A New Framework of Neurocomputational Model for Speech Production

Han Yan [1], *Jianwu Dang*[1,2], *Mengxue Cao*[3], *Bernd J. Kröger*[4,1]

[1] Tianjin Key Lab. of Cognitive Computing and Application, Tianjin University, Tianjin
[2] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3] Lab. of Phonetics and Speech Science, Institute of Linguistics, CASS, Beijing
[4] Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany

{yanhanwxf,dangjianwu}@tju.edu.cn, mengxuecao@outlook.com, bernd.kroeger@rwth-aachen.de

Abstract

Speech production is complex for the brain to control, since it involves many neural processes such as speech planning, motor control, auditory and somatosensory feedback. Those functions are thought to work both in cascaded and parallel, and the control signals are transformed from one brain area to others with "one-to-many" relations. To describe this situation, in this study, we developed a new framework for a neuro-computational model for speech production based on our previous studies. The proposed model is used to deal with dynamic properties of speech articulation for consonant-vowel (CV-) syllables. In our simulation, the neuronal groups (i.e., motor, auditory and somatosensory) were acquired by learning and stored in the self-organizing maps (SOMs), and those relations between the SOMs were investigated. The results show that the time-varying properties were represented properly. In the control signal flow, the model demonstrated "one-to-many" projections between the SOMs, where one neuron in an SOM on average was projected onto 1.64 neurons in another SOM.

Index Terms: speech production, neural computational framework, articulatory model, self-organizing map

1. Introduction

Modeling neurophysiological-based speech production is essential for exploring speech production and perception mechanisms in the higher levels of the human nervous system [1-3]. In the literature, two representative models have been developed, which are the DIVA (Directions into Velocities of Articulators) model [4-7] and the Kröger's neuro-computational model [8-10]. The DIVA model, proposed by Guenther, introduces feedforward and feedback control pathways, which simulate the speech production process by using the geometrically based articulatory model. The Kröger's neurocomputational model mainly follows the idea of DIVA model. In order to improve the feedback function, Kröger introduces a syllabic sensorimotor skill repository, which is modeled by a self-organizing map (SOM) [11]. In previous work, we combined the neurocomputational model with a physiological articulatory model [13-14] to develop a model that can replicate speech production processes from the central control to the peripheral movements [16].

The previous model included two motor states (or neuronal groups), the high-level motor state (for motor plan) and the low-level motor state (for motor control), which are connected to a core feature map modeled by the SOM. The SOM consists of a number of "model neurons", where each of the neuronal groups are formed in three different states regarding the articulation, acoustics, and perception, and those states are supposed to be paired using a fixed relation of "one-to-one" [16]. As well known, however, the relations between the different states should be "one-to-many" but not "one-to-one" [15]. Based on this common notion, in this study, we use four different SOMs to describe the motor plan state, motor control state, auditory state, and somatosensory state. We simulate CV-syllables using a physiological articulatory model, and train each of the SOMs individually. Based on the simulation results and trained SOMs, we explore the representation of the dynamic property, and investigate relation between the SOMs.

2. A neural computational framework for speech production

In this section, we propose a new framework of a neurocomputational model with four distinct SOMs, and briefly introduce the neural representation method using each state and the physiological articulatory model.

2.1. A framework of neurocomputational model

Figure 1 shows the new structure of a neurocomputational model, which is developed based on our previous work [16]. The proposed model has four SOMs including feed-forward mapping and sensory-based feedback pathways. The four SOMs describe the motor plan state, the somatosensory state, the auditory state, and the motor control state, where each state corresponds to the function of a certain brain area.

Through the feed-forward control pathway, the motor plan state (functionally corresponding to the Broca's area) describes the movement pattern of the articulators and transfers the control signals to the motor control state (primary motor cortex). To focus on the SOMs introduced in the model, we simplify the process from the motor control unit to speech organs, and suppose that muscle activation patterns are generated in the motor control unit to directly drive the physiological articulatory model. Thus, the physiological articulatory model generates articulatory movements according to the muscle activation patterns and synthesizes speech sound based on a time-varying vocal tract area function. Through the feedback control pathway, we use the auditory state (for auditory cortex) and somatosensory state (for sensory cortex) to monitor the peripheral execution in acoustic and articulatory aspects.

