Morphological personalization of a physiological articulatory model

Nana Nishimura*, Shin'ichi Kawamoto*, Jianwu Dang*† and Kiyoshi Honda†
*School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan.
E-mail: {n-nishimura, kawamoto, jdang}@jaist.ac.jp
†School of Computer Science and Technology, Tianjin University, Tianjin, China.
E-mail: khonda@sannet.ne.jp

***Abstract*—A physiological articulatory model is capable of simulating movements of the articulatory organs together with their morphology, without limitation of the observation approaches and ethics. The model can be a tool for investigating an aspect of individual characteristics of speech, if it is adjusted to the form a particular speaker's organs. This study reports our effort to personalize an articulatory model to multiple speakers. To do so, we propose a method for constructing personalized articulatory models based on the adaptation of a prototype model by transformation. Accordingly, new models were built for three speakers, which successfully reflected their morphological features. Also, the models were found to be able to simulate typical tongue movements in the same way as the prototype model performed.**

I. Introduction

Speech signals convey linguistic and non-linguistic information, and individual difference belongs to the latter. To investigate aspects of individual differences in speech signals, it is necessary to consider the effect of the morphology and movement of the articulatory organs on speech production. A physiological articulatory model is a powerful tool to investigate the issue, since it is equipped with all the necessary components such as anatomical structure or muscular functionality, and the model basically has no limitation for observation approaches and ethics. Also, it could be a good tool for clinical evaluation of articulation disorders, if its form is adjusted to the articulatory organs of a particular patient. For instance, Fujita et al. [1] successfully simulated the changes in movement of the tongue before and after a partial glossectomy based on his physiological model.

A number of articulatory models have been constructed for speech production studies [2]–[8]. Most of these studies have focused on investigating basic articulatory mechanisms of speech production, and they have not examined speaker-to-speaker differences in articulation process. To discuss the individual difference using a physiological articulatory model, it is necessary not only to reproduce speech sounds from the model but also to adapt the shapes of the speech organs for multiple subjects. Recently, Winkler et al. [9], [10] investigated the articulatory individual difference by constructing two speaker-specific models. Since they employ a 2D model, their study is limited to the model's midsagittal adaptation to speaker's organs, and the description of speaker's acoustic variations is largely speculative.

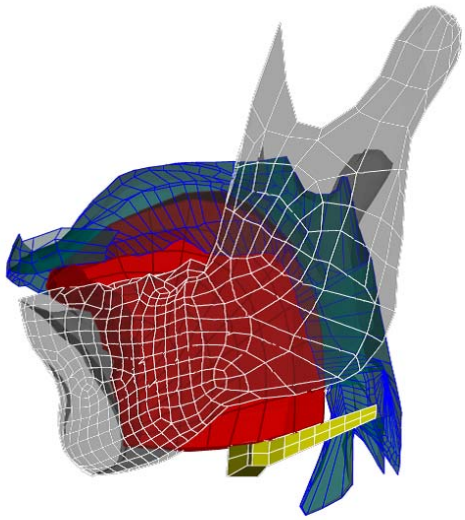

Fig. 1. Prototype model.
3D physiological articulatory model expand the model proposed by Dang et al. [11], [12]. The model consists of the tongue, jaw and vocal tract wall, etc., based on MR images.

Among the models, we employ the more elaborate model proposed by Dang et al. [11], [12], which is capable of realizing articulation and speech sounds via combinations of muscle activation signals. This model also enables us to investigate muscle activation patterns in speech production [13] or one-to-many relationship between speech sound and articulation [14]. However, the model-based analysis of the individual difference of articulation faces a technical problem for constructing new models for multiple speakers. Since the model has complicated structures, it would require time and effort to build a model for arbitrary speakers based their morphological data. To expand the use of the physiological articulatory model to the new application, it would be necessary to develop a technique to build personalized models with minimal effort. In this study, we propose a method to efficiently adapt a prototype physiological articulatory model to a specific speaker based on his/her morphological data.

