A Neural Understanding of Speech Motor Learning

Xi Chen[1], Jianwu Dang[1, 2], Han Yan[1], Qiang Fang[3] and Bernd J. Kröger [4, 1]
[1] Tianjin Key Laboratory of Cognitive Computing and Application,
School of Computer Science and Technology, Tianjin University, China
[2] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
[4] Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany
E-mail: chenxiyh@tju.edu.cn, bernd.kroeger@rwth-aachen.de

***Abstract*—Speech motor learning is still an under-discussion process in neural computational modeling. In this paper we focus on the relationship between vowel articulation and its muscle activation patterns, propose a neural understanding of speech motor learning and elucidate the neural strategy for speech learning of infants. An existing physiological model including speech articulator organs which has successfully replicated the biomechanical articulatory movement has been used. Self-organizing map related to the contour positions of control points and muscle activation patterns was established during speech motor learning. Experimental result refer to the one-to-many problem in the mapping between the high-level to the low-level motor states, which indicates that quite different muscle activation patterns can lead to similar articulatory positions.**

I. INTRODUCTION

Neural control model of speech production is important for exploring the speech production mechanism of human being. Computational model of speech production has been formulated since the middle of the twentieth century, when the first model was proposed [1]. Speech motor learning is an important part of neural-computational modeling, which states the relationship between target position of articulatory movements and muscle activation patterns acting on articulators.

According to the DIVA (Directions into Velocities of Articulators) model [2-5], speech production needs feedforward and feedback control pathways. The feedforward pathway directly controls the articulator movements. The approach proposed by Kröger [6-8] postulates a mental syllabary, where auditory states, motor plan states, and somatosensory states of all frequent syllables are stored using a self-organizing map (c.f. [9]), which is called "phonetic map" in that approach.

The approach postulated in this paper is based on the Kröger model and constructs a specific modeling of motor execution: The motor plan of a speech item is processed by a self-organized motor execution map, which generates specific muscle activation patterns for each speech gesture at the level of the primary motor map. Thus, we are using a physiological articulatory model, which is based on the morphological and physiological characteristics of the human speech organs, and driven by muscle activation patterns [10-11].

II. BACKGROUND FOR SPEECH PRODUCTION NEURAL CONTROL STRATEGY

In this section, we will give a brief overview concerning the whole structure of our neural model and its interface towards the physiological articulatory model [10-11].

A. Structure of Neural Control Model

In speech production, auditory, somatosensory, and motor information are integrated [3]. There are intricate interactions among different levels, starting from an intention, over selection of lexical items, generation of a phonological form towards the generation of articulatory movement patterns and the acoustic speech signal [12]. Since there are a number of important functions that could not be taken into account in this study, we deal with them based on certain hypotheses.

As feedback information, beside an acoustic (auditory) representation, a somatosensory representation is needed in order to control the correct execution of a motor plan. Somatosensory signals can be subdivided into tactile and proprioceptive signals. As feedforward information we use two levels of motor states: high-level motor state describes the movement pattern of articulators (i.e. contour points at surface of model articulators; red points in Fig. 1); low-level motor state describes the muscle activation patterns which lead to specific articulatory movements and positions.

Fig. 1 gives an overview concerning the whole structure of the neural control model, including the motor map, the auditory and somatosensory map, and the interconnecting phonetic and execution map. The term "map" is used to describe a set of model neurons, capable of representing a specific state of a speech item (e.g. motor, auditory, or somatosensory state) or speech knowledge (e.g. within motor execution and phonetic map). The later type of map is implemented as self-organizing map (SOM) [8].

Phonetic map is associated with high-level motor state, somatosensory state, and auditory state. In addition now, in this approach a further training is needed for association of high-level and low-level motor state by motor execution map (see the next chapter).

