How to Learn Proto Speech Patterns Using a Physiologically Based Vocal Tract Model

Bernd J. Kröger[1,2], *Xi Chen*[2], *Cornelia Eckers*[1], *Stefan Heim*[3,4,5,6]

[1]*Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany*

[2]*Cognitive Computation and Applications Laboratory, School of Computer Science and Technology, Tianjin University, Tianjin, P.R.China*

[3]*Section Structural Functional Brain Mapping, Department of Psychiatry, Psychotherapy, and Psychosomatics; Medical School, RWTH Aachen University, Aachen, Germany*

[4]*Research Centre Jülich, Institute of Neuroscience and Medicine (INM-1 and INM-3), Jülich, Germany*

[5]*Section Clinical and Cognitive Neurosciences, Department of Neurology, Medical School, RWTH Aachen University, Aachen, Germany*

[6]*JARA – Translational Brain Medicine, Jülich and Aachen, Germany*

bernd.kroeger@rwth-aachen.de

Abstract: For modeling early phases of speech acquisition (babbling and imitation) we used a geometrical (non-muscle based) vocal tract model so far. But especially in order to differentiate higher level and lower level motor representations it is essential to incorporate a *physiological vocal tract model* controlled by muscle force activation patterns. In this paper we will discuss, why higher and lower level motor representations should be separated and why these different representations are important already during early phases of speech acquisition. First simulation results are reported. In these simulation experiments a physiological (muscle based) vocal tract model is used for learning *proto speech patterns*, i.e. for learning prelinguistic vocalic babbling patterns.

1 Introduction

Speech production is a cognitive and sensorimotor process. If an utterance is intended to be produced, lexical concepts and subsequently phonological forms are activated at the semantic and phonological level and a further syntactic and phonological processing leads to a concrete but still cognitive symbolic representation of the utterance [1]. Subsequently the utterance can be executed using the vocal tract system. Whereas the cognitive processes of speech production are well investigated, this is not the case for the sensorimotor part. But during the last decade, few approaches focusing on these sensorimotor processes have been proposed [2, 3, 4]. In this paper we try to elucidate the sensorimotor part of speech production from the viewpoint of quantitative models and focus on early phases of speech acquisition.

2 Higher level and lower level motor representations

From the viewpoint of a physiological model of speech production (e.g. [5]) a lower level motor representation (lower control level) comprises neuromuscular activation patterns, directly controlling speech articulator muscles. But there are no simple relations between speech articulator positions and muscle activation patterns. Furthermore, it can be shown that vocal tract shapes, leading to comparable acoustic outputs (formant patterns), can result from an abundance of different muscle activation patterns as well as different speech articulator positions. For example an [i], which normally is produced with high tongue position in cooperation with a high positioning of the lower jaw can be produced as well with a more lowered jaw position, if the tongue compensates for this jaw lowering by increasing tongue height relative to lower jaw. Moreover bite block experiments indicate that a speaker – if for example his lower jaw is fixed to a permanently low position (bite-block condition) – is capable of producing understandable speech despite an articulatory perturbation [6]. Even if speaking with a bite block becomes easier (e.g. the case that if the bite block is inserted into the mouth over a longer time period [7]), the behavioral results given in [6, 7] suggest that after inserting a bite block, the speaker directly produces understandable speech. This experimental finding can only be explained by assuming a higher level motor representation, often called *motor plan*. This motor plan is capable of describing the production of speech items in terms of *temporally varying vocal tract shapes* rather than in terms of specific articulator movements (e.g. specific movements of lower jaw and tongue and lips relative to lower jaw). The idea of focusing on vocal tract shapes and linguistically relevant vocal tract constrictions (i.e. vocal tract constrictions representing different speech sounds) as an invariant level of speech production has been developed quantitatively in the Haskins Labs and is called task dynamics approach [8].

Based on the task dynamics approach a speech action representation has been introduced which defines the temporal organization of speech action units (or vocal tract action units [9]) within a syllable or word. These action units define the linguistically relevant vocal tract constrictions which are needed for the production of speech sounds. These actions need not to be specified with respect e.g. to the degree of lower jaw contribution to an action, i.e. with respect to the detailed realization of an action at the level of movements of the contributing articulators. An example is given here: in the case of a labial closure (as occurs in the case of a [b], [p], or [m]) the full closure of the lips needs to occur but the contribution of a lower jaw heightening may be more pronounced in an [i]- than in an [a]-context.