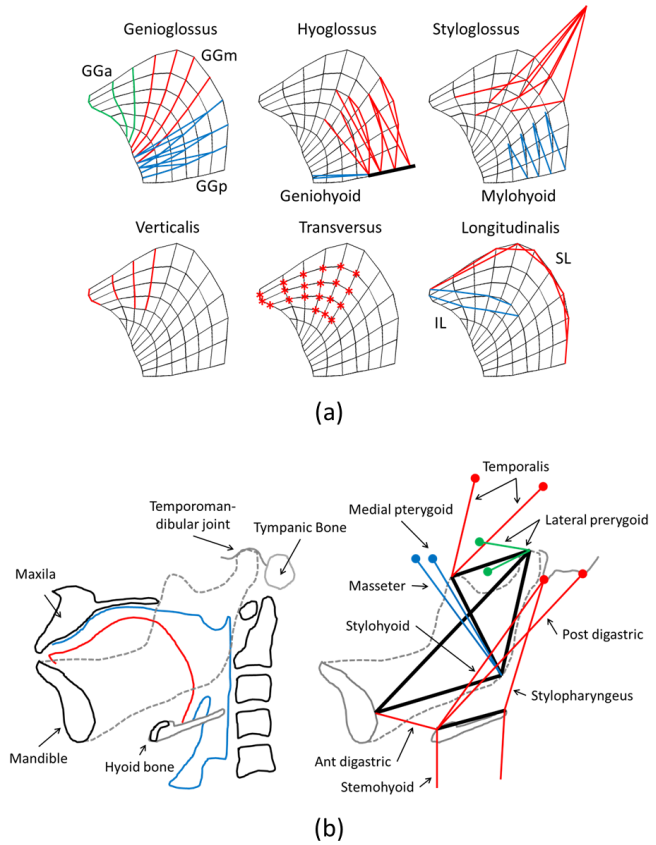

Fig. 2. The arrangement of the muscles [12].
(a): arrangement of the intrinsic/extrinsic muscles of tongue, (b): arrangement of the jaw muscles.

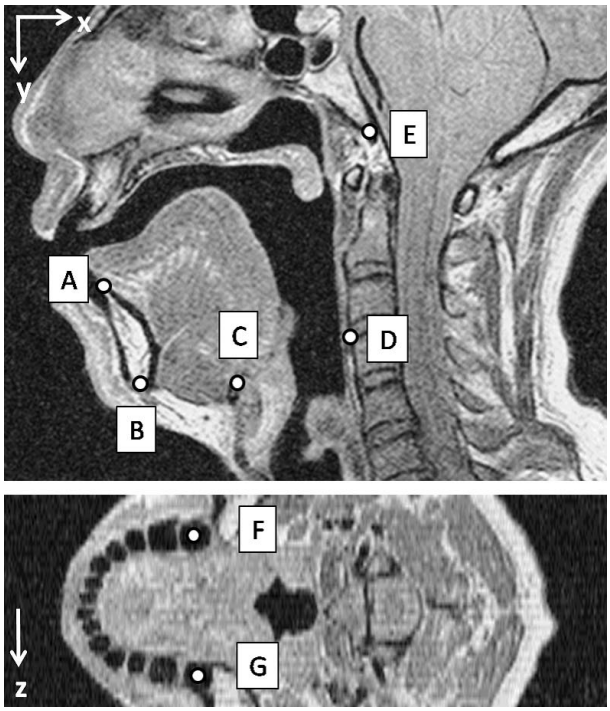

Fig. 3. Anchor points for linear transformation.
A: upper end of the bone marrow of the jaw, B: lower end of the bone narrow of the jaw, C: highest point of the hyoid bone, D: anterior edge of the cervical disc at C3-C4, E: lower end of the clivus, F: right second mandibular molar, G: left second mandibular molar.

II. Morphological Personalization

The prototype model used in this study is based on the one originally proposed by Dang et al. [11], [12]. The model consists of the tongue, jaw, hyoid bone, larynx, and vocal tract wall based on MR images of a Japanese male speaker. The articulatory organs were inter-connected by muscles and ligaments (Fig. 2), and the muscles are driven by muscle activation signals. The model's articulatory organs were defined 2.5 mm from midsagittal plane to right and left. In this study, we expand this model to full 3D shape (Fig. 1). Prototype model partially employs the function in ArtySynth [15] for computation and display.