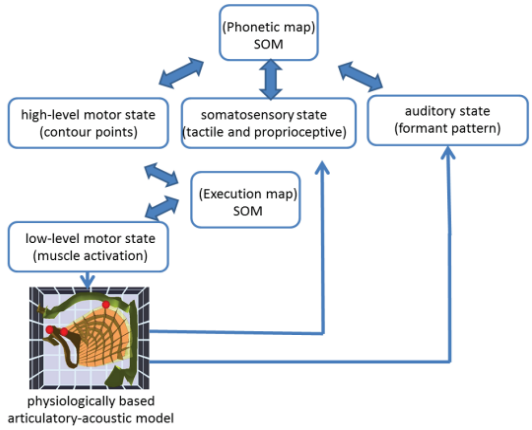

Fig. 1 Structure of whole neural control model. Two SOM maps are used: (i) Phonetic map is associated with high-level motor state, somatosensory state, and auditory state map. (ii) Execution map is associated with low-level motor state and high-level motor state.

During speech learning, these neural associations via self-organized phonetic map and execution map are adjusted [8]. The model gains some knowledge concerning the sensory-to-motor associations in phonetic map and different level motor associations in execution map, concerning typical realizations of some language-specific speech items. During speech production, the knowledge stored in the execution map generates muscle activation patterns leading to specific target-directed articulatory movements and consequently to the generation of acoustic speech signals. This process is simulated by the physiological articulatory model. Feedback signals such as somatosensory and auditory signals are used for learning in combination with motor signals.

B. Physiological Articulatory Model

The physiological articulatory model, which is used in this approach, was proposed by Dang and Honda (2004). The model is a semi-3D physiological articulatory model that consists of the tongue, jaw, hyoid bone, and rigid vocal-tract walls. The morphology of the speech organs was measured based on volumetric MRI data obtained from a male speaker and 11 major muscles are included in the tongue model [13]. The model is driven by the muscle activation pattern to produce speech movements, and then synthesizes speech sound using the transmission line model via the vocal tract shape.

C. Muscle combinations for motor learning

The most effective way to form a tongue shape by muscle contraction, according to the minimal energy principle [11], is for two agonist muscles to work together to achieve a given target. For example, the GGp and HG show diametrically opposing directions in experiments [11]. This indicates that the GGp and HG work antagonistically in governing the tongue dorsum movements. Similarly, the GGm and SG form another antagonist muscle pair. Our simulations indicate that the two muscles of the muscle pair worked in synergy for one articulator, while functioning as an antagonist muscle pair for another one. Furthermore, to simulate co-contractions between agonist and antagonist muscles, several three-muscle combinations have been designed. A three-muscle group consists of an independent muscle and a muscle pair, in which the activation of the independent muscle corresponds to the co-contraction level and governs a main part of the tongue, while the muscle pair manipulates the other part (e.g. tongue tip) via the mechanism of co-contraction of the agonist and antagonist muscles. This property can be used within our model to achieve multiple spatial targets (targets of different articulators) simultaneously, based on a strategy that accurately guarantees to reach a crucial target [11].

III. EXPRERIMENT

High-level motor learning is involved in our neural computational model experiments but was already described in previous publications [8]. It has been elucidated there that speech production knowledge can be learned (i) by neural associations between the self-organizing map and high-level motor state, somatosensory state, and auditory state map, and (ii) by the organization of this map itself.

In this paper, we mainly focused on execution motor learning, which connects the high-level motor states and the low-level motor states by introducing a self-organizing motor execution map.

A. Training Set for Motor Learning

In order to explore the low-level motor learning (i.e. organization of the execution map), corresponding muscle forces and articulatory contour points were generated as training items. As a starting point of babbling, we defined three ‘extreme proto-vocalic tongue states’ (high-front, high-back and low-back) forming palatal, velar, and pharyngeal proto-vocalic constrictions (Fig. 2).

Typical tongue movements have been generated as training set using major muscle combinations as mentioned in the previous section. 10 pairs of most efficient muscle combinations have been chosen [13]. The muscle groups include [GGm, HG, T-SL], [HG, GGm] for low tongue states, [GGp, MH] for high-front tongue states, [SG, MH], [SG, T-SL, MH], [GGp, SG] for high-back tongue states, [GGm, GGp], [GGm, GGp, SL] for mainly front and [SG, HG], [HG, SG, T-SL] for mainly back tongue states.