In the process of speech production based on the physiological articulatory model, in term of given articulatory targets, a set of muscle activation patterns are generated base on Equilibrium Position map (EP-map) [17] to be used to drive the articulatory model to produce articulatory movements, and then speech sounds (syllables) are

978-1-4799-4219-0/14/$31.00

synthesized according to the model based vocal tract area function. Based on the simulation data including the targets, muscle activation, movements, and sound, we can obtain the knowledge involved in the speech production processes using the SOM algorithm [12]. Each SOM used in this study consists of $n \times n$ "neurons", where each neuron contains the typical features of the corresponding level, and n is the dimension of the SOM. During the learning process, the typical speech production patterns are learned from the simulation data, and they are arranged by their inherent architecture. In the trained architecture, the extreme patterns are located in the vertex neurons of the map, while the middle neurons reflect the pattern transitions between the extreme patterns.

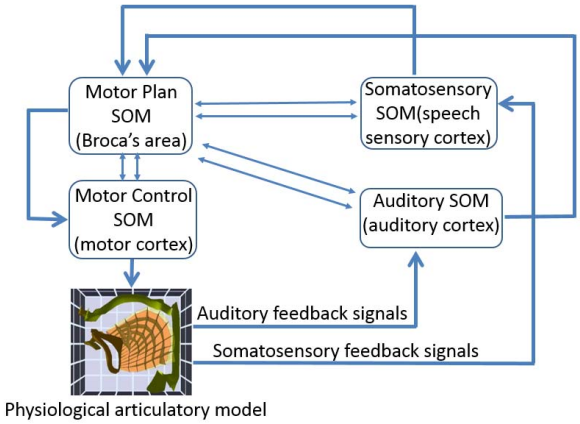

Figure 1: *Framework of the proposed neuro-computational model. The four SOMs describe the functions of corresponding brain areas and work at different levels of the speech production processes.*

After the training, the knowledge will be stored in the SOMs. The neurons in the SOM represent the learned relation of the speech production process in different levels. The mapping functions between the SOMs are learned using the model simulations. As shown in Figure 1, the four SOMs are linked by the bidirectional arrows, and the signal flows are shown by the thick arrows in speech production process.

2.2. Neural representations for CV-syllable

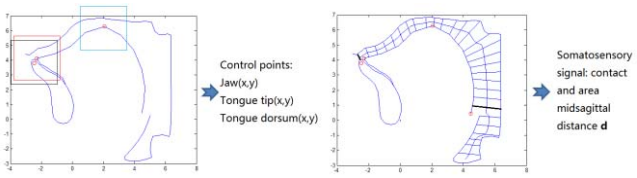

Figure 2: *The target consisting of three control points in the midsagittal view of the vocal tract (left panel); The contact region defined by the first 23 gridlines from lips to lower pharynx (right panel).*

The motor plan map is represented by articulatory parameters. Following our previous work, three control points are chosen in the midsagittal plane of the vocal tract, including the tongue tip, tongue dorsum and jaw. The three control points are shown in the left panel of Figure 2. In order to facilitate our training process, the moving scope of each control point is limited in a 3 cm × 3 cm square area, where the central point is the rest position of this control point. Meanwhile, the wall contact is considered in the simulation. We divide the articulation of CV-syllable into 23 time frames that simulates the dynamic process. The articulatory target in the neuron is represented by a normalized deviation from the central point for each control point in vertical and horizontal directions. The normalized deviation is calculated by Equation (1), where d stands for the actual displacement away from the central point for a control point in each frame, and the value 1.5 is the limited deviation of the control point. For each syllable, the map contains 6×23 cells.

$$cp(d) = \begin{cases} 0, d \in (-\infty, -1.5) \\ \dfrac{d+1.5}{3}, d \in [-1.5, 1.5] \\ 1, d \in (1.5, +\infty) \end{cases} \quad (1)$$

In the motor control map, we use 14 typical articulatory muscles (12 for the tongue and 2 for the jaw) to represent the muscle activation patterns. The normalized muscle activation is calculated for each neuron as in Equation (2), where f represents the activation force for each muscle. The force is set from 0 N to 6 N and normalized from 0 to 1. The value of f is treated into two parts: a linear part and a non-linear part. The f between 0 and 0.1 N is normalized linearly, while it is transformed to a logarithmic scale for f larger than 0.1 N.

$$ma(f) = \begin{cases} \dfrac{10}{7} \times f, f \in [0, 0.1) \\ 0.6236 + 0.4836 \times \log_{10}(f), f \in [0.1, 6] \end{cases} \quad (2)$$

In the auditory map, we use spectrographic information to represent the auditory characteristics of the speech signal [19]. The spectrogram is calculated by the short-time Fourier transform (STFT), which gives the information in both frequency domain and time domain. For speech sounds generated by the physiological articulatory model, its spectrogram is calculated using a 2048-point FFT with the hamming window. The amplitude ranges from 60dB to 100dB. and then it is normalized into the interval between 0 and 1. In the neural representation, the frequency scale is converted into 22-Bark scales. In the time dimension, activation of each neuron is calculated by the mean of 5 sample points, and 23 frames are obtained.

