Study of Control Strategy for a Physiological Articulatory Model by Mimicking Speech Motor Learning

Xiyu Wu[1], Jianguo Wei[2] and Jianwu Dang[1,2]

[1] School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan
E-mail: {xiyuwu, jdang}@jaist.ac.jp

[2] School of Computer Science and Technology, Tianjin University
No. 92 Weijin Road, Nankai District, Tianjin, China
E-mail: jianguo.fr@gmail.com

Abstract

A full 3D physiological articulatory model was developed to simulate the mechanism of human speech production. In order to control the physiological articulatory model to generate speech sound, we construct a mapping from acoustic objective to muscle activation patterns by simulating the babbling process in speech acquisition. After the babbling phase, the mapping from acoustic objective to muscle activation patterns was built up using a two layer neural network. In order to evaluate the mapping, muscle activation patterns were estimated from a known speech sound, and then were used to drive the physiological articulatory model to synthesize speech. The comparison results of the acoustic space and articulatory space showed that the proposed mapping-based control strategy was feasible for model control.

1. Introduction

Generating speech sound based on the mechanism of human speech production is a long term issue. For years, many researchers have endeavored to construct physiological articulatory models and control those models to generate speech [1][2]. A straightforward control parameter for the physiological articulatory model is the muscle activation pattern. The schematic of the control strategy is shown in Figure 1. Accordingly, speech production using by the physiological articulatory model arrives at the estimation of muscle activation patterns according to a given speech. So far, the estimations implemented in previous models usually consist of two steps. . As shown in Figure 1, arrow ① and arrow ② represent these two estimations. Step 1: Obtain articulatory target according to the intended speech by speech planning [3] or from empirical observation [4]. Step 2: Estimate muscle activation patterns from articulatory target by EP map [2] or heuristic method [4]. The estimation methods from articulatory target to muscle activation patterns have two main disadvantages: 1) The object of the EP map is to realize the articulatory target by muscle activation patterns without considering whether they obey the rules of humans or not. 2) It is difficult to automatically estimate muscle activation patterns for all the phonemes according to the heuristic method. From the application point of view, these two-step estimations may result in accumulated errors.

According to the motor learning theory [14], the coordinated muscle activation patterns are the result of task specific motor learning. When a movement is repeated over time, a long-term muscle memory is created for that task, eventually allowing it to be performed without conscious effort. After the motor learning process the relations among the acoustics, articulatory configurations and muscle activation patterns will be constructed. In this study, we clarify the relations by simulating the babbling process of infant.

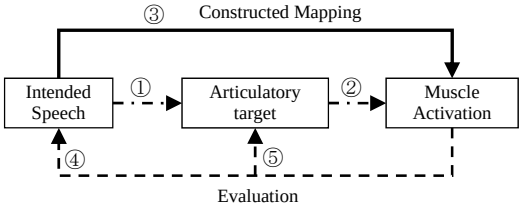

Figure 1. Schematic of the control strategy and evaluation

Other than the two-step estimations [2][4], it is much more likely that when human speak without any interference the direct mapping from acoustics to muscle activation patterns is used. In this study we construct a mapping (arrow ③ in Figure 1) from acoustic objectives to muscle activation patterns using a two layer feedforward neural network.

The motivation of constructing the mapping is to control the physiological articulatory model to generate desired speech. Therefore, the evaluation was done in the acoustic space (arrow ④ in Figure 1) to verify whether the acoustic objective can be realized or not. It seems that there is no guarantee in the articulatory space using the direct mapping from acoustic objective to muscle activation patterns to generate speech. For this reason, we carried out the evaluation to check the correctness in the articulatory space (arrow ⑤ in Figure 1) when using the constructed mapping to generate desired speech, where the correctness refers to that whether the articulatory configurations are the same as babbling process, when generating the same acoustic objective using the constructed mapping.