Coming back to the bite block experiment we assume that higher level motor representations of syllables are stored in a mental syllabary [3, 10, 11] (see also Fig. 1) while different lower level motor activation patterns may occur for a sound or syllable due to the different demands on articulators resulting from temporal overlap of speech action units (coarticulation) or due to external articulatory perturbations as introduced in bite-block experiments. This abundance of lower level motor activation patterns has been trained during speech acquisition and beside the mental syllabary we in adition assume a *motor execution module* for performing speech actions (Fig. 1).

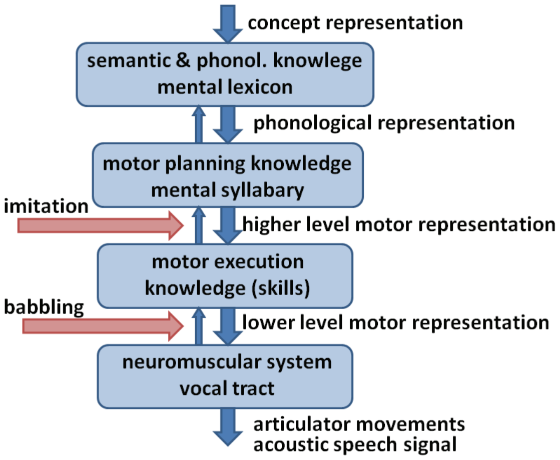

Figure 1 – Levels of neural representations and knowledge repositories and/or processing modules in speech production. Babbling and imitation starts at different levels and generates lower as well as higher level motor knowledge (see text).

3 From Proto action scores to language specific action scores

While a neuroantomical organization of the brain and its connections towards the vocal tract (as well as to all other parts of the body) is mainly predefined genetically, our predefined "knowledge reservoirs" as defined in Fig. 1 need to be "filled" during speech acquisition. The knowledge or skill acquisition is realized during first years of lifetime. Two basic learning phases can be separated within early speech acquisition, i.e. babbling and imitation [3].

If we assume that babbling starts with random neuromuscular motor activation patterns (activation of lower level motor representations, see Fig. 1), this leads to specific articulator movements and to a specific time series of vocal tract shapes and subsequently to higher level motor activation patterns via somatosensory feedback. Thus it can be assumed that the newborn within its first year of lifetime develops higher level action score representations comprising so called *proto speech actions* (also called "gross gestures" [12]), e.g. a proto opening or closing action, which is mainly driven by moving the lower jaw or like a proto lip rounding action. Later on during the first and second year of lifetime, when the toddler starts to imitate caretakers speech items, proto action scores can be activated via auditory stimulation (see the feedforward feedback model proposed in [3]) and these proto actions are now "fine tuned" with respect to language specific demands. Here the proto opening actions may be fine tuned in order to be capable of producing vowels while the proto closing action may be fine tuned with respect to consonants (Fig. 2). Thus a proto closing action – as it occurs during consonant production – may be tuned with respect to manner and place. In addition the temporal coordination of closing actions together with vocalic and/or velopharyngeal opening or closing actions is learned in order to produce voiced vs. voiceless or nasal vs. non-nasal speech sounds (Fig. 2).

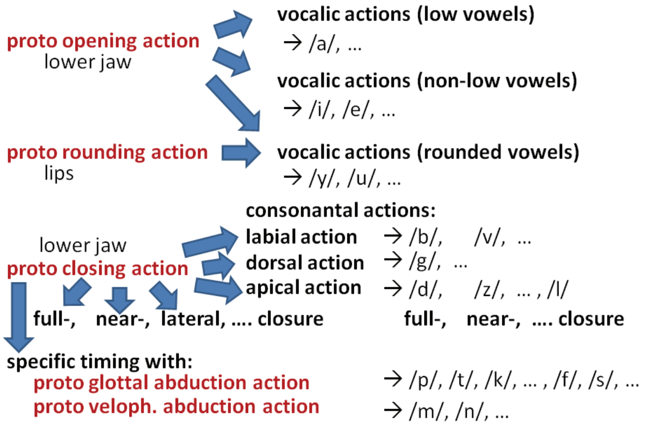
vocalic actions (low vowels)
→ /a/, ...
proto opening action
lower jaw
vocalic actions (non-low vowels)
→ /i/, /e/, ...
proto rounding action
lips
vocalic actions (rounded vowels)
→ /y/, /u/, ...
consonantal actions:
lower jaw
labial action → /b/, /v/, ...
proto closing action
dorsal action → /g/, ...
apical action → /d/, /z/, ... , /l/
full-, near-, lateral, ... closure
full-, near-, ... closure
specific timing with:
proto glottal abduction action
proto veloph. abduction action
→ /p/, /t/, /k/, ... , /f/, /s/, ...
→ /m/, /n/, ...