The somatosensory map is represented by the tactile signal, which is used to detect whether the collision occurs in the vocal tract during the speech production. We use the distance between the inner surface and the outer surface (the tract wall) of the vocal tract to represent the proprioception for the configuration of the vocal tract. The gridlines represent the d distance between the inner and outer surfaces of the vocal tract, as shown in the right part of the Figure 2. Here we select the first 23 gridlines from the lips to the lower pharynx to represent the somatosensory state in this study. Based on the previous experiment [16], we assume that the maximum distance of d is 2 cm, and this value is used for the normalization. In order to highlight the tactile condition, we use a squared function of d. Equation (3) gives the normalized amplitudes of d, where 1 represents that the collision occurs. The purpose to introduce this parameter is not only to display the collision but also to describe the situation approaching the collision, which avoids an overshoot at the collision.

$$tac(d) = \begin{cases} (1-\frac{d}{2})^2, d \in (0,2] \\ 0, d \in (2,+\infty) \end{cases} \quad (3)$$

3. Experiment and results

In this study, we focus on neural representation of the dynamic properties of articulation and acoustics for a CV-syllable, and construct a neuro-computational model to generate articulatory movement for the syllables.

3.1. Training sets of CV-syllables

We generated 1146 articulatory targets for vowels to train the neurocomputational model in the previous study. Three extreme "tongue states", high-front, high-back and low-back, were used to form the palatal, velar and pharyngeal constrictions. Based on these vocalic articulatory targets, we add the information of these three control points for certain typical consonants (e.g., /d/), and generate the articulatory targets set for /dV/, where V is the vowels with 1146 different targets [16].

Articulatory movements of /dV/ are simulated using the 1146 articulatory targets to drive the physiological articulatory model, and the time-varying cross-sectional area functions of the vocal tract are estimated accordingly. Speech sounds are synthesized using the transmission line model with the area functions.

From this simulation, we obtained four kinds of data: articulatory targets, muscle activation patterns, articulatory movements, and synthetic speech sound. The data at different stages are converted to the neural representations. The four training sets are used to train the four states SOM.

3.2. Construction of the SOMs

In this study, each SOM consists of 15 × 15 "model neurons". In order to obtain stable estimates, the number of iterations of the training process is set as 600.

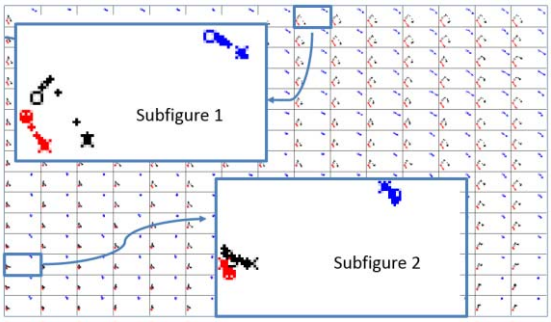

Figure 3: *The trained SOM for the motor plan map. The frontal boxes enlarge two neurons. In each neuron, the three control points, jaw, tongue tip and tongue dorsum, are plotted in red, black and blue, respectively.*

As described in the previous section, the deviation of the control points presents the neuron activation. Although the deviation is a good representation, it is difficult to give a spatial image for the readers. For this reason, we convert the deviation representation with their equilibrium positions to spatial locations for display. Figure 3 shows the trained SOM for the motor plan map in the spatial images, where the frontal boxes enlarge two neurons. In each neuron, the three control points, jaw, tongue tip and tongue dorsum, are displayed in red, black and blue, respectively. Each control point has 23 samples corresponding to 23 frames of the articulatory movement. The circles represent the rest positions, and the crosses represent the equilibrium positions. From the enlarged neurons, one can see that the trajectories of the three control points move from the rest positions (circle) to the equilibrium positions (cross) during the CV-syllable articulation. In subfigure 1, the control point of the tongue tip (black) first moves backward and upward to pronounce /d/, and then rapidly moves backward and downward to pronounce /V/. The tongue dorsum (blue) moves low-back in subfigure 1, and moves high-front in subfigure 2, respectively.