2. Method

In the following, we first briefly describe the construction of the physiological articulatory model and the muscle functions, and then introduce the mapping construction from acoustic objective to muscle activation patterns through the babbling process.

2. 1 Construction of Articulatory Model

A full 3D continuum physiological articulatory model was constructed based on ArtiSynth — A 3D Biomechanical Modeling Toolkit [5]. The morphological structure of the 3D tongue was obtained based on the volumetric MR images when producing the Japanese vowel /e/. The contours of the jaw and the vocal tract wall were carefully extracted from the MR images of vowel /e/ superimposed with the lower and upper teeth at the interval of 0.4cm in the transverse dimension. Definition of the muscles was based on their anatomical orientation. For the details of morphological data are described in [4]. Figure 2 shows the lateral view of the physiological articulatory model. From this figure we can see the configurations of the tongue, vocal tract wall, jaw and hyoid bone. The lines with a bundle shape represent the definition of different articulatory muscles. For simplicity, the larynx part was not used in this research although its structure has been constructed in this model.

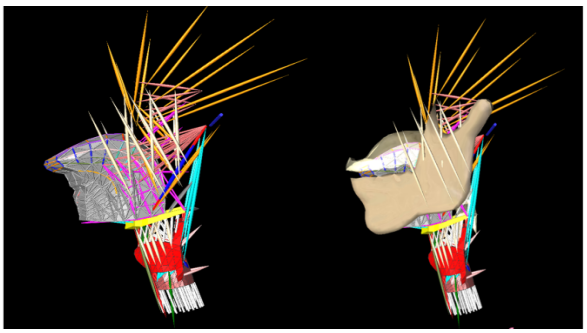

Figure 2. Physiological articulatory model

When articulatory muscles are activated in certain activation levels the articulators move to specific positions and maintain an equilibrium position. The activation level of the muscles defined in the physiological articulatory model was normalized within [0, 1], where 0 means that the muscle is not activated, and 1 means that the muscle is fully activated and generates maximum force. Figure 3 shows two examples of equilibrium positions of articulators driven by single muscle activations. Left panel and right panel show the midsagittal plane of the articulators driven by HG (Hyoglossus) and SG (Styloglossus) respectively with 0.2 activation level. In this figure, blue lines show the initial positions and black lines represent equilibrium positions.

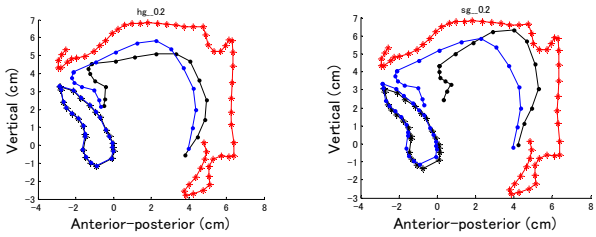

Figure 3. Equilibrium positions of the model driven by Hyoglossus (left panel) or Styloglossus (right panel) with an activation level of 0.2

Different muscle activation patterns can drive the articulators to different positions which results in the variance of vocal tract shape. The cross-sectional area function of a vocal tract can be calculated directly from the vocal tract shape, because the model is full 3D. The transmission-line model was implemented in the vocal tract area function to generate speech sound. The details of speech generation can be found in [6].

2.2 Babbling Simulation

To generate desired speech specific articulatory muscles are activated coordinately with specific activation levels. The emergence of articulatory muscle coordination is the consequence of task specific motor learning. In this babbling stage we did not impose any synergies and gave the largest possible number of degrees to the muscles.

During the babbling of infant, the formant distribution of vowels varies from language to language [7]. This evidence implies that the babbling is language dependent. In our study, Japanese is used as the language in the babbling simulation. Note that in the motor learning for speech production, auditory feedback is an important channel, as well as the somatosensory channel. In this babbling simulation, we just focused on the somatosensory channel. Because the articulation manners of consonants are complicated, for simplicity, in the motor learning process only vowels were involved.