Figure 2 – Language specific speech actions (bold black letters) result from few basic proto speech actions (bold red letters). One or more language specific speech actions are needed in order to produce a realization of a phoneme (non-bold black letters).

It should be noted here that we do not share the view that speech actions are by definition phonological units as is postulated in Articulatory Phonology [12, 13] but in accordance with [12, 13] it is assumed in our approach that speech actions are fine tuned with respect to the needs of language specific distinctiveness during speech acquisition.

4 The model

Higher and lower level motor representations are implemented in our current version of a quantitative model of speech production, perception and acquisition. At the lower control level, 15 muscle groups are defined (see [5] and Fig. 3) and muscle force patterns (neuromuscular activation patterns) can be set. Muscle activation patterns for a cardinal [a], [i], and [u] are displayed in Fig. 3. These vocalic neuromuscular activation patterns are based on babbling knowledge which has been included in an "EP-Map" as part of the physiological model [5]. Thus the EP-Map represents a knowledge repository which is capable of generating neuromuscular activation patterns for specific model articulator positions (lower jaw, tongue tip, tongue dorsum, tongue root; see Fig. 4). This EP-Map is comparable with a part of the "motor execution knowledge repository" as defined in Fig. 1. The relation between neuromuscular activation and resulting muscle force is defined by a quasi-logarithmic function (Fig. 5)

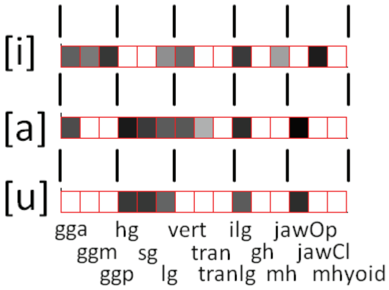

Figure 3 – Neuromuscular activation patterns for realizations of cardinal vowels [i], [a], and [u]. 15 model motoneurons are generating the muscle force pattern for 15 different muscle groups (White = no activation; black = full activation of motoneuron); names of muscle groups (see also [5]): gga = genioglossus anterior; ggm = middle portion of genioglossus; ggp = genioglossus posterior; hg = hyoglossus; sg = styloglossus; lg = superior part of longitudinalis; vert = verticalis; tran = inferior part of transversus; tranlg = inferior part of longitudinalis and superior part of transversus; ilg = inferior longuitudinalis; gh = geniohyoid; mh = mylohyoid; jawOp = muscle bundles for lower jaw lowering; jawCl = muscle bundles for lower jaw raising; mhyoid = mylohyoid (superior part). Each *model motoneuron* displayed here represents a bundle of real motoneurons associated with a muscle group.

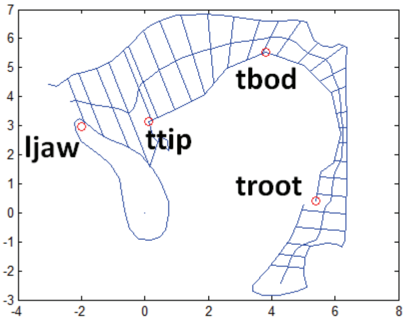

Figure 4 – Red circles indicate articulator positions of lower jaw (ljaw), tongue tip (ttip), tongue dorsum or tongue body (tbod), and tongue root (troot) for a realization of cardinal [a]. Blue lines indicate (i) surface of articulators, (ii) midline for airflow and (iii) distance lines between articulators and vocal tract walls (palate, velum, pharyngeal wall).

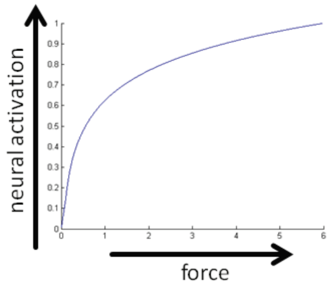

Figure 5 – Neuromuscular activation (0 = no activation; 1 = full activation of a model motoneuron) as function of muscle force (0 … 6 N). Neuromuscular activation generates muscle force.

At the higher motor level, a neural representation of the model articulator positions as displayed in Fig. 4 is used [14]. This has to be augmented towards a neural representation of speech action scores if our approach goes beyond modeling of single (static) vowels.

