A model based investigation of activation patterns of the tongue muscles for vowel production

Qiang Fang, Satoru Fujita, Xugang Lu, Jianwu Dang

Japan Advanced Institute of Science and Technology
{fangqiang, sfujita, xugang, jdang}@jaist.ac.jp

Abstract

Muscle activations in speech production are important for understanding speech control. To overcome the problems of previous methods, we proposed a physiological articulatory model based approach to explore the muscle activations in the production of the five sustained Japanese vowels through an optimization procedure which minimizes the morphological differences between the model simulations and MRI observations. The general findings were consistent with the observations obtained using EMG. In addition, we found that the Transversus and Verticalis actively participate in vowel production. This implies the proposed method is appropriate for estimating the muscle activation patterns during vowels production.

Index Terms: muscle activation, speech production, vowel production, 3D physiological articulatory model

1. Introduction

To understand how speech is controlled, it is important to investigate muscle activation patterns recruited in speech production. For years, speech scientists developed a number of methods to measure/estimate the muscle activations [1-4].

Tagged MRI based estimation methods have been extensively studied [1, 4]. Those methods, however, have several inherent drawbacks. First, it is difficult to define and evaluate the reference configuration. Second, since the tongue consists of volume preserving tissues and interweaved muscles, the muscle fibers may be shortened by either active contraction or passive deformation of the tongue, while only the active shortening of the muscles is associated with muscle activation. However, those two kinds of shortening can not be differentiated by tagged-MR images alone.

Electromyography (EMG) is the technique that directly measure muscle activations. The amount of the electrical activity measured by EMG depends on the size and function of the muscles as well as the location where the electrode attached [5]. It has been implemented to measure the muscle activation in speech production [2, 3]. However, it is difficult to measure the activations of the tiny intrinsic muscles, such as transversus and verticalis, by EMG.

As pointed out above, the previous methods have a variety of flaws in exploring the tongue muscle activations. To overcome those problems, in this study, we proposed a physiological articulatory model based method to estimate muscle activations in vowel production. The physiological articulatory model faithfully replicates the morphological and muscular structure of the supra-glottal articulatory system of a male subject. Since the model is equipped with human mechanisms of speech production, we can consider that the muscle activations of the model will achieve the “true values” of human articulation if the model generates the same articulatory posture as the target observation. An optimization procedure is conducted according to this criterion in the estimation. To increase accuracy, the initial setting for optimization is carefully selected to make the model-generated posture as closed as possible to the objective articulation. In this way, the activations of major extrinsic muscles and intrinsic muscles can be directly estimated without suffering from the uncertain information.

2. Configuration of the Physiological articulatory model

2.1. Configuration of the 3D tongue

Following the method proposed by Dang [6], the mesh structure of the tongue is based on the arrangement of the Genioglossus that fans out from attachment on the mandible. The initial shape of the tongue is obtained based on the volumetric MR image while producing the Japanese vowel /e/, which is close to a neutral position among the Japanese vowels. The configuration of the tongue is adapted to the realistic shape of the MRI-based tongue tissue geometry using five layers in the left-right dimension, which has a maximum width of 5.5 cm in the posterior portion, although the lateral-most regions of the tongue root are not included. The mesh structure of the tongue model in the lateral view consists of eleven layers, at nearly equal intervals, fanning out to the tongue surface from the attachment on the mandible, and seven layers in the perpendicular direction. Figure 1(a) shows the configuration of the 3D tongue in the physiological articulatory model.

2.2. Jaw and vocal tract wall

Figure 1: (a) The oblique view of the 3D tongue involved in the 3D physiological articulatory model. (b) The profile of the 3D physiological articulatory model

In order to produce speech sounds, we carefully extract contours of the jaw and vocal tract wall from the MRI images of vowel /e/ on which the lower and upper teeth are superimposed at intervals of 0.4cm in the left-right dimension. The sagittal images start from the outer surface of the molar on the left side and end at the outer surface of the molar on the right side. Therefore, the vocal-tract wall and jaw can form a completely enclosed space when the jaw clenches with
the maxilla. Finally, both the jaw and the vocal-tract wall consist of 15 layers, which spans over an interval of 5.6cm.