The vocal tract shapes of five Japanese vowels [8] are used as the motor learning targets, which were the average positions from X-ray microbeam observation data. In the X-ray microbeam data, five pellets were attached on the jaw and tongue surface of the subject. The X-ray microbeam data and the morphological prototype of the physiological articulatory model were obtained from the same subject, so the observed X-ray microbeam data and the positions of the simulated articulatory model can be compared directly. In Figure 4, five squares show the target position of vowel /a/; red points represent the rest positions; asterisks are the realized positions through motor learning process based on the physiological articulatory model.

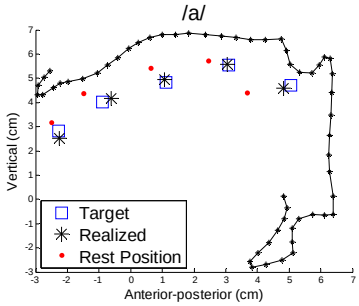

Figure 4. Articulatory target of vowel /a/

In babbling process, a number of different muscle activation patterns may tentatively input into the articulatory model in order to find the muscle activation pattern that can drive the articulatory model to reach the target. As the basic constraints, we implemented two principles in the process: 1) At each step, the movement must approach to the target, and

get closer step by step. 2) The selected muscle activation patterns must obey the economy and efficiency principles [9].

As a result, 9703 cases were obtained for five Japanese vowels in the babbling simulation. Because the lip part has not been constructed in the articulatory model, when calculating the area function, the typical lip area and lip length of five Japanese vowels were added to the area functions. The distributions of the first three formants are shown in Figure 5. Left panel shows the distribution of the first and second formants; right panel is for the first and third formants. From this figure, one can see that the formants have covered almost all the acoustic regions.

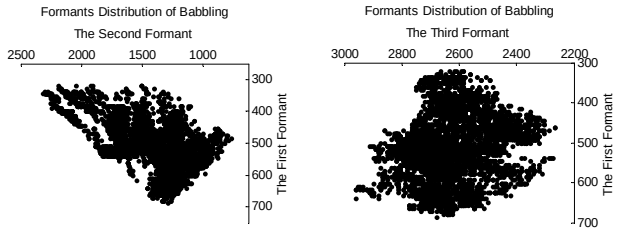

Figure 5. Distributions of the first three formants

2.3 Mapping Construction

The 9703 babbling cases obtained above consist of the parameters involved in three hierarchic levels of speech production: motor commands (muscle activation patterns), articulatory positions and acoustics. For the reason mentioned earlier, we constructed a mapping from acoustic objective to muscle activation, instead of the two-step estimation. In the previous neurocomputational model of speech production [10][11], mapping was constructed from auditory (acoustics) maps to articulatory position maps by neural network. The neural connections were adjusted during babbling training where these connections were used to simulate the synapse connecting neurons in brain cortex. In this study, a two layer neural network was used to construct this mapping.

The input vector was the first three formants and the output vector was 20 muscles and 2 lip parameters that are lip area and lip protrusion. All the parameters for the muscles were normalized between 0 and 1. We randomly selected 6947 simulations as the training data set and the left 2756 simulations as the testing set. A set of experiments were conducted uisng a number of neural network with different configurations. The neuron number was set as 5, 10, 15, 20 and 25, respectively. different combination of transfer functions. The best configurations were obtained by choosing the smallest prediction error in the opening test. As a result, the following configuration was used for the neural network to achieve the best performance. Transfer function of tansig (hyperbolic tangent sigmoid transfer function) and purelin (Linear transfer function) were used for hidden layer and output layer, respectively, and the neural number was set as 20 in the hidden layer.

In the opening test, estimating muscle activation patterns from the first three formants using the trained neural network the error rate was 0.068. In order to assess whether this mapping can be used to synthesize desired speech or not, we have to input the estimated muscle activation patterns into the physiological articulatory model.