5 Simulation of vowel babbling

As a starting point for modeling babbling by taking into account lower and higher level motor representations, we used 1146 linear combinations of three basic proto vocalic actions, i.e. a cardinal [i], a cardinal [a], and a cardinal [u]. These proto actions were generated on the basis of a babbling knowledge already included in EP-Map. Linear combinations of lower level muscle force patterns as well as of higher level articulator positions were used in order to generate babbling stimuli which can be interpreted as "interpolations" between the three cardinal vowels. These babbling stimuli thus covered the whole vowel space spanned by the three vowels. Lower and higher level motor representations as well as a neural representation of the resulting auditory formant pattern for each of the 1146 babbling training items served as input data for training a 20x20 self-organizing map. This self-organizing map organizes and associates the motor and auditory patterns given by the 1146 training items (this map is also called "phonetic map", cf. [3]). The ordering of phonetic states and the association of articulator positions and auditory information (formant pattern) is displayed in Fig. 6. The training comprised 360 training cycles (x 1146 training items = 412560 training steps) where training items were applied in random order within each training cycle.

From Fig. 6 it can be seen that phonetic states are ordered with respect to front-back and low-high, where front high vowels ([i]-like phonetic states) occur in the top right corner and where low vowels ([a]-like phonetic states) occur in the bottom left corner of the self-organizing map. Thus the model neurons within the phonetic map represent different vocalic states. Muscle force activation patterns within these corners of the self-organizing map are similar to those displayed in Fig. 3 for cardinal [i] and cardinal [a].