In the current model, the vocal-tract wall is composed of the upper teeth, hard palate, velum, pharyngeal wall, and larynx tube. The position of the velum can be adjusted according to the size of the velopharyngeal port, which is used for generating nasal sounds. Other parts of the vocal tract wall are treated as rigid components. Figure 1(b) shows the 3D physiological articulatory model.

When the jaw is open, the cheek is considered to form the lateral boundary of the vocal tract just outside the lower and upper teeth immediately. Therefore, the area function of the vocal tract can be directly calculated from the cross-sectional planes that intersect with the vocal tract. In this way, the 3D physiological articulatory model can depicts the area function of the vocal tract with any specific configurations and/or with dynamical variations more accurately.

2.3. Muscular structures

![Muscular structure images](image)

Figure 2: The muscular structure of the 3D physiological articulatory model. (Unit: cm)

To drive the physiological articulatory model according to human mechanism, associated muscles are included in the model. There are 9 muscles included in the tongue model. Three extrinsic muscles, Genioglossus (GG), Styloglossus (SG), and Hyoglossus (HG), are arranged mainly based on MRI analysis [7]. The intrinsic muscles, Superior Longitudinal (SL), Inferior Longitudinal (IL), Transversus (T), Verticalis (V), are modeled based on the anatomical data from Takemoto [8]. Tongue floor muscles, Mylohyoid (MH) and Geniohyoid (GH), are arranged based on anatomical literature. All the muscles are arranged in bilateral symmetry. According to the function of the different parts of GG [3], GG is divided into three portions: GG anterior (GGa), GG middle (GGm), and GG posterior (GGp). In addition the muscle groups response for jaw closing (jawCl) and jaw opening (jawOp) are included. Figure 2 shows the details of the muscular structure in the model.

3. Exploring the muscle activation patterns

It is reasonable to assume that the activations of the tongue muscles resulted from the estimation are the ones which can produce the observed posture of the tongue. Accordingly, the “true” muscle activation can be obtained when the 3D physiological articulatory model is driven to approach the postures of the observed articulations, by adjusting the activation of associated muscles, via an optimization procedure. To minimize the difference between the simulated and observed postures in 3D, the cost function is defined as in Eq.1. The right side of Eq.1 is the summation of two terms: the first term accounts for the difference of the tongue contours in the midsagittal plane, while the second term is related to the difference in the transversal dimension. Using this equation, the 3D tongue shape difference between simulation and observation can be taken into account to some extent.

\[
D = \frac{1}{N_s} (s - s_o)^T (s - s_o) + \frac{1}{N_w} (w - w_o)^T (w - w_o)
\]

where \( s_o \) and \( s \) are the vectors that represent the midsagittal contours of the tongue in the observation and simulation, respectively; \( w_o \) and \( w \) are the tongue width vectors obtained from observation and simulation, respectively; \( N_s \) and \( N_w \) are the number of dimensions of vector \( s \) and \( w \), respectively. Both \( s \) and \( w \) are functions with regard to muscle activation \( f \).

To calculate the difference between \( s_o \) and \( s \), we define a semi-polar coordinate system as shown in Figure 3. The contour of the tongue surface in the midsagittal plane is depicted by a sequence of lengths of the line segments that are nearly perpendicular to the tongue surface \( (s_1, s_2, ..., s_m) \). In this way, we can avoid defining corresponding nodes between observation and simulation, which is not easy to be determined.

To measure the difference in the transversal dimension, the tongue width of the model is defined as the average distance between the 2nd and 4th layers, which approximates the anatomical landmarks [9], in the left-right dimension.

Finally, estimating the muscle activations leads to looking for the optimal activation vector \( f \) that minimizes the difference between the observation and simulation. Nonetheless, it is difficult to explicitly define analytic functions for \( s \) and \( w \) with regard to \( f \), due to the complicated interactions between the tongue and surrounding structures. Hence, the gradient descent algorithm, which is widely used in various optimization tasks, is implemented to estimate the underlying muscle activation, as defined in Eq.2.

\[
f_s = f_{s+1} - \frac{0.05}{\lambda} \nabla f_s; \quad \lambda = \max \{ |d_i|, i = 1...N_f \}
\]

Where \( d_i \) is the \( i \)th element of vector \( \nabla f \), and \( N_f \) is the number of element in vector \( \nabla f \).