A. Material

To obtain speaker's morphological information, we used 3D MR images during sustained vowel productions recorded by the phonation-synchronized imaging method [16], [17]. In this study, the MR images of three Chinese speakers were used as target speakers for model adaptation. Since the Chinese vowel system differs from that of Japanese, we choose a vowel similar to Japanese vowel [e] that was used to build the prototype model. Three subjects (subject A to C, 1 male and 2 female speakers) have no history of speech disorders. The MRI data was obtained at ATR Brain Activity Imaging Center (ATR-BAIC) using a clinical MRI scanner, the MAGNEX ECLIPSE 1.5T Power Drive 250 (Shimadzu-Marconi). The scan parameters were as follows: 3.4 ms TE, 2200 ms TR, sagittal slice plane, 1.5 mm slice thickness, 1.5 mm slice gap, 256 256 mm field of view (FOV), and 512 512 pixel image size. The images were converted from DICOM to TIFF, together with other necessary preprocessing [18]. The voxel size was normalized to 0.5 0.5 0.5 mm. In addition, the tooth superimposition method [19] was applied before extracting contours of the vocal tract wall.

B. Personalization Method

In the personalization of the model, we adapt the prototype model to the morphological data of a given new subject, and the muscle structure of the prototype model was adjusted piecewise-proportionally to the morphology of the new model. This method enables us to obtain a new model by transforming the contour of articulatory organs while keeping the original model structure. This approach is much more efficient than that to newly build a model without the prototype model as a reference.

The method for personalization used in this study consists of three procedures: 1) transforming the tongue, vocal tract wall, and jaw; 2) transforming the rigid structures; and 3) tuning the attachment point of the styloglossus.

1) Tongue, vocal tract wall, and jaw: The adaptation of the prototype model to a new model is a mapping from one

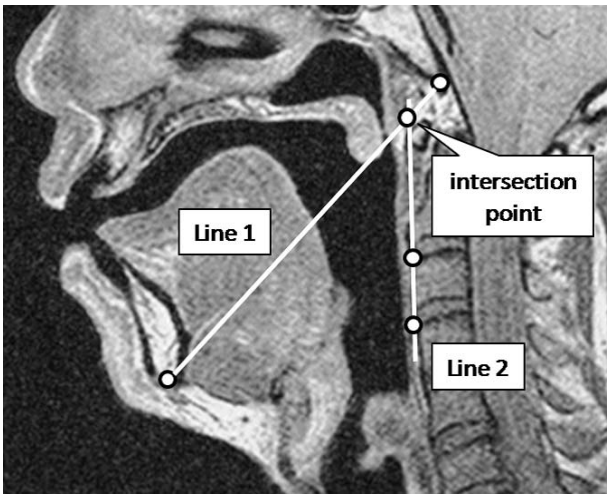

Fig. 4. Estimation of SG attachment point.
The two-line intersection was measured to estimate SG's attachment point for target speaker's narrow MRI volumes lacking styloid process regions.

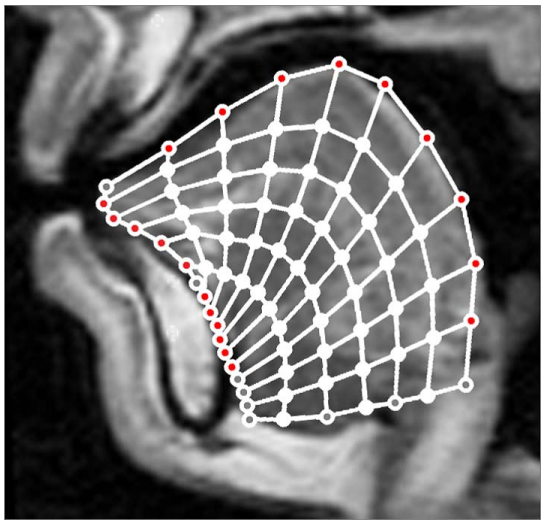

Fig. 5. Feature points of the tongue on the midsagittal plane.
red: target points for optimization, gray: base points, and white: other mesh points. In the model, 3D shape of the tongue is formed by five sagittal layers.

to another based on the morphological landmarks (i.e., feature points of articulatory organ shape). The radial basis function (RBF) [20] is employed in the mapping with

$$s_{i,k} = (A(\vec{p}_i))_k + \sum_{j=1}^{N} w_{j,k}\, g(||\vec{p}_i - \vec{f}_j||),\ i = 1, 2, \cdots, M \quad (1)$$

$$g(r) = \begin{cases} r^2 \log(r) & r \neq 0 \\ 0 & r = 0 \end{cases} \quad (2)$$

