapping. The pathway for infrequent syllables, also called *subsyllabic encoding route* (Levelt et al. 1999; Varley and Whiteside 2001), is assumed to assemble motor plans from smaller subsyllabic units. While these subsyllabic units are assumed to be segmental in the approaches of Levelt et al. (1999) and Varley and Whiteside (2001), ACT assumes a subsyllabic assembly by vocal tract action units. It should be noted that infrequent syllables are assembled in the motor planning module of ACT by looking for phonetically similar frequent syllables from the phonetic map (see also Section 2.3.3 of this chapter). In contrast to Varley and

Whiteside (2001), ACT does *not* assume an independence of frequent and infrequent pathways (see arrow from phonetic map to motor planning map in Fig. 2.1). A quantitative spell-out of the motor programming module and the production of frequent and infrequent syllables are topics of our current research activities.

Acknowledgment

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

References

Aichert I, Ziegler W (2004) Syllable frequency and syllable structure in apraxia of speech. *Brain and Language* 88: 148–59

Bernstein N (1967) *The Coordination and Regulation of Movement*. (Pergamon, London)

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. 873–6

Birkholz P, Jackel D, Kröger BJ (2007) Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech, and Language Processing 15: 1218–25

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

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

Browman C, Goldstein L (1989) Articulatory gestures as phonological units. *Phonology* 6: 201–51

Browman CP, Goldstein L (1992) Articulatory phonology: an overview. *Phonetica* 49: 155–80

Dell GS, Chang F, Griffin ZM (1999) Connectionist models of language production: lexical access and grammatical encoding. *Cognitive Science* 23: 517–41

Goldstein L, Fowler CA (2003) Articulatory phonology: a phonology for public language use. In: Schiller NO, Meyer AS (eds) *Phonetics and Phonology in Language Comprehension and Production*. (Mouton de Gruyter, Berlin, NewYork), pp. 159–207

Goldstein L, Byrd D, Saltzman E (2006) The role of vocal tract action units in understanding the evolution of phonology. In: Arbib MA (ed.) *Action to Language via the Mirror Neuron System*. (Cambridge University Press, Cambridge), pp. 215–49

Goldstein L, Pouplier M, Chen L, Saltzman L, Byrd D (2007) Dynamic action units slip in speech production errors. *Cognition* 103: 386–412

Guenther FH (2006) Cortical interaction underlying the production of speech sounds. *Journal of Communication Disorders* 39: 350–65

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

Indefrey P, Levelt WJM (2004) The spatial and temporal signatures of word production components. *Cognition* 92: 101–44

Kelso J, Saltzman E, Tuller B (1986) The dynamical perspective on speech production: data and theory. *Journal of Phonetics* 14: 29–59

Kohonen T (2001) *Self-organizing Maps*. (Springer, Berlin, New York)

Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2006a) Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer. Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP), pp. 565–8

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

Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2006c) Spatial-to-joint coordinate mapping in a neural model of speech production. DAGA-Proceedings of the Annual Meeting of the German Acoustical Society (DEGA, Braunschweig, Germany), pp. 561–2 (or see http://www.speechtrainer.eu)

Kröger BJ, Birkholz P (2007) A gesture-based concept for speech movement control in articulatory speech synthesis. In: Esposito A, Faundez-Zanuy M, Keller E, Marinaro M (eds) *Verbal and Nonverbal Communication Behaviours*, *LNAI 4775* (Springer Verlag, Berlin, Heidelberg), pp. 174–89. http://dx.doi.org/10.1007/978-3-540-76442-7_16

Kröger BJ, Kannampuzha J, Lowit A, Neuschaefer-Rube C (2008) Phonetotopy within a neuro-computational model of speech production and speech acquisition. In: Fuchs S, Loevenbruck H, Pape D, Perrier P (eds) *Some Aspects of Speech and the Brain*. (Peter Lang, Berlin), pp. 59–90

Kröger BJ, Kannampuzha J, Neuschaefer-Rube C (in press) Towards a neurocomputational model of speech production and perception. Speech Communication. http://dx.doi.org/10.1016/j.specom.2008.08.002

Levelt WJM, Wheeldon L (1994) Do speakers have access to a mental syllabary? *Cognition* 50: 239–69

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

Van Lieshout PHHM, Hulstijn W, Peters HFM (1996) From planning to articulation in speech production: what differentiates a person who stutters from a person who does not stutter? *Journal of Speech*, *Language*, *and Hearing Research* 39: 546–64

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: 223–45

Saltzman EL (1985) Task dynamic coordination of the speech articulators: a preliminary model. Haskins Laboratories Status Report on Speech Research SR-84: 1–18

Saltzman EL, Kelso JAS (1983) Skilled actions: a task dynamic approach. Haskins Laboratories Status Report on Speech Research SR-76: 3–50

Saltzman EL, Munhall KG (1989) A dynamic approach to gestural patterning in speech production. *Ecological Psychology* 1: 333–82

van der Merwe A (1997) A theoretical framework for the characterization of pathological speech sensorimotor control. In: MR McNeil (ed.) *Clinical Management of Sensorimotor Speech Disorders*. (Thieme, New York), pp. 1–26

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