3.1. Procedure of estimating muscle activations

Figure 4 shows the estimation procedure used in this study in detail. In the first step, we choose the muscle combinations for producing the vowels according to the EMG observations.
in Baer’s experiments as the initial muscle combinations (see Table 1 for details), and set the initial activations of the muscles by referencing the relative amplitude of the EMG signal. In the second step, the optimization procedure is implemented to look for the optimal muscle activations that minimize the difference between simulation and target MRI observation with current muscle combination. If the difference between simulation and target MRI observations is greater than a predefined threshold, we refine the muscle combinations by activating additional muscle as necessary and/or deactivating the muscles which contribute little, then go back to the optimization step. Otherwise, the estimation procedure terminates, and outputs the optimal muscle activations.

Figure 4: The procedure for estimating the muscle activations during vowel production

Table 1. The muscles used in the first step which are derived from the Baer’s EMG experiment [2]

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Muscle combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>SG, HG</td>
</tr>
<tr>
<td>/i/</td>
<td>GGm, GGp, MH</td>
</tr>
<tr>
<td>/u/</td>
<td>GGp, HG, SG</td>
</tr>
<tr>
<td>/e/</td>
<td>GGm, HG, MH, SG</td>
</tr>
<tr>
<td>/o/</td>
<td>HG, SG</td>
</tr>
</tbody>
</table>

Figure 5 shows an example of estimating the muscle activations for producing vowel /a/ based on the above procedure, where the observed tongue contour is indicated by the curve with square markers. When activating the muscles observed in EMG experiment alone, the simulation result is obtained by optimizing the amplitude of the muscle activation and is shown in Figure 5 by the curve with open circles. One can see that there is an obvious difference in the anterior portion, although the simulated contour of the tongue is consistent with the observation in the posterior portion. Especially, the length of the contour in the anterior portion in the simulation is much shorter than in the observation. It implies that we should activate some additional muscles that can elongate the anterior part of the tongue. Unfortunately, there is no muscle directly responsible for the elongation. Since the tongue body is volume preserving, muscle contraction compresses the soft tissue in the longitudinal dimension of the muscle fibers, and expands the tissue in the perpendicular dimension. Hence, the co-contraction of T and V will shrink the cross-sectional area of the tongue, while elongating the tongue in the longitudinal direction. For this reason, we take the muscles T and V into account in the estimation. After considering the T and V in simulation, the tongue contour becomes closer to the observation, as indicated by the curve with diamond markers in Figure 8. However, there is still some difference between the simulation and the corresponding observation at tongue blade.

A certain muscle is required to be activated to depress this part. As known, by referencing the anatomical structure and experiment data, GGa is an appropriate candidate. After taking GGa into account, the final optimization result with inverse triangles becomes coincident with the observation.

Figure 5: An example of estimating activations of tongue muscles in the production of vowel /a/. The curve with the squares is the observed outline of the tongue in the midsagittal plane. The curve with circles is the outline of the simulated tongue in the midsagittal plane by considering HG and SG only. The curve with diamonds is the outline of the simulated tongue by considering HG, SG, T and V. The curve with triangles is the outline of the simulated tongue by considering HG, SG, T, V, and GGa. (Unit: cm)

3.2. Evaluation of the proposed method

In this section, we evaluate the proposed method regarding the following two aspects. First, the evaluation is carried out from the morphological point of view, where the difference between the simulation and target articulations is examined. Second, we investigate consistency of the estimated muscle activations with the EMG observations.

3.2.1 Morphological evaluation

Figure 6: The results obtained by optimization. Panels (a)-(e) are the contours of tongue in the midsagittal plane for the observations (dashed curves) and simulations (solid curves) of vowel /a/, /i/, /u/, /e/ and /o/ respectively. Panel (f) is the corresponding tongue width pattern of simulated vowels. (Unit: cm)

The target articulations are simulated using the above procedure shown in Figure 4. The morphological comparisons between simulations and observations are shown in Figure 6 for the five vowels. The solid curves indicate the
final configuration in the midsagittal plane of the simulations, and the dashed curves display the target MRI observations of the five vowels /a/, /i/, /u/, /e/, and /o/ in Figure 6(a)-6(e), respectively. One can see that the simulations coincide with the target MRI observations when the model is driven by the optimized muscle activation patterns. The average errors of the simulations are 0.10, 0.11, 0.12, 0.15, and 0.06cm for vowels /a/, /i/, /u/, /e/, and /o/ respectively. The errors at the constrictions, which are crucial for vowel production, are 0.09, 0.09, 0.03, 0.05cm for /a/, /i/, /u/, and /o/ respectively. Figure 6(f) shows the tongue width evaluation, which changes in producing the five vowels /e/, /a/, /i/, /u/, and /o/.