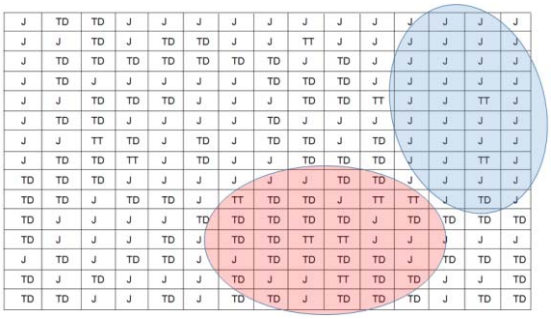

Figure 4: The motor control map. The letter in each neuron is displayed which organ of muscle plays the greatest role for the vowel part of the articulation. Jaw (J); tongue tip (TT); tongue dorsum (TD).

Figure 4 shows the motor control map based on the motor plan SOM. For each neuron in the motor plan SOM, we calculate the resultant forces of the muscles that effect on tongue tip, the tongue dorsum and the jaw to reach the corresponding control point positions of the vowel part, and the neurons in Figure 4 displays the maximum value of the three to demonstrate which articulatory muscle plays more important roles. We can observe in the figure, in some area like the top right corner, the jaw muscle is the key muscle, and the tongue muscle plays the greatest effect in some other areas. This distribution is somewhat similar to the measurement using the electrocortigraphic (ECoG) array over the left hemisphere of the brain [18].

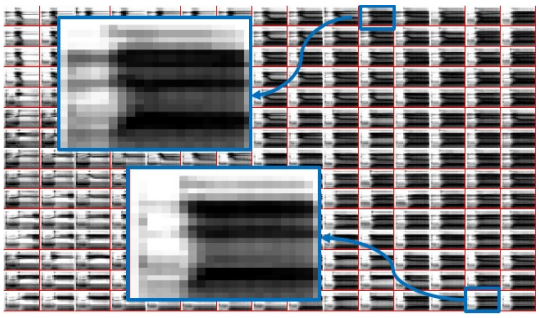

Figure 5: The SOM training result for the auditory map. Each neuron displays the neural representation of auditory state.

Figure 5 shows the trained SOM with 15 × 15 neurons for the auditory map. Each neuron in the SOM is displayed as a

spectrogram that reflects both the temporal and frequency characteristics of a speech signal. In each neuron, we can clearly observe the consonant segment, vowel segment and the transition between them.

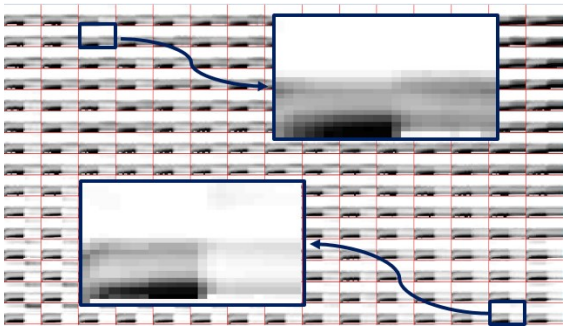

Figure 6: The SOM training result for the somato-sensory map. Each neuron displays the neural representation of the somatosensory state.

Figure 6 displays the SOM with 15 × 15 neurons trained for the somatosensory map. Each neuron in the SOM reflects the tactile condition in the vocal tract, which is calculated as in Equation (3). The representation becomes darker as the vocal tract becomes narrower. The dark parts indicate that the tongue and the tract wall are closely approximated, and that the collision is about to occur. In each neuron, the potential collisions mainly occur in the front part of the vocal tract when pronouncing the consonant /d/, since the tongue tip will move forward and upward to approach the alveolar region.

3.3. Mapping method for the SOMs

After the training processing, we obtained three SOMs for production of the CV-syllables, where the motor control SOM has not been obtained yet because more training data are needed. In this section, we propose the mapping method for the projection between different SOMs.

To investigate the relation between the three SOMs, we first labeled the neurons in each SOM as No.1 to No.225 by a left-to-right scan, and then generated 225 articulatory targets based on the control point information in the motor plan SOM, and then we drove the physiological articulatory model. By analyzing the articulatory and acoustic data, two temporal maps were obtained for the auditory and somatosensory features that correspond to the input of the motor plan SOM.

The relations are investigated between the temporal map and the SOM for the auditory and somatosensory states, respectively. In the human ears, human's perceptual resolution of formants is about 5%. Accordingly, if the difference between a neuron in the temporal map and the one in the corresponding SOM is less than 5%, we can consider these two neurons working the same task so as to construct a link between them. The difference between two neurons is calculated using the Euclidean distance. For example, the difference between the neuron No.8 in the temporal somatosensory map and the neuron No. 203 in the somatosensory SOM is 3.62%, so we think that there is a projection between these two neurons.