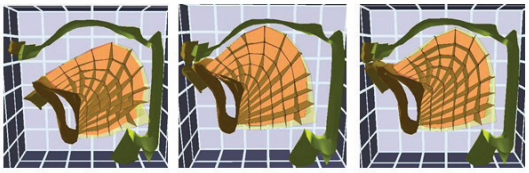

Fig. 2 ‘Extreme proto-vocalic tongue states’ (low, high-front, high-back) muscle pairs effect.

The 10 muscle combinations almost include all the tongue movements, so the set of these 10 muscle combinations were used as a basis for generating a set of whole proto-vocalic tongue babbling movements. And according to these muscle combinations we defined the ‘extreme proto-vocalic tongue

states' as the starting point of the training set. Because this vowel space is six dimensional (two dimensions for each of the three tongue contour points), a hyper plane is generated in the six-dimensional space. The training set is visualized in two dimensions for tongue dorsum contour point displacement in Fig. 3.

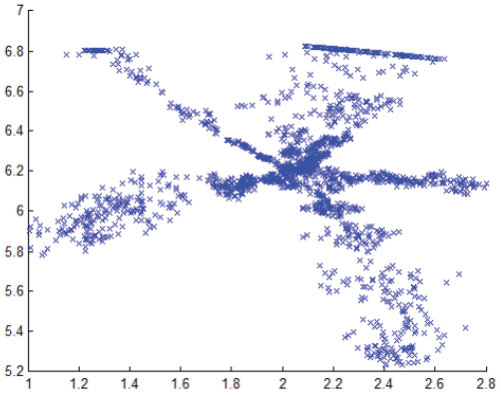

Fig. 3 Training set for motor execution map in the view of tongue contour point for tongue dorsum.

B. *Neural representation of motor states*

The neural representations used for the motor states including high-level motor states and low-level motor states are shown in Fig. 4.

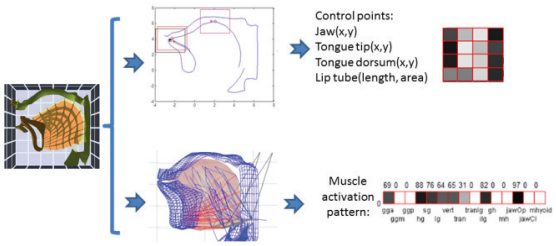

Fig. 4 Example of neural representation for motor states. Model neuron activation patterns are displayed as box-arrays. Degree of neural activation is displayed as grey scale (black refers to full activation).

For high-level motor state, we used neural representations of three tongue contour points [10]. m represents the relative location of current control point position according to its rest position. Contour point displacement of 1.5 mm and above leads to maximum neural activation.

$$hm\_n1(m) = \begin{cases} 0, & m \in (-\infty, -1.5) \\ \frac{m+1.5}{3}, & m \in [-1.5, 1.5] \\ 1, & m \in (1.5, +\infty) \end{cases} \quad (1)$$

$$hm\_n2(m) = 1 - hm\_n1(m) \quad (2)$$

For low-level motor state, muscle activation patterns are generated for 15 articulator muscles. Activation forces f for each muscle were set from 0 to 6N [10]. Neuron activation for muscles is represented using a logarithmic function.

$$ma(f) = \begin{cases} \frac{10}{7} \times f, & f \in [0, 0.1) \\ 0.6236 + 0.4836 \times log_{10}(f), & f \in [0.1, 6] \end{cases} \quad (3)$$

C. *Training Result of Execution Map*

As training result, an execution map and its synaptic links to the two motor state maps were trained using a 15×15 SOM. The number of training items was 1800. Synaptic link weights become stable after round about 600 iterations of training. After SOM training, each model neuron of SOM represents a learned (proto-vocalic) state. The articulatory positions of contour points inside each model neuron are mapping to a corresponding muscle activation pattern.