The tongue widths obtained show the same tendency as observed from MRI measurements: the tongue is wider in the vowels /i/ and /e/ than in vowels /a/, /u/, and /o/. Among the back vowels, the tongue width shows the same order as the corresponding observation (/u/>/a/>/o/). A quantitative analysis indicates that the differences of the width between the simulations and observations for each vowel are less than 0.07cm. This indicates that, by using the proposed physiological articulatory model based method, the morphological details of the target MRI observations are correctly represented by the optimal simulations.

3.3.2 Evaluation of muscle activations

For a correct estimation, the muscles discovered in the result should include the muscles whose activation was detected in the EMG experiment. Moreover, the activations of that muscle should show similar pattern as what was shown in EMG experiments in producing the vowels.

Table 2. The activation of tongue muscle obtained according to the above procedure (Unit: Newton)

<table>
<thead>
<tr>
<th></th>
<th>/a/</th>
<th>/i/</th>
<th>/u/</th>
<th>/e/</th>
<th>/o/</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGa</td>
<td>2.4</td>
<td>0.7</td>
<td>1.2</td>
<td>0.3</td>
<td>1.0</td>
</tr>
<tr>
<td>GGM</td>
<td>0.0</td>
<td>0.2</td>
<td>0</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>GGP</td>
<td>0.0</td>
<td>0.9</td>
<td>0.4</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>HG</td>
<td>6.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>4.0</td>
</tr>
<tr>
<td>SGa</td>
<td>0.0</td>
<td>0.8</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>SGp</td>
<td>1.0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.1</td>
<td>1.0</td>
</tr>
<tr>
<td>SL</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>T</td>
<td>0.6</td>
<td>0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>1.6</td>
</tr>
<tr>
<td>V</td>
<td>0.3</td>
<td>0.1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>IL</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>GH</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>MH</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>jawOp</td>
<td>6.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>4.0</td>
</tr>
<tr>
<td>jawCl</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Although the model based estimation starts from the EMG observations, the muscles activated at the beginning are possibly excluded in the optimization procedure, since the muscles would be deactivated if they make little contribution. The muscle activations corresponding to the above optimal simulations are obtained by the propose method, as shown in Table 2. In the table, the zero force means that the muscles are not activated in production of the vowels. Comparing the muscles activated for each vowel as estimated by the proposed method with the observations of corresponding vowels in Table 2, one can see that all of the muscles detected by EMG remain in estimation results.

In addition, as shown in Table 2, GGM (corresponds to GGa in EMG experiment [2]) and MH are activated with relatively higher activation level for vowels /i/ and /e/ than for other vowels; GGP showed strong activations in front vowels /i/ and /u/; HG and SG showed stronger activations in producing back vowels /a/, /o/, and /u/. All of these show muscle activations among the five vowels are consistent with the EMG observations [2].

The above results imply that the proposed method can reproduce the “true” muscle combinations in producing the five Japanese vowels as well as their muscle activations.

4. Conclusions

This paper proposed a physiological articulatory model-based method to estimate the muscle activations in the production of five sustained Japanese vowels. By using the proposed method, the extrinsic muscles observed in EMG experiments [6] appear robustly in the simulation results. In addition, the model based estimation method revealed for the first time that muscle T, V, and GGa are activated in producing all five Japanese vowels. T and V play important roles in manipulating the length of the tongue surface in the longitudinal direction, and GGa has an important function in active control of the anterior portion of the tongue body. This implies the proposed method is powerful for estimating the activations of tongue muscles in speech production.

5. Acknowledgements

This research is conducted as a program for the “21st Century COE Program” for promoting Science and Technology by Ministry of Education, Culture, Sports, Science and Technology. This study is also supported in part by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan.

6. References