After projecting the temporal maps to the SOMs, we can establish the mapping relation between the three SOMs. Figure 7 shows the mappings between the motor plan SOM and auditory SOM. The numbers indicate how many projections each neuron has, while "1" was omitted and shown as blank lattices. In this figure, we plot the links for partial of the neurons with four or five projections.

In the mapping result, one can clearly see that "one-to-many" relations take place frequently in the SOM projection. On average, one neuron can project to 1.64 neurons in other SOM, while the maximum one has five projections between the SOMs. As shown in Figure 7, the "one-to-many" relation between the two SOMs is bidirectional, which work for the learning process and control process, respectively.

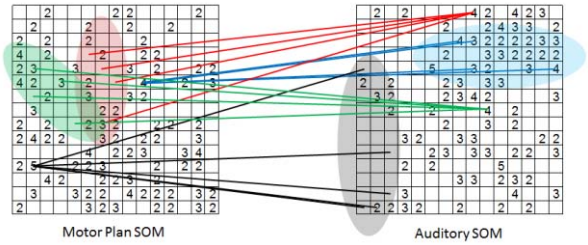

Figure 7: The projections between the motor plan SOM and the auditory SOM.

4. Conclusions

In the study, we proposed a new framework of the neurocomputational model for speech production, which was developed based on our previous work: different neuron states had a unique relation because they were placed in one SOM. In the new framework, we use three separate SOMs to describe the articulation, acoustics, and somatosensory states. This representation is capable of describing the realistic situation of the "one-to-many" relation between the different levels in the brain control processes for speech production. In our simulation, one neuron in a SOM on average can map onto 1.64 neurons in the neighbor SOM.

In the proposed model, we dealt with the time-varying properties of speech articulation for generating CV-syllables. The new representations were also designed to describe the dynamics in the different neuron states for CV-syllables. The motor plan SOM was represented using the deviations of the three control points from the rest positions during the CV-syllable articulation. The motor control map reflects the articulators with more activated muscles rather than the other articulators. The auditory SOM reflected both the temporal and frequency characteristics of speech sounds using spectrograms. The somatosensory SOM represented the tactile condition in the vocal tract, where the proposed parameter not only displays the collision but also shows the situation approaching the collision, which could avoid an overshoot at the collision.

In future studies, we further investigate the SOM for the motor control state and address more consonants by using the proposed model.

5. Acknowledgements

Our current work is supported in part by the National Basic Research Program of China PR (No. 2013CB329301), and the National Natural Science Foundation of China PR (No. 61233009 and No.6117501). This study is also supported in part by a Grant-in-Aid for Scientific Research of Japan (No.25330190).

6. References

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

[2] E.L. Saltzman, K. G. Munhal, "A dynamical approach to gestural patterning in speech production," Ecological Psychology 1, pp. 333-382, 1989.

[3] S.E. Boyce, "Coarticulatory organization for lip rounding in Turkish and English," The Journal of the Acoustical Society of America 88, pp. 2584, 1990.

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

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

[6] 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.

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

[8] B. J. Kröger, J. Kannampuzha, E. Kaufmann, "Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception." EPJ Nonlinear Biomedical Physics 2:2(2014)

[9] 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.

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

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

[12] T. Kohonen, "Self-Organizing Maps," Springer, Berlin New York, 2001.

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

[14] J. Dang, K. Honda, "Estimation of vocal tract shape from sounds via a physiological articulatory model," Journal of Phonetics 30, pp. 511-532, 2002.

[15] M. Garagnani, T. Wennekers and F. Pulvermuller, "A neuroanatomically grounded Hebbian-learning model of attention-language interactions in the human brain," European Journal of Neuroscience dol: 10.1111

[16] X. Chen, J. Dang, H. Yan, Q. Fang and B.J. Kröger, "A neural understanding of speech motor learning", APSIPA ASC 2013

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

[18] K. Bouchard, N. Mesgarani, K. Johnson, E. Chang "Functional organization of human sensorimotor cortex for speech articulation" Nature(2013), Vol. 495, 327-332

[19] M. Cao, A. Li, Q. Fang, E. Kaufmann, B.J. Kröger, "Interconnected growing self-organizing maps for auditory and semantic acquisition modeling," Front. Psychol(2014). 5:236.