3. Evaluation

The motivation of the constructed mapping was to control the physiological articulatory model to generate intended speech. Therefore, we first evaluate the constructed mapping in the acoustic space (arrow ④ in Figure 1) by carrying out the following steps. 1) Input the first three formants from a vocal configuration that was not used in the training to estimate muscle activation patterns. 2) Input the estimated muscle activation patterns into the physiological articulaotry model to synthesize speech. 3) Compare the input formants and the formants of synthesized speech to evaluate the control strategy. In this evaluation, 2756 sets of formants were inputted into the mappings. The distribution of the tested formants also covers almost all the acoustic regions but with less density than the training data.

$$\text{Re}\,lative \quad Error = \left|\frac{(OF - DF)}{DF}\right| \times 100\% \qquad (1)$$

Table 1. Absolute and relative errors

	F1	F 2	F 3
Absolute Errors (Hz)	32	100	62
Relative Errors (%)	6.65	7.08	2.00

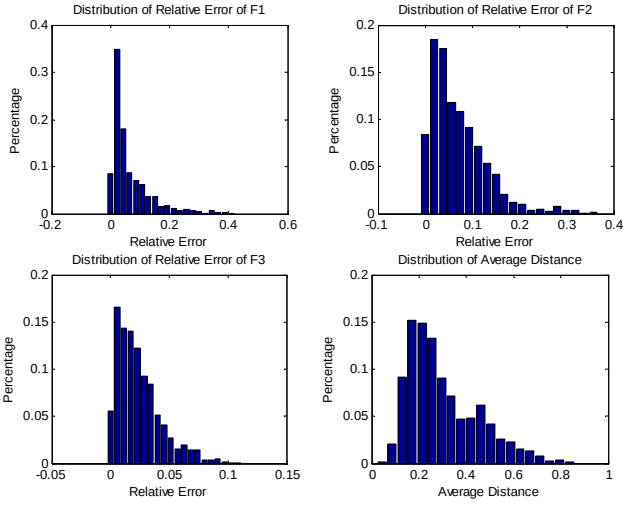

Figure 6. Distributions of relative errors of the first three formants and the average distance in the articulatory space (the vertical is a rate but not percentage)

According to the perception experiments [12][13], the difference limens were 5% for the first formant and 10% for the second formant. Accordingly, if the relative errors are within the region of the thresholds the synthesized speech can be regarded as achieving the goal. The relative error is defined in Equation (1), where OF and DF represent output formants and desired ones, respectively. The average errors over all cases are show in Table 1. From this table, one can see that the relative error of the first formant is little bit larger than the threshold and the second formant is within the threshold. Totally 55% of the simulation cases

meet the threshold of both the first formant and the second formant. The percentages of the relative errors for the formants are shown in Figure 6. The distribution of the relative error for the first formant, the second formant and the third formant are shown in the upper left, upper right and lower left panels, respectively. From this distribution, one can see that most of the relative errors are in a very small range. According to the evaluation, we can conclude that in most cases using the trained mapping the desired speech can be synthesized using the physiological articulatory model. Although there are still 45% of the desired acoustic targets that cannot achieve target so closely, although the relative errors are also in a small range. For identifying the vowel category, the tolerance of the formants is much larger than the difference limens, that is, 5% for the first formant and 10% the second formant. From this point of view, the errors introduced by using the mapping are acceptable.

Theoretically, there is a one-to-many problem from acoustic space to muscle activation space. For this reason, we clarify the situation by checking whether the articulatory configuration is the same as that obtained in babbling process, when generating the same acoustic objective using the mapping. There are no new articulaotry configurations generated for the same speech, if the articulatory space is consistent with the babbling process when produce the same acoustic objective.

In this evaluation we compared the articulatory configuration generated by using the direct mapping and the articulatory configuration generated during babbling process. The test set is the same as the evaluation in acoustic space. The five points in the midsagittal plane were used to represent the articulatory configurations (see Figure 4). The average distance of the five points was used as the parameter in the comparison. The lower right panel in Figure 6 is the distribution of this parameter (put the unit in the figure!). From this figure we can see that most of the differences are very small. Totally the average distance of all the test set is 0.30cm. The distance smaller than 0.30cm occupies more than 60% in the test set.