where $\boldsymbol{S} = (s_{i,k}) \in \mathbb{R}^{M\times 3}$ is a new model coordinates, $\boldsymbol{P} = (\vec{p}_1, \cdots, \vec{p}_M)^T \in \mathbb{R}^{M\times 3}$ is the prototype model coordinates, $\boldsymbol{F} = (\vec{f}_1, \cdots, \vec{f}_N)^T \in \mathbb{R}^{N\times 3}$ is the reference points that are selected feature points as inputs, $A(\cdot)$ is an affine transformation, $k = 1, 2, 3$ is the index number of the dimension, N is the number of the morphological landmarks, and M is the number of the prototype model coordinates. $||\cdot||$ means L_2 norm of the vector. $g(\cdot)$ is the thin-plane spline kernel, which we choose among others because it does not require optimization. This procedure enables us to reduce the number of reference points that are necessary to construct a model of the articulatory organs.

In the RBF method, the reference point configuration, such as the location and the number of input points, is relevant to the approximation accuracy of the target shape. For this reason, the reference point configuration is optimized to reduce the workload of the reference point extraction manually. The optimization criterion adopts the minimization of the approximation error of the tongue contour.

2) Rigid structures: The size information is used to fit the general shape of the prototype model to the observed data by proportional resizing, and the muscle structure is reformed in the same way simultaneously. We extract the anchor points (shown in Fig. 3) five times for each and use the average to calculate the following size parameters.

$$\text{Height} = \left|y_B - y_E\right| \quad (3)$$

$$\text{Width} = \left|z_F - z_G\right| \quad (4)$$

$$\text{Depth} = \left|\tfrac{x_A + x_B}{2} + \tfrac{x_F + x_G}{2}\right| \quad (5)$$

$$\text{TongueDepth} = \left|x_C - x_D\right| \quad (6)$$

After obtaining these size parameters, we calculate the magnification factor between the prototype model and the speaker model.

3) Attachment point of Styloglossus(SG): In our earlier work with linear model mapping, a few problems were noticed regarding the styloglossus (SG): unrealistic orientation of this muscle after the transformation and difficulty in realizing velar constriction. The problems seemed due to inadequate mapping of the SG, and re-definition of the styloglossus in personalized models was necessary.

The attachment point of the SG is the styloid process on the skull base, and its location influences the tongue movement. Since the styloid process could not be identified on MRI data sets using personalization, we first estimate the attachment point by its projection onto the midsagittal plane as the intersection point of the two lines on the midsagittal x- and y-axes as shown in Fig. 4. In the figure, Line 1 is defined from the lowest point of the clivus to the lower end of the bone mallow of the jaw, and Line 2 is between the edges of the cervical disc at C2-C3 and C3-C4. After obtaining the intersection, its location on the z axis is derived by (4).

The estimation error of the SG attachment was evaluated by using MRI data sets that include the styloid process. This error was evaluated on the midsagittal plane by the Euclidean distance between the estimate point and the manually extracted

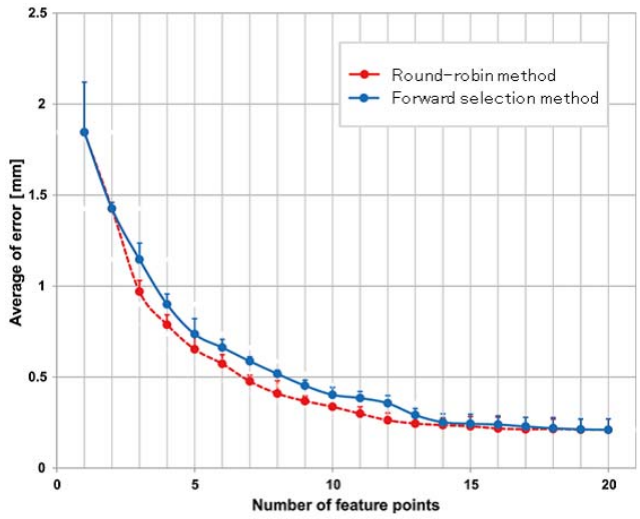

Fig. 6. Error of tongue contour.
X axis is the number of feature points (excluding the base points). Y axis is the minimum error in the three-dimensional contour of the tongue (average of the three subjects).

point. The mean error for the three subjects was 7.13 mm. Within our criteria for evaluation, this error did not influence the effect of the SG to make a contact between the tongue and the palate in simulation.