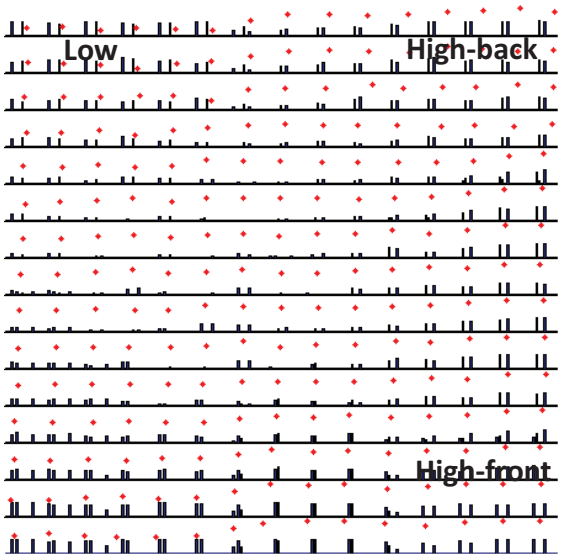

Fig. 5 Display of SOM link weights after training of execution map. The red point within the box, representing each model neuron within 15x15 SOM map, represents displacement of tongue dorsum contour point (high-level motor state). The vertical bars represent muscle force pattern for the 10 muscle combinations, as are defined in the text (low-level motor state).

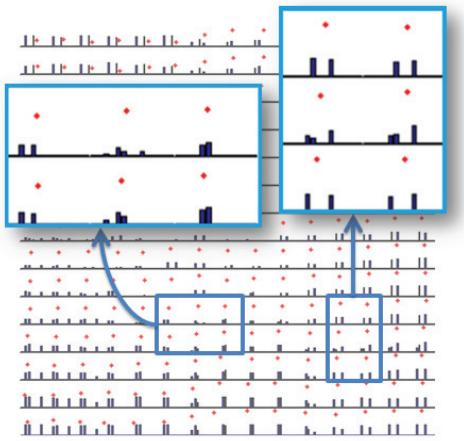

Fig. 6 Display of two groups of six example neurons, exemplifying the one-to-many problem, as it is solved in our SOM execution map. The two boxes in foreground give an example, how different muscle activation patterns can lead to a similar high-level motor state.

Fig. 5 displays the association of higher- and lower-level motor states represented in each proto-vocalic SOM state. Only the tongue dorsum position is shown here for high-level motor state (red point inside each model neuron). The vertical bars represent activation pattern for the major muscle of tongue positions (ordering according to [10]: GGa, GGm, GGp, HG, SG, SL, Vert, Tran, T-SL, IL, GH, MH, JawOp, JawCl). It can be seen from Fig. 5 that vocalic tongue positions continuously change from left-top corner (low tongue position) to right-bottom corner (high-front tongue position), which means that basic vocalic features like front-high, back-high, and low have been learned during the SOM training.

As shown in Fig. 6, the execution map exhibits areas with little variance of tongue positions, while the associated muscle activation patterns strongly vary. This is an exemplification of the "one-to-many" problem. For a given muscle activation pattern, a unique articulatory configuration can be found, while the mapping from spatial position to muscle activation patterns is not unique.

D. Evaluation

Fig. 7 gives visualization for our first evaluation experiment concerning the one-to-many problem. In the figure, there are 225 points, reflecting the 15x15 SOM neurons. Small circle or small fork with different color for each point reflects ten muscle combinations for tongue dorsum, the ten muscle combinations are [GGm, HG, T-SL], [HG, GGm], [GGp, MH], [SG, MH], [SG, T-SL, MH], [GGp, SG], [GGm, GGp], [GGm, GGp, SL], [SG, HG], [HG, SG, T-SL]. The horizontal axis displays the energy with respect to current muscle force. The vertical axis gives the distance between the contour point position and the rest position. Thus, points with short vertical distances, we can say that their contour point positions are almost the same. But these points can exhibit different colors (different muscle activation patterns) and furthermore the energies of them are different. The big oval areas give examples that some points with close vertical distance but with different colors or energies exhibit different muscle force patterns.

Evaluation experiment 2 is to evaluate the quality of the execution map SOM. We have known that the SOM training has found the muscle activation pattern for each tongue position defined by the tongue dorsum contour point by using a 15x15 execution map SOM. So we can compare the entire tongue contour defined by muscle activation patterns and defined by contour point position to evaluate the quality of this SOM for predicting correct muscle activation patterns for a given tongue contour as defined by three tongue dorsum contour points.