From the evaluations in both the acoustic space and articulatory space, one can see that the mapping based control strategy can drive the physiological articulatory model not only to generate desired speech but also to generate reasonable articulatory configurations with close distance to the target.

4. Conclusion

In this study, we constructed a direct mapping from acoustic objectives to muscle activation patterns and proposed a mapping-based control strategy for the physiological articulatory model. This mapping was constructed by simulating the speech motor learning process of humans. During the babbling training the articulatory positions of five Japanese vowels were used as the babbling targets. Through the babbling process, the mapping was constructed from acoustics to muscle activation by a two layer neural network. This mapping was evaluated by implementing it in the physiological articulatory model to generate a given speech sound, meanwhile the evaluation was also carried out in the articulatory space. These experiments proved that the mapping from acoustic objective to muscle activation patterns was effective.

From the perception point of view, the first three formants have different degree of contribution. Therefore, in the future work, individual formants will be weighted according to their perceptual importance . As mentioned earlier, both the auditory feedback and somatosensory feedback play important roles during the speech motor learning process. So far, only the somatosensory feedback has been implemented during the babbling simulation. The auditory feedback will be implemented to the motor learning process in the future work.

5. Acknowledgement

Thanks to MS. Stéphanie Buchaillard for her diligent work on constructing the 3D physiological articulatory model. This study is partially supported by 21st Century COE Program. This study is also supported in part by the National Thousand Talents Program of China, and in part by Grant-in-Aid for Scientific Research of Japan (No. 22500150).

References

[1] Y. Payan and P. Perrier, “Synthesis of V-V sequences with a 2D biomechanical tongue shape in vowel production” speech Communication. 22, 185-206, 1997

[2] J. Dang and K. Honda, “Construction and control of a physiological articulatory model,” J. Acoust. Soc. Am., 115, 853–870, 2004.

[3] BS. Atal, JJ. Chang, MV. Mathews and JW. Tukey “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique,” J. Acoust. Soc. Am., 63, 1535-1553, 1978.

[4] Q. Fang, S. Fujita, X. Lu and J. Dang, “A model-based investigation of activations of the tongue muscles in vowel production,” AST, 277-287, 2009.04

[5] http://www.magic.ubc.ca/artisynth/pmwiki.php?n=Main.HomePage

[6] X. Wu, Y. Wang and J. Dang., “Investigation of speech production using a 3D physiological articulatory model” ASJ Autumn Meeting, 335–338, 2009.

[7] S. Rvachew, K. Mattock and L. Polka “Developmental and cross-linguistic variation in the infant vowel space: The case of Canadian English and Canadian French” J. Acoust. Soc. Am., 120, 2250–2259, 2006.

[8] J. Dang, and K. Honda, “Investigation of the acoustic characteristics of the velum for vowels,” Proc. ICSLP 94, 603-606,Yokohama, 1994.09

[9] X. Wu, Q. Fang and J. Dang,. “Investigation of Muscle Activation in Speech Production Based on an Articulatory Model” Proc. ISCSLP, 330-334, Tainan, 2010.

[10] B. J. Kröger, J. Kannampuzha, C.Neuschaefer-Rube, “Towards a neurocomputational model of speech production and perception,” speech Communication. 51, 793-809, 2008.

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

[12] J.L. Flanagan., “Speech analysis synthesis and perception” New York /Berlin: Springerverlag. 1972

[13] T. Nakagawa., et al, “Tonal difference limens for second formant frequencies of synthesized Japanese vowels,” Ann. Bull. RILP., 1982 16: p. 81-88.
[14] R. G. Carson., “Changes in muscle coordination with training”, J. Appl. Physiol. 101: 1506–1513, 2006