III. Results

A. Optimization of the feature points

Fig. 5 shows the feature points of the tongue on the midsagittal mesh layer. The red circles are the object for optimization, the gray circles are the base points that are used in the RBF transformation constantly, and the white circles are other mesh points. We optimized the feature points to minimize the contour error by changing their location and the numbers by employing two methods, round-robin competition method and forward selection method. The round-robin competition method calculates all possible combinations for a given number of feature points, and then it selects one combination with which the error reached the minimum. The forward selection method adds a new point in the previously determined feature points stepwise, and it selects the one that can reduce the error further. After transformed by the RBF method, we calculate the error of the contour by the Euclidean distance between the correct surface and the feature points of the tongue contour.

The optimization of the feature points was conducted as described above for each model. Fig. 6 shows the relationship between the error in the tongue contour and the number of feature points for three subjects. In the figure, the horizontal axis is the number of feature points (excluding the base points) and the vertical axis is the minimum error in the three-dimensional contour of the tongue (average of the three subjects). The result shows that the error of the forward selection method is slightly larger than that of the round-robin competition method, where the maximum difference is about 0.18 mm. Referring to the voxel size 0.5 0.5 0.5 mm, both methods can keep the errors within one pixel by using ten and more feature points (excluding the base points) on each layer. The standard deviation (error bars in the figure) is within 0.09 mm among the three subjects, except in the case of a single feature point. This suggests that the optimized locations of the feature points are free from the effect of speaker's morphological variations.

B. Construction of personalized models

Fig. 7 shows the midsagittal MRI and articulatory organ contours (the vocal tract wall, jaw, and tongue, before and after the simulation) for the midsagittal plane of each personalized model. Comparing the MR images and models, it is confirmed that the personalized models kept the geometrical relationship of the articulatory organs and successfully reproduced the personal features. When a force is applied to each muscle, simulation results of tongue movement corresponded with those of the prototype model. Fig. 7 shows the results of simulation when SG is activated in the level of 40 %. In this study, we didn't use the collision decision method between the palate and tongue for investigate the tongue movement when used the same force. These tongue movements are also consistent with the results from the previous study [21].

IV. Discussion

A. Personalization method

We proposed a new method to adapt the prototype model to arbitrary speakers using the speaker's morphological data. The constructed models were capable of simulating typical tongue functions similar to those of the prototype model. This result indicates that our method is applicable to modeling speaker individuality of the articulatory organs. In addition, by adapting the features from the prototype model to the new speakers, the effort for model construction was significantly reduced.

Furthermore, we constructed three additional models for new subjects (three Chinese male speakers, D, E, and F) using the proposed method. The MRI data were obtained with the same conditions as those for the subjects A, B, and C. The personalized models for all six speakers performed similarly with no obvious gender differences.

B. Optimization of the feature points

In model optimization, the round-robin competition method and forward selection method made no significant difference for the contour errors after optimization. This result indicated that the forward selection method gives the similar result as that of the round-robin competition method in reducing the contour error. This suggests that we can freely select the number of feature points according to the desired accuracy by using result of the forward method, where the forward selection is convenient for users to operate.