The grid representation method (Fig. 8, using the distance d(i) (i=1,2,3,…,23) is used for calculating differences in tongue contour of initial tongue contours and tongue contours resulting from muscle activation patterns calculated from execution map SOM.

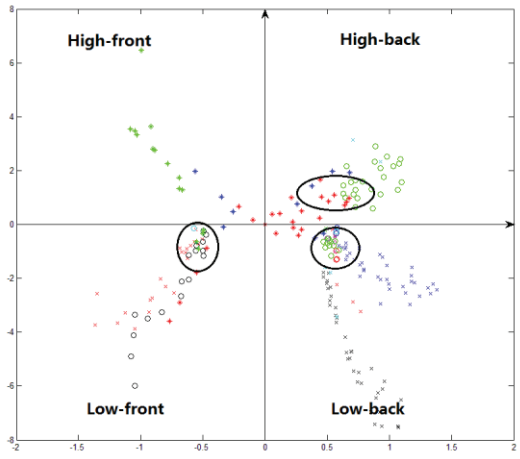

Fig. 7 Display of first evaluation experiment. Each quadrant represents a typical tongue movement direction (high-front, high-back, low-front, and low-back). Different muscle combinations are shown in different colorful marks: [GGm, HG, T-SL] is taken as black fork, [HG, GGm] is taken as cyan fork, [GGp, MH] is taken as green plus, [SG, MH] is taken as green circle, [SG, T-SL, MH] is taken as blue plus, [GGp, SG] is taken as red plus, [GGm, GGp] is taken as black circle, [GGm, GGp, SL] is taken as red fork, [SG, HG] is taken as blue fork, [HG, SG, T-SL] is taken as red circle.

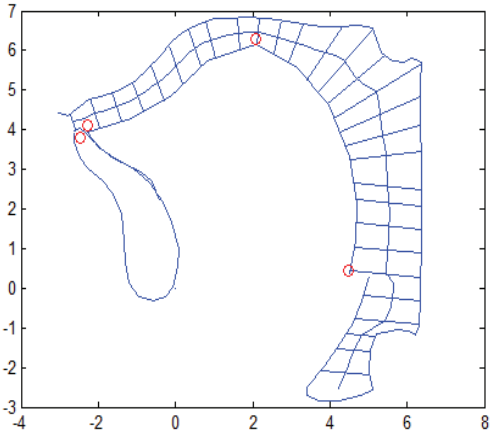

Fig. 8 Midsagittal view of the vocal tract in our model. The first 23 gridlines, reaching from lips to lower pharynx, were taken for calculating the vocal tract area and thus indirectly for calculating current tongue position within evaluation experiment 2.

The mean difference in millimeters of the d-distance for each tongue position over all neurons of the execution map (225 neurons in 15x15 size map) is 0.684mm and the standard deviation is 0.252mm. Fig. 9 displays the difference between the entire tongue contour defined by muscle activation patterns and those, directly defined by contour point position for the typical tongue positions of three proto-vocalic states: high-front, high-back and low. From the figure the two entire tongue contours are almost overlapped together.