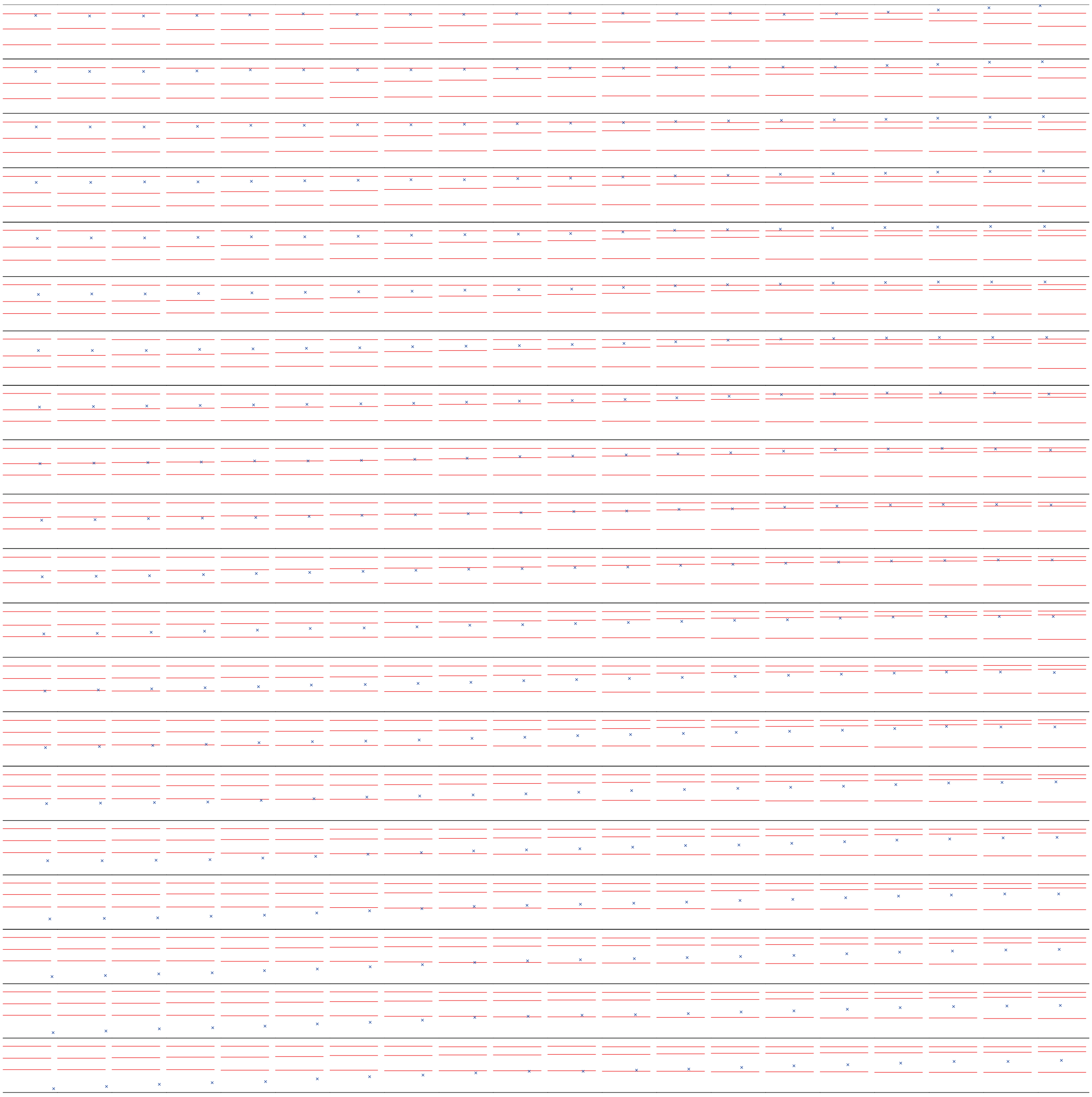

Figure 6 – Display of auditory formant pattern (horizontal red lines give F1, F2, and F3 in bark) and of articulator position of tongue dorsum for 20x20 model neurons representing a self-organizing phonetic map. Top right area represents [i]-like states; bottom left area represents [a]-like states (for more information concerning self-organizing maps in production models see [5]).

6 Preparing vowel imitation

From a perceptual inspection of babbling results (i.e. states, which are represented by the neurons in the phonetic map), it can be concluded that some neurons within the self-organizing map (Fig. 6) represent realizations of Mandarin /i/ and /a/; Mandarin was chosen as target language here. For starting imitation experiments, we assume that the production of vowels represented by these /i/- or /a/-states within the SOM are rewarded by a teacher or caretaker, and that the model (or child) thus is motivated to produce especially those states which occur in these "award regions" of the self-organizing map, in order to strengthen the representation of phonetic realizations for specific phonemic states.

Therefore it is planned to generate further training items, which are located within or near the "award regions" of the self-organizing map in order to refine vowel productions, which represent phoneme realizations of a target language.

7 Conclusions

It should be noted that we are just at the beginning of integrating lower and higher motor control levels within our neural model of speech production, perception, and acquisition. Even at the level of babbling, more training sets should be generated in order to identify possible "award regions" which could serve as a "gamete" for phoneme regions at the level of the self-organizing phonetic map (cf. [3] and see Fig. 6). First imitation experiments for Mandarin vowels are conducted in the moment in our labs in order to investigate the development of stable phoneme regions at the level of a self-organizing phonetic map during speech acquisition.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61233009)

Literature

[1] Levelt WJM, Roelofs A, Meyer AS (1999) A theory of lexical access in speech production. Behavioral and Brain Sciences 22, 1-75

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

[3] Kröger BJ, Kannampuzha J, Neuschaefer-Rube C (2009) Towards a neurocomputational model of speech production and perception. Speech Communication 51: 793-809

[4] Hickok G (2012) Computational neuroanatomy of speech production. Nature Reviews Neuroscience 13, 135-145

[5] Dang J, Honda K (2004) Construction and control of a physiological articulatory model. Journal of the Acoustical Society of America 115, 853-870

[6] Fowler CA, Turvey MT (1980) Immediate compensation in bite-block speech. Phonetica 37, 306-326

[7] McFarland DH, Baum SR (1995) Incomplete compensation to articulatory perturbation. Journal of the Acoustical Society of America 97, 1865-1873

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

[9] Goldstein L, Byrd D, Saltzman E (2006) The role of vocal tract gestural action units in understanding the evolution of phonology. In: Arbib M (ed.) Action to Language via the Mirror Neuron System (Cambridge University Press, Cambridge, MA, USA), pp. 215-249

[10] Kröger BJ, Birkholz P, Neuschaefer-Rube C (2011) Towards an articulation-based developmental robotics approach for word processing in face-to-face communication. PALADYN Journal of Behavioral Robotics 2, 82-93

[11] Eckers C, Kröger BJ (2012) Semantic, phonetic, and phonological knowledge in a neurocomputational model of speech acquisition. In: Wolff M (ed.) Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2012 (TUDpress, Dresden, Germany), pp. 244-251

[12] Browman CP, Goldstein L (1989) Articulatory gestures as phonological units. Phonology 6, 201-251

[13] Browman CP, Goldstein L (1992) Articulatory phonology: an overview. Phonetica 49, 155-180

[14] Chen X, Dang J, Wang Y, Wei J, Fang Q, Kröger BJ (in preparation) Towards a neural control concept for a physiological articulatory model of speech production.