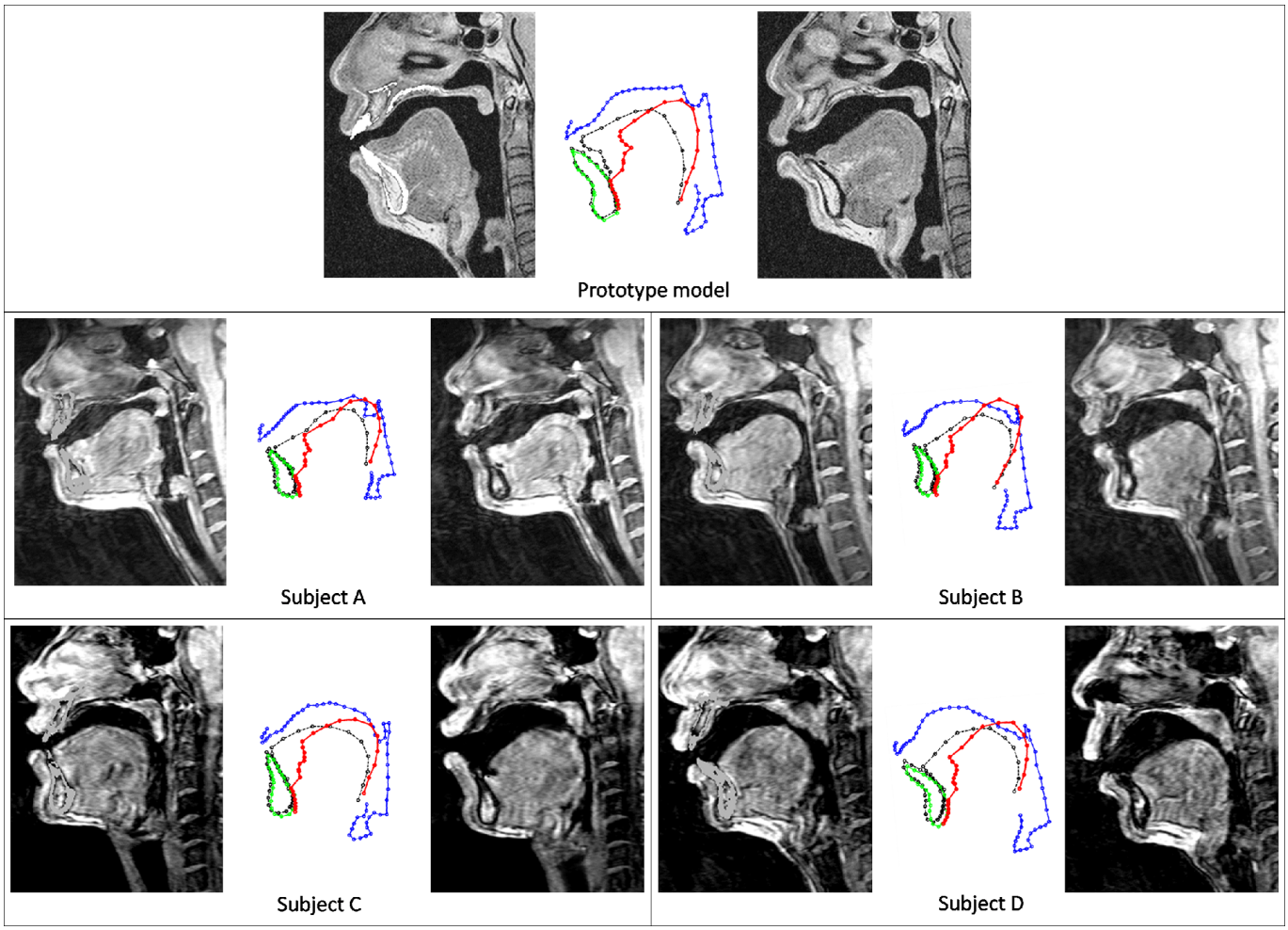

Fig. 7. Midsagittal MRI and simulation result.

Midsagittal MRI and model simulation are shown in each panel. left: Chinese vowel [ɣ] (the prototype model shown Japanese vowel [e]), middle: simulation result (the tongue: black dotted line shown the initial position and red solid line shown SG=40 %), right: Chinese vowel [o] (the prototype model shown Japanese vowel [o]). subject A, B: female, subject C, D: male. In this study, we didn't use the collision decision method between the palate and tongue (see text).

C. Simulation performance

In the simulation results (shown in Fig. 7), the movement of the tongue differed across subjects in direction or distance even if the same force was applied. The differences were also seen in the locations of constriction between the tongue and palate. One possible cause of this problem may be that the same articulatory configuration can be formed using different muscle combinations. In this simulation, we used the SG alone to simulate the tongue movement. However, it has been conjectured that the back-upward tongue movement could be realized by the intrinsic muscles alone [22], which suggests that further work may be necessary.

Another possible source of the problem may be the muscular structure of the tongue and its modeling. In the prototype model, the styloglossus runs linearly between the attachment and insertion points. However, recent studies suggest a new interpretation about this muscle [23], [24]. Although our current model is capable of performing tongue articulation for extreme back vowels, detailed muscle anatomy must be reflected in the model for further improvements.

V. Conclusion

In this study, we proposed a method for constructing personalized articulatory models much more efficiently than that for building ones without any prototype model. By using those personalized models, it will become possible to compare individual differences of articulatory structure or movement of the tongue. The feature point selection method used in this study can control the model accuracy by choosing the number of feature points. In the future, we will personalize the muscular structure of the tongue to study the details of their roles in speech production and other physiological activities.

ACKNOWLEDGMENT

This study was supported in part by JSPS KAKENHI Grant Numbers 22500150, 24700191, 25240026, 25330190 and in part by grants from the National Natural Science Foundation of China (Key Program No. 61233009, and General Program No. 61175016).