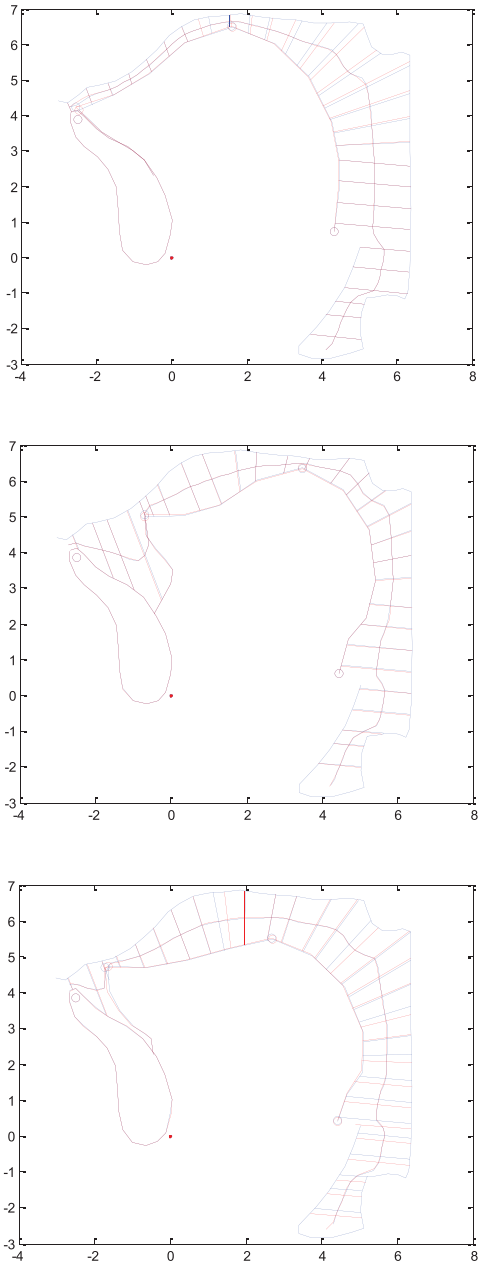

Fig. 9 The entire tongue contour defined by muscle activation patterns (red lines) and defined by control point (blue lines) for three proto-vocalic states: high-front, high-back and low.

IV. CONCLUSIONS

A neural control concept for controlling a physiological articulatory model of speech production with phonetic map and motor execution map is introduced. The training of motor execution map is described in detail in this paper. From the training results it can be concluded that the knowledge concerning the association of high-level motor information (geometric contour points) and low-level motor information (i.e. muscle activation pattern in relation to contour points) can be learned by using a self-organizing map approach. In particular it has been shown that the "one-to-many" problem can be addressed by using our SOM approach. In further studies we will address the neural processing of consonants in the context of syllables and first words.

ACKNOWLEDGMENT

This work is supported in part by the National Basic Research Program of China (No. 2013CB329301), and in part by the national natural science foundation of China under contract No. 61233009 and No.6117501. This study was also supported in part by the Ministry of Science and Technology of P.R. China with contract No. 2011BAH16B01-02.

REFERENCES

[1] W. L. Henke, "Dynamic articulatory model of speech production using computer simulation," *Ph.D. dissertation*, Massachusetts Institute of Technology, 1966.

[2] F. H. Guenther, "A neural network model of speech acquisition and motor equivalent speech production," *Biological Cybernetics.72*, pp. 43-53, 1994.

[3] F. H. Guenther, "Cortical interaction underlying the production of speech sounds," *J. Comm. Disorders* 39, pp. 350-365, 2006.

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

[5] F. H. Guenther, T. Vladusich, "A neural theory of speech acquisition and production," *Journal of Neurolinguistics* 25(5), pp. 408-422, 2012.

[6] B. J. Kröger, P. Birkholz, J. Kannampuzha, C. Neuschaefer-Rube, "Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer," *International Conference on Spoken Language Processing (Interspeech 2006–ICSLP)*, pp.565-568, 2006.

[7] B. J. Kröger, P. Birkholz, "A gesture-based concept for speech movement control in articulatory speech synthesis," *Verbal and Nonverbal Communication Behaviours*, Springer, Berlin, pp. 174-189, 2007.

[8] B. J. Kröger, J. Kannampuzha, C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication* 51, pp. 793-809, 2009.

[9] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin NewYork, 2001.

[10] J. Dang and K. Honda, "A physiological model of a dynamic vocal tract for speech production," *Acoustical Science and Technology* 22, pp. 415-425, 2001.

[11] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *Journal of the Acoustical Society of America* 115(2), pp. 853-870, 2004.

[12] W. J. M. Levelt, A. Roelofs, A. Meyer, "A theory of lexical access in speech production," *Behavioral and Brain Sciences* 22, pp. 1-75, 1999.

[13] Q. Fang, S. Fujita, X. Lu and J. Dang, "A model-based investigation of activations of the tongue muscles in vowel production," *AST*, pp. 277-287, 2009.

[This paper was published in APSIPA,2013]