REFERENCES

[1] S. Fujita, J. Dang, N. Suzuki and K. Honda, "A Computational Tongue Model and its Clinical Application," *Oral Science Internationa*, 4(2), pp. 97-109, 2007.

[2] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, 53, pp. 1070-1082, 1973.

[3] Y. Payan and P. Perrier, "Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis," *Speech Communication*, 22(2-3), pp. 185-205, 1997.

[4] V. Sanguineti, R. Laboissiere and D. Ostry, "A dynamic biomechanical model for neural control of speech production," *J. Acoust. Soc. Am.*, 103(3), pp. 1615-1627, 1998.

[5] P. Badin and A. Serrurier, "Three-dimensional linear modeling of tongue: Articulatory data and models," *7th liternational Seminar on Speech Production*, pp. 395-402, 2006.

[6] A. Serrurier and P. Badin, "A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data," *J. Acoust. Soc. Am.*, 123(4), pp. 2335-2355, 2008.

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

[8] B. Kroger, J. Kannampuzha and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Communication*, 51, pp. 792-809, 2009.

[9] R. Winkler, S. Fuchs, P. Perrier and M. Tiede, "Biomechanical Tongue Models: An Approach to Studying Inter-speaker Variability," *Interspeech Florence*, pp. 273-276, 2011.

[10] R. Winkler, S. Fuchs, P. Perrier and M. Tiede, "Speaker-specific biomechanical models: From acoustic variability via articulatory variability to the variability of motor commands in selected tongue muscles," *9th liternational Seminar on Speech Production*, pp. 219-226, 2011.

[11] J. Dang and K. Honda, "A physiological articulatory model for simulating speech production process," *Acoust. Sci. and Tech.*, 22(6), pp. 415-425, 2001.

[12] J. Dang and K. HondaC"Construction and control of a physiological articulatory model," *J. Acoust. Soc. Am.*, 115(2)Cpp. 853-870C2004D

[13] X. Wu, Q. Fang and J. dang, "Investigation of Muscle Activation in Speech Production Based on an Articulatory Model," *Proc. ISCSLP.*, pp. 330-334, 2010.

[14] A. Nishikido and J. Dang, "Model-based investigation on one-to-many relationship between speech sound and articulation," *J. Acoust. Sco. Jpn.*, 67(1), pp. 3-14, 2011. (in Japanese)

[15] ArtiSynth, "3D Biomechanical Modeling Toolkit," Online:http://www.magic.ubc.ca/artisynth/pmwiki.php.

[16] S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, Y. Nakamura and N. Ninomiya, "MRI-based speech production study using a synchronized sampling method," *J. Acoust. Sco. Jpn.*, 20(5), pp. 375-379, 1999.

[17] H. Takemoto, K. Honda, S. Masaki, Y. Shimada and I. Fujimoto, "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," *J. Acous. Soc. Am.*, 119(2), pp. 1037-1049, 2006.

[18] G. Wang, T. Kitamura, X. Lu, J. Dang and J. Kong, "MRI-based Study of Morphological and Acoustical Properties of Mandarin Sustained Steady Vowel," *J. Signal Processing*, 12(4), pp. 311-314, 2008.

[19] H. Takemoto, T. Kitamura, H. Nishimoto and K. Honda, "A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions," *Acoust. Sci. and Tech.*, 25(6), pp. 468-474, 2004.

[20] N. Arad and D. Reisfeld, "Image Warping Using few Anchor Points and Radial Functions," *Computer Graphics Forum.*, 14(1), pp. 35-56, 1995.

[21] Q. Fang, S. Fujita, X. Lu and J. Dang, "A model-based investigation of activations of the tongue muscles in vowel production," *Acoust. Sci. and Tech.*, 30(4), pp. 277-287, 2009.

[22] I. Stavness, J. E. Lloyd and S. Fels, "Automatic prediction of tongue muscle activations using a finite element model," *Journal of Biomechanics*, 45, pp. 2841-2848, 2012.

[23] S. Takano and K. Honda, "An MRI analysis of the extrinsic tongue muscles during vowel production," *Speech Communication*, 49, pp. 49-58, 2007.

[24] K. Honda, S. Takano and H. Takemoto, "Effects of side cavities and tongue stabilization: possible extensions of quantal theory," *Journal of Phonetics*, 38, pp. 33-43, 2010.